# 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 intervals. Classical kriging models use Euclidean (straight line) distance when modeling spatial autocorrelation. However, for estuaries, inlets, and bays, shortest-in-water distance may capture the system's proximity dependencies better than Euclidean distance when boundary constraints are present. Shortest-in-water distance has been used to krige such regions (Little et al., 1997; Rathbun, 1998); however, the variance-covariance matrices used in these models have not been shown to be mathematically valid. In this paper, a new kriging model is developed for irregularly shaped regions in $\mathbb{R}^2$. This model incorporates the notion of flow connected distance into a valid variance-covariance matrix through the use of a random walk on a lattice, process convolutions, and the non-stationary kriging equations. The model developed in this paper is compared to existing methods of spatial prediction over irregularly shaped regions 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 $\beta$ describe the mean (or fixed) structure of the model, whereas the random errors $\epsilon$ 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 uncorrelated (i.e. $\Sigma = \sigma^2 I$), but with spatial models, this assumption is relaxed so that the spatial dependencies among errors inform $\Sigma$.

Generally, $\Sigma$ is modeled through a data driven approach. However, the stream network model presented in Ver Hoef, Peterson, & Theobald (2006) can be seen to address the question of why near things are more closely related than distant things, which leads to a more mechanistic description of $\Sigma$. 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. Additionally, information related to hydrologic discharge is used to inform the error structure.

Kriging is a method of spatial prediction 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 et al., 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 $\Sigma$ has not been verified in these models though Rathbun (1998) recognizes the issue. 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. 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.

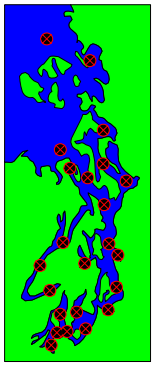
# 2 Background

## 2.1 Spatial Random Processes, Geostatistics, and Kriging

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 $s_1, s_2, \cdots, s_m \in D$ 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. In addition to predictions, kriging produces prediction intervals. By modeling the spatial autocorrelation within a system, the interval estimates produces by these models are spatially structured. That is, the prediction intervals constrict at locations near sampled sites. This is a useful feature, and when used well, kriging models can reduce prediction standard errors without sacrificing the performance of prediction intervals. Later on in section 4, water chlorophyll data collected from Puget Sound (State of Washington Department of Ecology, 2015) is used in comparing models. The sites where water chlorophyll was sampled and a few predictive maps produced by a kriging models are shown in Figure 1 to illustrate the spatial structure of the prediction intervals produced by kriging models.



(a) Sampled Sites


(b) Point Predictions


(c) PI Width


(d) Low 2.5


(e) High 2.5

Figure 1: (a) depicts the 22 sites where water chlorophyll was sampled over a 10 day period from 1/13/15 to 1/22/15; (b), the kriged predictions produced by a kriging model; (c), the width of the 95 % prediction interval; and (d) and (e), the lower and upper bounds of the 95% prediction interval. Code for plots (b) through (e) was supplied by Margarette Short (2018).

## 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 focus of the following discussion. Nevertheless, the model can be easily extended to the case of universal kriging.

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 $i = 1, 2, ..., m$ 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. When formulated in terms of 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. $C(h) = Cov[Z(s_i), Z(s_j)]$ exists and only depends upon $h = s_i - s_j$ for all $s_i$ and $s_j$ 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

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

(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 stationary, a simple relation exists between the variogram and the covariogram. Specifically, $\gamma(h) = C(0) - C(h)$. 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 isotropic if $\gamma(h) = \gamma(d(s_i, s_j))$ for every $s_i$ and $s_j \in D$ where $d(s_i, s_j)$ is the distance between sites $s_i$ and $s_j$ (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 modeling and estimation of $\Sigma$ possible.

After stationarity and isotropy assumptions are made, an estimate of the true variogram is made. This estimate is called the empirical variogram. Usually, the method of moments estimator (Matheron, 1962) or the robust estimator (Cressie & Hawkins, 1980) is used as the empirical variogram. After an empirical variogram has been constructed, a valid parametric family of variogram is chosen by "eyeballing" the empirical variogram. After the variogram family has been chosen, it is assumed to be the true variogram. Variograms are classically parametrized by three parameters

called the nugget, sill, and range. These parameters can then be estimated using a variety of least squares and maximum likelihood based techniques (Cressie, 1992, p. 90). The process of choosing a variogram family and estimating the model parameters is illustrated in Figure 2. Because the topic of variogram validity is of central importance in this paper, the topic is discussed in the next section.

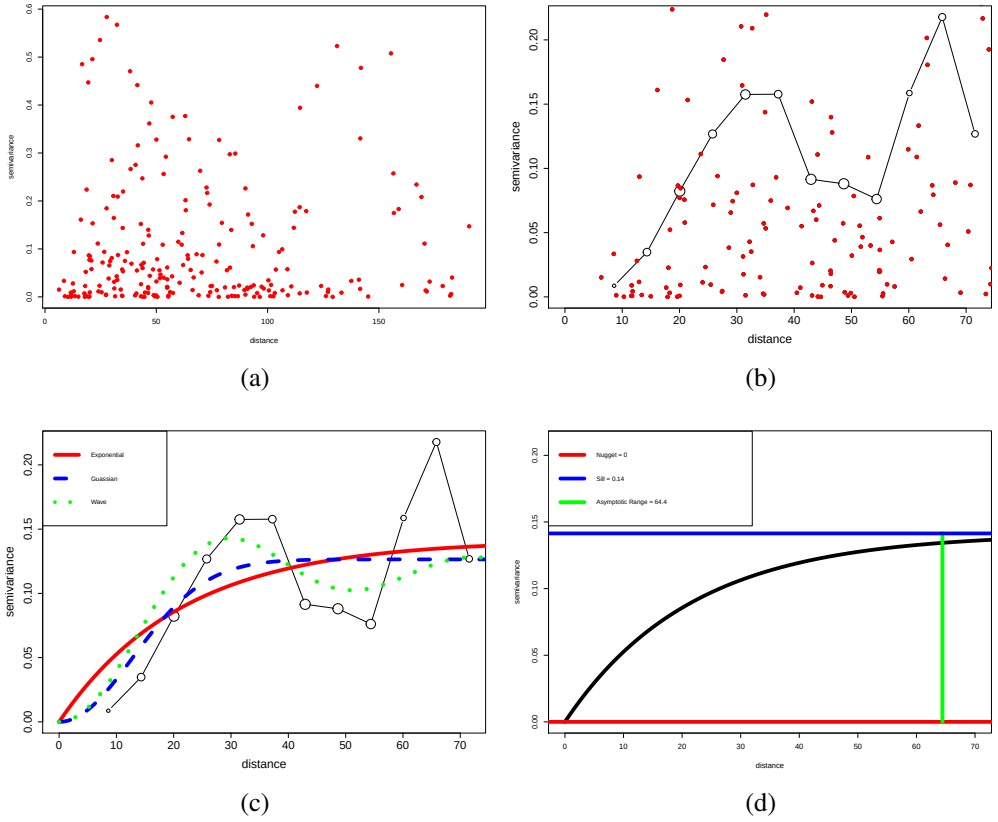



Figure 2: Using chlorophyll data from Puget Sound: (a) plots the variogram cloud (empirical semivariance against distance); (b) approximates the variogram using the robust estimator; (c) considers the exponential, Gaussian, and wave variogram families for potential use; and (d) estimates the model parameters after the true variogram is assumed to belong to the exponential family.

## 2.3 Error Structure Validity

### 2.3.1 Defining Variogram and Covariogram Validity

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

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

for any $s_1, s_2, ..., s_m \in D$ and any $a_1, a_2, ..., a_m \in \mathbb{R}$ where $\sum_{i=1}^{m} a_i = 0$ (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 ($\Sigma$) 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 $\Sigma$ are $4.23$ and $-0.24$. Because the second eigenvalue is negative, $\Sigma$ 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 embarrassing negative variance estimates.

Once a variogram's parameters have been estimated, predictions are then made by calculating the best linear predictor $Z(s_0) = \sum_{i=1}^{m} \lambda_i Z(s_i)$ subject to the constraint $\sum_{i=1}^{m} \lambda_i = 0$ which guarantees uniform unbiasedness of the predictor (Cressie, 1992, p. 120). The coefficients $\lambda = [\lambda_1, \lambda_2, ..., \lambda_m]^T$ 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 $\gamma = [\gamma(s_0 - s_1), \gamma(s_0 - s_2), ..., \gamma(s_0 - s_m)]^T$ and $\Gamma$ is a $m \times m$ matrix where $\Gamma_{ij} = \gamma(s_i - s_j)$. Additionally the minimized mean-squared prediction error $\sigma^2(s_0)$ is given by

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

(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 $Var[Z(s_i) - Z(s_j)]$ be a function of $s_i - s_j$. 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, $\gamma$ is re-defined as $\gamma = [\gamma(s_0, s_1), \gamma(s_0, s_2), ..., \gamma(s_0, s_m)]^T$ and $\Gamma$ 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 $D$.

### 2.3.2 Non-Euclidean Distance Metrics

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 $s_i = [x_i, y_i]^T$ and $s_j = [x_j, y_j]^T$ is given by $||s_i - s_j|| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. 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 straightforward 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 illustrated 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. When a Gaussian variogram's nugget, sill, and range parameters are 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 $\Sigma$ 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.

### 2.3.3 SWD

For inlets, estuaries, and bays, hydrologic connectivity can be seen to explain the presence of spatial autocorrelation. When complicated boundary constraints are present, SWD may capture the system's proximity dependencies better than Euclidean distance. In aquatic systems characterized by branching topologies such as Puget Sound, SWD is a better proxy than Euclidean distance for hydrologic connectivity (see Figure 3). 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 be noted, however, that the varigrams used in Rathbun (1998) pass Curriano's Test For Invalidity. Even so, the variograms are not guaranteed to be valid when used with SWD.

(a) Euclidean Distance (b) SWD

Figure 3: In (a), the Euclidean distance between two sites doesn't relate to the notion of hydrologic connectivity. In (b), the SWD between the sites relates to the notion of hydrologic connectivity quite well.

### 2.3.4 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 referred 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 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 $s \in D$ and where $W(x)$ is a white noise random process. The resulting variogram is then given by

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

provided that $2\gamma(h|\theta) < \infty$. An important result is that the variogram defined in Equation (4) is guaranteed to be valid Barry et al. (1996). When the process defined in Equation (3) is second-order stationary, a covariogram is given by

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

Ver Hoef et al. (2006). Similarly, $C(h)$ is guaranteed to be valid.

When 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

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

The condition expressed in Equation (4) (i.e. $2\gamma(h) < \infty$) is satisfied and $2\gamma(h)$ is a valid variogram. Because Equation (4) is critical to this paper it is illustrated further through a pictorial representation of Equation (6) in Figure 4.

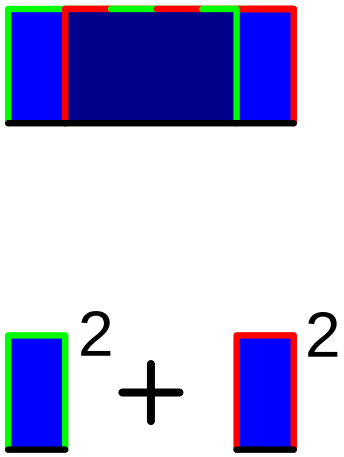



Figure 4: Here, $g(x)$ and $g(x-h)$ are plotted. $g(x)$ is outlined in green while $g(x-h)$ is outlined in red. To evaluate the variogram at a lag of $h$, the overlapping area (shaded in dark blue) is removed, the remaining area is squared and added together.

Because process convolutions guarantee variogram validity, the technique can be extremely useful in developing valid variance/covariance structures over non-Euclidean distance metrics. Ver Hoef et al. (2006), Peterson et al. (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.

The moving average function $g(\cdot)$ can be generalized to construct a non-stationary version of the variogram. 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

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

The function $g(\cdot, \cdot|\theta)$ will be referred to as a generalized moving average function. The variogram defined by $g(\cdot, \cdot|\theta)$ is not necessarily stationary. However, the construction defined in Equation (7) guarantees the validity of the resulting

variogram. Here, $g(\cdot, \cdot|\theta)$ can be seen as defining a regularity assumption to allow for parameter estimation. In the next section, a generalized moving average function is used to incorporate the notion of hydrologic connectivity into a valid error structure.

# 3 The Proposed Model

## 3.1 Random Walks to Approximate Diffusions

Random walks on lattices have been used to model spatial autocorrelation in intrinsic simultaneous autoregressive (SAR) models (Hanks, 2015). Additionally, random walks on a lattice can be used to approximate Brownian motion (or diffusion). This idea has been used to develop new solutions to problems with irregular boundary constraints (Barry & McIntyre, 2011; McIntyre & Barry, 2018). Here, the basic idea is analagous 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 after a region's topology. In the model proposed in this section a random walk constructed over a lattice is used to define a boundary respecting moving average kernal.

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 indices. Additionally, $n_i$ is chosen to represent the $i^{th}$ node; $B_i$, to represent the set of all locations in the $i^{th}$ grid-box; and $s^{(i)}$ and $u^{(i)}$, to represent arbitrary locations in $B_i$. If the context does not require a subscript or superscript, 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. In Figure 5, a lattice is constructed over Puget Sound.

To construct the random walk, a Markov chain is defined. 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 position of the random walk after $k \geq 0$ steps when originating from site $s$. A Markov chain is defined by

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

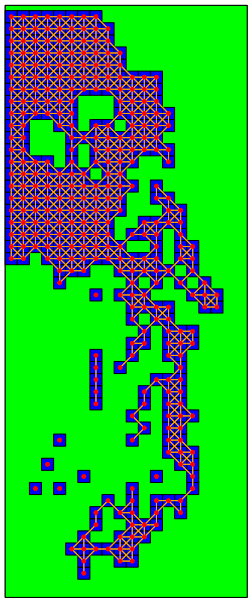


Figure 5: Neighbor relationships are defined over a mesh inscribed in the region

for all $i$ and $j \in I$ where $q_i$ is node $n_i$'s number of neighbors, and $0 < \theta_d < 1$. Here, the parameter $\theta_d$ governs the rate of diffusion. When $\theta_d$ decreases, the diffusion rate increases. Here, the transitional probability between adjacent neighbors, $\theta_d/8$, is constant which cooresponds to isotrophic diffusion. In the implementation taken in this paper, $\theta_d$ 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 a random walk can get from one node to any other through the neighboring relationships, the transitional probabilities will eventually become uniform (Rosenblatt, 2012). A diffusion is illustrated in Figure 6 as Equations (9) and (10) can be opaque.

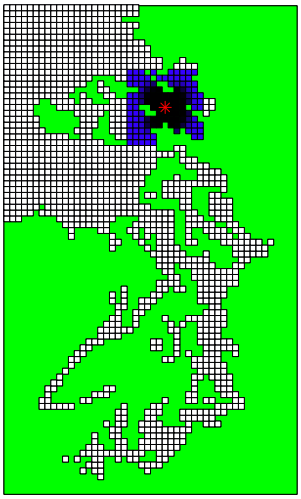


Figure 6: A diffusion of 7 steps originating from the point marked in red is pictured. Note: the plot is fat as a latitudinal correction hadn't yet been made at the time of plot construction.

## 3.2 Defining the Moving Average Function

To incorporate a flow connected proximity dependence into the model, a diffusion of parametrized length $\kappa$ is used as the kernel of the moving average function. When $D$ has boundary constraints and a diffusion of length $\kappa$ is used as the kernel of the moving average function, the random process defined by the construction is non-stationary. For this reason, a generalized moving average is used. To define this function formally, let

$$g(s, u|\theta) = \theta_s P[X_s(\kappa) = u]. \tag{11}$$

Here, the parameter $\theta_s$ 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 incorporate a nugget effect. The nugget effect represents small scale variability that is commonly found in mining applications. 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 $g(\cdot, \cdot|\theta)$ defined, the equivalent construction to Equation (8) is given by

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

and the corresponding covariogram is given by

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

In interpreting Equations (12) and (13) it will be useful to refer to Figures 4 and 6. Two diffusions of length $\kappa$ 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 $\gamma(s,u)$ need to be estimated. Different parameter estimation techniques could be used including least squared based methods. In this paper, residual maximum likelihood (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 structure is accommodated 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

### 4.1 The Data

The proposed model was compared to existing methods of spatial prediction using water chlorophyll data collected from Puget Sound (State of Washington Department of Ecology, 2015). The data and R code used in Section 4 can be found at `https://github.com/jordyBernard/mastersProject`. Four water quality metrics from 2015's survey were explored for potential use: dissolved oxygen, chlorophyll, salinity, and temperature taken as averages through the water column. A few preliminary diagnostics were used to explore the metrics for potential use. Ultimately, the chlorophyll data was used in the analysis because the data showed a clear spatial structure, a potential boundary influence, and the sample sites were separated far enough apart to get good parameter estimates.

### 4.2 Overview of Models

Three different types of models were compared to the proposed model in this study. Three ordinary kriging models (OK) that use Euclidean distance were tested. These models are differentiated by variogram family, and Exponential, Guassian, and Wave variograms (outlined in Table 4.2) are used in the analysis. Weighted Least Squares (WLS) were used to fit the three Euclidean models. 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-geostatistical model was also used in the comparison. There are a number of spatial smoothers designed for irregularly shaped regions. Among the most popular of these smoothers is the non-parametric spline-based Soap Film Smoother (Wood et al., 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.

| 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 4.2 contains a short list of some 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 $\tau^2$ is called the nugget, $\sigma^2$ is the partial sill, and $\phi$ is the range parameter (see Figure 2).

## 4.3 Computation Details

### 4.3.1 The Proposed Model

When constructing the lattice, the moving average kernal will better resemble a diffusion when the mesh size is scaled down. Additionally, the lattice better resembles the geographic area when the mesh is shrunk. There are, however, computational challenges associated with a fine mesh. To start, special attention needs to be given to how matrix operations are performed when the mesh is fine. 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 transition matrix) be the $m \times m$ matrix of transition 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. Additionally, 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 $a$ correspond to adding more dye to the diffusion. This change solves underflow issues without changing the fit of the model; however, the parameter $\theta_s$ from Equation (11) will be scaled.

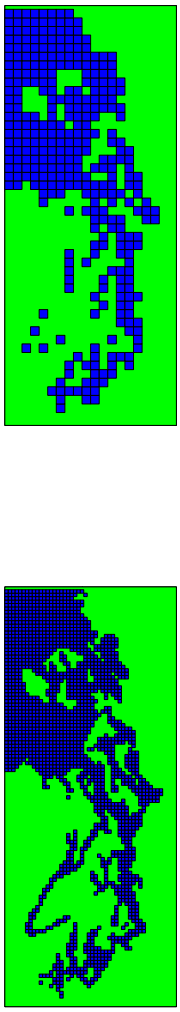


Figure 7: As the mesh size is scaled down, the lattice better represents the spatial region. When running the models, a mesh four times as the bottom plot was used (four boxes to every grid-box).

### 4.3.2 OK Euclidean, SWD, and SOAP Models

The three Euclidean OK models were analyzed using the R package *geoR* (Ribeiro Jr & Diggle, 2018). Additionally, the SOAP model was implemented using the *mcgv* package (Wood, 2011). A package hasn't been developed for SWD Kriging. Part of the reason for this is that there have been computational challenges associated with 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.

## 4.4 Results

In this study, four summary metrics were used to gauge 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. Chlorophyll 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, when MSPE is higher, the prediction intervals perform better. The relationship between MSPE and prediction interval performance is more consistent 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.

# 5 Future Work

The SOAP model performed best in the model comparison. However, there may be instances when a geostatistical model is preferred over the SOAP model. Geostatistical models give insight into a process's correlation structure. To understand the benefits that can be associated with a geostatistical parameterization, consider, for example, the range parameter. In the model developed in this paper $\kappa$ is seen as the range parameter. When a range parameter is estimated, it can be used to understand the distance or proximity relationship at which sites can be treated as independent. This is useful information and can be used to inform sampling practices.

The model developed in this paper may have practical utility, and an R package may be developed in the future. The package's foundations are in place; however, some tidying up will need to be done. Boundaries were hard coded in the analysis, and the program may benefit from a new optimization routine. In estimating the model parameters, a hand-coded grid search algorithm was used. 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 sped up through the use of other optimization techniques such as simulated annealing. Finally, the code needs to be build and documented for consumer, rather than personal, use.

Another interesting note is that this is the third study that has confirmed the utility of SWD kriging despite the method's mathematical shortcomings. This leads to an interesting philosophical question: is it permissible 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. It is also possible that the exponential variogram is valid when used with SWD but the result hasn't been shown yet. Additionally, variance estimates can always be inspected for negativity after a model is fit. For these reasons, a program that automates SWD kriging could also be of potential use. A function may be included in the ultimate R package to streamline SWD 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.

Margarette Short. (2018).

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.

Ribeiro Jr, P. J., & Diggle, P. J. (2018). geor: Analysis of geostatistical data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=geoR` (R package version 1.7-5.2.1)

Rosenblatt, M. (2012). *Markov processes, structure and asymptotic behavior: Structure and asymptotic behavior* (Vol. 184). Springer Science & Business Media.

State of Washington Department of Ecology. (2015). *Marine long-term monitoring survey [data file].* (`https://fortress.wa.gov/ecy/eap/marinewq/mwdataset.asp`)

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. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

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.