HHS Public Access

Author manuscript

Math Geosci. Author manuscript; available in PMC 2020 August 01.

Published in final edited form as:

Math Geosci. 2019 August ; 51(6): 767–791. doi:10.1007/s11004-019-09791-y.

Development and Evaluation of Geostatistical Methods for Non-Euclidean-Based Spatial Covariance Matrices

Benjamin J. K. Davis[1,2], **Frank C. Curriero**[1,2]

[1.]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

[2.]Spatial Science for Public Health Center, Johns Hopkins University, Baltimore, MD 21205, USA

Abstract

Customary and routine practice of geostatistical modeling assumes that inter-point distances are a Euclidean metric (i.e., as the crow flies) when characterizing spatial variation. There are many real-world settings, however, in which the use of a non-Euclidean distance is more appropriate, for example in complex bodies of water. However, if such a distance is used with current semivariogram functions, the resulting spatial covariance matrices are no longer guaranteed to be positive-definite. Previous attempts to address this issue for geostatistical prediction (i.e., kriging) models transform the non-Euclidean space into a Euclidean metric, such as through multi-dimensional scaling (MDS). However, these attempts estimate spatial covariances only after distances are scaled. An alternative method is proposed to re-estimate a spatial covariance structure originally based on a non-Euclidean distance metric to ensure validity. This method is compared to the standard use of Euclidean distance, as well as a previously utilized MDS method. All methods are evaluated using cross-validation assessments on both simulated and real-world experiments. Results show a high level of bias in prediction variance for the previously developed MDS method that has not been highlighted previously. Conversely, the proposed method offers a preferred tradeoff between prediction accuracy and prediction variance and at times outperforms the existing methods for both sets of metrics. Overall results indicate that this proposed method can provide improved geostatistical predictions while ensuring valid results when the use of non-Euclidean distances is warranted.

Keywords

Geostatistics; Kriging; Non-Euclidean distances; Positive-definite covariance matrices; Multidimensional scaling; Water salinity

Corresponding Author: Benjamin J.K. Davis: Bdavis64@jhmi.edu, +01.410.502.0143.
B.J.K.D.: 627 N. Washington Street, Room 2-A, Baltimore, MD 21205.
F.C.C.: 615 N. Wolfe street, Room E6541, Baltimore, MD 21205.

1 Introduction

Spatial prediction (i.e., spatial interpolation) encompasses a suite of techniques that use information at sampled spatial locations to predict at surrounding unsampled locations. Often predictions are generated at numerous locations so that results can be presented as a smoothed spatial prediction surface on a map. Spatial prediction has been used in a number of environmental and public health endeavors across soil, air, and water environments (Berman et al. 2015; Hengl et al. 2004; Henshaw et al. 2004; Jeffrey et al. 2001; Laaha et al. 2014). Kriging is a statistical regression-based method of spatial prediction developed in the field of geostatistics (Cressie 1993). Kriging involves estimating a parameterized, spatial covariance structure which is used to produce best linear unbiased predictions at unsampled locations. Kriged predictions are therefore weighted averages of sampled data with more weight generally given to data closer to the prediction location.

How closeness or proximity of locations can and should be measured is often an unaddressed issue. Kriging methods were developed with an underlying assumption that inter-point distances are a straight-line Euclidean metric (i.e., as the crow flies). This assumption ensures that the corresponding spatial covariance matrix for the variable of interest is positive-definite (Cressie 1993). However, real-world applications are common where a non-Euclidean measure is superior. Examples include road-network distances, veins and channels in geological deposits, ventilation systems in occupational settings, and air transport in metropolitan areas. Of interest in this paper is inter-point distances between monitoring stations in complex (e.g., non-convex) bodies of water. Spatial prediction in these settings could be applied to water quality measures that are necessary to determine ecosystem health and to develop environmental predictions of infectious microorganisms that reside in such waters (Davis et al. 2017; Murphy et al. 2015).

When a body of water is a non-convex polygon, Euclidean distances between locations may cross water-land boundaries making such a proximity metric less appropriate when compared to a non-Euclidean water distance (Rathbun 1998). This is prominently demonstrated in coastal estuaries such as the Chesapeake Bay in the Mid-Atlantic region of the United States, where winding tributaries make up a large portion of the surface area (Fig. 1). While non-Euclidean water distances may be more appropriate for spatial prediction, using them with existing geostatistical semivariogram/covariogram functions results in the positive-definite characteristic of the corresponding spatial covariance matrix no longer being ensured (Curriero 2006). Failing to recognize this fact when kriging with non-Euclidean distances can result in negative prediction variances, a clear indication that the underlying statistical theory is no longer valid; a situation that has sometimes been overlooked in previous studies (Gardner et al. 2003; Little et al. 1997; Rathbun 1998).

There have been several efforts to make the use of non-Euclidean distances valid for spatial prediction via kriging (Boisvert and Deutsch 2011; Del Castillo et al. 2015; Løland and Host 2003; Lu et al. 2011; Lu et al. 2014; Murphy et al. 2015; Sampson and Guttorp 1992; Ver Hoef 2018). However, most of these attempts choose to focus on transforming either the geographic area or the distance structure, and subsequently estimating spatial dependence from variogram functions. Ideally the general covariance structure estimated from a non-

Euclidean distance could be preserved, and instead relatively straight forward post-hoc corrections could be applied to ensure a positive-definite covariance function.

This paper describes such a method that maintains the overall covariance structure formed from a non-Euclidean distance for use in kriging models while ensuring that covariance matrices remain positive-definite. The method allows the spatial variation in the variable of interest to be first characterized using the suggested non-Euclidean distance followed by a procedure that ensures resulting covariance structures are positive-definite. After providing a brief overview of geostatistical nomenclature (Sect. 2.1), an existing method using multi-dimensional scaling (MDS) to approximate non-Euclidean distance as a Euclidean metric and the proposed method are defined (Sect. 2.2). Performance evaluation techniques to compare the two methods to the standard approach of using a Euclidean distance metric are described in Sect. 2.3, while Sect. 2.4 details a number of experimental data settings developed to apply the evaluation techniques. Particulars of geostatistical analyses and distance enumeration are described in Sect. 2.5. Prediction performance evaluation in these experiments is reported in Sect. 3, while Sect. 4 discusses the evaluation findings. Computational limitations to the non-Euclidean methods, and potential solutions, are discussed in Sect. 5. Finally, a brief conclusion is presented in Sect. 6.

2 Methods

2.1 Distance-Based Spatial Covariance Structure

The general framework for geostatistical prediction, known as universal kriging, can be written as a linear regression model

$$Y(s) = \beta_0 + \beta_1 X_1(s) + \cdots + \beta_p X_p(s) + \epsilon(s); \quad \epsilon(s) \sim N(0, \Sigma), \tag{1}$$

where s represents a spatial location, $Y(s)$ is the (possibly transformed) outcome variable of interest at location s, $X_1(s), \ldots, X_p(s)$ are covariates indexed by location s, and $\beta_0, \ldots, \beta_p$ their associated parameters. The residual error term $\epsilon(s)$ has a covariance/correlation structure Σ to represent residual spatial variation: spatial dependence in outcome Y not accounted for by the included covariates. When Eq. (1) is used for spatial prediction, Σ is of primary focus and routinely modeled using a distance-based parametrized covariogram/semivariogram function, chosen from a set of known valid positive-definite/conditionally negative-definite functions and estimated based on the data (Cressie 1993). These functions are frequently parameterized via the range, ϕ, partial sill, σ^2, and nugget, τ^2. For example, the exponential semivariogram function can be written as

$$\begin{aligned} \gamma(h;\theta) &= \tau^2 + \sigma^2\left\{1 - \exp\left(-\frac{\|h\|}{\phi}\right)\right\} \\ \theta &= \left(\tau^2, \sigma^2, \phi\right), \tau^2 \geq 0, \sigma^2 \geq 0, \phi \geq 0, \end{aligned} \tag{2}$$

where γ is the semivariance, and $||h||$ corresponds to Euclidean distance. Note the representations and proposed methods that follow could apply similarly to kriging with non-

Gaussian data distributions such as those based on the binomial and Poisson distributions (Diggle and Ribeiro 2007).

Let d_{ij} represent the non-Euclidean distance (e.g., water distance) between any two locations s_i and S_j with $[d_{ij}]$ denoting the matrix of all inter-point distances and $[\| s_i - s_j \|]$ the corresponding matrix of Euclidean inter-point distances. Use $C(\cdot)$ and $\gamma(\cdot)$ to denote the pool of commonly used parametric isotropic covariogram and semivariogram functions (e.g., the exponential and Gaussian functions) for characterizing spatial variation with stationarity assumptions so that $C(\cdot) = C(0) - \gamma(\cdot)$ (Cressie 1993). These functions were developed to be valid when used with a Euclidean distance that is positive-definite and conditionally negative definite for covariogram and semivariogram functions respectively. However, as previously demonstrated by Curriero (2006), there are no guarantees that the functions $C([d_{ij}])$ will result in a valid positive-definite covariance matrix. The methodology proposed here is to develop a valid adequate approximation to $C([d_{ij}])$.

2.2 Proposed Method

A previous approach based on MDS approximation seeks to find a Euclidean distance approximation to the set of non-Euclidean distances, $\left[\mathrm{d_{ij}}\right] \approx \left[\left\|s_i^* - s_j^*\right\|\right]$, where transformed locations s^* are usually in a much higher dimension up to a maximum of $k = N - 1$, where N is the number of dimensions of $[\mathrm{d_{ij}}]$ (Boisvert and Deutsch 2011; Loland and Host 2003; Murphy et al. 2015). This occurs by approximately embedding the matrix of non-Euclidean distances into a k-dimensional Euclidean space derived via eigen decomposition (Mardia et al. 1979). Kriging proceeds globally across the system of equations (i.e., between all locations of interest) using $\mathrm{C}\left(\left\|s_i^* - s_j^*\right\|\right)$, which is valid since it is based on a Euclidean distance. In other words, this MDS approach finds a new transformed set of coordinates, s^*, so that the Euclidean distances between this new set of coordinates approximate the original set of non-Euclidean distances. A potential limitation of this MDS approach, referred to hereafter as the MDSdist method, is that it only considers the location information (via the non-Euclidean distance matrix). It does not explicitly consider spatial dependence of the outcome data, $Y(\cdot)$, via estimates of $C(\cdot)$ or $\gamma(\cdot)$. Even though MDSdist then estimates the covariogram/semivariogram based upon the scaled distances, the results from $C(\cdot)$ and $\gamma(\cdot)$ will be different than if based upon the non-Euclidean distances.

To address this potential limitation, an alternative approach proposes to first estimate the spatial covariance between all locations of interest based on the non-Euclidean distances, $C([d_{ij}])$. followed by an approximation to ensure positive-definiteness. The estimated covariance can be obtained by proceeding as though the non-Euclidean distances d_{ij} are valid to use, for example, by estimating the semivariogram, using weighted least squares to fit a semivariogram function and transforming that function to its corresponding covariogram. Note this last step needs to be performed not only for the non-Euclidean distances between the sample locations but also between the non-Euclidean distances from all locations for which spatial prediction is sought, hence building the full covariance structure of the observed and prediction locations. A similar process is carried out when applying the MDSdist method.

If the non-Euclidean-based covariance matrix $C([d_{ij}])$ is not positive-definite, it is then approximated with a close positive-definite matrix across the system of equations and as such hereafter is referred to as the ClosePD method. The matrix undergoes eigen decomposition

$$C\left(\left[d_{ij}\right]\right) = V\Lambda V', \tag{3}$$

where V is a square matrix whose i^{th} column is the eigenvector v_i, and Λ is the diagonal matrix of eigenvalues, $\Lambda_{j,j} = \lambda_j$, and are ordered in a decreasing fashion $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, where k is the dimension of V. In a positive-definite matrix all eigenvalues are positive, $\lambda_1, \lambda_2, \ldots, \lambda_k > 0$. Therefore, a threshold, ε is chosen

$$\varepsilon = \frac{\lambda_1}{\tau}, \tag{4}$$

where τ is a predetermined tolerance value. The threshold is applied to all eigenvalues as follows

$$\begin{cases} \lambda_i, & \text{if } \lambda_i \geq \varepsilon \\ \varepsilon, & \text{if } \lambda_i < \varepsilon \end{cases}, \tag{5}$$

and a new eigenvalue matrix, $\tilde{\Lambda}_{j,j} = \tilde{\lambda}_j$ is used to rescale the covariance matrix

$$\tilde{C}\left(\left[d_{ij}\right]\right) = V\tilde{\Lambda}V', \tag{6}$$

thus ensuring that the covariance matrix will be positive-definite (Cheng and Higham 1998; Lucas 2001). Prediction can then proceed based on the covariance formulation of the kriging system of equations.

2.3 Performance Evaluation

Leave-one-out cross-validation (LOOCV) is performed across a variety of simulated and real data experiments to compare spatial prediction performance of the proposed method with the original MDSdist approach and the standard use of Euclidean distance. Several cross-validation metrics are used to assess prediction performance and to evaluate comparisons across methods. These included the cross-validation R^2 (CV-R^2) statistic, mean error (ME), root mean squared error (RMSE), mean prediction standard error (MPSE), root mean squared standardized error (RMSSE), and the proportion of predictions that fell within the 95% prediction interval (PI95) (Congdon and Martin 2007; Cressie 1993; ESRI 2016). The CV-R^2 statistic provides an overall measure of the goodness of fit. A value of 1.0 for this statistic indicates a perfect fit, while a value below 0.0 signifies a model that performs worse than if one were to apply the mean value of the dataset (Berman et al. 2015). ME provides information on prediction bias; a negative value signifies systematic overestimation while a positive value signifies underestimation. RMSE is used to evaluate overall prediction error across methods with larger values indicating higher error. MPSE indicates the average prediction variance, while RMSSE can be used to evaluate the accuracy of prediction

variance: a value of 1.0 indicates errors are completely accurate, a value greater than 1.0 indicates the model is underestimating prediction variance, and a value less than 1.0 indicates an overestimation (Cressie 1993; ESRI 2016). Analogous results for RMSSE are observed for the PI95 statistic such that accurate prediction variance estimation would result in a value of 0.95, underestimation would result in values less than 0.95, and overestimation would result in values asymptotically approaching 1.00. These metrics are further defined in "Appendix A".

2.4 Experimental Methodology

In order to assess the performance of the proposed method as compared to existing methods, LOOCV is used in three distinct data experiments in the Chesapeake Bay (Fig. 2). First, data are simulated using non-Euclidean spatial covariance structures. Second, data are simulated using regression coefficients in piecewise linear regressions. Finally, real-world data are extracted from an open-source water quality database. Each experiment is discussed in more detail below. Note that for all experiments, spatial covariance estimation and subsequent prediction during LOOCV is performed using an ordinary kriging model

$$Y(s) = \beta_0 + \epsilon(s); \quad \epsilon(s) \sim N(0, \Sigma)\,. \tag{7}$$

This model requires that all variation of $Y(s)$ be estimated in the covariance matrix, Σ. Ordinary kriging is commonly used in the field of geostatistics when potential covariates are unknown or otherwise unavailable.

2.4.1 Non-Euclidean Spatial Covariance Simulations—In this experiment data are directly simulated using non-Euclidean spatial covariance structures. Unconditional Gaussian random field simulations are traditionally applied with fixed values for the range, sill, and nugget parameters (e.g., Eq. (2)), and are based upon Euclidean distances to ensure resulting covariance matrices are positive-definite. To test the proposed methodology, calculated inter-point water distances across the Chesapeake are used to populate the distance matrix, which is then used by an unconditional Gaussian random field to simulate data throughout the bay's spatial domain. Multiple semivariance functions (e.g., exponential, Gaussian, etc.) are considered in combination with several different fixed values for the range and sill parameters; the nugget parameter in this experiment is always equal to zero. This results in a large number of simulated datasets; a representative example is shown in Fig. 3. LOOCV is performed using 400 randomly selected points throughout each simulated random field.

2.4.2 Piecewise Regression Simulations—In this experiment data are simulated using large-scale regression coefficients. Specifically, a piecewise regression is used to simulate data that are a function of water distance from a pre-specified origin in the Chesapeake Bay's spatial domain. The piecewise regression introduces a statistical artifact so that at relatively short and long distances, no spatial trend is observed. This modification is necessary to develop an inter-point distance range at which data are no longer spatially dependent in the domain, an important consideration for estimating valid Matém

semivariogram functions (Matheron 1971). This piecewise regression model can be written as

$$\begin{aligned} Y(s) &= \beta_0 + \beta_1 X_1(s) + \epsilon(s); \quad \epsilon(s) \sim N\left(0, \sigma^2\right) \quad for\, a < X_1 < b \\ Y(s) &= \beta_0 + \epsilon(s); \quad \epsilon(s) \sim N\left(0, \sigma^2\right) \quad for\, X_1 < a \;\&\; X_1 > b, \end{aligned} \tag{8}$$

where X_1 is water distance measured from a starting location, a is a predetermined minimum water distance, above which the linear function of water distance, β_1 is simulated, and b is a predetermined maximum water distance. Note that Eq. (8), substantially differs from the universal kriging model (i.e., Eq. (1)) by assuming residual independence. Several piecewise regressions are developed, considering a number of different values for β_1. a, and b. Two starting locations are utilized: one at the mouth of the bay in Virginia and another at the head of the Potomac River tidal waters in the District of Columbia (Fig. 2). A representative example of data simulated from one of these piecewise regressions is displayed in Fig. 4. Note the clear lack of spatial variation in this figure at relatively near and far distances from the mouth of the bay. This simulated data setting is more robust to the experiment described in Sect. 2.4.1 as the non-Euclidean spatial covariance of the simulated data are not guaranteed to be embeddable when estimated using Eq. (7). Therefore, the proposed method ClosePD should be considered particularly relevant during the performance evaluation for this experiment. LOOCV is performed using 400 randomly selected points throughout each dataset simulated by piecewise regression.

2.4.3 Real-World Data—The final experiment uses water quality measurements from the Chesapeake Bay to compare the proposed geostatistical methods. Surface water quality data originates from a previously analyzed dataset on the bacterium *Vibrio parahaemolyticus* in the Chesapeake Bay (Davis et al. 2017). Data include salinity measurements from July 2009 at 148 monitoring stations (Fig. 5), a subset of the stations used in Murphy et al. (2015) to evaluate the performance of MDSdist on more historical Chesapeake Bay water quality data. The time range of the data is purposely selected to ensure that only one measure is available at each station and as few of the stations as possible are missing measurements. LOOCV is again applied for evaluation. Methods are also applied to the full spatial domain of the Chesapeake Bay to create maps of kriged predictions and variances.

2.5 Implementation Details

All non-Euclidean geostatistical methods are implemented in R statistical software v.3.5.1 by using altered functions from the geoR package (R Core Team 2016; Ribeiro and Diggle 2016) for semivariogram estimation, kriging and other related analyses. The posdefify function from package sfsmisc is used to perform the ClosePD method (i.e., Eqs. (3), (4), (5), (6)) (Maechler 2016). A number of additional software packages are used for simulation, prediction performance evaluation analyses, and for visualizing results, including: ggplot2, splancs, RandomFields, bigmemory, el071, matrixcalc, and rgdal (Bivand et al. 2016; Kane et al. 2013; Meyer et al. 2015; Novomestky 2012; Rowlingson and Diggle 2015; Schlather et al. 2015; Wickham 2009).

A two-dimensional surface water distance matrix for the Chesapeake Bay is populated with calculations of the shortest distance between any two given points that do not cross the geographic boundaries of the bay's tidal waters. The polygon shapefile of the bay's tidal waters (Fig. 2), available from the Chesapeake Bay Program (Chesapeake Bay Program 2017), is appended to include portions of the Manokin, Transquaking, and Chicamacomico Rivers, obtained from the United States Geological Survey's National Hydrography Dataset (USGS 2016). The appended spatial polygon is dissolved, simplified and then rasterized into pixels of dimension 1kmx 1km using the ArcGIS software v.10.3 and corrected using the open-source ArcMap Raster Edit Suite (ESRI 2011; Yu et al. 2015). The raster layer includes 11,112 pixels at which to predict across the Chesapeake Bay's spatial domain. The raster layer is brought into R and non-Euclidean cost-based distances are calculated using the gdistance package (Etten 2015). The applied cost function makes it impossible for water-distance paths to cross the boundary of the Chesapeake Bay tidal waters (Figs. 6 and 7).

3 Results

Note that when the methods MDSdist and ClosePD are discussed they are only referenced when using non-Euclidean distances. As shown previously (Murphy et al. 2015), the MDSdist method partially embeds non-Euclidean water distances in the Chesapeake Bay, requiring 214 (53.6%) of the N-1 potential eigenvalues for the simulated experiments (Sects. 2.4.1 & 2.4.2), 90 (60.8%) eigenvalues for the water quality sample data LOOCV experiment (Sect. 2.4.3), and 5,730 (50.9%) eigenvalues when kriging water quality at all locations in the Chesapeake Bay. Approximated scaled distances systematically overestimate the original non-Euclidean distances (RMSE=8.92 km; Fig. 8a). The ClosePD tolerances, τ from Eq. (4), are set to 1×10^{3}, 1×10^{2}. and 1×10^{4} for the simulation experiments, real-world data LOOCV, and kriging real-world data at all unsampled locations respectively. For all data experiments, semivariogram estimation using different distance structures consistently produce distinct characterizations of spatial variation (Fig. 9). Comparisons of estimated covariances reveal poor approximation of the original non-Euclidean spatial dependence when using the MDSdist method, with general overestimation of covariances across the range of distances (Fig. 8b). Similar comparisons show a superior approximation of non-Euclidean covariances when using the ClosePD method (Fig. 8c), although a slight underestimation occurs throughout, particularly when covariances are high (i.e., when interpoint distances are small).

For all non-Euclidean spatial covariance simulations (Sect. 2.4.1), the resulting covariance based on the non-Euclidean water distance yields positive-definite matrices. Hence application of the ClosePD method is no different than naïve use of the non-Euclidean water distance in this experiment. Cross-validation evaluation of this experiment across a representative sample of varying partial sill and range coefficients are summarized in Table 1. For example, range coefficients listed represent effective ranges at 50% (200.1 km), 33.3% (133.5 km) and 20% (80.1 km) of the maximum calculated water distance in the Chesapeake Bay. A step-wise improvement of RMSE is observed across covariance simulations in the following order: standard use of Euclidean distance, MDSdist, and use of non-Euclidean water distance. Further exploration of these simulations shows that when the range parameter of the semivariogram is smaller, the spread of prediction error results across

the methods increases, indicating that the use of non-Euclidean water distance is more optimal when the extent of spatial dependence is restricted. This is somewhat intuitive given that at large distances in the Chesapeake Bay's spatial domain, the potential discrepancy between Euclidean and non-Euclidean distances becomes less distinct, even when considering locations in tidal tributaries.

Representative results from piecewise regression simulated data (Sect. 2.4.2) are summarized in Table 2. In this experiment, the standard use of Euclidean distance outperforms methods using water distance in terms of overall prediction error (CV-R^2 and RMSE). However, the standard use of Euclidean distance results in substantial underestimation of prediction variance (RMSSE = 1.37). MDSdist, while slightly outperforming the ClosePD method in terms of overall prediction error, results in a large underestimation of prediction variance (RMSSE=4.70 and MPSE=29.86). ClosePD has the most accurate estimates of prediction variance, although its relatively poor prediction performance in this setting should not be ignored (CV-R^2 = 0.68). No systematic bias is seen across methods for over/underestimation of prediction errors as evaluated using the ME metric.

For the water quality data experiment (Sect. 2.4.3), seven monitoring stations are missing measures of salinity (n=141). Cross-validation results, as summarized in Table 3, reveal that ClosePD outperforms the other methods in terms of prediction accuracy, as measured by RMSE and CV-R^2, while MDSdist performs the worst. MDSdist also results in a small negative bias in prediction error, as evaluated via ME, while a similar but positive bias is seen for ClosePD. MDSdist continues to display substantial underestimation of prediction variance, with its MPSE being 4x smaller than for ClosePD and with a PI95=0.46. In contrast, standard use of Euclidean distance displays a slight overestimation, while ClosePD displays the least biased estimate of prediction variance ($RMSSE_{Euclidean}$= 0.63 and $RMSSE_{closePD}$=1.24).

The plots of kriged estimates for salinity reveals further differences across the methods (Fig. 10a). The MDSdist method displays high levels of salinity in some of the middle-western tributaries and up through most of the York River, although measurements at monitoring stations do not support such predictions (Fig. 5). The ClosePD method predicts relatively high salinities reaching into many of the rivers when compared to the standard use of Euclidean distance. This discrepancy is likely due to relatively short Euclidean distances to measurements in neighboring tributaries with less saline waters. While the distinction between these two methods is subtle, ClosePD predictions appears to best represent the sampled monitoring stations.

The standard use of Euclidean distance results in kriged variances that have almost no spatial variation, while MDSdist predicts large outliers near some tributary edges (Fig. 10b). ClosePD results in the most natural spatial variation, becoming higher when spatial areas are farther away from sampling locations, particularly in the middle of the Potomac River. MDSdist persists in its underestimation of kriged variance across the Bay's spatial domain.

4 Discussion

This report investigates a geostatistical method that re-estimates non-Euclidean-based covariance structures to ensure that spatial predictions via kriging are mathematically valid. Results, while encouraging, are somewhat mixed, with relative prediction performance varying across methods throughout different evaluation settings. These results may be due in part to the spatial domain in which simulations are developed, the Chesapeake Bay, specifically its many parallel tributaries, highlighted in more detail below. However, in terms of bias of prediction variance, the ClosePD approach consistently has the most accurate estimates overall, while a previously utilized non-Euclidean distance method, MDSdist, displays moderate to severe bias (usually an underestimation) of prediction variance.

The overall strong prediction performance for the standard use of Euclidean distance in the piecewise regression simulations is unexpected, as previous work shows the MDSdist approach consistently outperforming the standard use of Euclidean distance in non-convex spatial domains (Murphy et al. 2015). It is worth noting that for non-Euclidean spatial covariance simulations, both non-Euclidean geostatistical methods have improved prediction accuracy (lower prediction error) when compared to the standard use of Euclidean distance. Therefore, it is likely that the piecewise regression simulations presented here are not properly calibrated to make a non-Euclidean distance the preferred approach for kriging. Given the structure of the Chesapeake Bay's tidal waters, parallel relationships are likely created along some of the major western tributaries in the large-scale spatial data simulations (e.g., Fig. 4), which could have allowed a Euclidean metric that crosses over land to outperform other methods. These simulation results support the theory that in non-convex spatial domains such as the Chesapeake Bay, real-world data could be accurately estimated using Euclidean distance, a non-Euclidean distance, or a combination of the two (Ver Hoef 2018). Such an example could include water temperature, which is influenced by ocean water from the mouth of the bay and freshwater from the heads of tributaries. However, each individual dataset would have to be evaluated in a similar manner as seen in this work in order to determine which distance metric is most appropriate.

Future attempts to simulate spatial data as a function of non-Euclidean distances should consider alternative approaches for data simulation to ensure that non-Euclidean geostatistical methods are preferred for estimation and prediction. Alternate simulations could include multiple large-scale parameters in a single piecewise regression model that are based upon non-Euclidean distances originating from multiple points in the spatial domain. Autoregressive models could also be utilized and would remove the need for the statistical artifacts of piecewise regressions while still using large-scale spatial simulations. Other estuarine systems such as the Salish Sea in both Washington, United States, and British Columbia, Canada, may provide a more optimal spatial domain to develop such large-scale, non-Euclidean spatial data simulations. If adequate simulations can be developed, the idea of using small-scale covariance structures (i.e., ordinary kriging) to estimate simulated large-scale trends should be explored further. The distinction between large-scale and small-scale spatial variation is not well-defined and may be better understood along a spectrum. Future independent research can continue to parse out the relationship between the two model

structures and determine if there are any consistent trends across types and settings of data simulation.

In terms of prediction accuracy, the MDSdist and ClosePD approaches often vary in their relative performance. However, data simulations and water quality data analysis reveal that MDSdist produces substantially biased estimates for prediction variance, often underestimating the variance. Avoiding bias is often the underlying argument for investigating spatial dependence structures in statistical models, specifically with regard to the underestimation of standard errors of large-scale fixed-effect parameters that can occur if spatial dependence is ignored. Bias of the prediction variance however should be viewed with equal concern, as it indicates whether the uncertainty of the prediction estimate is inaccurate, and such accuracy is needed for appropriately-sized prediction intervals. Kriging is frequently used to create spatial predictions that are then utilized in subsequent inference/prediction models (Liu et al. 2016). Propagating the uncertainty of these spatial predictions is an important, though unfortunately often overlooked, aspect of building appropriate statistical models. One of the major benefits of using geostatistical models, as opposed to other spatial interpolation techniques, is that kriging produces unique values of uncertainty for every prediction location. Systematic underestimation of prediction variance affects uncertainty propagation by providing the end-user with more confidence in the original prediction values and in subsequent models than is truly deserved. The preferred balance between accuracy of prediction and accuracy of precision is one that will likely vary across kriging applications and by the overall goals of the modelers. However, the large prediction variance bias observed for MDSdist in this work indicates there may be previously unidentified concerns regarding the approach, and prediction metrics such as RMSSE should be considered before proceeding with such a method.

The specific reasons why MDSdist results in such a large bias for prediction variance are not well understood, but it is likely due in part to the fact that this method re-estimates spatial covariance after scaling distances to a high-dimensional space and are thus inevitably different from the water distance in the observed space. In this work MDS resulted in a poor approximation of original non-Euclidean distances (Fig. 8a), particularly at smaller distances, and this persisted even when considering a weighted MDS approach. Previous work also shows that the embedding estimation creates additional error relative to the original non-Euclidean distances (Boisvert 2010). Current real-world cross-validation results show MDSdist displaying a noticeable bias in prediction variance, which is in contrast to previously reported findings also conducted in the Chesapeake Bay (Murphy et al. 2015). However, if the substantial bias in prediction variance observed for the MDSdist approach in the current work is found to be generalizable, this may question the use of MDSdist as a method for non-Euclidean geostatistics. Alternative geographic embedding approaches, such as local linear embedding, may be more appropriate for the Chesapeake Bay and other similar spatial domains, and so should also be considered when such bias is observed (Roweis and Saul 2000).

The ClosePD method, which often displays the best accuracy for prediction variance in this work, as well as slight to moderately high prediction accuracy relative to other methods, appears to be a strong candidate for future uses of non-Euclidean distance measures in the

field of geostatistics. Its overall performance, especially in the water quality results, indicate that the method creates predictions that are more optimal than the standard use of Euclidean distance, while also providing a more balanced tradeoff of prediction accuracy and precision accuracy than the MDSdist approach. Additional work is still needed to standardize this method, as its performance is sensitive to the tolerance, τ from Eq. (4), that is set to ensure eigenvalue positivity. Initial sensitivity analyses indicate that if the tolerance is too large (i.e., positive eigenvalues are allowed to become exceedingly small), the resulting covariance matrix is too similar to the original covariance matrix based upon non-Euclidean water distances and can remain essentially non-positive-definite. Conversely, if the tolerance is too small (i.e., eigenvalues remain relatively large), the resulting covariance matrix poorly approximates the original non-Euclidean covariance matrix. Current work suggests that if the tolerance is set so that the smallest positive eigenvalue, λ_k, ranges between 1 and 5, the method performs adequately, as reported in the Sect. 3. Future work may also consider a similar function that applies an arguably more sophisticated algorithm to determine a near positive-definite covariance matrix (Higham 2002). Finally, the ClosePD method should be compared to a recently developed method that ensures a positive-definite covariance matrix when considering non-Euclidean distances for a mixed-effect model (Ver Hoef 2018).

In the piecewise regression simulations, many realizations result in a non-Euclidean covariance matrix that is already positive-definite. As described previously, the concern of using a non-Euclidean distance for geostatistical modeling is not that there always be an invalid covariance structure but rather that validity is no longer ensured (Curriero 2006). The concerns about the spatial data simulation experiments detailed above may have contributed in part to the high observation rate of positive-definite spatial covariance structures based on non-Euclidean distances. However, it may also be that the frequency of observing invalid covariance structures from non-Euclidean distances is somewhat rare. While examples of positive-definite covariance matrices created by using non-Euclidean distances have been shown in the literature (Jensen et al. 2006), it would be dangerous to therefore assume that all future spatial interpolation applications will also be mathematically valid in a given setting. Applying the method proposed in this report will circumvent any issue of mathematically invalid predictions, while still returning optimal results if covariance matrices estimated using non-Euclidean distances are already positive-definite.

5 Computational Limitations

As has been indicated throughout, the proposed method is currently implemented in a global framework and attempts to use all sample information to identify best linear unbiased predictions at all locations in the sample domain. This is appropriate for the simulated and real-world experiments utilized in this work (datasets up to 400 entries and interpolation grid locations ~11,000). However, for research investigations that consider substantially larger datasets or spatial domains, full estimation of the semivariance and covariance matrix either through the MDSdist method or the ClosePD method may prove computationally impractical.

To circumvent this issue, a number of procedures can be considered. For the MDSdist method, an alternative approach can be implemented using several landmark locations in the

spatial domain to approximate the higher-dimensional embedded distances (Boisvert and Deutsch 2011). It should be noted however that this simplification can result in additional error in distance estimation, which can lead to increased smoothing in kriging (Boisvert 2010). This technique also cannot be directly adapted to the ClosePD method, which requires additional computational time to decompose the covariance matrix, recalculate eigenvalues below the pre-determined threshold, and then compose a rescaled covariance matrix. A more recent approach reduces computational demands by considering a reduced-rank method, which applies location-based knots to estimate spatial covariances in a mixed-effect model (Ver Hoef 2018). To circumvent computational bottlenecks with large data while preserving the ClosePD methodology, one could instead consider a localized form of kriging in which only a subset of the closest sampling locations are used to populate semivariance and covariance matrices (Datta et al. 2016). The original characterization of the non-Euclidean variogram via weighted least squares can still be conveniently used even when N is considerably large, allowing for an appropriate characterization of the covariance structure based upon nearby sampled data. Future work should attempt to directly apply such a local kriging approach to warranted examples.

6 Conclusions

This report proposes a geostatistical method that re-estimates non-Euclidean-based covariance structures to ensure that spatial predictions via kriging are mathematically valid. Experimental results indicate that the method performs adequately across a range of settings, while also highlighting potential prediction uncertainty bias for a previously developed method that transforms non-Euclidean distances prior to covariance estimation. Overall results indicate that the proposed method can provide improved geostatistical predictions while ensuring valid results when the use of non-Euclidean distances is warranted. This report also proposes future experiments that can further test the robustness and generalizability of this method. Finally, adjustments to the method are suggested for experimental settings where the number of observations or the size of the spatial domain may lead to computational bottlenecks.

Acknowledgements

This work was supported by the National Institutes of Allergy and Infectious Diseases (grant no. 1R01AI123931-01A1 to F.C.C. [principal investigator]). Additional support for B.J.K.D. was provided in part by the Johns Hopkins' Environment, Energy, Sustainability & Health Institute Fellowship and the Center for a Livable Future-Lerner Fellowship, as well as The National Science Foundation's Water, Climate, and Health Integrative Education and Research traineeship. The authors would like to thank Tim Shields for helping to develop the schematic maps displayed in this paper.

Appendix A:: List ofEquations for Cross-Validation Metrics

Equations for the cross-validation metrics are provided below

$$\text{CV-R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}; \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{9}$$

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i, \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{11}$$

$$\mathrm{MPSE} = \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}_i, \tag{12}$$

$$\mathrm{RMSSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{\hat{\sigma}_i}\right)^2}, \tag{13}$$

$$\mathrm{PI95} = \frac{1}{n}\sum_{i=1}^{n} I_i \quad ; \text{where } I_i = \begin{cases} 1 & \text{if } \hat{y}_i - 1.96\hat{\sigma}_i \le y_i \le \hat{y}_i + 1.96\hat{\sigma}_i \\ 0 & \text{if } y_i < \hat{y}_i - 1.96\hat{\sigma}_i \text{ or } y_i > \hat{y}_i + 1.96\hat{\sigma}_i \end{cases}, \tag{14}$$

where y_i is an observed outcome for location i, $\hat{y}_i$ is the predicted (i.e., kriged) value, and $\hat{\sigma}_i$ is the kriging standard error.

References

Berman JD, Breysse PN, White RH, Waugh DW, Curriero FC (2015) Evaluating methods for spatial mapping: Applications for estimating ozone concentrations across the contiguous United States. Environmental Technology & Innovation 3:1–10

Bivand R, Keitt T, Rowlingson B (2016) rgdal: Bindings for the Geospatial Data Abstraction Library, R package version 1.1–10 edn,

Boisvert JB (2010) Geostatistics with Locally Varying Anisotropy. University of Alberta

Boisvert JB, Deutsch CV (2011) Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances. Computers and Geosciences 37:495–510

Cheng SH, Higham NJ (1998) A modified Cholesky algorithm based on a symmetric indefinite factorization. SIAM Journal on Matrix Analysis and Applications 19:1097–1110

Chesapeake Bay Program (2017) Data Hub: CBP GIS Datasets. Chesapeake Bay Program. ftp://ftp.chesapeakebay.net/pub/Geographic/. Accessed September 17, 2015 2017

Congdon CD, Martin JD On using standard residuals as a metric of kriging model quality. In: Proceedings of the 48th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, Honolulu HI, 2007.

Cressie NAC (1993) Statistics for Spatial Data. Revised edn John Wiley & Sons,

Curriero FC (2006) On the use of non-euclidean distance measures in geostatistics. Mathematical Geology 38:907–926

Datta A, Banerjee S, Finley AO, Gelfand AE (2016) On nearest-neighbor Gaussian process models for massive spatial data. Wiley Interdisciplinary Reviews: Computational Statistics 8:162–171 [PubMed: 29657666]

Davis BJ, Jacobs JM, Davis MF, Schwab KJ, DePaola A, Curriero FC (2017) Environmental determinants of Vibrio parahaemolyticus in the Chesapeake Bay. Appl Environ Microbiol 83:e01147–01117 [PubMed: 28842541]

Del Castillo E, Colosimo BM, Tajbakhsh SD (2015) Geodesic gaussian processes for the parametric reconstruction of a free-form surface. Technometrics 57:87–99 doi: 10.1080/00401706.2013.879075

Diggle PJ, Ribeiro PJ (2007) Model-based Geostatistics Springer Series in Statistics. Springer, New York, NY

ESRI (2011) ArcGIS Desktop: Release 10.3. Environmental Systems Research Institute, Redlands, CA

ESRI (2016) Cross Validation. esri http://desktop.arcgis.com/en/arcmap/10.3/tools/geostatistical-analyst-toolbox/cross-validation.htm. Accessed June 27, 2016 2016

Etten Jv (2015) gdistance: Distances and Routes on Geographical Grids, R package version 1.1–9 edn,

Gardner B, Sullivan PJ, Lembo AJ Jr (2003) Predicting stream temperatures: Geostatistical model comparison using alternative distance metrics. Canadian Journal of Fisheries and Aquatic Sciences 60:344–351

Hengl T, Heuvelink GB, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120:75–93

Henshaw SL, Curriero FC, Shields TM, Glass GE, Strickland PT, Breysse PN (2004) Geostatistics and GIS: tools for characterizing environmental contamination. J Med Syst 28:335–348 [PubMed: 15366239]

Higham NJ (2002) Computing the nearest correlation matrix—a problem from finance. IMA Journal of Numerical Analysis 22:329–343

Jeffrey SJ, Carter JO, Moodie KB, Beswick AR (2001) Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ Model Software 16:309–330

Jensen OP, Christman MC, Miller TJ (2006) Landscape-based geostatistics: a case study of the distribution of blue crab in Chesapeake Bay. Environmetrics 17:605–621

Kane MJ, Emerson J, Weston S (2013) Scalable strategies for computing with massive data. Journal of Statistical Software 55:1–19

Laaha G, Skøien J, Blöschl G (2014) Spatial prediction on river networks: comparison of top-kriging with regional regression. Hydrological Processes 28:315–324

Little LS, Edwards D, Porter DE (1997) Kriging in estuaries: as the crow flies, or as the fish swims? Journal of experimental marine biology and ecology 213:1–11

Liu R, Young MT, Chen J-C, Kaufman JD, Chen H (2016) Ambient air pollution exposures and risk of Parkinson disease. Environ Health Perspect 124:1759 [PubMed: 27285422]

Løland A, Host G (2003) Spatial covariance modelling in a complex coastal domain by multidimensional scaling. Environmetrics 14:307–321 doi:10.1002/env.588

Lu B, Charlton M, Fotheringham AS Geographically Weighted Regression using a non-Euclidean distance metric with a study on London house price data. In: Procedia Environmental Sciences, 2011 pp 92–97. doi:10.1016/j.proenv.2011.07.017

Lu B, Charlton M, Harris P, Fotheringham AS (2014) Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house price data. International Journal of Geographical Information Science 28:660–681 doi:10.1080/13658816.2013.865739

Lucas C (2001) Computing nearest covariance and correlation matrices. M.S, Thesis, University of Manchester

Maechler M (2016) sfsmisc: Utilities from “Seminar fuer Statistik” ETH Zurich, R package version 1.1–0 edn,

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. Academic Press, London, U.K.

Matheron G (1971) The theory of regionalized variables and its applications. Les Cahiers de Morphologie Mathematique 5:218

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2015) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, R package version 1.6–7. edn,

Murphy R, Perlman E, Ball WP, Curriero FC (2015) Water-Distance-Based Kriging in Chesapeake Bay. Journal of Hydrologic Engineering 20:0501403

Novomestky F (2012) matrixcalc: Collection of functions for matrix calculations, R package version 1.0–3 edn,

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rathbun SL (1998) Spatial modelling in irregularly shaped regions: Kriging estuaries. Environmetrics 9:109–129

Ribeiro PJ, Diggle PJ (2016) geoR: Analysis of Geostatistical Data, R package version 1.7–5.2 edn,

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326 doi:10.1126/science.290.5500.2323 [PubMed: 11125150]

Rowlingson B, Diggle P (2015) splancs: Spatial and Space-Time Point Pattern Analysis, R package version 2.01–38 edn,

Sampson PD, Guttorp P (1992) Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association 87:108–119

Schlather M, Malinowski A, Menck PJ, Oesting M, Strokorb K (2015) Analysis, simulation and prediction of multivariate random fields with package RandomFields. Journal of Statistical Software 63:1–25

USGS (2016) The National Hydrography Dataset. https://nhd.usgs.gov/index.html. Accessed December 3, 2016 2016

Ver Hoef JM (2018) Kriging models for linear networks and non-Euclidean distances: Cautions and solutions. Methods in Ecology and Evolution doi:10.1111/2041-210X.12979

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York

Yu H, Wang X, Qing J, Nie H (2015) ArcMap Raster Edit Suite (ARES), 0.2.1 edn,

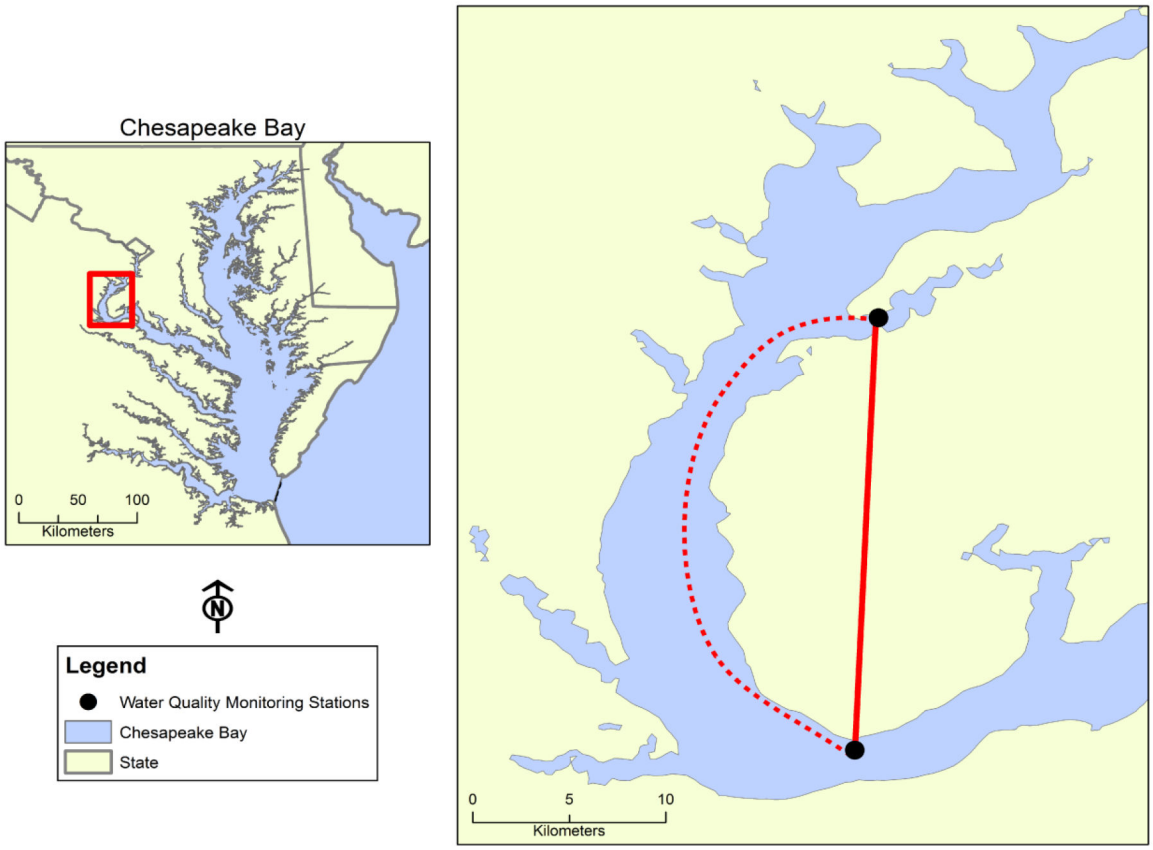


Fig. 1.

Two Chesapeake Bay water quality monitoring stations in the Potomac River showing Euclidean distance (solid line) and a non-Euclidean water distance (dashed line) between stations demonstrating Euclidean distance as inappropriate and intersecting land

Fig. 2.
Chesapeake Bay and surrounding U.S. states labelled with locations of interest

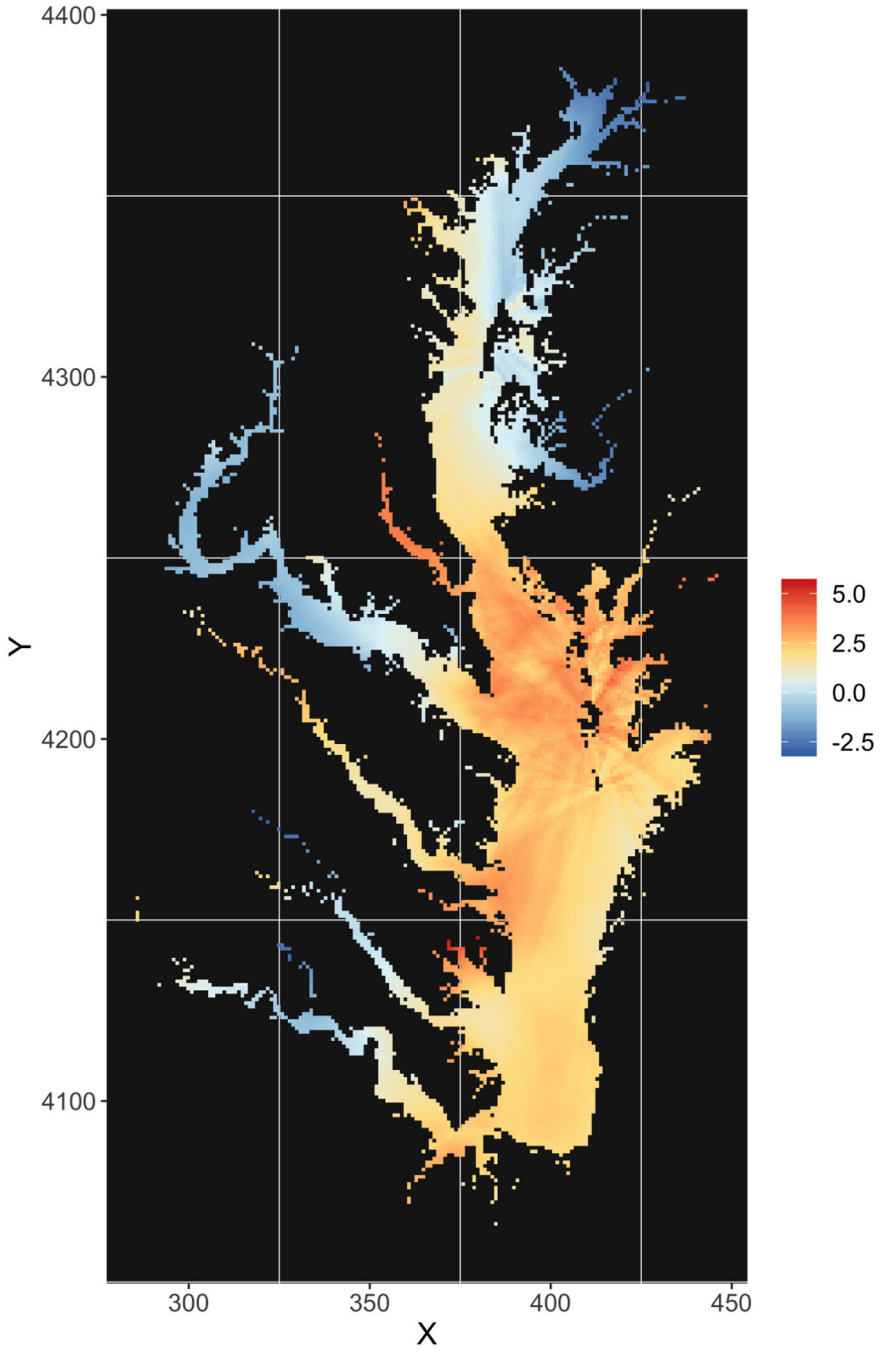


Fig. 3.

Representative simulation using non-Euclidean spatial covariance: Gaussian semivariogram model with parameters $\sigma^2 = 4$, $\phi = 66.7$km, geographical coordinates are projected into Universal Transverse Mercator zone 18N with units of kilometers

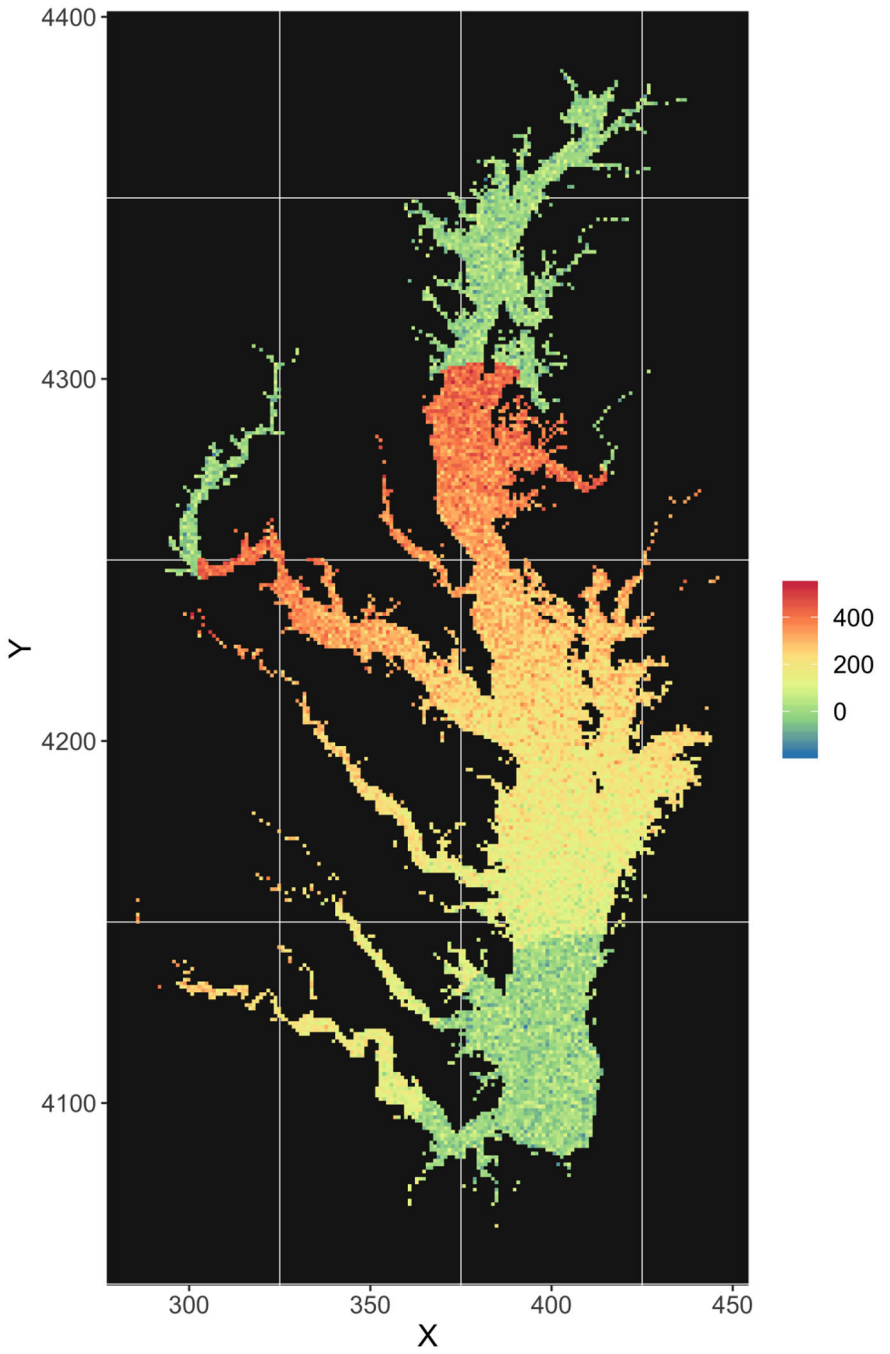


Fig. 4.

Representative simulation using piecewise regression: linear function of water distance from the mouth of the bay ($\beta = 2.0$), thresholds set at <50km and > 210km with random error $\epsilon \sim \mathcal{N}(0, 50)$ applied throughout, projected into Universal Transverse Mercator zone 18N

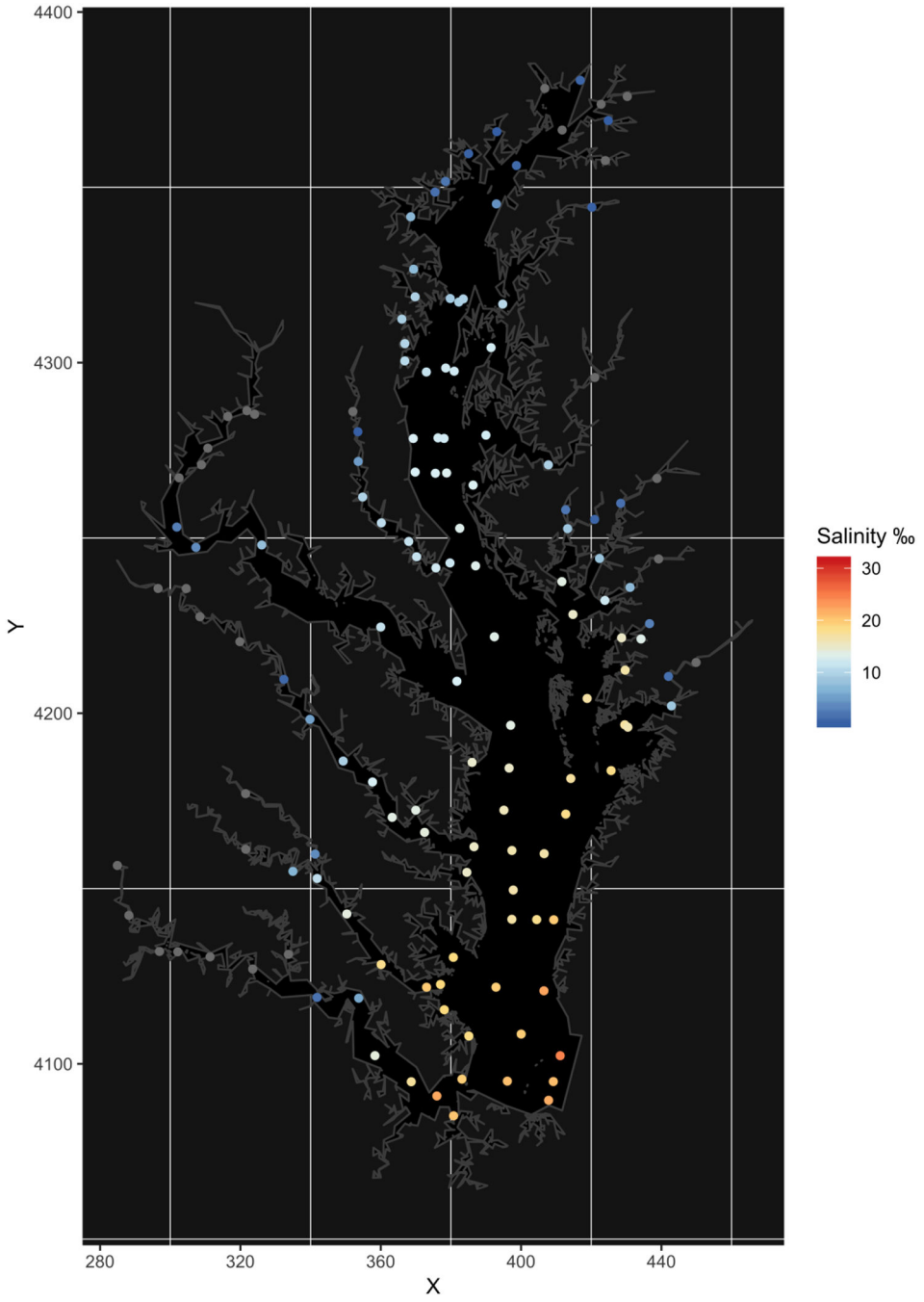


Fig. 5.

Measurements of salinity at water quality monitoring stations during July 2009 projected into Universal Transverse Mercator zone 18N

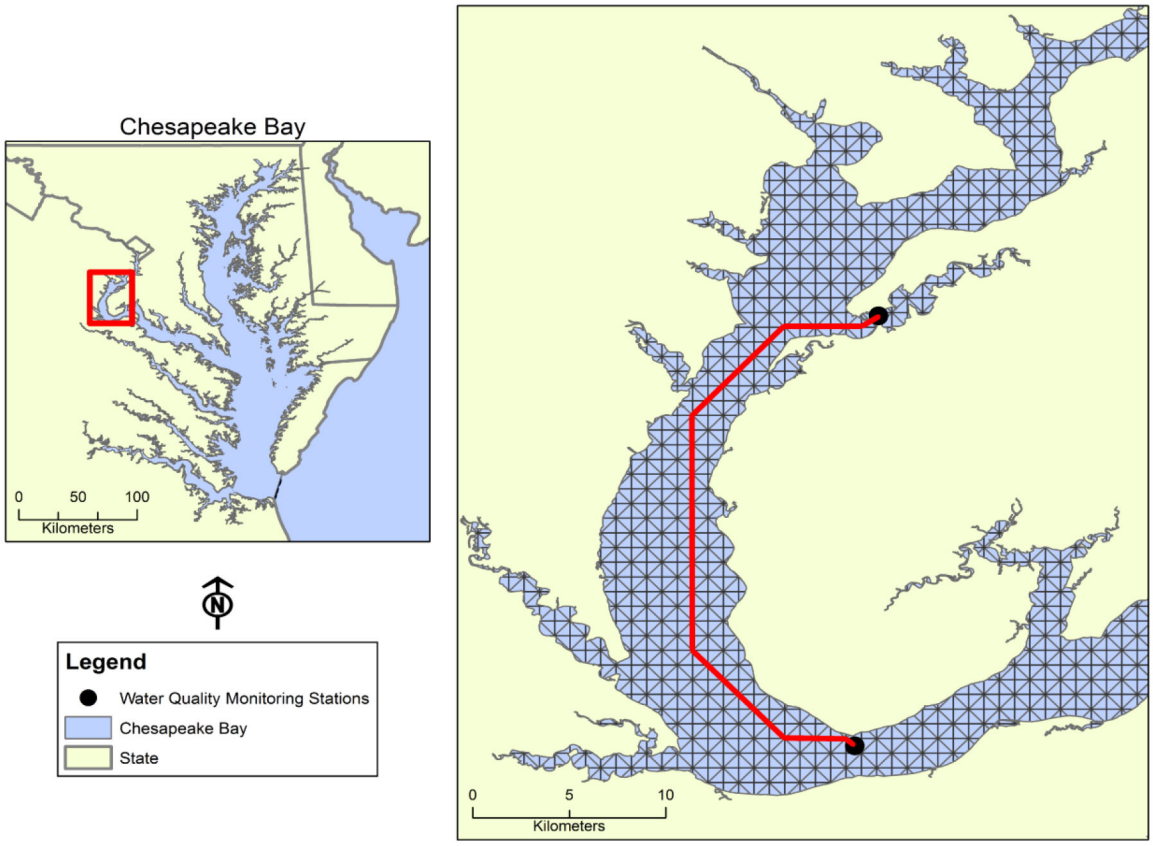


Fig. 6.

Schematic of water distance computation using raster cells to create multi-directional piecewise linear paths to enumerate water distances

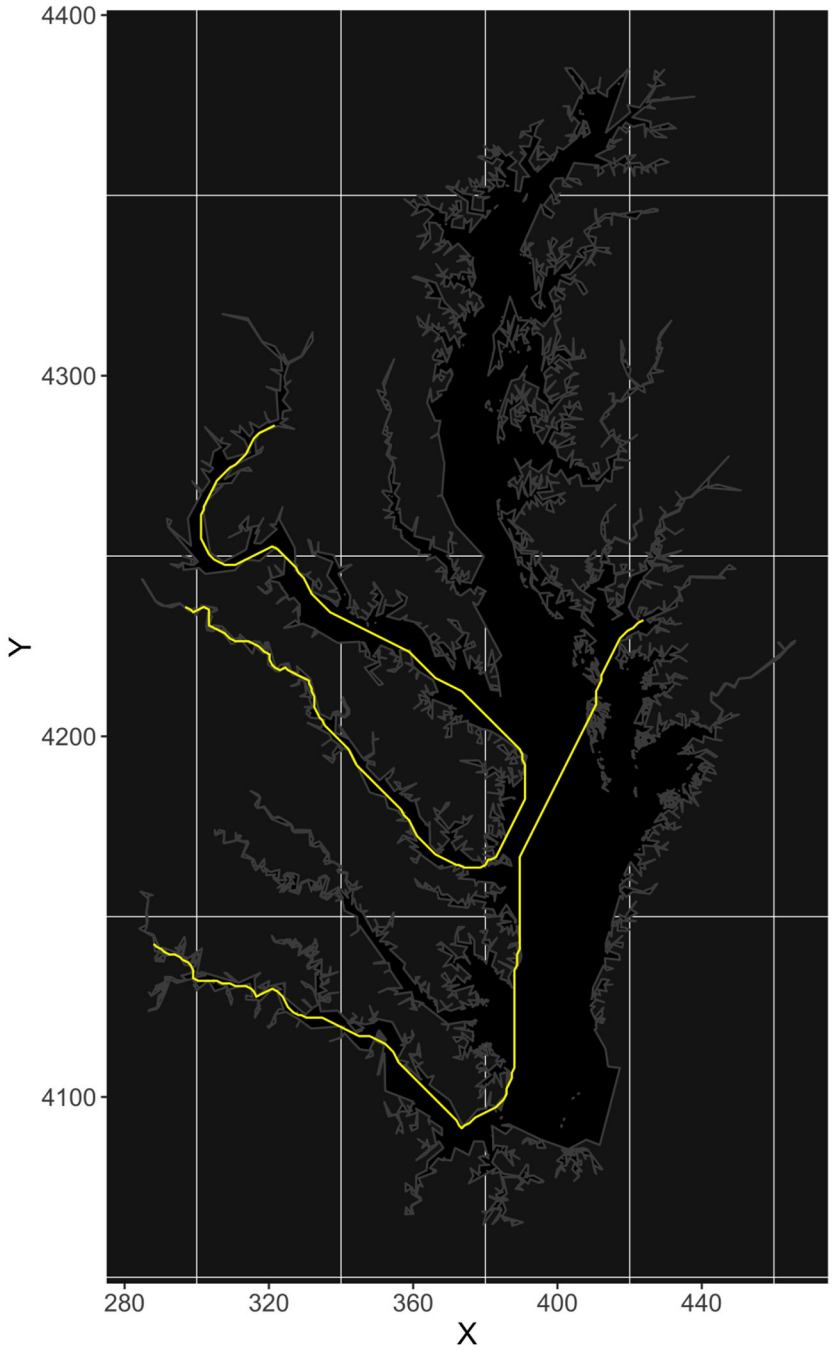


Fig. 7.
Two water distance paths computed using the gdistance package in R

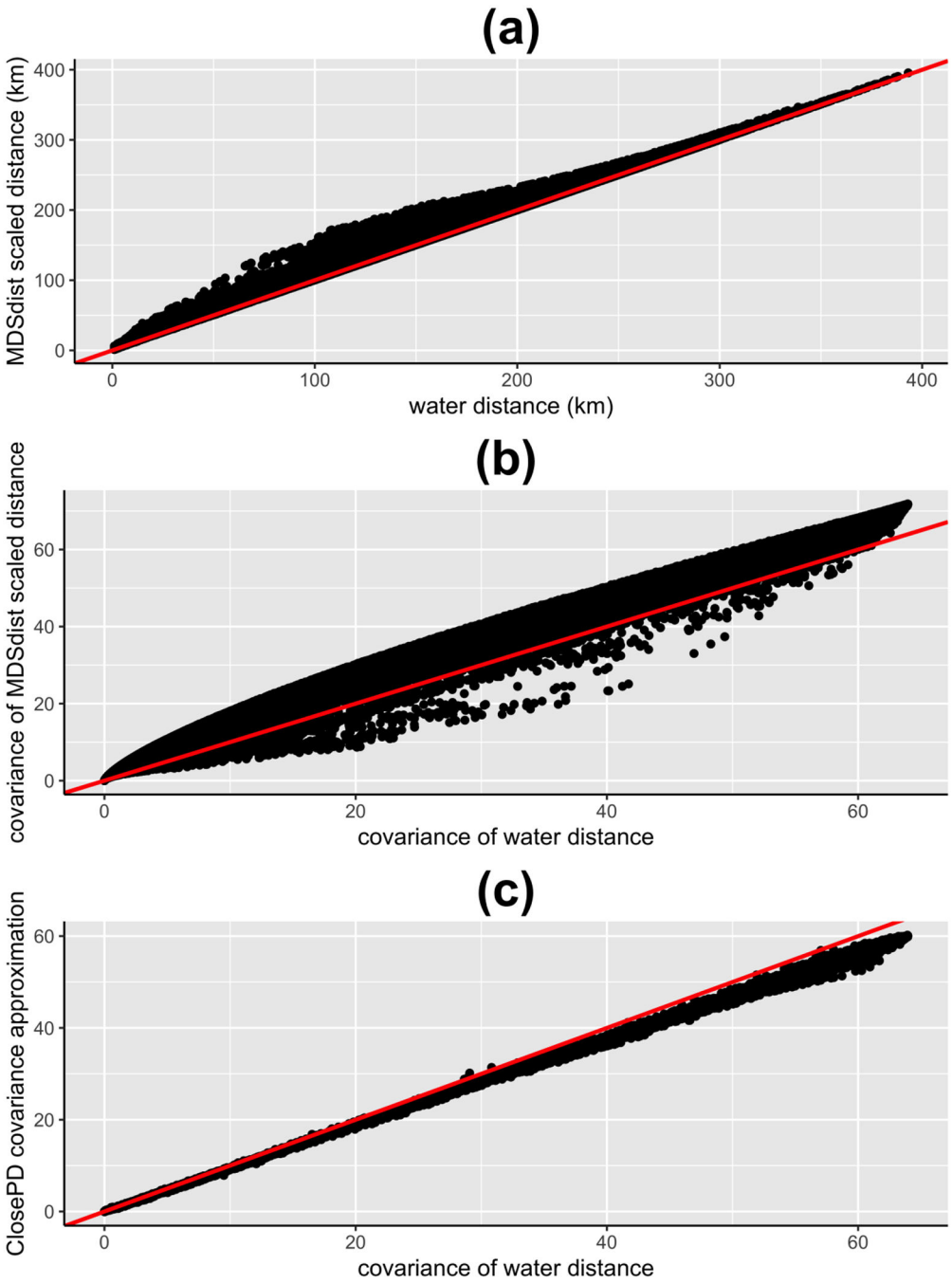


Fig. 8.

Representative plots with 45-degree red lines comparing **a** distances from MDSdist $\left(\left\|s_i^* - s_j^*\right\|\right)$ and from non-Euclidean (d_{ij}) **b** covariances from MDSdist $C\left(\left\|s_i^* - s_j^*\right\|\right)$ and from non-Euclidean $C([d_{ij}])$ **c** covariances from ClosePD $\tilde{C}\left(\left[d_{ij}\right]\right)$ and fromnon-Euclidean

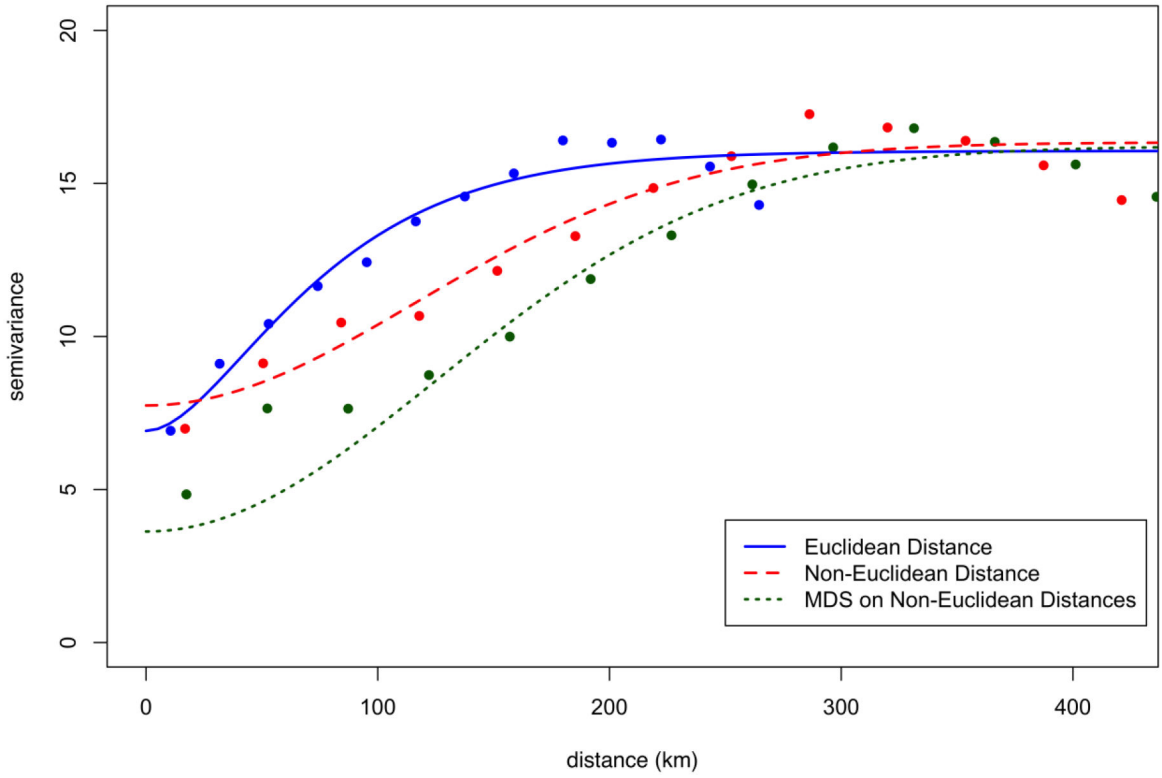
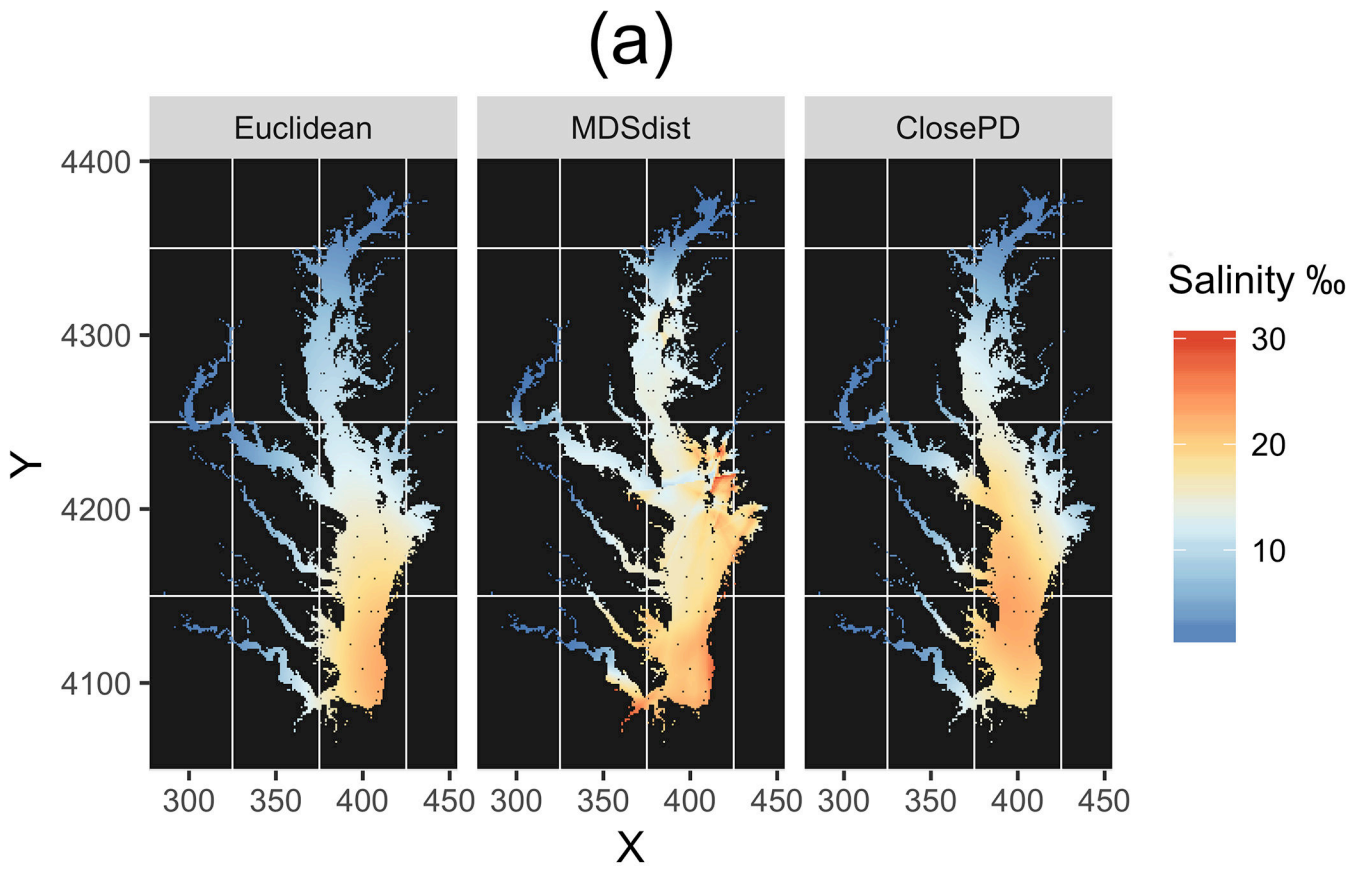


Fig. 9.

Representative plot of empirical semivariograms from a piecewise regression model and estimated Matém semivariogram functions using Euclidean distances, non-Euclidean distances and non-Euclidean distances approximated using MDS



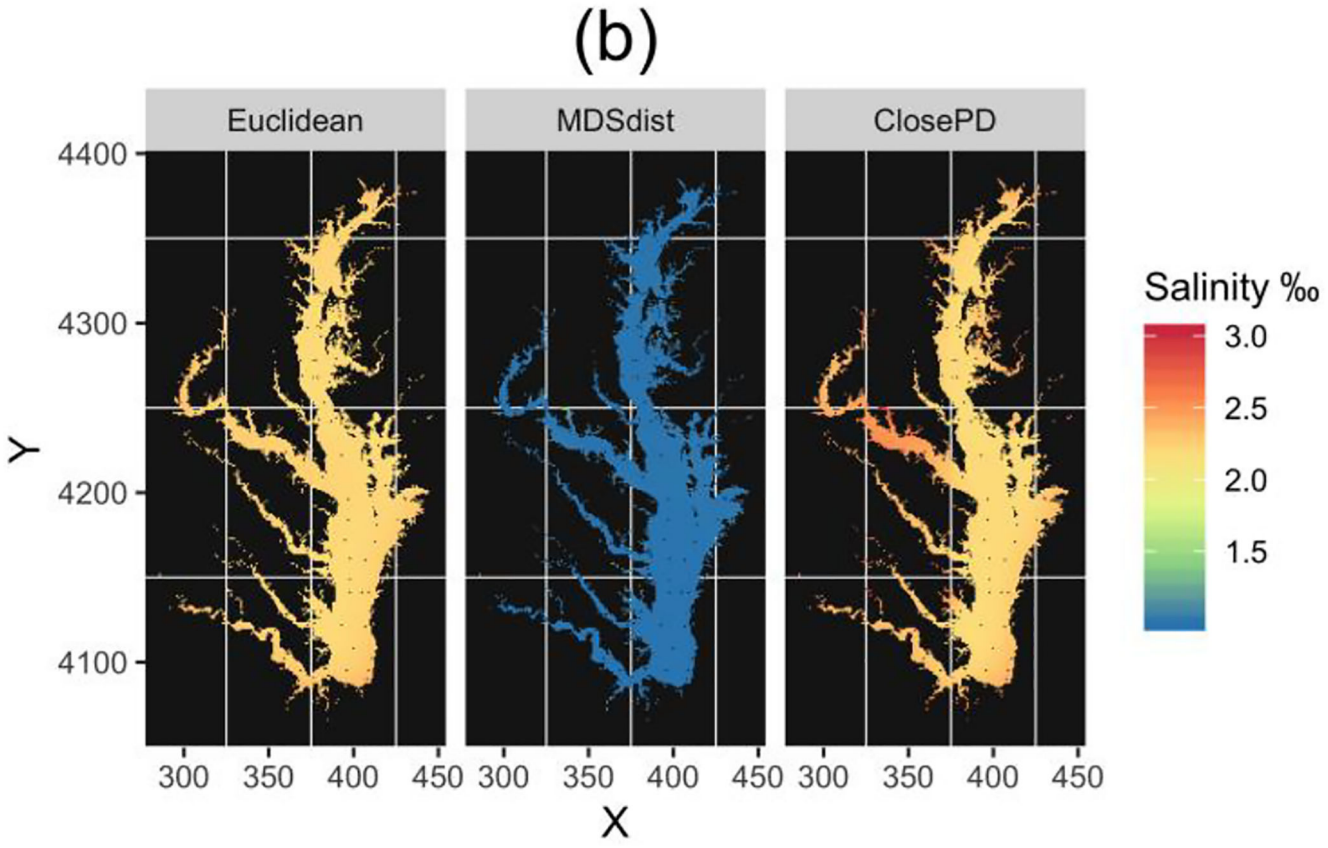


Fig. 10.

Results of salinity kriging across the Chesapeake Bay projected into Universal Transverse Mercator zone 18N **a** kriged (prediction) estimates **b** kriged (prediction) variances

Table 1.

Representative LOOCV results for the non-Euclidean spatial covariance simulations. Reported semivariograms are estimated using the exponential function.

Covariance Parameters (σ^2, ϕ)[a]	Method	RMSE[b]	MPSE[c]	RMSSE[d]
(4, 26.7)				
	Euclidean[e]	0.94	1.03	0.88
	MDSdist	0.93	1.06	0.83
	Non-Euclidean[f]	0.91	1.01	0.86
(50, 44.5)				
	Euclidean	2.65	2.83	0.86
	MDSdist	2.62	3.07	0.80
	Non-Euclidean	2.57	2.81	0.85
(625, 66.7)				
	Euclidean	7.83	8.24	0.87
	MDSdist	7.67	8.98	0.80
	Non-Euclidean	7.56	8.16	0.85

[a] σ^2 = partial sill, ϕ = range,

[b] root-mean-squared error,

[c] mean prediction standard error,

[d] root-mean-square standardized error,

[e] standard Euclidean distance,

[f] naïve use of non-Euclidean distance.

Table 2.

Representative LOOCV results for data simulated using a piecewise regression originating at the mouth of the Chesapeake Bay and with no-trend thresholds at less than 50km and greater 210km. Semivariograms are estimated using the Matérn function.

Method	CV-R^2 [a]	ME [b]	RMSE [c]	MPSE [d]	RMSSE [e]	PI95 [f]
Euclidean [g]	0.79	0.02	66.11	47.93	1.37	92.23
MDSdist	0.70	0.00	72.81	29.86	4.70	79.49
ClosePD	0.68	0.01	81.47	65.8	1.15	96.16

a) cross-validation R^2.

b) mean error,

c) root-mean-squared error,

d) mean prediction standard error,

e) root-mean-square standardized error,

f) proportion of prediction estimates that fell within the 95% prediction interval,

g) standard Euclidean distance.

Table 3.

LOOCV results of (n=141) water quality monitoring stations in the Chesapeake Bay for measures of salinity (‰) from July 2009. Semivariograms were estimated using the Matérn function.

Method	CV-R^2 [a]	ME [b]	RMSE [c]	MPSE [d]	RMSSE [e]	PI95 [f]
Euclidean [g]	0.82	0.02	2.96	4.70	0.63	0.99
MDSdist	0.72	−0.07	3.77	0.45	10.36	0.46
ClosePD	0.89	0.05	2.38	1.92	1.24	0.93

a) cross-validation R^2.

b) mean error,

c) root-mean-squared error,

d) mean prediction standard error,

e) root-mean-square standardized error,

f) proportion of prediction estimates that fell within the 95% prediction interval,

g) standard Euclidean distance.