Mapping Oregon Groundwater: a Geo-Statistical Analysis in Spatial Interpolation

---

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 Y. Kim

Acknowledgements

It often seems like the most important people are also the most under-appreciated. For this reason I'd like to thank my amazing parents for their unconditional love and support during my time at Reed and always. Parents like these are hard to come by, and literally none of this would be possible without them.

Second, to my patient and thoughtful adviser, Albert, for being just the right amounts of encouraging, demanding, and understanding. My frequent crises could very well have sabotaged my entire thesis, and working with Albert made the process a manageable and, yes, rewarding task.

Also, to Abigail for being the most consistently supportive and loving friend anyone could ask for. You are my inspiration.

Finally, to my attractive and talented friends Mark, Maren, Kiki, Will, and everyone else in the Reed community who has remained positive and determined during the rough-and-tumble of college and life. Go team.

Preface

Before my Senior year, I thought of the Thesis as being the encapsulating representation of the Reed experience. Every ounce of knowledge from every class seemed to coyly suggest its candidacy for a topic or focus during the final year. Of course, mathematics itself is a cumulative subject, with each theorem building upon its predecessors, albeit with branching concentrations and applications. Similarly, the process of learning mathematics is not a simple matter of memorizing formulas but of constantly building an understanding of a complex and vivid language. Because of this, I expected my thesis to epitomize my compound knowledge of my time at Reed – a final theorem, of sorts – and in a way it does. But ascribing such importance to a relatively small portion of my education possibly did more harm than good. Yes, some of my classmates found great joy in pouring their souls into their theses, and I wanted to be one of them, but the self-expectation to perform at the highest of my abilities usually left me suspended in a state of emotional overload. If I had simply relaxed and treated my thesis as a fun project and not an encapsulation of my abilities, both the process and the final result could have been much cleaner.

This isn't to say I'm not proud of my thesis. It's easily the most extensive project I've ever worked on[1], and I learned a lot while researching and coding my topic. However, my knowledge of mathematics is almost a bi-product of my four years at Reed. My thesis too is a corollary to a couple of classes in probability and statistics I took during my Junior year, classes I did relatively poorly in but which gave me an opportunity to explore the world of data analysis and practical programming.

So, to the handful of people who will ultimately read this thesis, this is not my great manifesto. This is not groundbreaking research. This does not advance any particular body of knowledge. This is an exposition of a relatively obscure but potentially useful segment of statistics. This is a (semi-) organized retelling of a story already known to many academics and researchers. However, there are a lot of pretty pictures.

[1] with the possible exception of planning Paideia during Fall 2013

Table of Contents

List of Tables

List of Figures

Introduction

Perhaps the most common statistical problem is inferring about some unknown truth using incomplete data. Constraints such as time and resources often make gathering every piece of information difficult or impossible, so the gaps must be filled with a best guess. With statistics, it is possible to rigorously define exactly what "best" means, so that the only way to possibly be more confident is to collect more data. This thesis will explore a geological statistical application in spatial interpolation, in which a finite number of observations over a two-dimensional surface are quilted together as to form a complete representation of the surface. For example, if a water well is dug somewhere in the state of Oregon, it may indicate the depth at which water is first reached. This singular well gives no information about water depth a kilometer away, but a cluster of wells may provide some insight into the average depth in that area and even allow one to guess the water depth at a particular location without having to drill another well.

This thesis is an exposition of a geostatistical method called *Kriging*, formulated by statisticians to predict mineral and ore distributions and expanded by Noel. A. C. Cressie in the groundbreaking book *Statistics for Spatial Data* [3]. It uses creative optimization methods to determine the best way to interpolate between points, and is far more extensive than the brief example presented here. However, this thesis provides a comprehensive overview of some of the crucial components of the Kriging method and applies it to an Oregon-sourced dataset to provide an interesting glimpse of the subterranean world.

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 [3], 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, previously developed 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 $\mu(\mathbf{s})$, where $\mathbf{s}$ is a coordinate vector with entries corresponding to the dimensions of S, and an i.i.d. random error process with zero mean $e(\mathbf{s})$. Together these specify that

[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.

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

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

Separating $\mu(\mathbf{s})$ from $e(\mathbf{s})$ allows the variance of the error process to be calculated separately from the mean. 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 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 (BLUP) for geostatistical random fields [4], with "best" meaning the function that minimizes the mean squared error of the prediction points across all possible linear models.

The purpose of the mean function is to help provide an estimate for the residuals $e(\mathbf{s})$, denoted 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

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

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

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

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 [4] and higher-order polynomials such as the quadratic allow for omnidirectional prediction calculations.

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 $\mathbf{Y} = [Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)]^T$. [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 ([4], 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 [4]. 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].

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

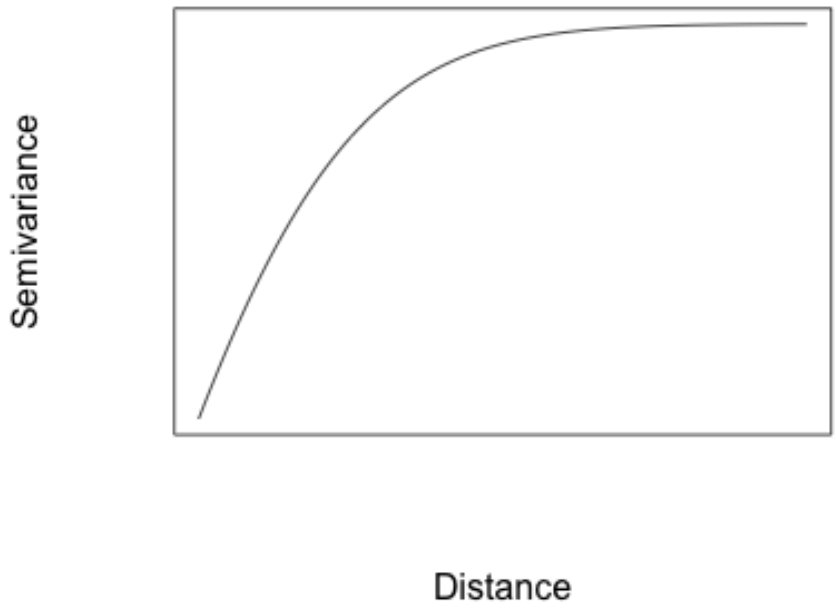

Figure 1.1: An exponential semivariogram

The distance between any two points x_i and x_j can be used to define a new set $H = \{x_i - x_j : x_i, x_j \in B\}$ of 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 correlation 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. The right number of bins depends on the specific needs of the researcher. 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 optimized semivariogram.

Fitting a smoothed parametric curve 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 [4]. 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 the model to minimize inaccuracies.

1.4 Spatial Prediction: Kriging

Given a prediction point $\mathbf{s}_0$[3], 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

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

predictors that are both (1) linear, and (2) unbiased:

1. $\hat{Y}(\mathbf{s}_0) = \boldsymbol{\lambda}^T\mathbf{Y}$, where $\boldsymbol{\lambda}$ 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, $\boldsymbol{\lambda}\mathbf{X} = \mathbf{X}(\mathbf{s}_0)$.

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 $Y(\mathbf{s}_0)$. Recall from Section 1.2 that "best" means having the smallest mean squared error within the class of linear unbiased predictors. There are several ways of solving this. Cressie [3] gives a proof using differential calculus and Lagrange multipliers, while Zimmerman and Stein [4] 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 $\mathbf{x}_0 = \mathbf{X}(\mathbf{s}_0)$.

From this equation and the empirical variogram it is then possible to calculate the prediction error associated with each point. Minimizing this prediction error variance then gives us the Kriging variance which can be expressed as

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

This should not be confused with the prediction error variance itself, however, which is our confidence in the prediction due to the error term. The Kriging variance instead indicates how reliable the prediction is at a specific prediction 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 [7]. 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.

2.1.1 The Data

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 from 2005 to 2015 was selected for the sake of accuracy.

In addition to this, a few other error-correction edits were made to the data. Several lat/long locations placed points off the coast and these were removed from the sample. The presence of these points may indicate a larger trend of recording errors

among inland points, but since determining this error would be nearly impossible, the recorded values were taken at face value in the initial stages. The Kriging process, however, does give some options for accounting for error and this is explored later.

Another edit made to the data was removing co-located points. Several points (n=15) had the same lat/long entries but different recorded depth values. While these points could be used to calculate the variogram's nugget effect (described in Section 2.2), their relatively rare frequency taken with the already established recording errors found in the sample warranted their exclusion. In addition to this, Oregon's records impose an upper bound of 1000 feet to the depth of wells drilled.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	30.0	82.0	134.4	172.0	1000.0

Table 2.1: Summary of groundwater depths across all observations

Table 2.1.1 gives a few preliminary stats on the data. The longitude and latitude variables correspond loosely with Oregon's bounding perimeter, and the depth values range from zero to 1000 feet. Figure 2.1 shows the histogram of recorded depths. Of particular note is the high frequency of shallower (< 100 feet) wells across the state.

Finally, Figure 2.2 plots all observations over a map of Oregon, with color corresponding to depth. The clustering along the West side of the state is due to statewide population densities, with more Oregon residents living near the I-5 highway that runs North/South from Washington to California. This clustering becomes important later when we look at the predicted Kriging variances and compare it to the the density map.

2.2 The Variogram

As described in section 1.3, the variogram plots the degree of correlation between points at varying distances. This correlation is then used to predict a depth at an unsampled location ($\mathbf{s}_0$) by comparing its distance to known locations to the lags, or specific variances, associated with those distances. The degree to which each lag agrees on the predicted value of $\mathbf{s}_0$ then determines the Kriging variance.

Several choices go into the process of plotting the variogram. First, an effective range must be established in order to determine the maximum separation distance that will be used for the calculation. As shown in Table 2.2, the average distance between data points is 237.8 kilometers while the maximum is 719 kilometers. For this reason, setting the maximum variogram distance to 300 kilometers as in Figure

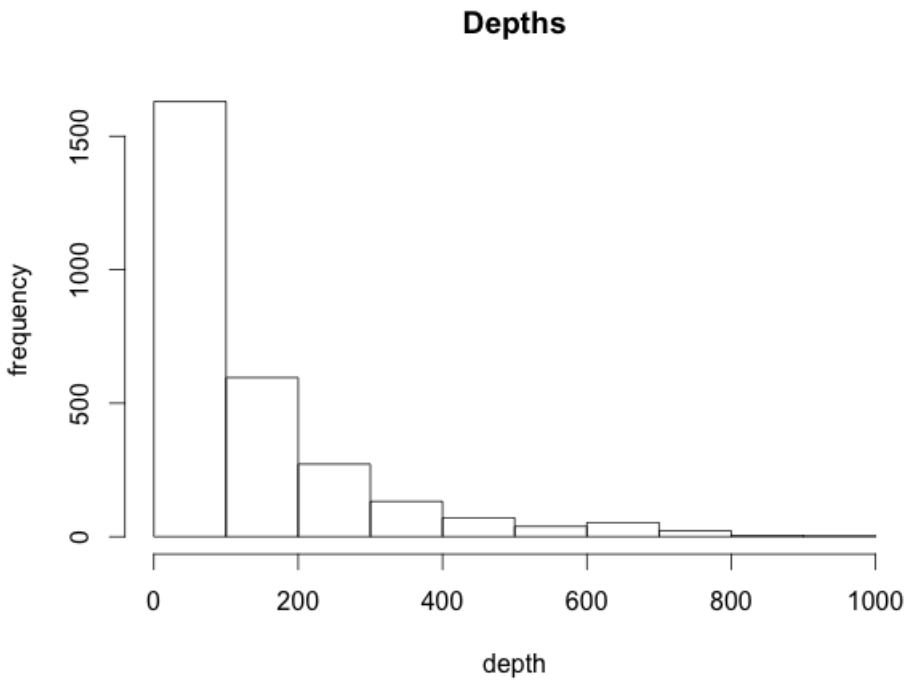

Figure 2.1: Depth frequency across data observations

2.3 cuts out a lot of noise from distant and unrelated pairs while still including as many pairs as possible.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	118	209	237	330	719

Table 2.2: Cartesian distances across data in kilometers

The second main choice is how many bins to use. Again, the tradeoff between more or fewer bins is accuracy vs. precision. More bins means a greater number of distances is represented in the variogram, and fewer bins means each bin has a lower sampling variation. In most geostatistical applications, anywhere between 10 and 30 bins is typically used [4]. For the purposes of the Oregon groundwater data, a binning number of 15 allows for 20 kilometers of separation between lags and between 115,000 and 240,000 pairs per bin.

The third necessary decision is whether to modify the data to represent accurate spatial distance. Since latitude and longitude are a two-dimensional projection on a curved surface, latitudes are equidistant, but the distances between longitudes varies [2]. Because of this, distances between two points, (x_1, y_1) and (x_2, y_2), will not be accurate with respect to Euclidian distance, $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Convert-

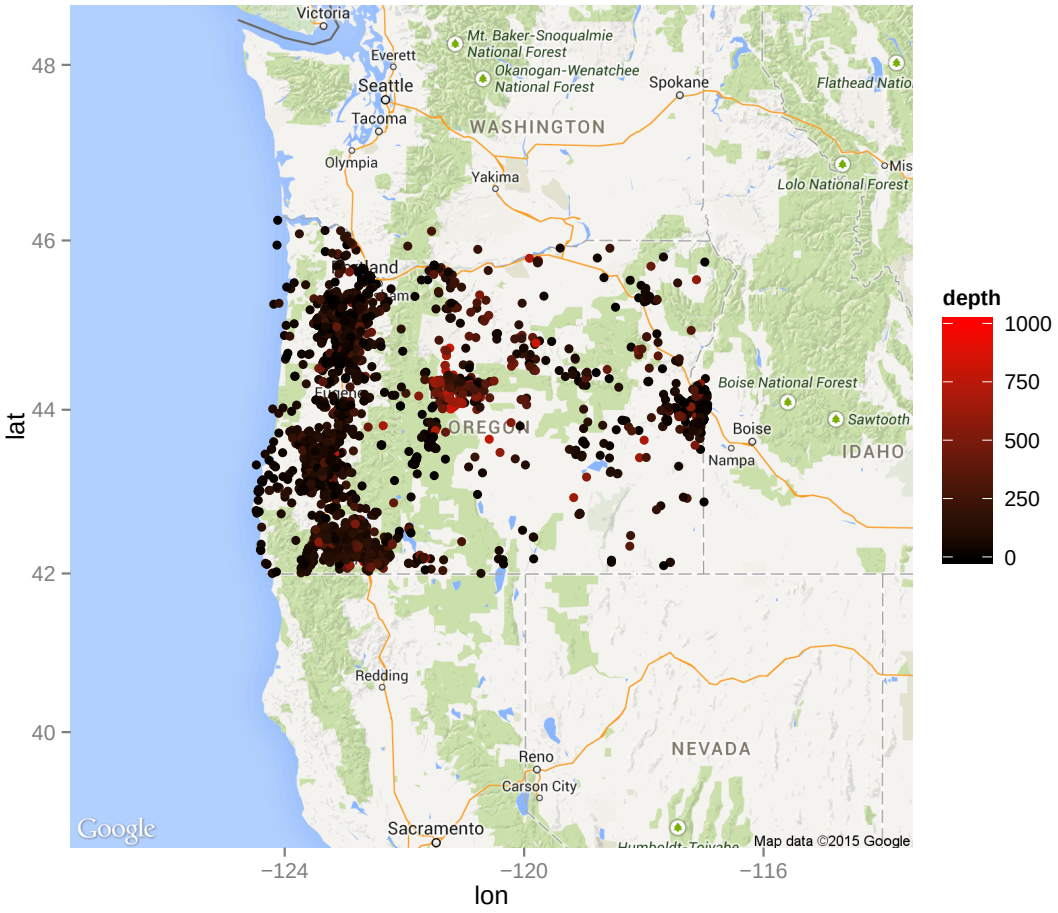

Figure 2.2: Data observations with recorded depth values

ing to a grid-based coordinate system such as the Universal Transverse Mercator projection solves any problems this may cause.

Table 2.3 shows the total number of pairs and variances for each bin. When plotted as in Figure 2.3, we see a trend similar to that in Figure 1.1. This is the empirical semivariogram and is the foundation for the fitted model.

2.2.1 Fitting a model

Variograms in general tend to increase roughly with distance. Because of this, the most common fitted models are ones that increase monotonically, but this is not a necessary condition. As described in Chapter 1, a smoothed parametric curve provides a continuously lagged, nonpositive definite model with lower variance than the empirical variogram. Any smoothing model must meet the three conditions from Chapter 1, namely 1) vanishing at zero, 2) evenness, and 3) conditional negative

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

Table 2.3: Variogram binning

definiteness. The most common tends to be the Matérn model, given by

$$\gamma(h) = \theta_1 \left(1 - \frac{(h/\theta_2)^\nu \kappa_\nu(h/\theta_2)}{2^{\nu-1}\Gamma(\nu)}\right)$$

where κ_ν is the modified Bessel function of the second kind of order ν [4] and (θ_1, θ_2) is a parameter vector to be optimized.

Figure 2.4 shows the empirical variogram from Figure 2.3 fitted with several common models, including the Matérn. The fit was calculated using Weighted Least Squares (WLS) given by

$$\hat{\theta} = \text{argmin} \sum_{h \in H_u} \frac{n(\mathbf{h}_u)}{[\gamma(\mathbf{h}_u)]^2} [\hat{\gamma}(\mathbf{h}_u) - \gamma(\mathbf{h}_u)]^2$$

with all variables defined as in Chapter 1. Note that either a large $\gamma(\mathbf{h}_u)$ or a small $n(\mathbf{h}_u)$ corresponds to smaller weights. This means that larger lags are given less weight than smaller lags.

When choosing a model, a few particular attributes play an important role. Notably the *sill*, *range*, and *nugget* of the model will significantly affect the Kriging calculation:

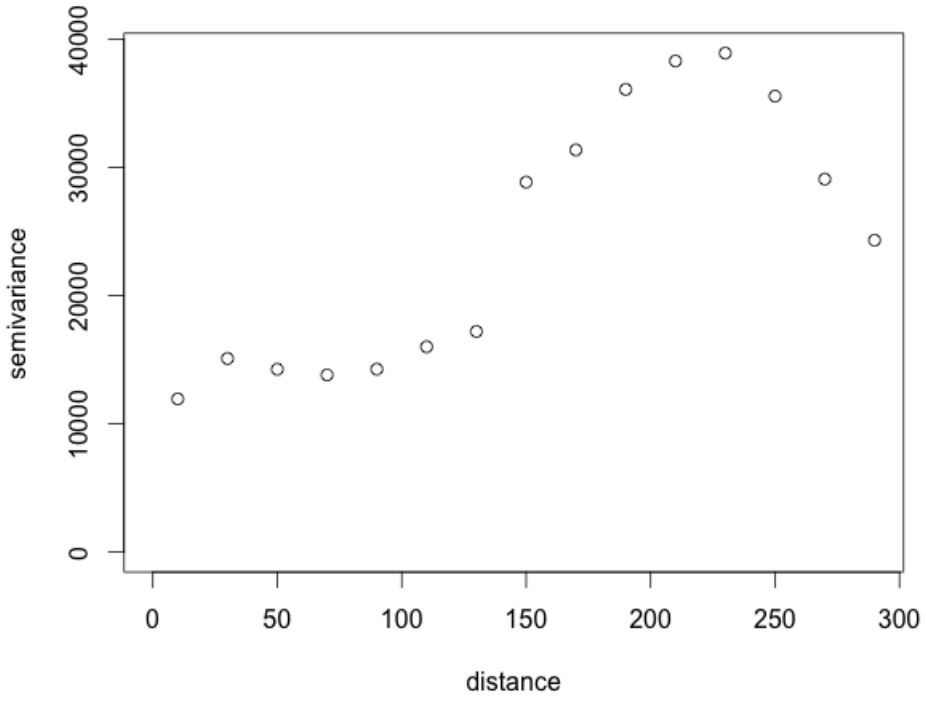

Figure 2.3: Empirical variogram with 300 km maximum distance

- The sill, denoted σ^2, is effectually the highest point in the variogram, or the maximum variance across all pairs of data points (i.e. $\lim_{h\to\infty}\gamma(h)$ provided the limit exists). (For the Matérn model this corresponds with θ_1)

- The range, denoted ϕ, is the smallest value of $\mathbf{h}$ for which the variogram equals its sill.

- The nugget, with variance denoted by τ^2, is defined as $\lim_{h\to 0}\gamma(h)$, or the y-intercept, and can be thought of as a measurement error that accounts for differing data values at very close or co-located points, even though theoretically there should be no variability.

These three attributes are generally considered just as important as closeness-of-fit when choosing a variogram model.

Table 2.4 gives the optimized parameter vectors for each model as fitted to the empirical variogram from this section. Though all are good fits, the Kriging step in the following section uses the spherical model, given by

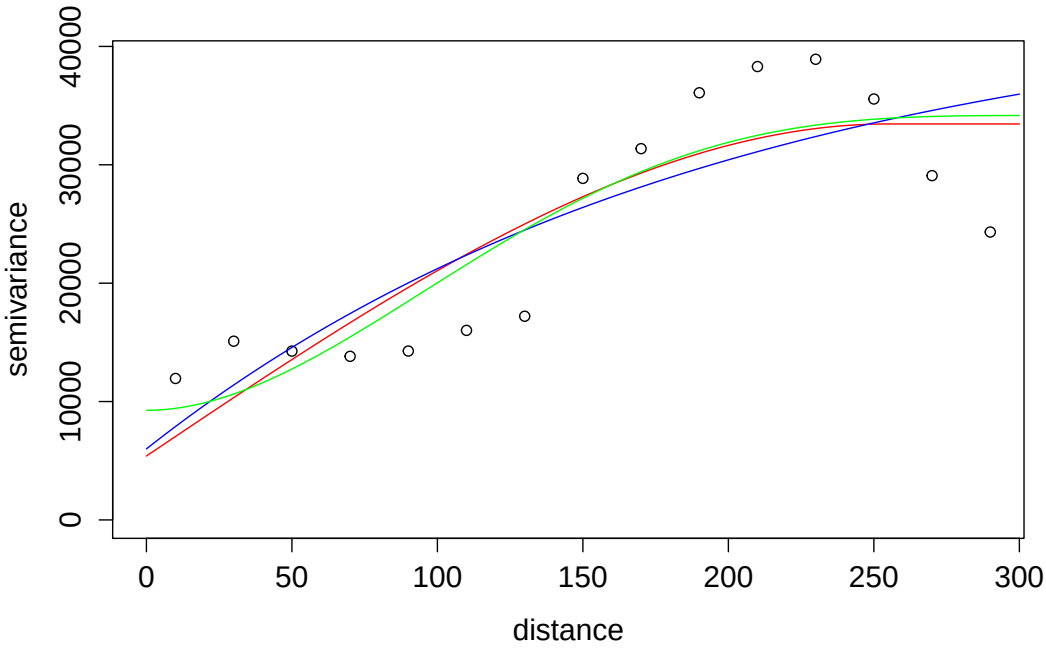

Figure 2.4: Empirical variogram with Matérn (blue), spherical (red), and cubic (green) models.

$$\gamma(h) = \theta_1 \left(\frac{3h}{2\theta_2} - \frac{h^3}{2\theta_2^3} \right)$$

for $0 \leq h \leq \theta_2$ and 0 for $h > \theta_2$.

Spherical	τ^2: 5402	σ^2: 28048	ϕ: 254
Matérn	τ^2: 6003	σ^2: 38428	ϕ: 198
Cubic	τ^2: 9260	σ^2: 24906	ϕ: 316

Table 2.4: Optimized parameters for several fitted variogram models

2.3 Prediction

Once a working model has been established for the variogram, the prediction step is relatively straightforward. Using the equation from Chapter 1, and the variogram model from the previous section, we can define a grid of points over a prediction

region and use it to build a topographical map.

Figure 2.5 shows section of a grid of points constrained to the boundary of Oregon. These will be the prediction locations for the Kriging function. By further dividing the region into prediction blocks rather than points as in Figure 2.6, it is possible to color each block with the prediction value associated with that block.

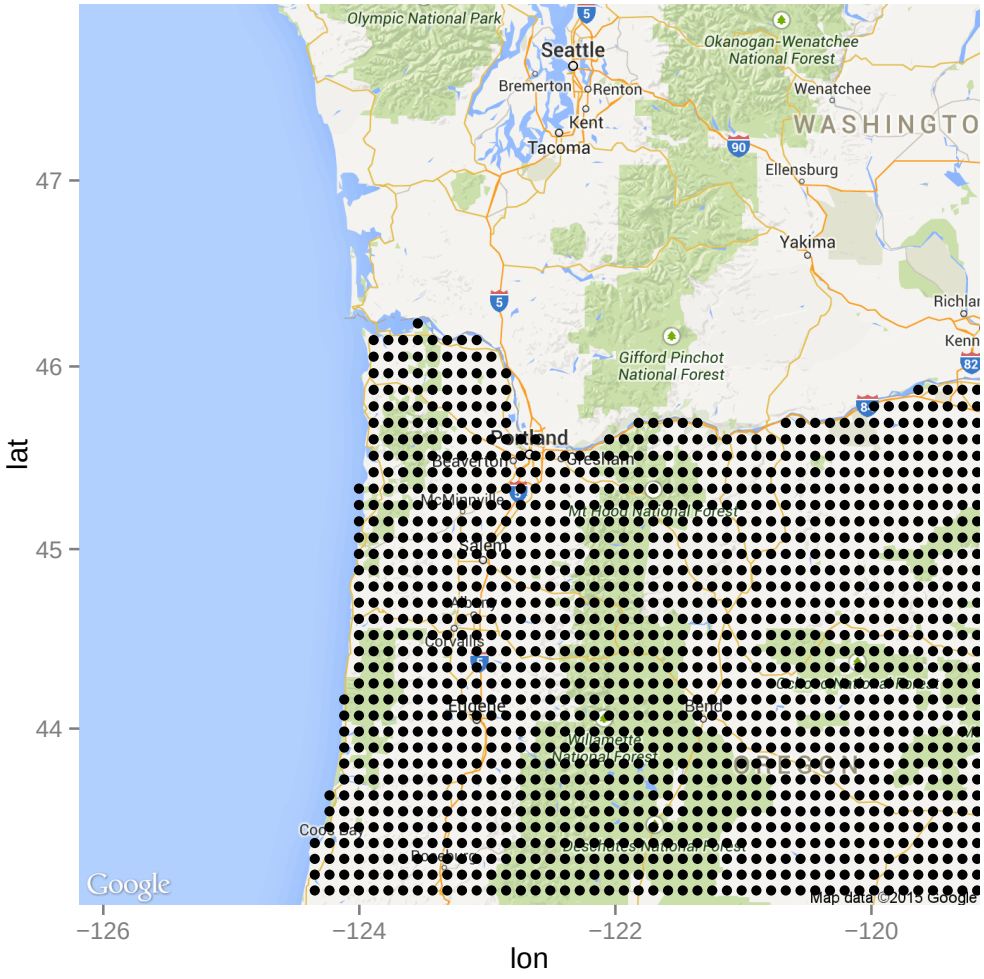

Figure 2.5: Close-up of gridded prediction points

Finally, mapping out contours as in Figure 2.7 culminates the process by giving a comprehensive visual estimate of the groundwater throughout the state of Oregon, with lighter colors corresponding to deeper water.

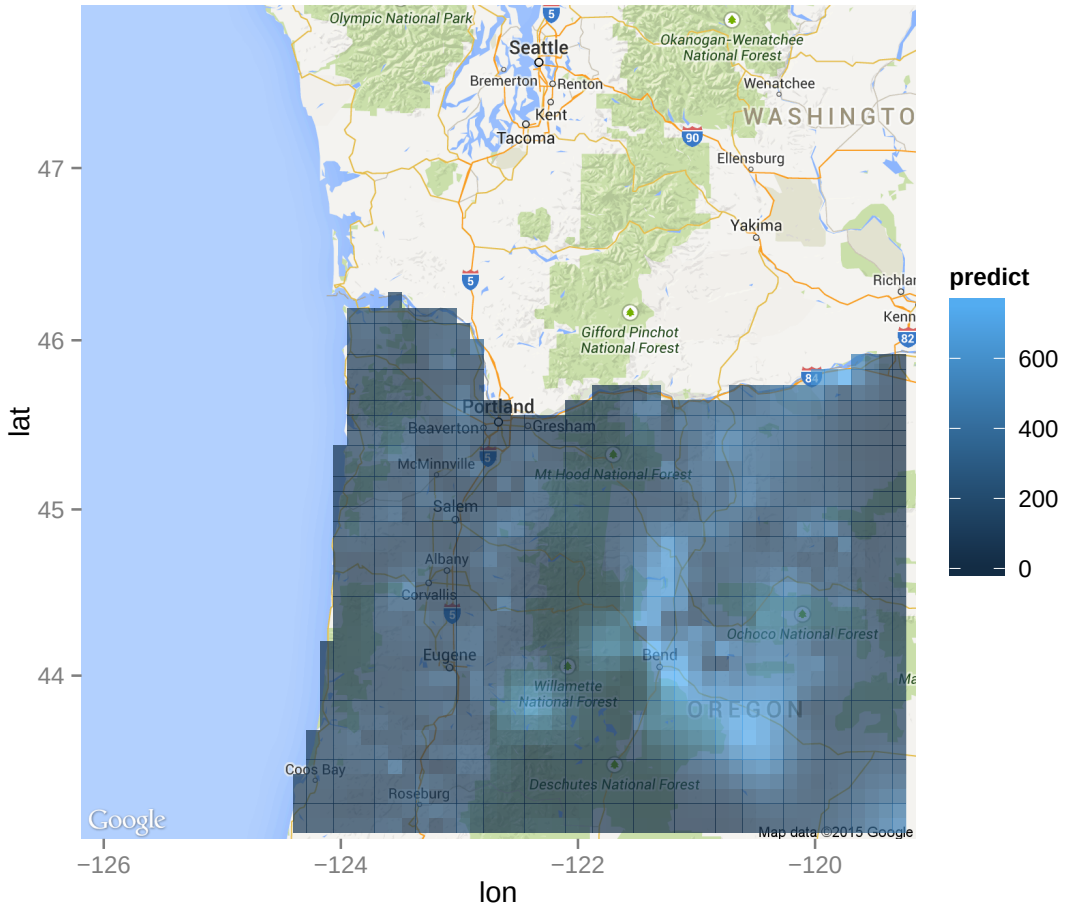

Figure 2.6: Close-up of tiled grid with prediction values of first depth

2.3.1 Prediction variance

As described in Chapter 1, Kriging also provides a variance for each prediction point based on agreement between each lag of the semivariogram. Taking the square root of the variance then gives the standard deviation which can then be plotted similarly to the prediction region itself. Figure 2.8 maps the relative certainty of the prediction points produced by the Kriging function. Comparing this to Figure 2.2 demonstrates how the clustering along the Westmost edge of the state provided the most prediction confidence in that region, with higher variances occurring in subregions populated by fewer points. In other words, the prediction variance roughly corresponds to the inverse of the density of observations, similarly to the standard error of the sample mean $\frac{\sigma}{\sqrt{n}}$.

An interesting effect found in Figure 2.8 is the fringe effects around the border of the prediction region. Kriging is inherently an interpolation, rather than extrapola-

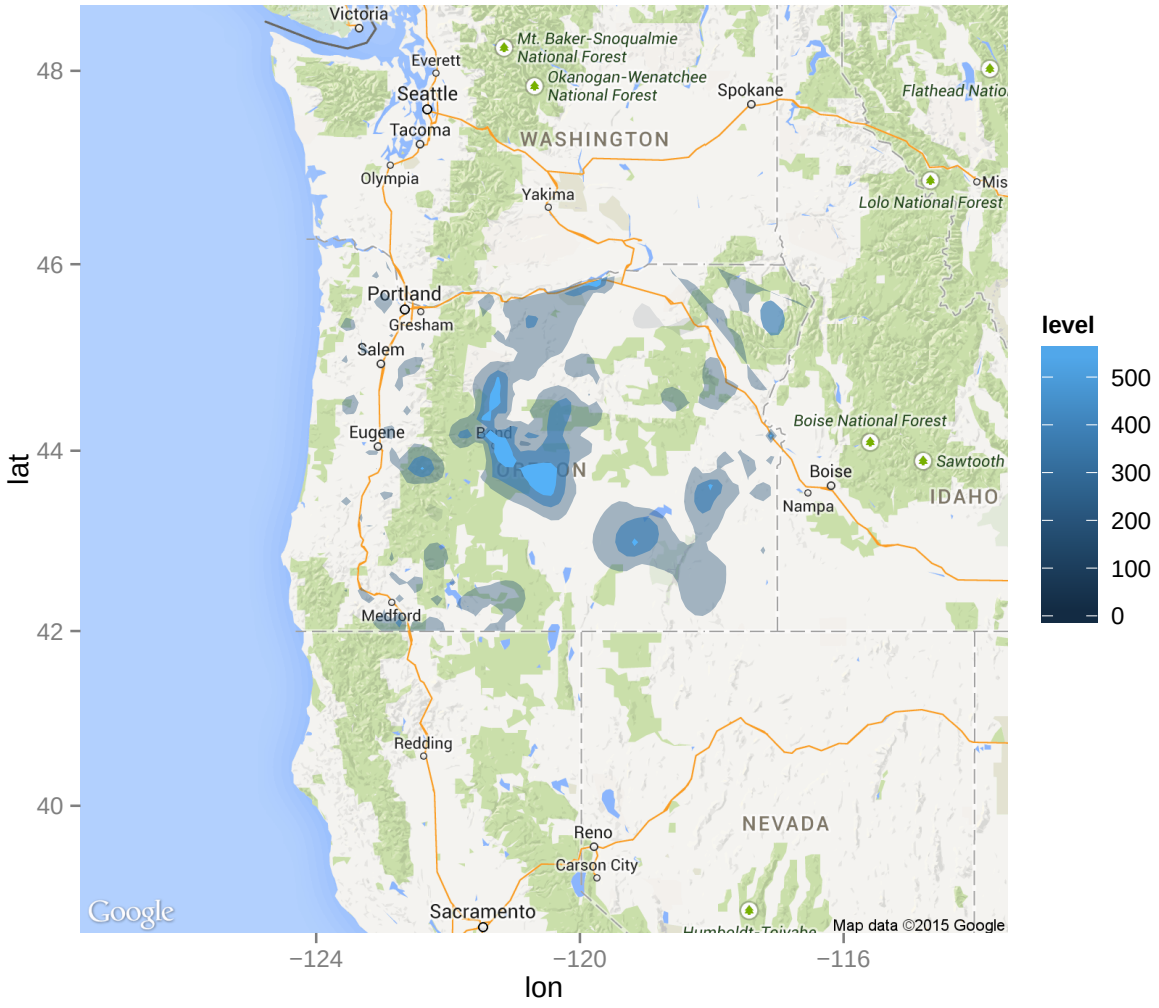

Figure 2.7: Prediction map with contour regions of first-water depth

tion method. Points outside the original bounding region will experience significantly more variance than points further inside the prediction region, even when compared to areas of lower density. For this reason, cropping the border provides a greater contrast to the grid and allows the density effects to be more apparent.

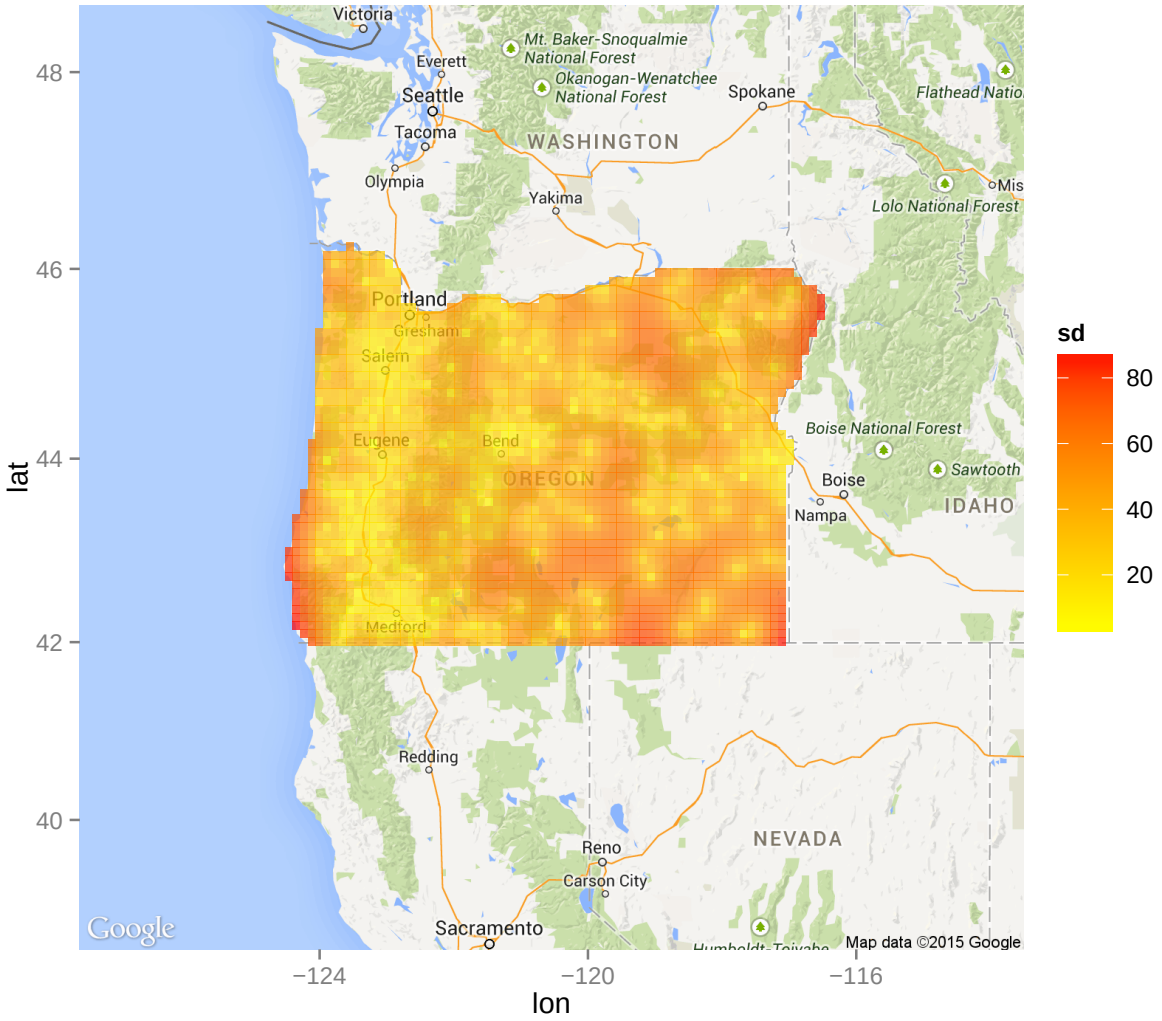

Figure 2.8: Standard deviations plot of prediction certainty at each prediction point

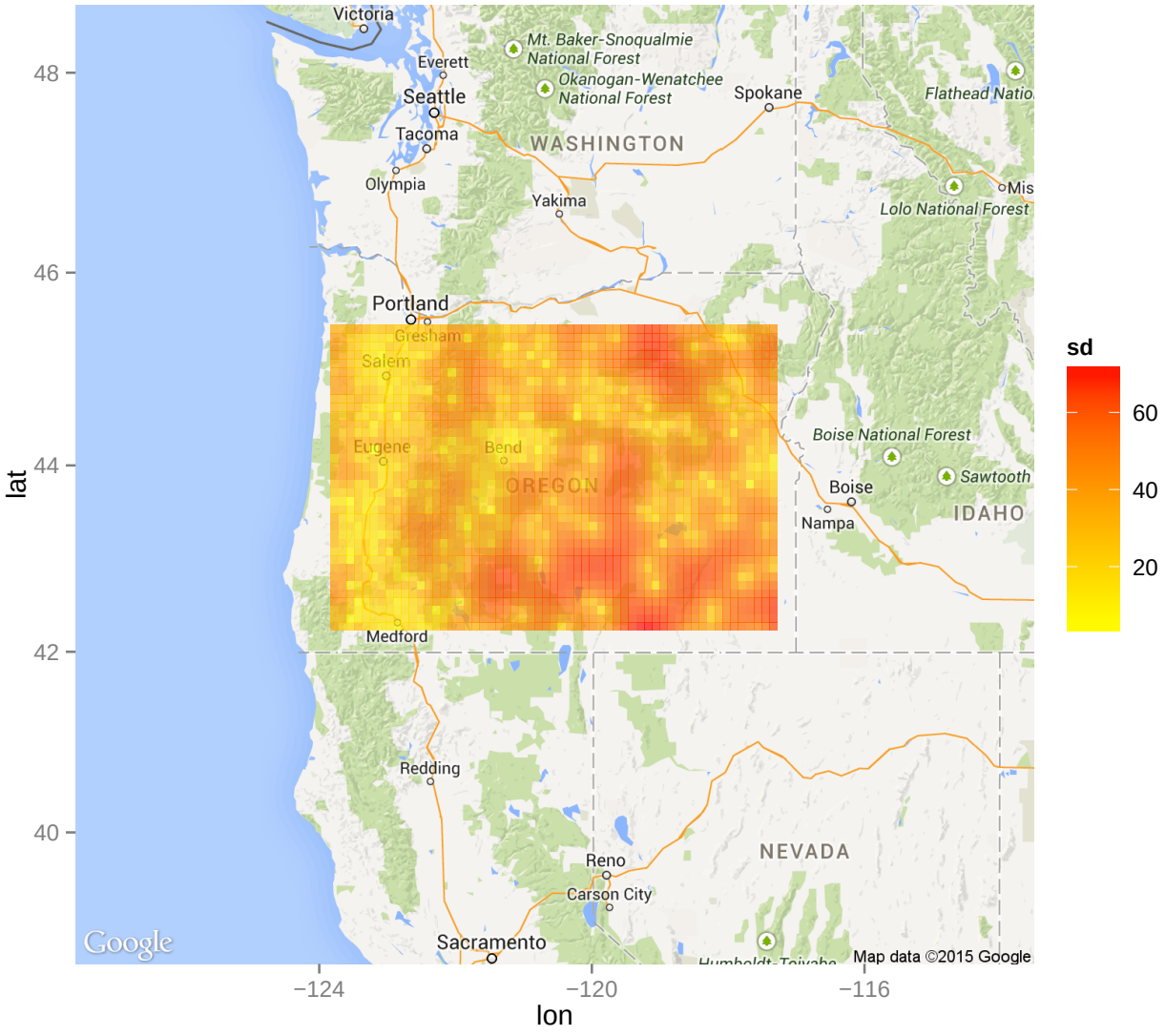

Figure 2.9: Cropped standard deviation plot showing prediction certainty at prediction locations

Conclusion

Groundwater is an essential part of any natural ecosystem. As part of the hydrological cycle of evaporation and precipitation, groundwater represents a relatively constant measure of total water within a region because of its relative immunity to the frequent changes occurring at the surface level [7]. Because of this, it is an excellent candidate for spatial interpolation and prediction. By measuring and mapping the distribution of groundwater throughout the state of Oregon, water quality regulators may accurately assess the health of surface vegetation and other factors. Since groundwater is also closely related to the type of surrounding soil, with certain soil types being more or less conducive to water storage, water maps also may provide a way of determining geological factors across a region. From an even more pragmatic standpoint, Kriging with well data provides useful information when digging new wells – someone drilling near Bend, Oregon may expect to drill several hundred feet before hitting water.

This data also indicates that the center of Oregon is by far the driest portion of the state. This thesis looked specifically at the depth at which water was first recorded. The resulting map, therefore, actually depicts Oregon's *water table* – a two-dimensional sheet of subterranean water. Other variables that could have been subjected to the same process include total depth or flow rate, though these are usually correlated with the water table itself and would have likely resulted in a similar map.

Overall, Kriging proved to be an effective process for analyzing this type of data. While further analyses of different variograms, alternate optimization methods, or more complex mean functions could be explored, the standard Kriging process born from the South African mining industry provides an accurate and informative depiction of Oregon groundwater.

Appendix: R code and packages

The majority of the work for this thesis was done with R, an open source software program for statistical programming. The bulk of the Kriging process was completed with the help of the **geoR** package [5] and its programmed variogram, Kriging, and variance functions. For the maps and graphics, **ggmap** [6] and **ggplot2** [9] provided a clean and functional way to visually depict the results over a map of Oregon. The original dataset from the Oregon.gov Water Resources Department contained nearly 500,000 data points and a significant amount of extraneous information. For this, the R package **dplyr** [10] proved to be an invaluable tool for paring down the original dataset to something manageable. Below is a partial log of many of the important steps taken to produce the results found in this thesis; the complete R code can be found at this project's GitHub repository `https://github.com/rosenbl/Thesis` (file name = kriging.R) along with the original data files and edited data.

```
library(ggmap)
library(ggplot2)
library(SpatialEpi)
library(geoR)
library(dplyr)
```

```
latlong <- read.csv("latlong.csv")
grid <- read.csv("grid2.csv")
oregonborder <- read.csv("oregonborder3.csv")

Map <- get_map(location = "Oregon",
               color = "color", # or bw
               source = "google",
```

```
               maptype = "roadmap",
               zoom = 6)

gm <- ggmap(Map,
      extent = "panel",
      ylab = "Latitude",
      xlab = "Longitude",
      legend = "right")

gm + geom_point(data = latlong,
             #color = "red",
             aes(x = x, y = y, color = depth))
             + scale_colour_gradient(low="black", high="red")

# summary of data
plot(grid$x, grid$y)
dim(grid)

dists <- dist(grid[,1:2])
summary(dists)

# set up data for variogram
geogrid <- as.geodata(grid)
geogrid$kappa <- 0.5
geogrid$lambda <- 1
geogrid$cov.model <- "spherical"

breaks <- seq(from = 0, to = 300, by = 300/15)

# make variogram
variogram <- variog(geogrid, breaks = breaks, option = "bin")

plot(variogram, main = "Variogram")


v.summary <- cbind(cbind(breaks[2:16]), variogram$v, variogram$n)
```

```
colnames(v.summary) <- c("lag", "semi-variance", "# of pairs")

v.summary
plot(v.summary[,1], v.summary[,3], main="Lag distances",
xlab="lag", ylab="# of pairs")

# fit model
fit <- variofit(variogram,
                ini.cov.pars=c(40000,225),
                cov.model="spherical",
                fix.nugget=FALSE,
                max.dist=300)
summary(fit)
lines(fit)


# choose prediction regions based on coordinate mins/maxes
summary(grid)

loc <- expand.grid(seq(-10885, -10070, by=10), seq(4668, 5200, by=10))

border <- latlong2grid(oregonborder)

ins <- locations.inside(loc, border)

# Kriging step, provide sigmasq/phi vector from fitted variogram
krige <- krige.conv(geogrid, loc=ins,
krige=krige.control(cov.pars=c(28048.6894, 254.8803)))


resultsxy <- data.frame(
  x = ins$Var1,
  y = ins$Var2
  )
```

```
resultslatlong <- grid2latlong(resultsxy)

results <- data.frame(
  x = resultslatlong$x,
  y = resultslatlong$y,
  predict = krige$predict,
  var = krige$krige.var
)

# build the maps

# prediction points only
gm + geom_point(data=results, alpha=0.5, aes(x,y))

# prediction points with predicted values
gm + geom_point(data=results, alpha=0.7, aes(x, y, color=predict))

# tiled prediction map
tile <- gm + geom_tile(data=results, alpha=0.5, aes(x, y, fill=predict))
tile

# with contour plot
contour <- tile + geom_contour(data=results, aes(x,y, z=predict))
contour

# polygon plot
poly <- gm + stat_contour(data=results,
                          geom="polygon",
                          bins=4,
                          aes(x,y, z=predict, fill=..level.., alpha = ..level..))

poly + guides(alpha="none")

# variance map
gm + geom_tile(data=results, alpha=0.7,
               aes(x,y, fill=var)) + scale_fill_gradient(low="yellow", high="red")
```

```
# standard deviation map
sd <- sqrt(results$var)

gm + geom_tile(data=results, alpha=0.7, aes(x,y, fill=sd))
+ scale_fill_gradient(low="yellow", high="red")

# cropped variance map
crop <- filter(results, x > -123.8) %>% filter(x < -117.3)
%>% filter(y > 42.2) %>% filter(y < 45.5)

gm + geom_tile(data=crop, alpha=0.7,
               aes(x,y, fill=var))
               + scale_fill_gradient(low="yellow", high="red")

# cropped sd map
sd <- sqrt(crop$var)
gm + geom_tile(data=crop, alpha=0.7, aes(x,y, fill=sd))
+ scale_fill_gradient(low="yellow", high="red")

# some more data visualization
hist(latlong$depth, main="Depths", xlab="depth", ylab="frequency")
```

Bibliography

[1] Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer, New York, NY, 2008.

[2] Cici Chen, Albert Y. Kim, Michelle Ross, and Jon Wakefield. *SpatialEpi: Methods and Data for Spatial Epidemiology*, 2014. R package version 1.2.1.

[3] Noel A.C. Cressie. *Statistics for Spatial Data*. John Wiley and Sons, Inc, New York, NY, 1993.

[4] Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes, and Peter Guttorp. *Handbook of Spatial Statistics*. Chapman and Hall, Boca Raton, FL, 2010.

[5] Paulo J. Ribeiro Jr and Peter J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):14–18, June 2001. ISSN 1609-3631.

[6] David Kahle and Hadley Wickham. *ggmap: A package for spatial visualization with Google Maps and OpenStreetMap*, 2013. R package version 2.3.

[7] Oregon Department of Environmental Quality. Water quality: Groundwater protection. 2015.

[8] Michael Stein. *Interpolation of Spatial Data*. Springer, New York, NY, 1999.

[9] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[10] Hadley Wickham and Romain Francois. *dplyr: a grammar of data manipulation*, 2014. R package version 0.2.

[11] Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.