A GEOSTATISTICAL MODEL BASED ON BROWNIAN MOTION TO KRIGE REGIONS IN $\mathbb{R}^2$ WITH IRREGULAR BOUNDARIES AND HOLES

By

Jordy Bernard

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Statistics

University of Alaska Fairbanks

May 2019

APPROVED:

Julie McIntyre, Committee Chair

Ron Barry, Committee Chair

Scott Goddard, Committee Member

Anthony Rickard, Chair

Department of Mathematics and Statistics

Abstract

Kriging is a geostatistical interpolation method that produces predictions and prediction standard errors. Classical kriging models use Euclidean (straight line) distance when modelling spatial autocorrelation. However, for estuaries, inlets, and bays with shortest-in-water distance can capture the proximity dependencies within the system better that Euclidean distance when boundary constraints are present (Little, Edwards, & Porter, 1997; Rathbun, 1998). However, the variance-covariance matrix that results in kriging models that use the shortest-in-water distance metric hasn't been shown to be mathematically valid. In this paper, a new kriging model is developed for irregularly shaped regions in $\mathbb{R}^2$ where the notion of flow connectivity is encorporated into a valid error structure through the use of a random walk on a lattice, process convolutions, and the non-stationary kriging equations. Finally, the model is compared to other geostatistical and non-spatial smoothers using water quality data from Puget Sound.

Contents

1 Introduction

It is almost convention to begin a manuscript in spatial statistics by stating Tobler's first law of geography. It goes "Everything is related to everything else, but near things are more closely related than distant things." (Tobler, 1970). This phenomenon, referred to as spatial autocorrelation, lies at the heart of spatial statistics and distinguishes the field from ordinary statistical theory. Consider the usual general linear model $Y = X\beta + \epsilon$. Here, the design matrix X together with the parameters β describe the mean (or fixed) structure of the model, whereas the random errors ϵ along with the variance covariance matrix $var(\epsilon) = \Sigma$ describe the error (or random) structure of the model. With ordinary statistical models, the random errors are assumed to be independent (i.e. $\Sigma = \sigma^2 I$), but with spatial models, this assumption is relaxed so that the spatial dependencies among errors inform Σ.

Generally, Σ is modeled through a purely emperical (data driven) approach. However, the stream network model presented in (Ver Hoef, Peterson, & Theobald, 2006) addresses the question of why near thing are more closely related than distant things which leads to a more mechanistic description of Σ. In the context of stream networks, a fundamental mechanism (hydrologic flow) explains the presence of spatial autocorrelation, and this mechanism is used in determining the relatedness of random errors within the system. In this model, basic properties, rather than a complete mathematical description, of streamflow are incorporated into the error structure. Nearby reaches of stream located on the same stream segment have a higher proportion of common water molecules flowing through them than reaches of stream that are far apart, flow disconnected, or separated by intermediate confluences. Therefore, water samples taken from the nearby common reaches should, in general, be more related. To incorporate the notion of hydrologic flow into the error structure, flow connected river distance is used in lieu of classical Euclidean (as-a-crow-flies) distance to capture proximity dependencies within the system.

Kriging is a method of spatial prediction, or smoothing, where prediction standard errors are produced. Non-Euclidean distance metrics have been useful in bringing the theory of kriging to new systems characterized by an irregular topology. In addition to stream networks, system specific geostatistical models have been developed for spheres (Gneiting et al., 2013) and for road networks (Zou, Yue, Li, & Yeh, 2012). Attempts have been made to develop kriging models for regions contained in $\mathbb{R}^2$ with irregular boundaries and holes such as estuaries, lakes, and sounds. (Little et al., 1997) and (Rathbun, 1998) use shortest-in-water distance (SWD) to krige inlets, estuaries, and bays respectively. However, the mathematical validity of Σ has not been verified in these models though (Rathbun, 1998) recognizes the issue. Additionally, another non-Euclidean distance metric may capture the notion of a flow connected proximity dependence better than SWD. For these reasons, the kriging of irregularly shaped regions in $\mathbb{R}^2$ is an open problem and is the subject of this paper.

In what follows, the tactics of (Ver Hoef et al., 2006) are followed to develop a new geostatistical model for irregular shaped regions in $\mathbb{R}^2$ where flow can be seen to account for autocorrelation within the system. This document

is organized with the intention of highlighting this methodology, as the framework could be useful in the development other system specific models. In Chapter 2, necessary background material and literature is summarized. Afterwards in Chapter 3, the mathematical framework of the proposed model is presented. Finally, in Chapter 3, the model is compared to existing methods of spatial prediction for irregularly shaped regions using water quality data from Puget Sound.

2 Background

2.1 Spatial Random Processes and Geostatistics

The primary concern of spatial statistics is in modeling data as a realization of a spatial stochastic process. A basic description of spatial stochastic processes is provided here following (Cressie, 1992). To start, let $s \in \mathbb{R}^d$, $d \in \mathbb{Z}$ represent an arbitrary data location which is typically represented by a column vector in $\mathbb{R}^d$ of coordinate locations. Later on, the concern will be the modeling of two-dimensional spatial regions (i.e. $d = 2$). Now let $Z(s)$ be a random vector located at spatial location $s \in D \subset \mathbb{R}^d$. Here, $Z(s)$ can be seen to represent the value of a variable sampled at site s, and the set D is referred to as the index set and represents a general spatial region of interest. Now, $\{Z(s) : s \in D\}$ defines a random process (i.e. a set of random variables indexed over a spatial domain and defined over a common probability space). With the final requirement that s be allowed to vary continuously throughout D, data can be categorized as geostatistical.

Continuously distributed spatial data is commonly found in the natural world, and it many times can be seen as arising from a spatial stochastic process. For this reason, a broad range of problems can be modeled geostatistically. Since the 1980's, geostatistical models have become popular in fields of atmospheric science, soil science, and hydrology. Because geostatistics has a wide range of applications, the prefix "geo" referring to the earth, can be misleading (the prefix is explained by the field's historical development, as geostatistics was originally developed as a method of ore-reserve estimation (Cressie, 1992, p. 29)). Many times, the primary goal of geostatistical analyses is prediction. That is, samples are collected from sites $s_1, s_2, \cdots, s_n \in D$ with the intention of predicting $Z(s_0)$ for all $s_0 \in D$. To interpolate these estimates, a geostatistical technique called kriging is classically employed.

(a) Kriged Surface

(b) Kriged PI Width

(c) Kriged Low 2.5

(d) Kriged High 2.5

(e) NS Low 2.5

(f) NS High 2.5

Figure 1: Using the chloryphyll data from Puget Sound; (a) plots kriged predictions resulting from the use of a Euclidean distance metric; (b) gives the kriged prediction interval width; (c) and (d) show lower and upper bounds of the kriged 95% prediction interval; and (e) and (f) depict the constant prediction interval bounds produced by non-spatial models.

2.2 Ordinary Kriging

There are different kriging methods differentiated by the assumptions made about a random process's underlying mean structure. Ordinary kriging, which assumes a constant mean, is the focused of the following discussion. Nevertheless, the model can be easily extended to the case of universal kriging. The main benefit to kriging over non-spatial methods of prediction lies in the spatial structure of the prediction intervals. With non-spatial smoothers, prediction

intervals do not possess a spatial structure in that they are constant across the region. However, the prediction intervals for kriging models are constricted at locations nearby sampled sites (see Figure 1). When used appropriately, kriging models reduce the variability of random errors without sacrificing the reliability of the associated prediction intervals.

In providing a broad overview of ordinary kriging, let $Z(s_i)$ be a set of measurements taken from $s_i \in D \subset \mathbb{R}^d$ for $i = 1, 2, ..., n$ where each s_i is allowed to vary continuously throughout the domain. These measurements are seen as a realization of the random process $Z(\cdot)$. With slightly different stationarity assumptions, ordinary kriging can be formulated in terms of a variogram or a covariogram. With the covariogram, $Z(\cdot)$ is assumed to be second order (or weakly) stationary. $Z(\cdot)$ is defined to be second order stationary when

1. $\mathbb{E}[Z(\cdot)] = \mu$
2. $C(h) = Cov[Z(s_1), Z(s_2)]$ exists and only depends upon $h = s_1 - s_2$ for all s_1 and s_2 in D

(Cressie, 1992, p. 53). Here, function $C(h)$ is called the covariogram. Ordinary kriging can, however, be carried out under slightly weaker conditions than second order stationarity. Specifically, $Z(\cdot)$ need only be intrinsically stationary. Intrinsic stationarity is defined through the first difference. That is, $Z(\cdot)$ is defined to be intrinsically stationary when the second condition listed above is replaced with the condition

$$2\gamma(h) = Var[Z(s_1) - Z(s_2)] \text{ exists and only depends upon } h = s_1 - s_2 \text{ for all } s_1 \text{ and } s_2 \text{ in } D$$

(Cressie, 1992, p. 40). Here, the function $2\gamma(h)$ is called the variogram and $\gamma(h)$ is called the semivariogram. When a process is second order stationarity, a simple relation exists between the variogram and the covariogram. Specifically, $\gamma(h) = C(0) - C(h)$. Here, $C(0)$ is called the sill (see Figure 2). However, when a process is weakly stationary, the covariogram may not be defined. For this reason, in the following discussion, the variogram may be used in lieu of the covariogram to speak in general terms so that the less restrictive conditions are addressed. In addition to the regularity assumption of stationarity, it is often times assumed that $Z(\cdot)$ is isotropic. A weakly stationary process is defined to be isotropic if $\gamma(h) = \gamma(d(s_1, s_2))$ where $d(s_1, s_2)$ is the distance between sites s_1 and s_2 (Cressie, 1992, p. 53). This assumption essentially states that there isn't a directional dependence in a random processes error structure. With these assumptions, restrictions are placed on the error structure which make the modelling and estimation of Σ possible.

After stationarity and isotropy assumptions are made, an estimate of the true variogram is made. This estimate is called the emperical variogram. Usually, the method of moments estimator (Matheron, 1962) or the robust estimator (Cressie & Hawkins, 1980) is used as the emperical variogram (see Figure 2). After an emperical variogram has been constructed, a valid parametric family of variogram is chosen by "eyeballing" the empirical variogram (Barry, Jay, & Hoef, 1996). After the family of variogram has been chosen, it is assumed to be the true variogram and its

parameters are estimated (Barry et al., 1996). Parameter estimation can be carried out using a variety of least squares and maximum likelihood based techniques (Cressie, 1992, p. 90). Because the topic of a variogram validity is of central importance in this paper, the topic is discussed in its own section in the next paragraph.

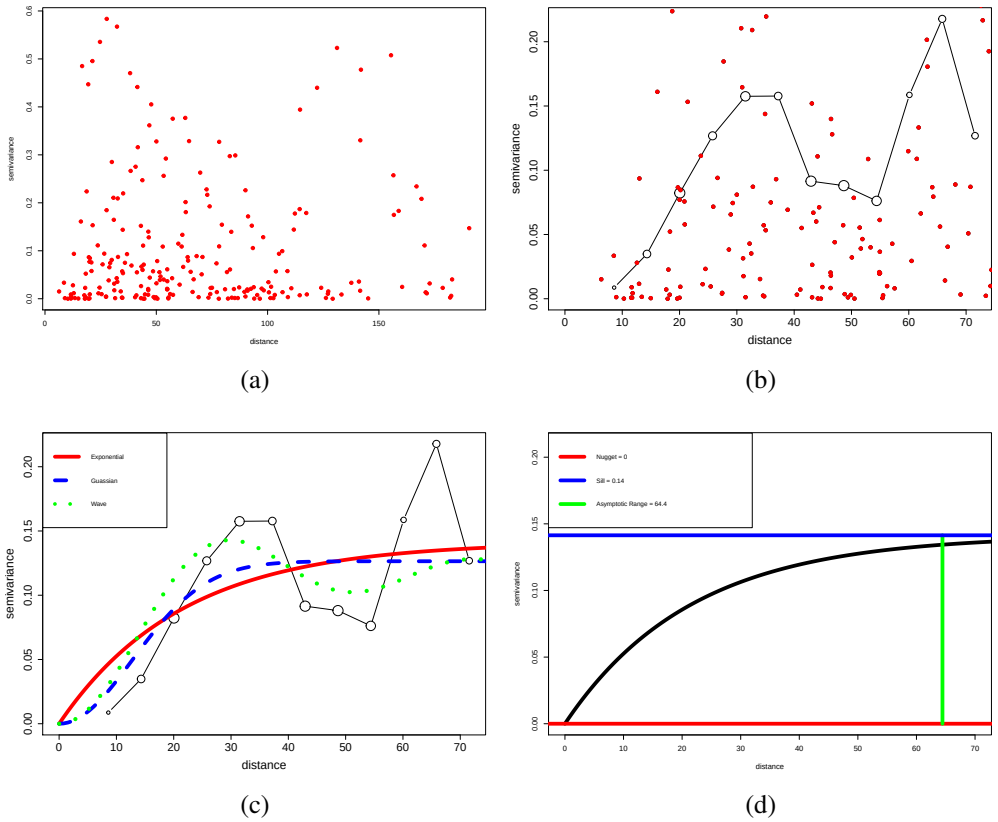

Figure 2: Using chloryphyll data from Puget Sound: (a) plots the variogram cloud (emperical semivariance against distance); (b) overlays the robust estimate of the variogram; (c) fits an exponential, Gaussian, and wave variogram to the emperical variogram using weighted least squares (WLS); and, in (d) the exponential variogram is used to depicts the usual parameterization for classically used variogram families.

A variogram family is said to be valid when every function in a parametric variogram family is conditionally negative definite. That is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(h) \leq 0$$

for any $s_1, s_2, ..., s_n \in D$ and any $a_1, a_2, ..., a_n \in \mathbb{R}$ where $\sum_{i=1}^{n} a_i = 0$ (Cressie, 1992, p. 60). To re-express this condition in terms of the covariogram, a parametric covariogram family is valid when every function in the family is positive definite (Cressie, 1992, p. 86). The condition of positive/negative definiteness guarantees that the variance covariance matrix that results from the variogram/covariogram (Σ) is positive semidefinite. A matrix is positive semidefinite when the matrix is symmetric with non-negative eigenvalues. A valid variance covariance matrix guarantees that negative variances will not result from linear combinations of the associated random variables (Cressie,

1992, p. 84). To illustrate with an example, consider the simple variance covariance matrix for the pair of random variables (Y_1, Y_2)

$$\Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 1 \end{bmatrix}$$

The eigenvalues of Σ are 4.23 and -0.24. Because the second eigenvalue is negative, Σ isn't positive semi-definite and doesn't represent a valid error structure. To see how negative variances can result from an invalid matrix, compute

$$Var(Y_1 - 2Y_2) = Var(Y_1) + 4Var(Y_2) - 4Cov(Y_1, Y_2) = -1$$

When an invalid variance covariance matrix is used in a kriging model, the model can produce embarrasing negative variance estimates.

Once a variogram's parameters have been estimated, predictions are then made by calculating the best linear predictor $Z(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$ subject to the constraint $\sum_{i=1}^{n} \lambda_i = 0$ which guarantees uniform unbiasedness of the predictor (Cressie, 1992, p. 120). The coefficients $\lambda = [\lambda_1, \lambda_2, ..., \lambda_n]^T$ are given by

$$\lambda^T = \left(\gamma + 1^T \frac{1 - 1^T \Gamma^{-1} \lambda}{1^T \Gamma^{-1} 1}\right)^T \Gamma^{-1} \tag{1}$$

where $\gamma = [\gamma(s_0 - s_1), \gamma(s_0 - s_2), ..., \gamma(s_0 - s_n)]^T$ and Γ is a $n \times n$ matrix where $\Gamma_{ij} = \gamma(s_i - s_j)$. Additionally the minimized mean-squared prediction error $\sigma_k^2(s_0)$ is given by

$$\sigma_k^2(s_0) = \gamma^T \Gamma^{-1} \gamma - \frac{(1^T \Gamma^{-1} \gamma - 1)^2}{1^T \Gamma^{-1} \gamma} \tag{2}$$

(Cressie, 1992, p. 122). From these two equations predictions can be made for measurements $Z(s_0)$ for every $s_0 \in D$ along with an associated prediction interval.

Now, an important note that will be used later on is that the ordinary kriging equations do not necessarily require that $Var(Z(s_i) - Z(s_j))$ is a function $s_i - s_j$. In other words, the ordinary kriging equations can be written in terms of variances that do not have to be stationary. In the broader formulation, γ is re-defined as $\gamma = [\gamma(s_0, s_1), \gamma(s_0, s_2), ..., \gamma(s_0, s_n)]^T$ and Γ as $\Gamma_{ij} = \gamma(s_i, s_j)$ where $\gamma(s_i, s_j) = \frac{1}{2} Var[Z(s_i) - Z(s_j]$ (Cressie, 1992, p 123). The non-stationary kriging equations can be formulated in terms of covariances when $Cov(s_i, s_j)$ is known for every s_i and s_j in D.

2.3 Kriging with Non-Euclidean Distance Metrics, Valid Error Structures, and the Irregular Boundaries Problem

Euclidean distance refers to straight-line distance. The Euclidean distance between two sites is given by the usual Pythagorean Formula. That is, in two dimensions the distance between sites $s_1 = [x_1, y_1]^T$ and $s_2 = [x_2, y_2]^T$ is given by $||s_1 - s_2|| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. In certain systems, a non-Euclidean distance metric may do a better job at capturing the system's proximities than the classic Euclidean distance metric. For this reason, new models with non-Euclidean distance metrics have been effective in kriging regions with unusual topologies. Unfortunately, kriging with a non-Euclidean distance metric is more complicated than simply picking a new distance metric, fitting a variogram, and plugging into the kriging equations.

The reason that non-Euclidean kriging isn't a straight forward problem is that a valid variogram or covariogram may become invalid when used with a non-Euclidean distance metric. This is an important point and will be illustrate with an example following (Rathbun, 1998) and (Curriero, 2006). The Manhattan distance between two sites s_i and s_j is given by $|x_i - x_j| + |y_i - y_j|$. Now suppose that four sites are located in $\mathbb{R}^2$ where the sites form a square with unit legs. Let the coordinates of the four sites, s_1, s_2, s_3, s_4, be given by $(x_1, y_1) = (0, 0), (x_2, y_2) = (1, 0), (x_3, y_3) = (0, 1)$, and $(x_4, y_4) = (1, 1)$ respectively. The Gaussian variogram is an example of a variogram family that is valid when used with the Euclidean distance metric (see Figure 2). When a Gaussian variogram's nugget, sill, and range parameters set to 0, 20, and 4 respectively, the variogram is given by $\gamma(h) = 20[1 - exp(-h^2/4)]$. When Manhattan distance is used in lieu of Euclidean distance, the variance covariance matrix that results from the Gaussian variogram is

$$\Sigma = \begin{bmatrix} 20.0 & 15.58 & 15.58 & 7.36 \\ 15.58 & 20.00 & 7.36 & 15.58 \\ 15.58 & 7.36 & 20.00 & 15.58 \\ 7.36 & 15.58 & 15.58 & 20.00 \end{bmatrix}$$

The eigenvalues of Σ are given by 58.5,12.6,12.6, and -3.8. This means that the Gaussian covariance function isn't valid when used with Manhattan distance as the eigenvalue -3.8 is negative. Later on, this method (computing the eigenvalues of four sites arranged in a square to screen for invalidity) will be referred to as Curriano's Test for Invalidity.

Variogram Family	$\gamma(h)$ with $h \neq 0$
Exponential	$\tau^2 + \sigma^2[1 - exp\{-\|h\|/\phi\}]$
Gaussian	$\tau^2 + \sigma^2[1 - exp\{-h^2/\phi\}]$
Wave	$\tau^2 + \sigma^2\Big[1 - \frac{\sin \|\|h\|\|/\phi}{\|\|h\|\|/\phi}\Big]$

Table 2.3 contains a short list of variogram families that are valid when used with a Euclidean distance metric. For the listed variograms listed in the table, $\gamma(0) = 0$. Here, the parameter τ^2 is called the nugget, σ^2 is the partial sill, and ϕ is the range parameter (see Figure 2).

When a Euclidean distance metric is used to krige such systems, boundaries are ignored which results in poor prediction resulting from smoothing across boundaries. As mentioned, (Rathbun, 1998) and (Little et al., 1997) use SWD in place of Euclidean distance to address this problem. In these papers, variograms designed for Euclidean distance are used with SWD when kriging the region. It should, however, be noted that the varigrams used in (Rathbun, 1998) pass Curriano's Test For Invalidity. Even so, the variograms are not guarenteed to be valid when used with SWD.

(a) Euclidean Distance (b) SWD Distance

Figure 3: Distance metrics: (a) The Euclidean distance between two sites is represented by the red line. The distance metric violates (cuts across) the boundary.

2.4 The Use of Process Convolutions to Develop Valid Error Structures

There are two approaches to show that a variogram is valid when used with a particular distance metric. The first approach is to propose a variogram and then verify that it is valid using direct methods (Cressie, 1992, 86). The drawback to this approach is that when a distance metric is complicated, showing variogram validity is a difficult task. The other, more constructive, approach is to employ process convolutions (also refered to as moving average functions) to develop valid variogram families.

A large class of valid variograms can be constructed by convolving a moving average function across a white noise random process. To construct a valid variogram using a moving average approach, a parametrized function $g : D \to \mathbb{R}$ (called a the moving average function) is chosen, and a random process $Z(\cdot)$ is defined by

$$Z(s|\theta) = \int_{x \in D} g(x - s|\theta) W(x) dx \tag{3}$$

for $s \in D$, where $W(x)$ is a white noise random process, and where

$$2\gamma(h|\theta) = \int_{x \in D} [g(x|\theta) - g(x-h|\theta)]^2 dx < \infty \tag{4}$$

(Barry et al., 1996). This final condition guarentees the validity of the variogram and thus the validity of the coore-sponding covariogram and variance-covariance matrix. When the process $Z(\cdot)$ is second-order stationary the covari-ogram is then given by

$$C(h|\theta) = \int_{x \in D} g(x|\theta) g(x-h|\theta) dx \tag{5}$$

(Ver Hoef et al., 2006).

It will now be illustrated how process convolutions can be used to construct valid variograms through a simple example. We begin by defining a moving average function $g : \mathbb{R} \to \mathbb{R}$ as $g(x) = I\Big(-\frac{1}{2} < x < \frac{1}{2}\Big)$ where I is the usual indicator function. The moving average function is un-parametrized for illustrative simplicity. A random process $Z(\cdot)$ is now defined using Equation (3) as

$$\begin{aligned} Z(s) &= \int_{x \in \mathbb{R}} I\Big(-\frac{1}{2} < x - s < \frac{1}{2}\Big) W(x) dx \\ &= \int_{x \in \mathbb{R}} I\Big(s - \frac{1}{2} < x < s + \frac{1}{2}\Big) W(x) dx \end{aligned}$$

Using Equation (4), the variogram is then given by

$$\begin{aligned} 2\gamma(h) &= \int_{x \in \mathbb{R}} \Big[I\Big(-\frac{1}{2} < x - s < \frac{1}{2}\Big) - I\Big(-\frac{1}{2} < x - s - h < \frac{1}{2}\Big)\Big]^2 dx \\ &= \int_{x \in \mathbb{R}} \Big[I\Big(s - \frac{1}{2} < x < s + \frac{1}{2}\Big) - I\Big(s + h - \frac{1}{2} < x < s + h + \frac{1}{2}\Big)\Big]^2 dx \\ &= \int_{x \in \mathbb{R}} \Big[I\Big(-\frac{1}{2} < x < \frac{1}{2}\Big) - I\Big(h - \frac{1}{2} < x < h + \frac{1}{2}\Big)\Big]^2 dx \\ &= \begin{cases} 0 & \text{if } h = 0 \\ 2|h| & \text{if } 0 < |h| < 1 \\ 1 & \text{if} |h| > 1 \end{cases} \end{aligned} \tag{6}$$

Because the condition expressed in Equation (4) (i.e. $2\gamma(h) < \infty$) is satisfied, the variogram is guarenteed to be valid. For this reason, process convolutions can be extremely in developing models valid variance/covariance structures for non-Euclidean distance metrics. (Ver Hoef et al., 2006), (Peterson, Theobald, & ver Hoef, 2007), and (Ver Hoef & Peterson, 2010) use a process convolution based approach to krige river networks; (Zou et al., 2012), to krige road networks; and (Gneiting et al., 2013), to krige spheres. Because Equation (4) is critical to this paper it is illustrated further through a pictoral representation of Equation (6) in Figure 4.

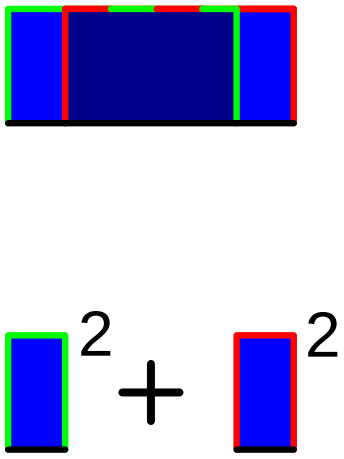

Figure 4: A pictoral representation of 1: Here, $g(x)$ and $g(x-h)$ are plotted. $g(x)$ is outlined in green while $g(x-h)$ is outlined in red. To evaluate the variogram at a lag of h, the overlapping area (shaded in dark blue) is removed, the remaining area is squared and added together.

The moving average function $g(\cdot)$ can be generalized to construct non-stationary covariance functions. In this extension, a more general function $g : D \times D \to \mathbb{R}$ is chosen. A random process $Z(\cdot)$ is then defined by

$$Z(s|\theta) = \int_{x \in D} g(s, x|\theta) W(x) dx \tag{7}$$

Where $W(x)$ is a white-noise random process and where the condition

$$2\gamma(s, u|\theta) = \int_{x \in D} [g(s, x|\theta) - g(u, x|\theta)]^2 dx < \infty \tag{8}$$

is satisfied. When a non-stationary variogram is constructed from a general moving average function $g(\cdot, \cdot)$ a regularity assumptions is needed to make parameter estimation possible. In many cases, a stationary variogram defined over a non-Euclidean distance will be equivalent to a non-stationary variogram. This is the case with the model proposed later on. A non-stationary variogram is used as the equivalent distance metric under the stationarity condition would be abstract. This distance metric is equivalent to a regularity assumption which allows the parameters of the non-stationary variogram to be estimated.

3 The Proposed Model

This chapter develops the mathematical framework of the proposed model. The chapter begins by defining a moving average function to construct a valid non-stationary covariogram. The kernal of the moving average function will be defined with a random walk on a lattice to approximate a boundary respecting version of the Gaussian kernal. Though the language of "distance" is not used, the primary aim of the proposed model is to encorporate the notion of flow connectivity into the error structure.

3.1 Random Walks to Approximate a Gaussian Kernal

Random walks on lattices have been useful in modeling spatial autocorrelation for intrinsic simultaneous autoregressive (SAR) models (Hanks, 2015). Additionally, Gaussian variograms are commonly used in geostatistical analyses. The Gaussian variogram can be constructed with a process convolution where the moving average function has a Gaussian kernal. A boundary respecting approximation of the Gaussian kernal can be constructed by defining a random walk on a lattice, and the construction has been useful in developing new solutions to problems with boundary constraints (McIntyre & Barry, 2018) and (Barry & McIntyre, 2011). The basic idea is analogized to the outward diffusion of ink that, when dropped in water, respects and works around boundaries. This behavior is mimicked when a random walk is defined on a lattice shaped by a regions topology.

To construct a lattice for the random walk following (Barry & McIntyre, 2011), begin by letting the spatial index $D \subset \mathbb{R}^2$ represent the geographic region of interest. The only restriction placed on D is that the set must be of positive volume in $\mathbb{R}^2$: the region allows for irregular boundaries, holes, and non-connected regions. D is now overlaid with a fine mesh. The mesh is then trimmed so that every grid-box contains at least one $s \in D$. The trimmed mesh defines a partition of D. A node is now placed at the center of each grid-box, and N is used to denote the total number of nodes. Each node is indexed, and I is used to represent the set of all indexes. Additionally, n_i is chosen to represent the i^{th} node; B_i, to represent the set of all locations in the i^{th} partition; and s_i and u_i, to represent arbitrary locations in B_i. If the context does not require the subscript, it may be dropped. As an example, n may be used in lieu of n_i to denote an arbitrary node. Now, a lattice isn't complete without defined neighbor relationships. In the implementation taken in this paper, the node itself along with the directly adjacent nodes in the N, S, E, W, NE, NW, SE, and SW directions are considered to be neighbors. The lattice construction is illustrated in (Figure 5).

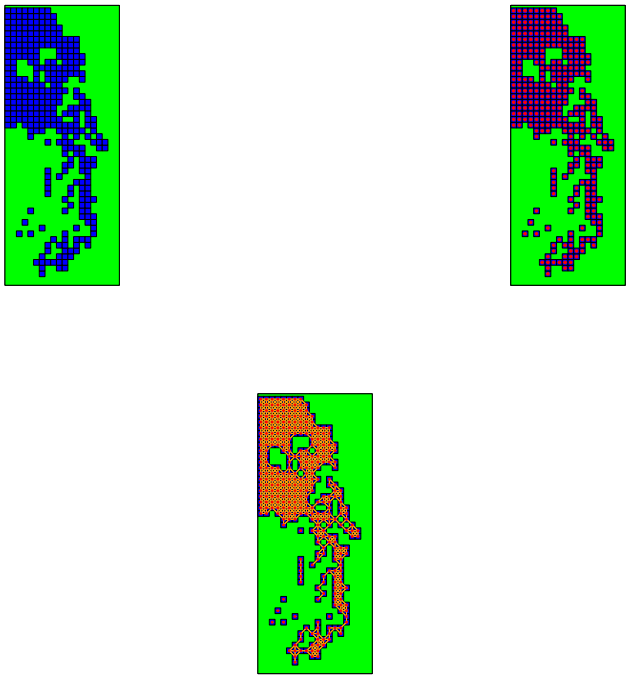

Figure 5: Constructing a lattice: (a) mesh is inscribed; (b) nodes are placed in the center of each gridbox; and, (c) neighbor relations are defined.

A random walk will now be constructed on the lattice by defining a Markov chain. Markov chains describe a sequence of possible events where each state depends on, and only on, the previous state. To begin, let $X_s(k)$ represent the location of position of the random walk after $k \geq 0$ steps when originating from site s. A Markov chain is defined by

$$P[X_{s_i}(1) = u_j] = \begin{cases} 0 & \text{if } n_i \neq n_j \text{ aren't neighbors} \\ \dfrac{\theta_d}{8} & \text{if } n_i \neq n_j \text{ are neighbors} \\ 1 - \dfrac{\theta_d(q_i - 1)}{8} & \text{if } n_i = n_j \end{cases} \tag{9}$$

for all i and $j \in I$ where q_i is n_i's number of neighbors, and $0 < \theta_d < 1$. Here, the parameter θ_d governs the rate of diffusion. When θ_d decreases, the diffusion rate increases. In the implementation taken in this paper, θ_d was fixed with $\theta_d = 1$.

To define the initial state of the chain, let

$$P[X_{s_i}(0) = u_j] = \begin{cases} 1 & \text{if } s_i = u_j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Later on, $P[X_s(k) = u]$ will be referred to as a diffusion of length k originating from s. An interesting note is that

when all of the nodes are connected in the sense that you can get from 1 node to any other through the neighboring relationships, the transitional probabilities will eventually become uniform (Barry & McIntyre, 2011). A diffusion of length k can be used to approximate a boundary respecting version of the Gaussian kernal. Equations (9) and (10) do little to clarify this point, so to make the concept less opaque, refer to Figure 6.

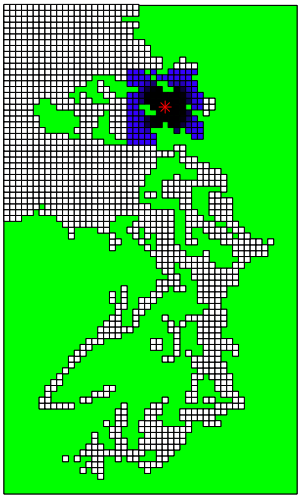

Figure 6: Random walks to approximate a boundary respecting Gaussian kernal: a diffusion of 7 steps originating from the point marked in red is pictured. Note: the plot is fat as a latitudidinal correction hadn't yet been made at the time of the plots construction

3.2 Defining the Moving Average Function

To encorporate a flow connected proximity dependence into the model, a diffusion of parametrized length θ_f was used as the kernal of the moving average function. To define the moving average function $g : D \times D \to \mathbb{R}$ formally, let

$$g(s|\theta) = \theta_s P[X_s(k) = u] I(k = \theta_f) \tag{11}$$

Here, the parameter θ_s controls the scale of the moving average function. This parameter can be interpreted in different ways: it can be seen as governing the usual sill, or equivalently, as governing the quantity of ink in the diffusion. An additional parameter could be included in Equation (11) to encorporate a nugget effect. The nugget effect represents small scale variability that is commonly found in mining applications (see Figure 2 (d) and Table 2.3). The decision to omit a nugget effect is due to its likely negligible relevance in water quality applications. In this context, a small change in location is not anticipated to cause material changes in water chemistry.

With the moving average function $g(\cdot|\theta)$ defined, the equivalent construction to Equation (7) is given by

$$2\gamma(s,u) = \theta_s^2 \int_{x \in D} \{P[X_s(k) = x] - P[X_u(k) = x]\}^2 I(k = \theta_f) dx \tag{12}$$

Similarly, the covariogram from Equation (8) becomes

$$C(s,u) = \theta_s^2 \int_{x \in D} P[X_s(k) = x] P[X_u(k) = x] I(k = \theta_f) dx \tag{13}$$

In interpreting Equations (12) and (13) it will be useful to refer to Figures 4 and 6. Two diffusions of length θ_f are started from sites s and u. The diffusions' overlapping area is removed, and $2\gamma(s,u)$ is given by squaring and scaling the remainder. Similarly, $C(s,u)$ is given by the area of the pointwise product of the diffusions.

3.3 Parameter Estimation and the Kriging Step

At this point the parameters of $\gamma(s,u)$ are estimated. Different parameter estimations techniques could be used including least square based methods (Cressie, 1992, p.). In this paper, residual maximum likelyhood (REML) was used. The choice was based on ease of model extention. When the model is generalized to the universal case (with a non-constant mean) the more complicated mean stucture is accomidated by REML.

Once the model parameters have been estimated, predictions and prediction standard errors are obtained using the non-stationary versions of Equations (1) and (2).

4 Model Comparison

In this project, the proposed model is compared to existing smoothers using water chloryphyll data collected by the State of Washington Department of Ecology and downloaded from https://fortress.wa.gov/ecy/eap/marinewq/mwdataset.asp. R was used to develop and test the model.

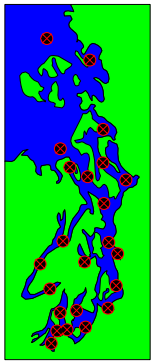

Figure 7: Average chloryphyll was sampled at 22 sites throughout Puget Sound during a ten day period from 1/13/15 to 1/22/15. The sample locations are depicted in the figure above.

4.1 The Data

To test the model, emperical data was used to gauge the performance of the proposed model. Fortunately a snapshot of water quality is taken each year in Puget Sound. Four water quality metrics from 2015's survey were explored for potential use: dissolved oxygen, chloryphyll, salinity, and temperature taken as averages through the water column. A few preliminary diagnostics were used to explore the metrics for potential use (See Figures 8 and 9). Ultimately, chloryphyll (taken as an average through the water column) was used in the analysis.

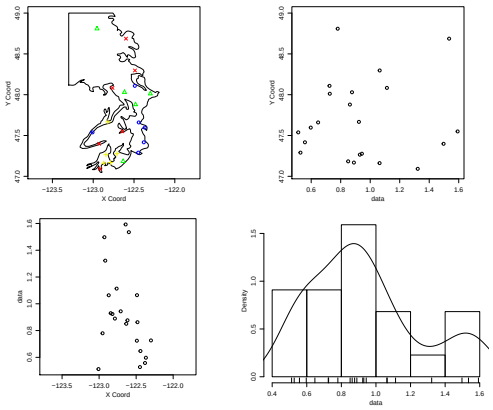

Figure 8: In the upper left panel, chloryphyll sample sites are marked. Sites marked in red, blue, and green denote high, intermediate, and low chloryphyll value respectively. The clustering of colors shows a clear spatial structure.

. .

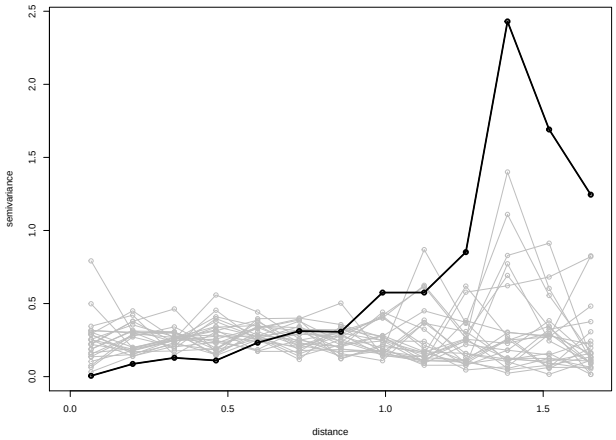

Figure 9: Emperical variograms envelope plot. The grey lines depict emperical variograms resulting from shuffled data while the black line denotes is the emperical variogram from the actual data. The plot possesses a clear spatial trend as the actual emperical variogram looks unlike the rest. However, the plot reveals that temperature may not be the most appropriate metric for model comparison. The maximum variance is found at large distances and there are few samples located this far apart, so it is not clear if the sill has been reached. This can cause imprecise parameter estimates. For this reason, Temperature and salinity were rejected as metrics.

4.2 Overview of Models

Three different types of models were compared to the proposed model in this study. Three ordinary krigers (OK) that use Euclidean distance tested. These models are differentiated by variogram family, and an Exponential, Guassian, and Wave variogram (outlined in Table 2.3) was used in this analysis. Wave variograms are somewhat unconventional. Figure 10 explains why this variogram was included in the analysis. Weighted Least Squares (WLS) were used to fit the three Euclidean models.

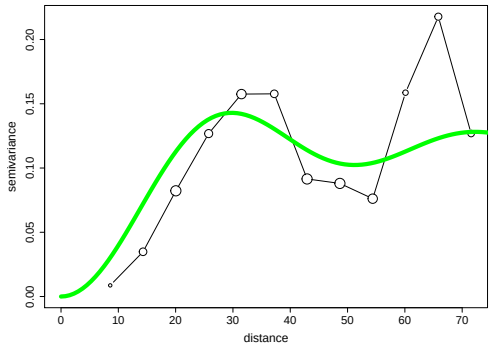

Figure 10: The emperical variogram for chloryphyll is plotted and a wave variogram is overlaid in green. The emperical variogram is bimodal which could be the result from Puget Sounds irregualar topology. Because the wave variogram captures this structure, it was included in the analysis.

The second type of model used in the analysis is an OK that uses (SWD) rather than Euclidean distance. Here, an exponential variogram was used because it passes Curriano's Test for Invalidity (Rathbun, 1998). A non-spatial smoother was also used in the comparison. There are a number of non-spatial smoothers designed for irregularly shaped regions. Among the most popular of these smoothers is the spline-based Soap Film Smoother (Wood, Bravington, & Hedley, 2008). The Soap Film Smoother (SOAP) is relatively easy to implement and has been shown to perform well compared to other similar models. For these reasons, it was included used in the analysis.

4.3 Computation Details

4.3.1 The Proposed Model

The proposed model takes a rather computational approach to kriging as the variogram is defined over a complex lattice. In constructing the lattice, when the size of the mesh is scaled down, the moving average function better approximates a Gaussian kernal. There are, however, computational challenges associated with a fine mesh.

When the mesh is fine, special attention needs to be given to how matrix operations are performed. The expression $P[X_s(k) = u]$ defined in Equations (9) and (10) needs to be evaluated for large values of k. To accomplish this, following (Barry & McIntyre, 2011), $P[X_s(k) = u]$ is re-expressed in matrix form. Let T (called the transitional matrix) be the $n \times n$ matrix of transitional probabilities defined by $T_{ij} = P[X_{s_i}(1) = u_j]$. Additionally, let p_s and p_u be location vectors that contain a one in the i^{th} and j^{th} row respectively and zeros elsewhere. Now, $P[X_s(k) = u] = p_s^T T^k p_u$. When n and k are large, special attention needs to be given to how matrix multiplication is performed. Specifically, multiplying $T^k p_u$ from right to left (i.e. $T(T(T(\cdots(Tp_u))$ will substantially decrease computation times. Additionally, because the matrix T^k contains mostly zeros (at least for small values of k), the use of sparse matrices will increase computing performance.

Finally, a fine mesh can cause underflow issues. To fix the problem, $T^k(ap)$ can be computed instead of $T^k(p)$ where $a > 1$ is a fixed and arbitrary constant. Here, larger values of a coorespond to adding more dye to the diffusion. This change solves underflow issues without changing the fit of the model; however, the parameter θ_s from Equation (11) will be scaled.

(a) (b)

Figure 11: Adjusting mesh size: As a mesh is scaled down, it approximates a region with irregular topologies. When running the models, a mesh that was four times as fine (b) was used (four boxes to every grid-box).

4.3.2 OK Euclidean Models

The *geoR* package was used in the analysis the three Euclidean Ordinary Krigers.

4.3.3 SWD Model

Currently, a package doesn't exist for SWD Kriging. Part of the reason for this is that there have been computational challenges associated calculating SWD distances. Generally, labor intensive manual path tracing techniques are used (Rathbun, 1998). A different approach was taken here. The R package *gdistance* calculates least cost paths. A least cost path can, for instance, be used to determine the cheapest route for a pipeline. By rasterizing the region and assigning a high cost to cells containing land, a least cost path can be used to calculate SWD automatically.

4.3.4 Soap Film Smoother

The SOAP model was implemented using the R package *mcgv*. As mentioned, SOAP is a spline-based technique. The inclusion of islands results in a loss of degrees of freedom. Interestingly, the best fitting SOAP models (as determined by AIC) were obtained with the removal of all islands.

4.4 Results

In this study, four summary metrics were used to guage each model's predictive ability and prediction interval reliability: leave-one-out cross validation was used in calculating mean squared prediction error (MSPE) to measure predictive ability, and the percentages of left-out observations contained in 75%, 85%, and 95% prediction intervals (%75, %85, %95) were used to measure the reliability of the intervals. The results are summarized in the following table:

Model	MSPE	% 75	% 85	% 95
Proposed	0.0998	0.7273	0.7727	0.8636
OK Exponential	0.1602	0.7273	0.7727	0.7727
OK Gaussian	0.1261	0.6364	0.7273	0.7273
OK Wave	0.1192	0.2727	0.3182	0.5910
SWD Exponential	0.1165	0.8182	0.9545	0.9545
SOAP	0.0915	0.7273	0.8182	0.9545

Table 4.4 contains results to the model comparison. Chloryphyll concentration was measured in (ug/l) which explains the small scale found in the MSPE column.

Table 4.4 reveals a few interesting things. For the Euclidean models, a tradeoff between predictive ability and prediction interval reliability exists: when MSPE is higher, the prediction intervals perform better. The relationship between MSPE and and prediction interval performance appears to be more consistant for the non-Euclidean kriging models however. Further, for the OK models, MSPE was lowest for the two models that used a non-Euclidean metric. Additionally, these models outperformed the Euclidean models in terms of prediction interval performance. From this, we can infer that the non-Euclidean distance metrics capture the proximity dependencies within the system better that the Euclidean distance metric.

The SWD and proposed model performed similarly in terms of prediction interval performance. For the 75% prediction interval the SWD and Brownian model were off by 7% and 2% respectively; for the 85% interval, they were off by 10% and 8%; and for the 95% interval, they were off by 0% and 9%. However, in terms of MSPE, the proposed model performed slightly better than the SWD model (the difference in MSPE was 0.0167).

Finally, the non-spatial model (SOAP) performed best in terms of MSPE. The difference in MSPE between SOAP and the proposed model was 0.0083. Additionally, the prediction intervals of the SOAP model performed best in terms of reliability. It is tempting to conclude that the non-spatial prediction intervals produced by the SOAP model are the best. This conclusion would, however, be inaccurate. Geostatistical prediction intervals have a spatial structure which is not shared by non-spatial models (see Figure 1). When used appropriately, geostatistical models reduce the variability of random errors without without sacrificing the reliability of prediction intervals. For this reason and it performed well in terms of predictive ability and prediction interval reliability when compared to the other geostatistical models, the proposed model proposed in this paper is of potential utility.

5 Conclusion

The ultimate aim of this project is to develop an R package for public use. The program's foundations are in place; however, some tidying up needs to be done. Boundaries were hard coded in the analysis, and the program may benefit from a new optimization routine. A hand-coded grid search algorithm was used in the analysis. While this worked (even for leave-one-out cross validation with 22 sites, a non-parallel implementation, and an overly fine mesh), the computation could be spead up through the used of other optimization techniques such as simulated annealing. Finally, the code needs to build and documented for consumer, rather than

personal, use.

Another interesting note is that this is the fourth study that has confirmed the utility of SWD kriging dispite the method's mathematical shortcomings. This leads to a very interesting philosophical question: is it permissable to use a model that has been shown to work in practice when its mathematical validity hasn't been demonstrated? There isn't an easy answer to this question. Now, variance estimates can always be inspected for negativity after a model is fit. It is also possible that the exponential variogram is valid when used with SWD but the result hasn't been shown yet. Additionally, other variograms that are valid with SWD may be developed in the future. For this reason, it is my personal belief that an R program that streamlines SWD kriging could be of potential use. The R package I hope to ultimately develop will have functions for both methods of kriging.

References

Barry, R. P., Jay, M., & Hoef, V. (1996). Blackbox kriging: spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics*, 297–322.

Barry, R. P., & McIntyre, J. (2011). Estimating animal densities and home range in regions with irregular boundaries and holes: A lattice-based alternative to the kernel density estimator. *Ecological Modelling*, *222*(10), 1666–1672.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, *4*(5).

Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, *12*(2), 115–125.

Curriero, F. C. (2006). On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology*, *38*(8), 907–926.

Gneiting, T., et al. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, *19*(4), 1327–1349.

Hanks, E. M. (2015). A constructive spatio-temporal approach to modeling spatial covariance. *arXiv preprint arXiv:1506.03824*.

Little, L. S., Edwards, D., & Porter, D. E. (1997). Kriging in estuaries: as the crow flies, or as the fish swims? *Journal of experimental marine biology and ecology*, *213*(1), 1–11.

Matheron, G. (1962). *Traité de géostatistique appliquée. 1 (1962)* (Vol. 1). Editions Technip.

McIntyre, J., & Barry, R. P. (2018). A lattice-based smoother for regions with irregular boundaries and holes. *Journal of Computational and Graphical Statistics*, *27*(2), 360–367.

Peterson, E. E., Theobald, D. M., & ver Hoef, J. M. (2007). Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater biology*, *52*(2), 267–279.

Rathbun, S. L. (1998). Spatial modelling in irregularly shaped regions: kriging estuaries. *Environmetrics: The official journal of the International Environmetrics Society*, *9*(2), 109–129.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*(sup1), 234–240.

Ver Hoef, J. M., Peterson, E., & Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, *13*(4), 449–464.

Ver Hoef, J. M., & Peterson, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, *105*(489), 6–18.

Wood, S. N., Bravington, M. V., & Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 931–955.

Zou, H., Yue, Y., Li, Q., & Yeh, A. G. (2012). An improved distance metric for the interpolation of link-based traffic

data using kriging: a case study of a large-scale urban road network. *International Journal of Geographical Information Science*, *26*(4), 667–689.