My Final College Paper

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Blake Rosenthal

May 2015

Approved for the Division
(Mathematics)

_____

Albert Kim

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

Kriging

1.1 Spatial statistics

Kriging is a method utilized in the field of geostatistics to model spatial data. Originally developed from the South African mining industry in the 1950's **?**, kriging provided a way to predict ore-grade distributions based on a limited empirical sample. Though the name comes from mining engineer D. G. Krige, methods for optimal spatial linear prediction from Wold (1938), Kolmogorov (1941b), and Wiener (1949) all include the crucial covariance component of spatial interpolation, realizing that points closer to the prediction point should be given greater weights than further points. This is the cornerstone of the kriging method and is explored in detail in section 3.

Given a spatially continuous random process $Y(x)$ over some two-dimensional region B, a data sample $S_i : i = 1, \ldots, n$ is obtained from Y at locations $x_i : i = 1, \ldots, n$. [1]From a practical perspective, Y can be thought of as an underlying but unknown distribution of a variable of interest over B, be it ore-density, mineral concentrations, elevation, etc. S is therefore a set of vectors containing an independent spatial component and a dependent variable or variables. Since S is only a small and incomplete realization of the field Y, the standard geostatistical approach is to impose an underlying structure to the field consisting of a mean function $\mu(\mathbf{s})$ and a random error process with zero mean $e(\mathbf{s})$. Together these specify that

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + e(\mathbf{s}),$$

where $\mu(\mathbf{s}) = E[Y(\mathbf{s})]$.

The goal is to make some predictions regarding the underlying random process Y. Kriging at its simplest is a matter of predicting a value of $Y(x_i)$ at an arbitrary point within the region B. *Simple kriging* assumes Y to have a constant mean which is

[1]A note on notation: here, x will be used to specify a generic point in Y, while $\mathbf{s}$ will be used to indicate the vector of spatial coordinates or other dependent variables that make up the sample S.

estimated from the sample mean of S. *Ordinary kriging* uses the estimated covariance structure of Y to replace the sample mean with the generalized least squares estimate of μ. Finally, *universal kriging* uses a trend surface model for the mean.

1.2 Estimation of the mean function

Simple, ordinary, and universal kriging all differ in their approach to estimating the mean function. Simple kriging, which assumes a constant mean, is typically dismissed by most statisticians since it usually fails to accurately describe any naturally occurring random process. Here we go over universal kriging since it is the best linear unbiased prediction model for geostatistical random fields **?**.

The purpose of the mean function is to help provide an estimate for the residuals $e(\mathbf{s})$, given by $\hat{e}$. This estimate is then used to calculate the semivariogram, described in the following section, which is then used in the universal-kriging equations. The mean function is given by $\mu(\mathbf{s}) = E[Y(\mathbf{s})]$ and is modeled as the linear equation

$$\mu(\mathbf{s}; \beta) = \mathbf{X}(\mathbf{s})^T \beta$$

where $\mathbf{X}(\mathbf{s})$ is a vector of covariates observed at $\mathbf{s}$ and β is an unrestricted parameter vector. These variables could be simply latitude and longitude coordinates, but may also include such information as elevation, slope, windspeed, etc. If using only latitude and longitude, for example, a first order trend surface model is given by

$$\mu(\mathbf{s}; \beta) = \beta_0 + \beta_1 s_1 + \beta_2 s_2$$

where $\mathbf{s} = (s_1, s_2)$ are latitude and longitude. This definition, however, is not invariant to the choice of origin or orientation of the coordinate system **?** and higher-order polynomials such as the quadratic loosen the restriction of independent latitude/longitude effects. [2]

$$\mu(\mathbf{s}; \beta) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_{11} s_1^2 + \beta_{12} s_1 s_2 + \beta_{22} s_2^2.$$

At this point the provisional linear mean function is then fitted to the available data. There are many ways to do this, but the ordinary least squares method is typically used. This method yields an estimator $\hat{\beta}_{OLS}$ given by

$$\hat{\beta}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

where $\mathbf{X} = [X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_n)]^T$ and $\mathbf{Y} = [Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)]^T$. [3]

[2] check with Albert

[3] Equivalently, and perhaps easier to work with, $\hat{\beta}_{OLS} = \operatorname{argmin} \sum_{i=1}^{n} [Y(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)^T \beta]^2$.

It is possible to stop the analysis here, but once we have the second-order dependency structure of the semivariogram from the following section we can reestimate the mean function using estimated generalized least squares. A method given by Zimmerman and Stein (**?**, p. 40) involves estimating a covariance matrix to include in the mean estimation. The updated mean function can be then used to recalculate the residuals $\hat{e}(\mathbf{s})$ for the semivariogram.

1.3 Covariance and the variogram

Part of the effectiveness of the kriging method comes from the recognition that the data from a spatial sample are correlated based on proximity. Points closer together are expected to be more highly correlated than points with greater spatial separation. The variogram, or semivariogram, plots this correlation as a function of distance, and the empirical semivariogram is the observed covariance structure of the data **?**. Given this, the semivariogram is defined by $\gamma(x_i - x_j) = \frac{1}{2}\text{var}\{e(x_i) - e(x_j)\}$, for all $x_i, x_j \in B$. Intuitively, the semivariogram provides a way to visualize the correlative effects of distance on the sampled data. For example, given a set of locations in the Cartesian plane, points with no separation distance could be expected to have zero variation in their dependent variables, while the variance between very distant points can be expected to be much higher [Fig. 1.1].

The distance between any two points x_i and x_j can be can be used to define a new set $H = \{x_i - x_j : x_i, x_j \in B\}$ for the continuous distribution of distances, or lags, in B. Elements of H can be grouped into bins $H_1, H_2, \ldots, H_k$. A representative lag for the entire bin $\mathbf{h}_u$ can be used to define the unbiased estimator of $\gamma(\mathbf{h}_u)$ by

$$\hat{\gamma}(\mathbf{h}_u) = \frac{1}{2n(H_u)} \sum_{x_i - x_j \in H_u} [\hat{e}(x_i) - \ \hat{e}(x_j)]^2 \quad (u = 1, \ldots, k)$$

where $n(H_u)$ is the number of lags within the bin H_u and $\hat{e}(x_i)$ is the residual at point x_i after estimating the mean. This assumes that covariance between data points is a function of spatial distance only, and not location or other factors. This estimation also requires a subjective choice in binning – since any exact distance, or lag, between two points is unlikely to occur frequently within a sample, it is necessary to group distances into representative intervals, or bins. A common way to do this is to make this binning choice up front, perhaps grouping the data into thirty or so bins then choosing $\mathbf{h}_u$ to be the average of all the lags that fall into a given bin. Therefore, unless the data is taken on a rectangular or polar grid, the accuracy of the semivariogram will always be dependent on the binning choices. What's the right number of bins? There's a trade-off – more bins means that $\mathbf{h}_u$ is a better estimation of its representative bin H_u, yet there are fewer lags to any particular bin and a smaller sample size and therefore a greater sampling variation. This is an interesting optimization problem on its own, but the data itself may impose binning restrictions depending on the sample size and other factors. This means that there is therefore no uniquely

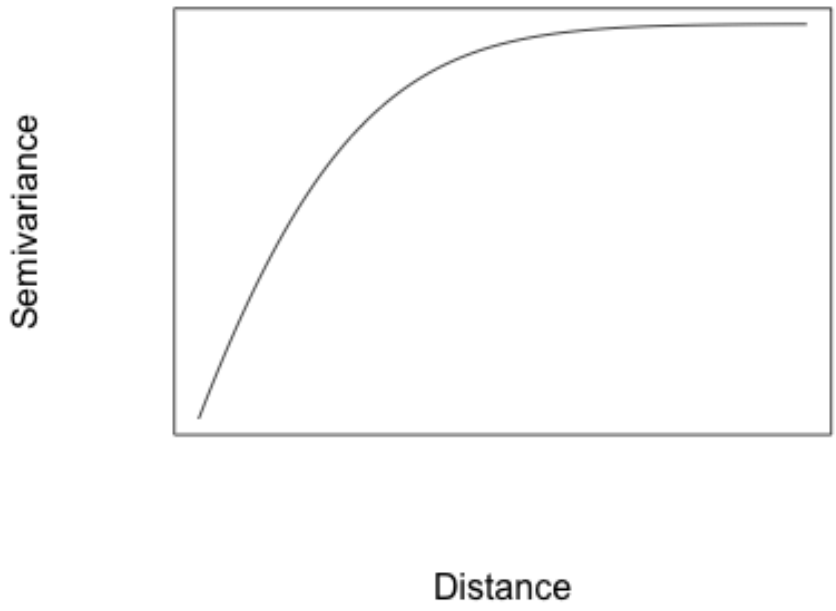

Figure 1.1: An exponential semivariogram

optimized semivariogram.

Fitting a parametric model to the empirical variogram gives a convenient equation to work with for several reasons – first, the empirical semivariogram will often have a high variance, and a smoothed version will have a lower variance that is easier to work with. Second, the empirical semivariogram usually fails to be conditionally nonpositive definite. This is a necessary condition when choosing predictors at later stages since the prediction error variance must be nonnegative at every point in the field. Third, predicting locations at lags not represented by the chosen bins requires a continuous function, something only a smoothed variogram can accomplish. This smoothed version must satisfy the following necessary and sufficient conditions to be a valid semivariogram:

1. Vanishing at 0: $\gamma(\mathbf{0}) = 0$

2. Evenness: $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$ for all $\mathbf{h}$

3. Conditional negative definiteness: $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(x_i - x_j) \leq 0$ for all n, all $s_1, \ldots, s_n$ and all $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$

A crucial assumption is that a "true" semivariogram exists for the entire region. By modeling the empirical semivariogram and fitting it to a curve we are guessing at the underlying model that represents the entire process. In a way, this describes the entire study of statistics in general: using incomplete data to make an educated guess about the underlying, inherently unknowable, system and adjusting to the model to minimize inaccuracies.

1.4 Spatial Prediction: Kriging

Given a prediction point $\mathbf{s}_0$ [4], the goal of kriging is to find a predictor $\hat{Y}(\mathbf{s}_0)$ for $Y(\mathbf{s}_0)$ that minimizes the prediction error variance $\text{var}[\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0)$ of all possible predictors that are both (1), linear, and (2) unbiased:

1. $\hat{Y}(\mathbf{s}_0) = \lambda^T \mathbf{Y}$, where λ is a vector of fixed constants and $\sum \lambda_i = 1$.
2. $E[\hat{Y}(\mathbf{s}_0)] = E[Y(\mathbf{s}_0)]$, or equivalently, $\lambda \mathbf{X} = \mathbf{X}(\mathbf{s}_0)$.

Here λ can be thought of as a vector of weights applied to the sample data. Since the value of Y at $\mathbf{s}_0$ depends solely on the empirical data, optimizing this linear predictor with respect to the given restraints gives a unique solution. If λ is a solution to this problem, then $\lambda^T \mathbf{Y}$ is a best linear unbiased predictor (BLUP) for $Y(\mathbf{s}_0)$. Here "best" means having the smallest mean squared error within the class of linear unbiased predictors. There are several ways of solving this. Cressie **?** gives a proof using differential calculus and Lagrange multipliers, while Zimmerman and Stein **?** give a geometric proof. Both give the following solution:

$$\hat{Y}(\mathbf{s}_0) = [\gamma + \mathbf{X}(\mathbf{X}^T \Gamma^{-1} \mathbf{X})^{-1}(\mathbf{x}_0 - \mathbf{X}^T \Gamma^{-1} \gamma)]^T \Gamma^{-1} \mathbf{Y}$$

where $\gamma = [\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0)]^T$, Γ is the $n \times n$ symmetric matrix with ijth element $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ and $\mathbf{x}_0 = \mathbf{X}(\mathbf{s}_0)$.

Minimizing the prediction error variance then gives us the kriging variance which can be expressed as

$$\sigma^2(\mathbf{s}_0) = \gamma^T \Gamma^{-1} \gamma - (\mathbf{X}^T \Gamma^{-1} \gamma - \mathbf{x}_0)^T (\mathbf{X}^T \Gamma^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \Gamma^{-1} \gamma - \mathbf{x}_0).$$

[4] Usually this is an unknown point in B, but can also be a known point.

Chapter 2

Oregon Water Data

2.1 Background

A wide variety of data can be used for kriging, provided it meets a few necessary conditions. Most importantly, the data must be a sample from a spatially continuous random process. Since kriging provides a point prediction for any location within a region, a discrete or discontinuous random process cannot be used.

This thesis will analyze Oregon ground water depth using data from Oregon's Water Resources Department. The data itself is a collection of logs recorded by Oregon-bonded well drillers and includes such information as the drilling date, the depth of the well, the depth of the first occurrence of water, and flow rate. The Water Resources Department uses this data to monitor water quality throughout the state of Oregon. For the purposes of this thesis, the data represents a partial sampling from a mostly continuous supply of subterranean water. By applying kriging to the available sample, it is possible to make predictions for unsampled locations in Oregon.

Oregon's records contain nearly 500,000 wells in the state. Since the individual contractors are responsible for recording their own observations, much of the data is incomplete. Only a handful of the observations include the latitude and longitude coordinates necessary for spatial prediction. Of this subsample, a ten-year date window was selected for the sake of accuracy.

	x	y	depth
	Min. :-124.5	Min. :42.00	Min. : 0.0
	1st Qu.:-123.2	1st Qu.:42.94	1st Qu.: 30.0
	Median :-122.9	Median :43.97	Median : 82.0
	Mean :-122.1	Mean :43.81	Mean : 134.4
	3rd Qu.:-121.5	3rd Qu.:44.67	3rd Qu.: 172.0
	Max. :-116.9	Max. :46.23	Max. :1000.0

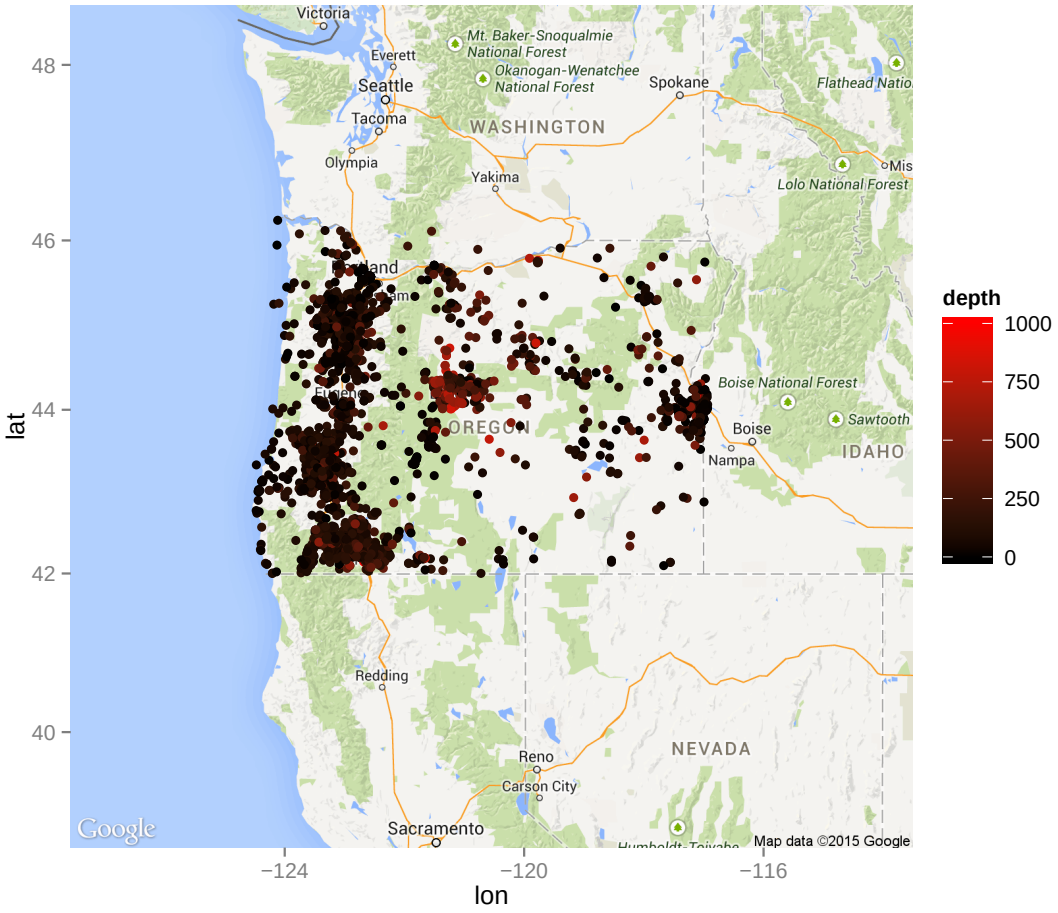

Figure 2.1: Data observations with recorded depth values

Lag distances	lag	# of pairs
20	11948.97	114487
40	15098.10	171869
60	14260.73	175836
80	13820.19	166287
100	14267.44	175306
120	16011.45	168228
140	17206.47	183247
160	28851.95	200061
180	31363.58	234284
200	36084.79	226972
220	38299.19	199070
240	38914.69	174760
260	35559.17	179645
280	29084.78	172046
300	24320.69	153338

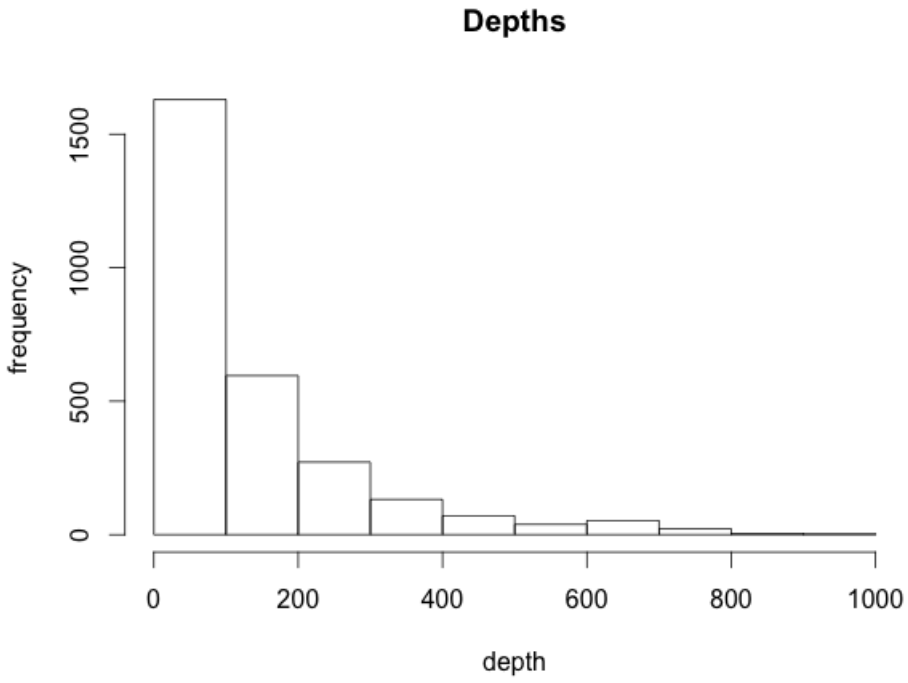

Figure 2.2: Depth frequency across data observations

Spherical variogram
tausq sigmasq phi 5402.2780 28048.6894 254.8803

Matern variogram
tausq sigmasq phi 6003.337 38428.555 198.255
Exponential variogram
tausq sigmasq phi 6003.337 38428.555 198.255
Cubic variogram
tausq sigmasq phi 9260.4886 24906.8104 316.9988

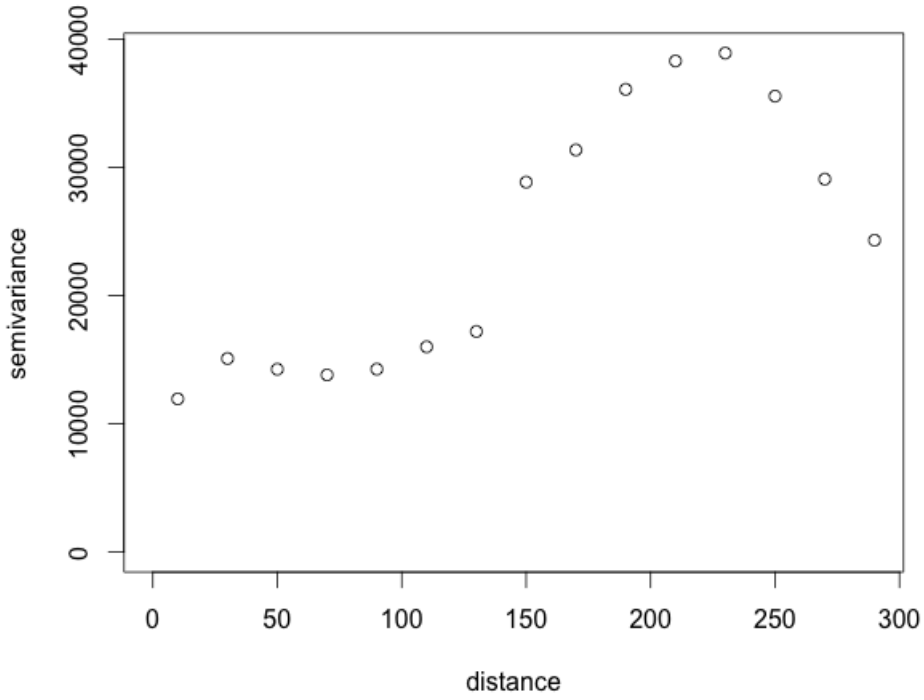

Figure 2.3: Empirical variogram with 300 km maximum distance

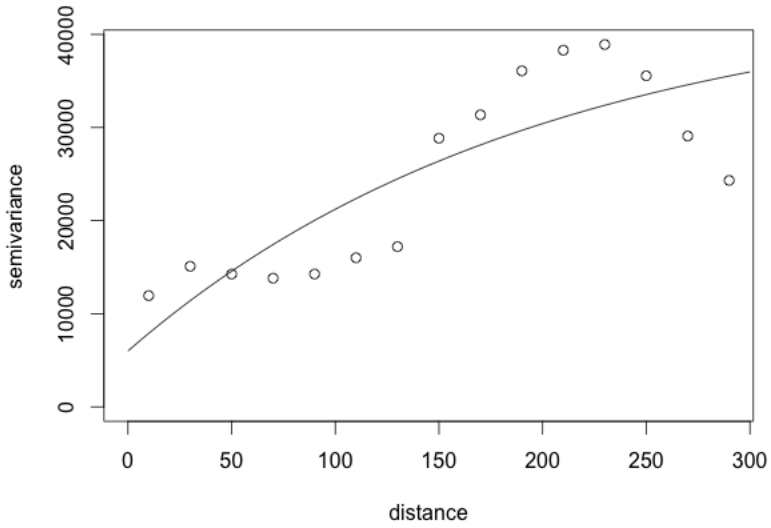

Figure 2.4: Empirical variogram with matern model

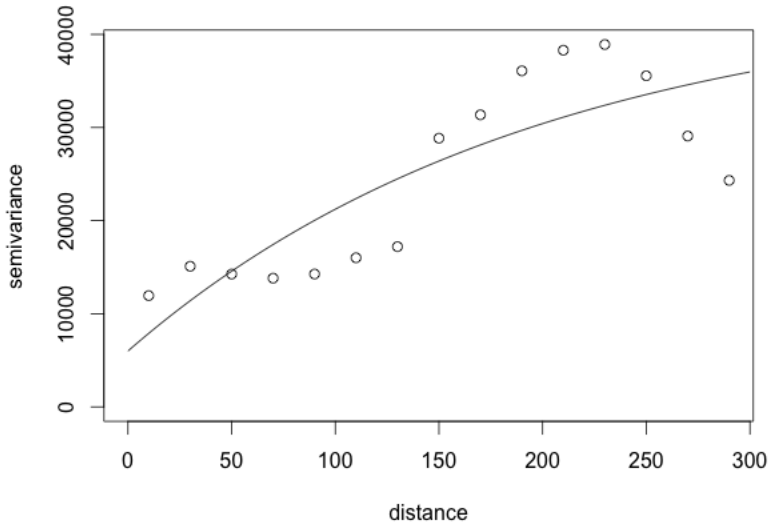

Figure 2.5: Empirical variogram with exponential model

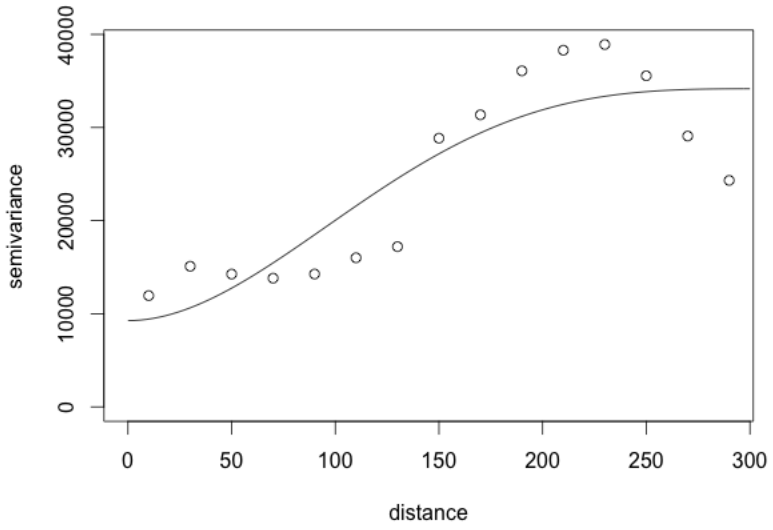

Figure 2.6: Empirical variogram with cubic model

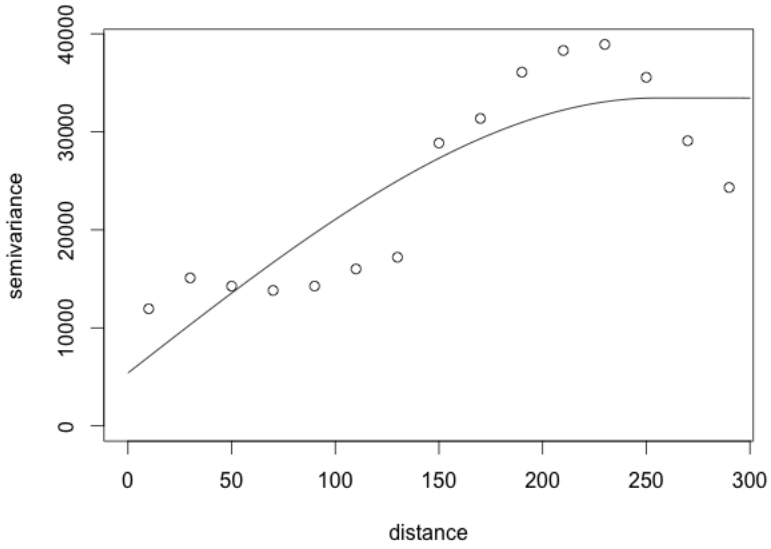

Figure 2.7: Empirical variogram with spherical model

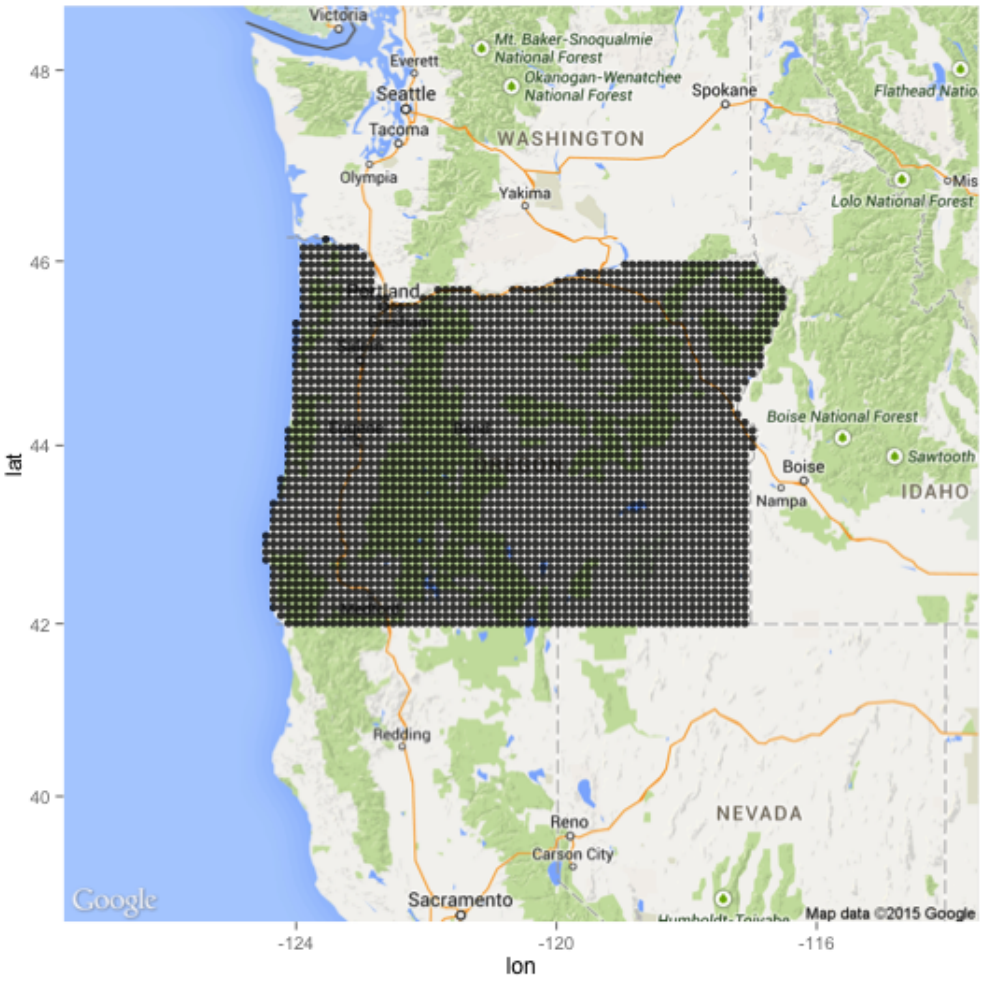

Figure 2.8: Gridded prediction points

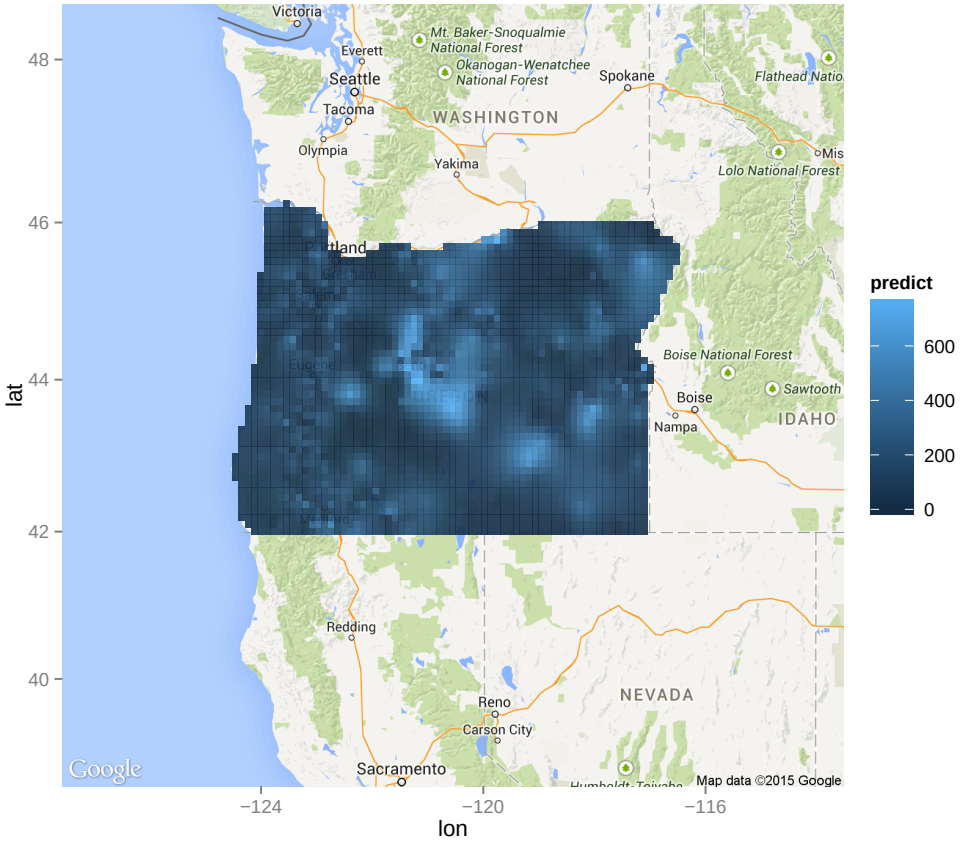

Figure 2.9: Tiled prediction grid

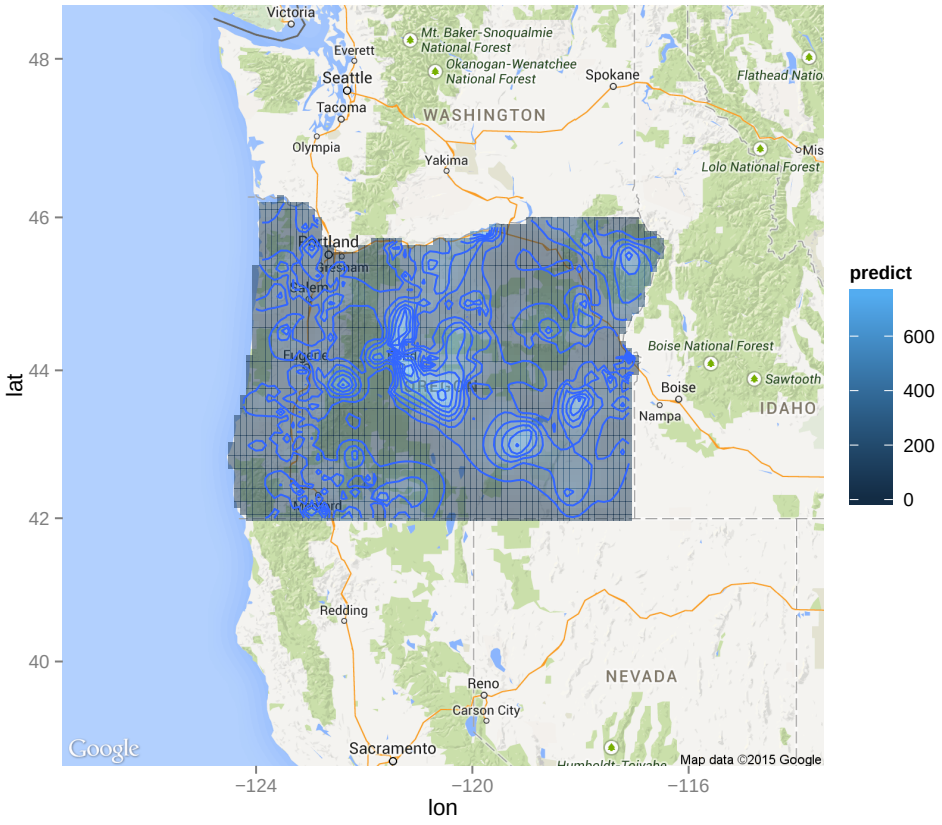

Figure 2.10: Prediction grid with contour lines

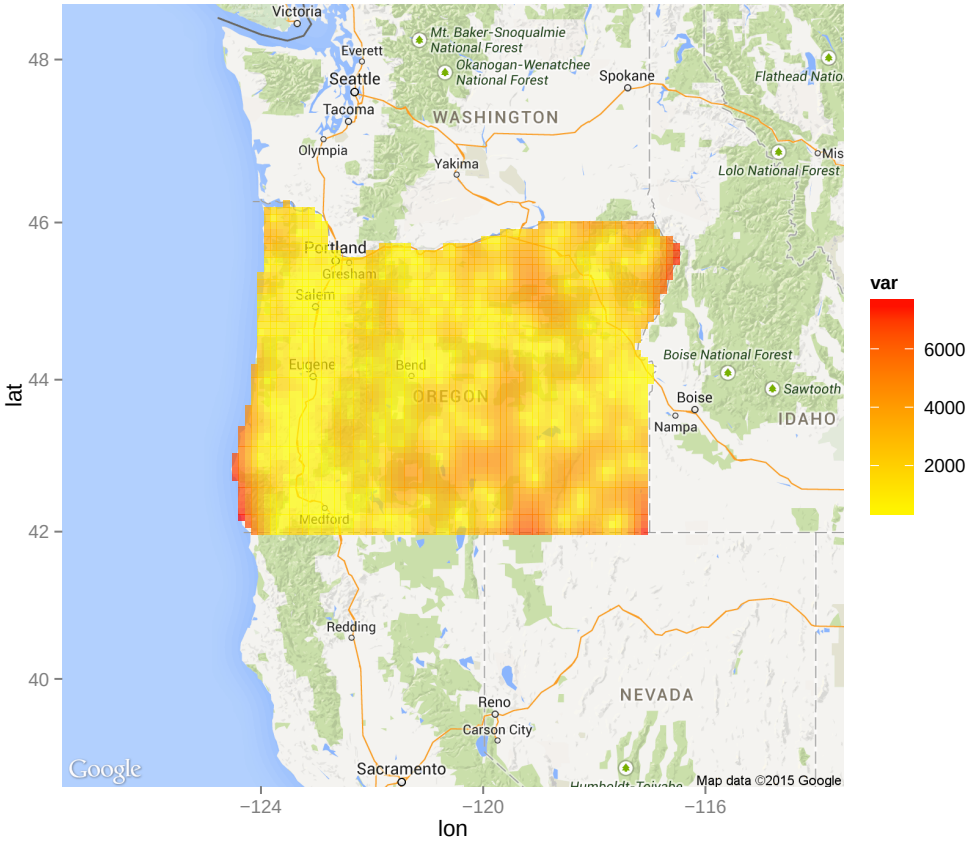

Figure 2.11: Variance grid

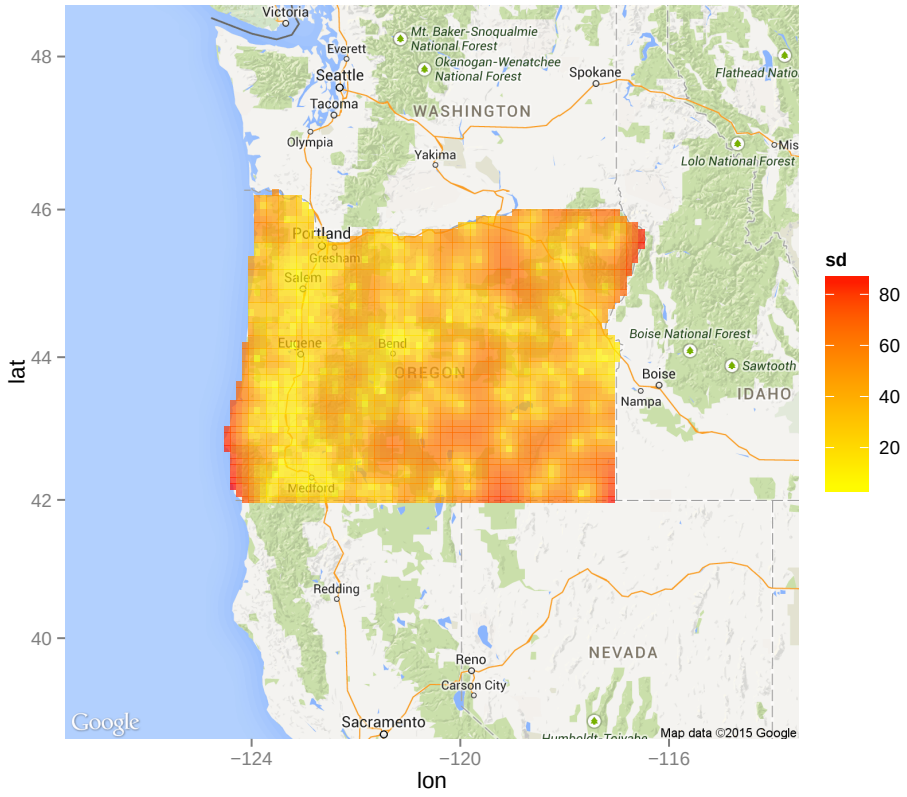

Figure 2.12: Standard deviation grid

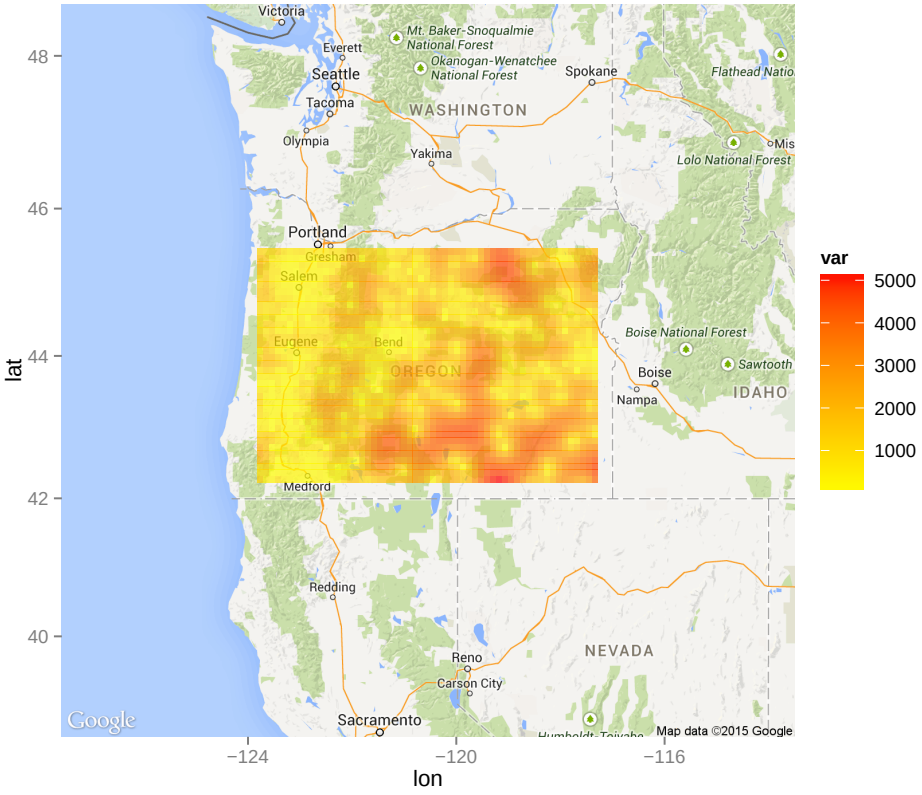

Figure 2.13: Cropped variance plot

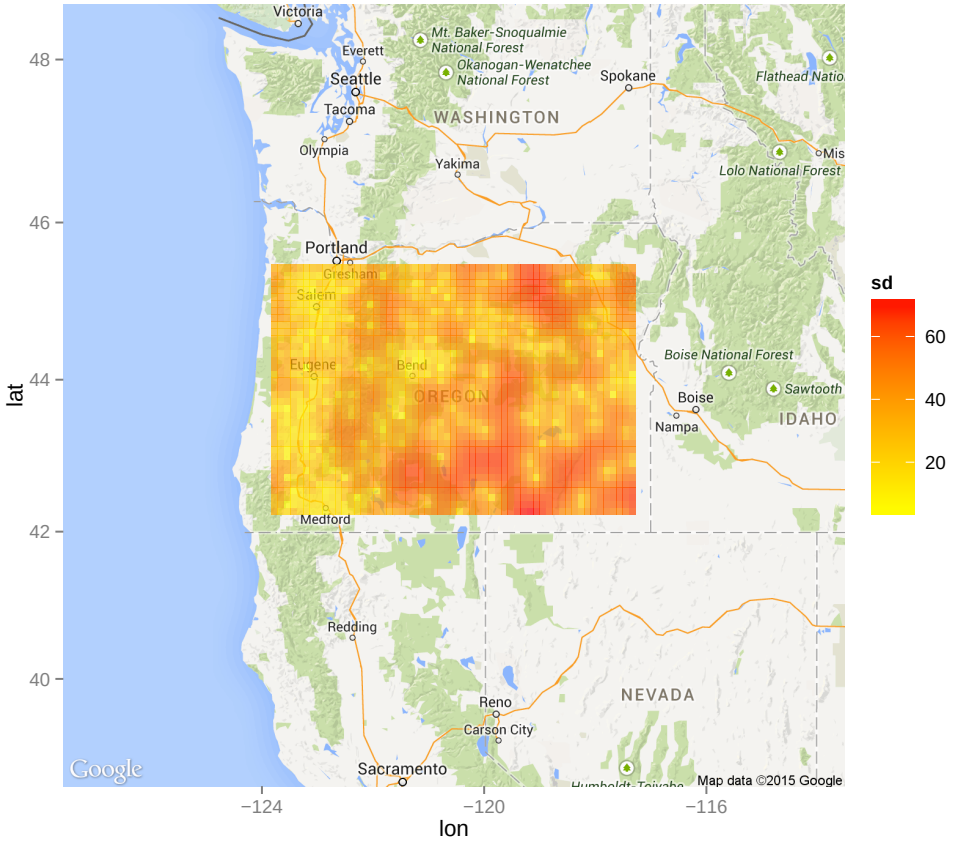

Figure 2.14: Cropped standard deviation plot

Conclusion

Appendix A

The First Appendix

Appendix B

The Second Appendix, for Fun

Bibliography

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. New York, NY: Springer.

Cressie, N. A. (1993). *Statistics for Spatial Data*. New York, NY: John Wiley and Sons, Inc.

Gelfand, A. E., Diggle, P. J., Fuentes, M., & Guttorp, P. (2010). *Handbook of Spatial Statistics*. Boca Raton, FL: Chapman and Hall.

Stein, M. (1999). *Interpolation of Spatial Data*. New York, NY: Springer.