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 Cressie (1993), 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 Gelfand et al. (2010).

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 Gelfand et al. (2010) 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

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

Zimmerman and Stein (Gelfand et al. (2010), 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 Gelfand et al. (2010). 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].

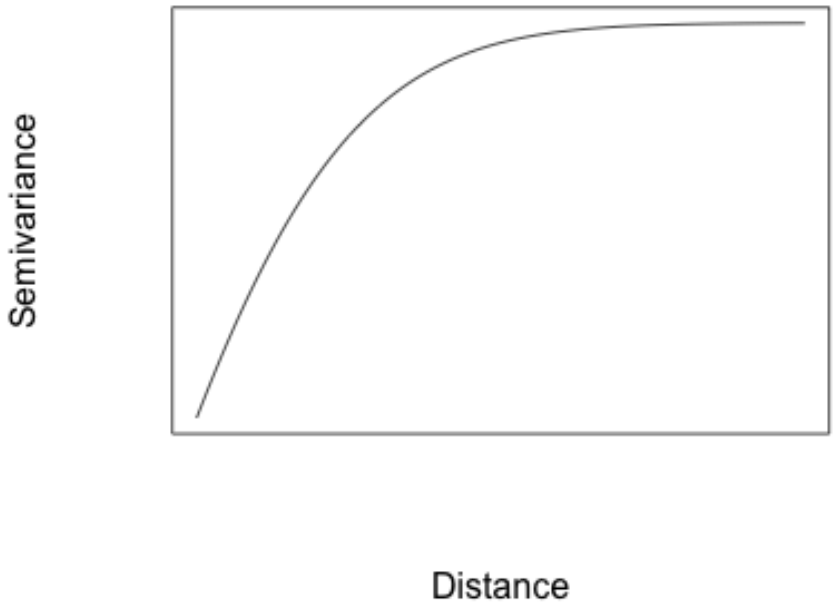

Figure 1.1: An exponential semivariogram

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 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 $\mathbf{s}_0$ [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 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 Cressie (1993) gives a proof using differential calculus and Lagrange multipliers, while Zimmerman and Stein Gelfand et al. (2010) 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

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

[3] 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.

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 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 theses 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 for 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, their relatively rare frequency taken with the already established recording errors found in the sample warranted their exclusion.

Longitude	Latitude	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

Table 2.1: Summary of groundwater data variables

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 frequency of recorded depths across all observations. 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 the superimposed standard deviation 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 associated with those distances. The degree to which the relative lags agree 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 2.3 cuts out a lot of noise from distant and unrelated pairs while still including as many pairs as possible.

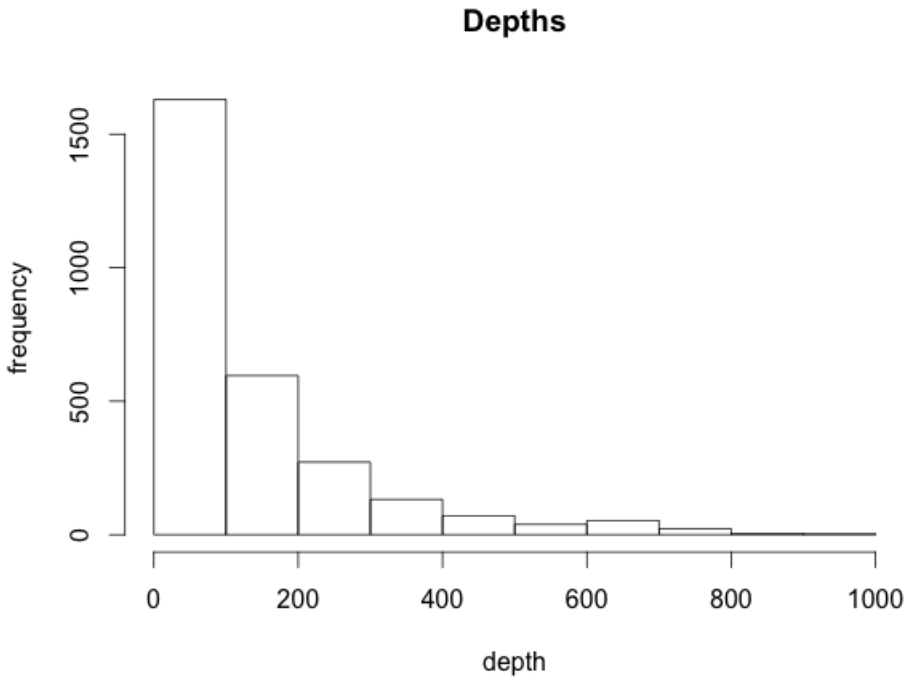

Figure 2.1: Depth frequency across data observations

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0007	118.3000	209.1000	237.8000	330.0000	719.0000

Table 2.2: Cartesian distances across data in kilometers

The second main choice is how many bins to choose. Again, the tradeoff between more or fewer bins is accuracy vs. variance. 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. 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[1]. 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, $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Converting to a grid-based coordinate system 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

[1]Chen et al. (2014)

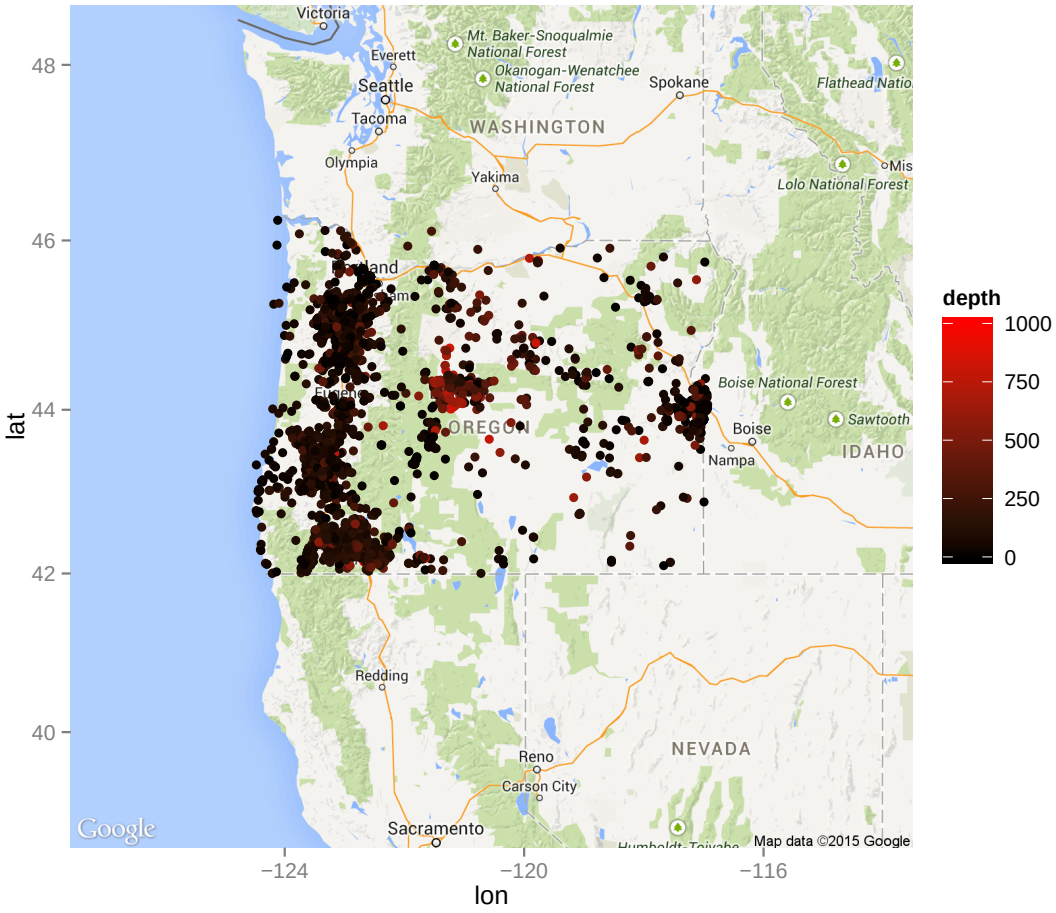

Figure 2.2: Data observations with recorded depth values

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. Any model must meet the three conditions from Chapter 1, namely 1) vanishing at zero, 2) evenness, and 3) conditional negative 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 (Gelfand et al. (2010)) and (θ_1, θ_2) is a parameter vector to be optimized.

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

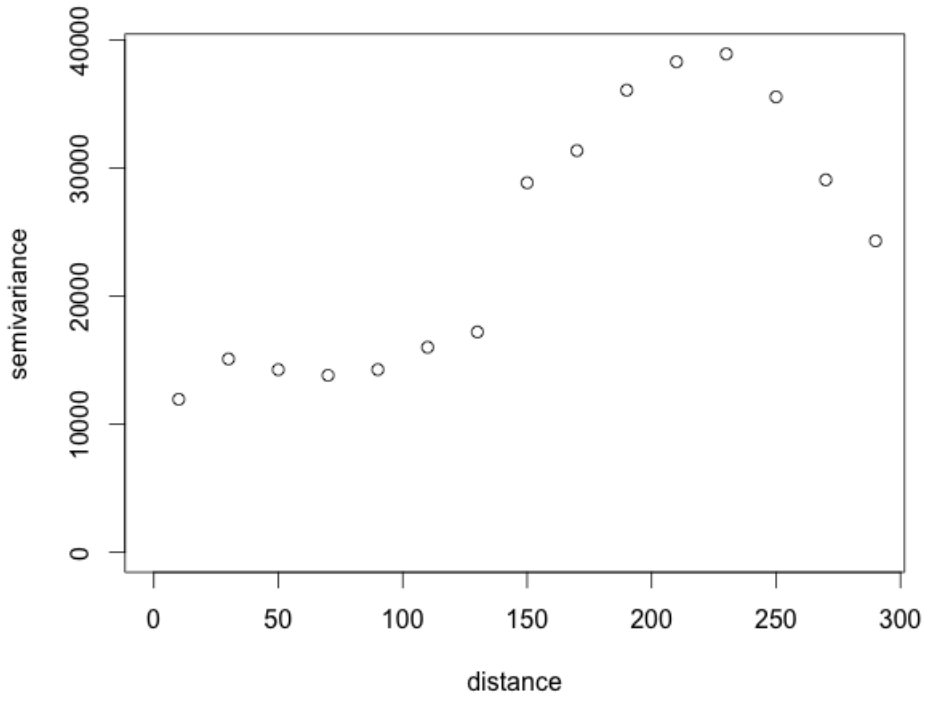

Figure 2.3: Empirical variogram with 300 km maximum distance

Figure 2.4 shows the empirical variogram from Figure 2.3 fitted with a Matérn

model. 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. The sill 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 is the smallest value of $\mathbf{h}$ for which the variogram equals its sill. The nugget 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. These three attributes are generally considered just as much as closeness-of-fit when choosing a variogram model.

In addition to the Matérn model, several common functions include the exponential (Figure 2.5), the power (Figure 2.6), and the spherical (Figure 2.7). Table 2.4 gives the optimized parameter vectors for each model as fitted to the empirical variogram from this section, with tau squared being the nugget variance, sigma squared the sill, and phi the range. Though all are good fits, the kriging step in the following section uses the spherical model, given by

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

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

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 2.8 shows a grid of points constrained to the outer perimeter 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 2.9, it is possible to color each block with the prediction value associated with that block.

Figure 2.9

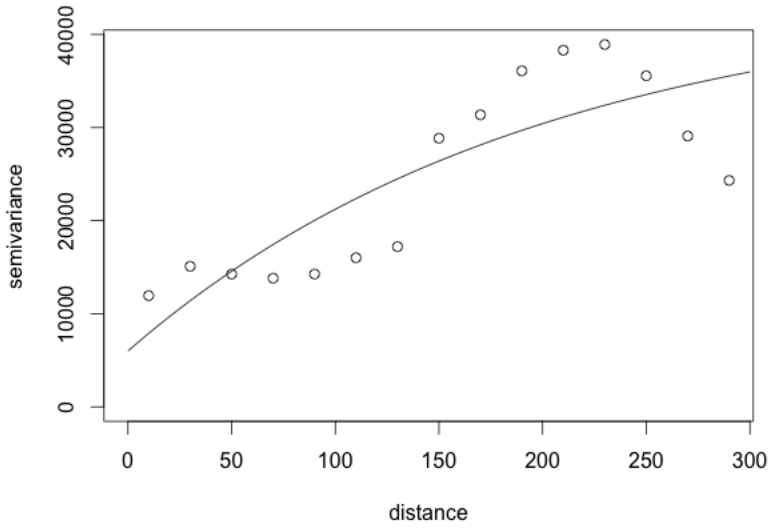

Figure 2.4: Empirical variogram with matérn model

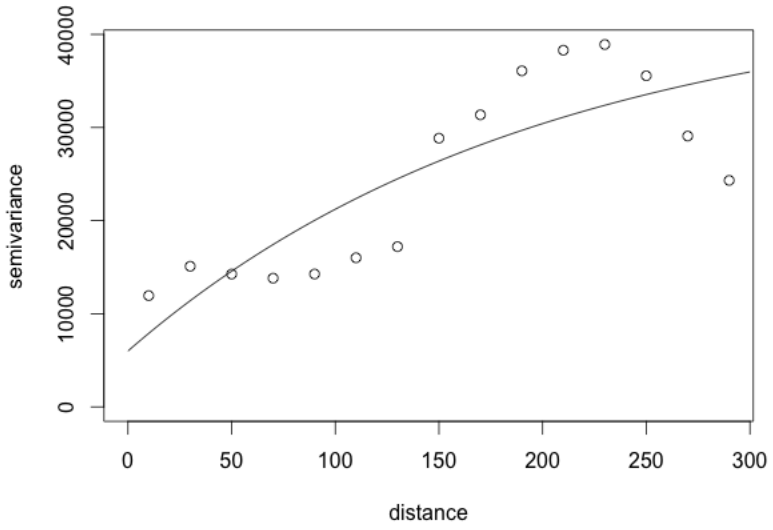

Figure 2.5: Empirical variogram with exponential model

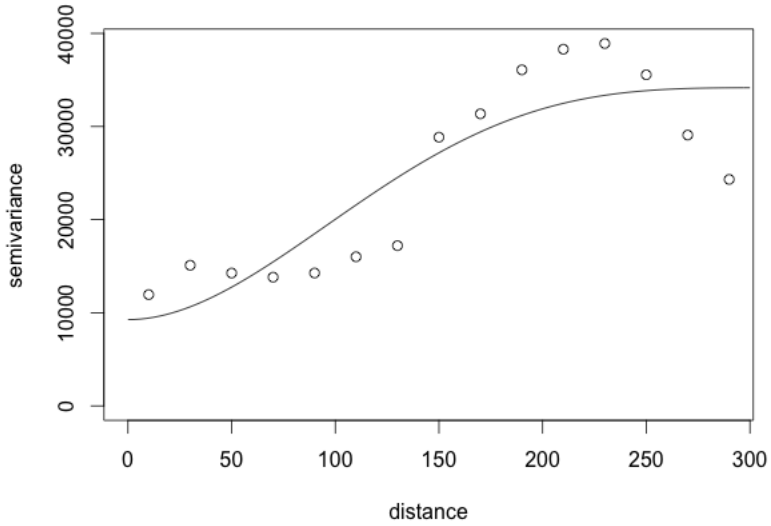

Figure 2.6: Empirical variogram with cubic model

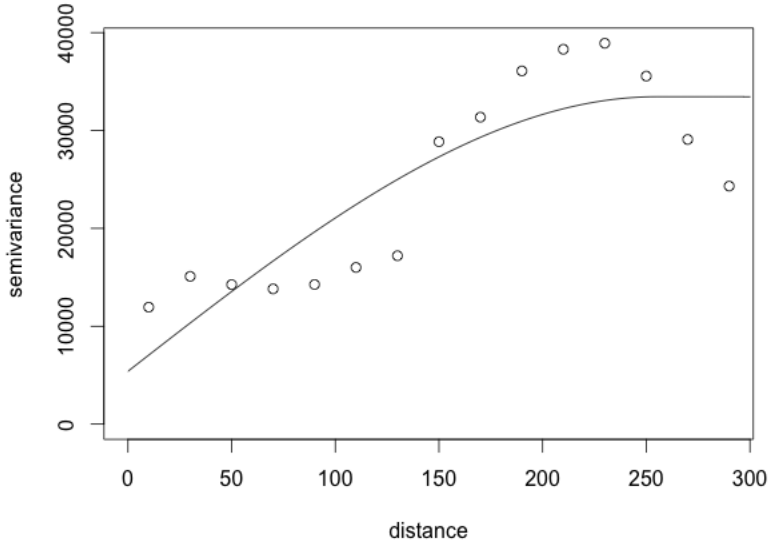

Figure 2.7: Empirical variogram with spherical model

Spherical variogram		
tau squared	sigma squared	phi
5402.2780	28048.6894	254.8803
Matérn variogram		
tau squared	sigma squared	phi
6003.337	38428.555	198.255
Exponential variogram		
tau squared	sigma squared	phi
6003.337	38428.555	198.255
Cubic variogram		
tau squared	sigma squared	phi
9260.4886	24906.8104	316.9988

Table 2.4: Optimized parameters for several fitted variogram models

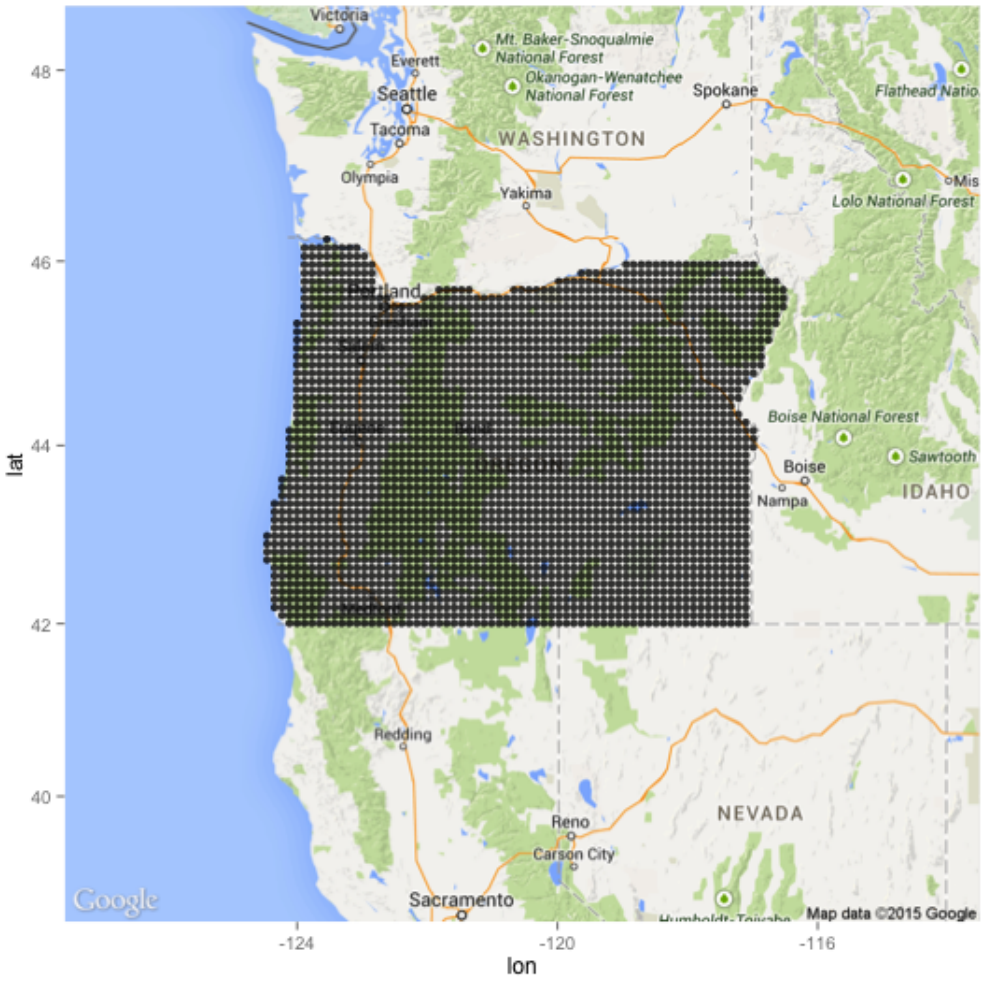

Figure 2.8: Gridded prediction points

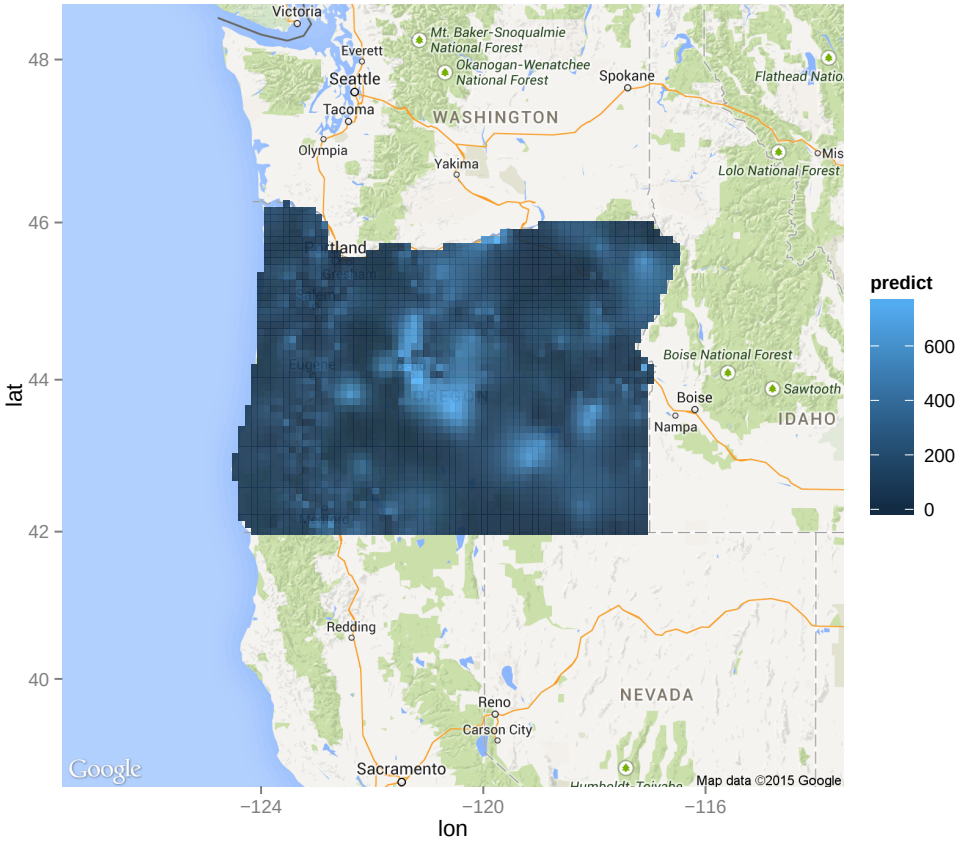

Figure 2.9: Tiled prediction grid

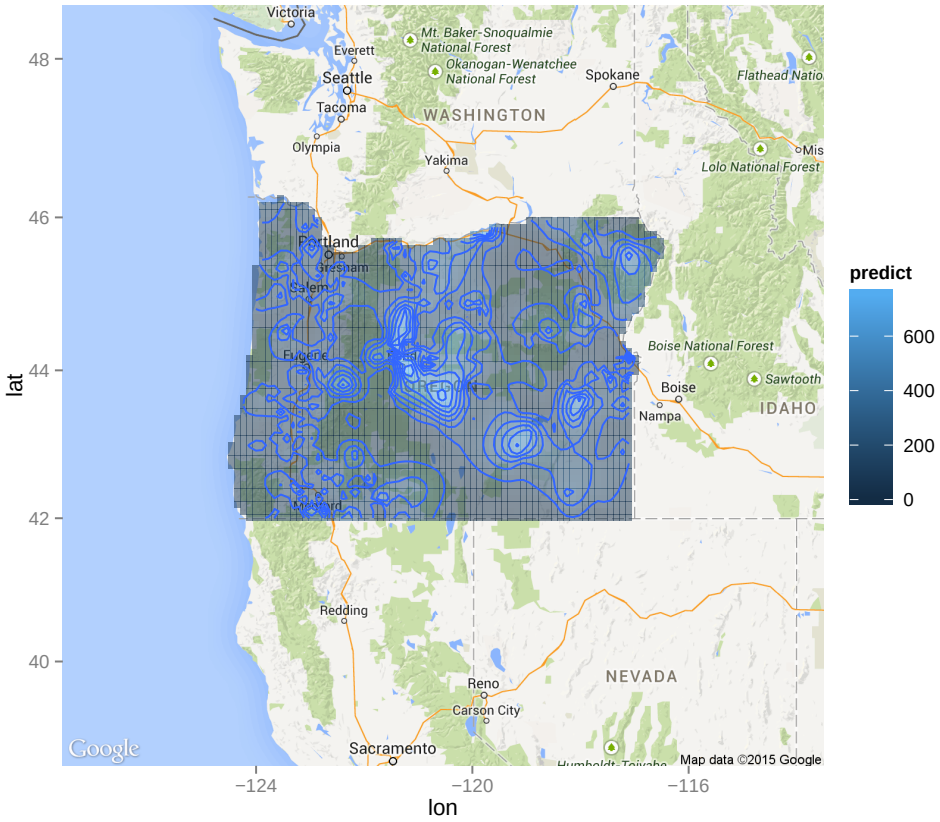

Figure 2.10: Prediction grid with contour lines

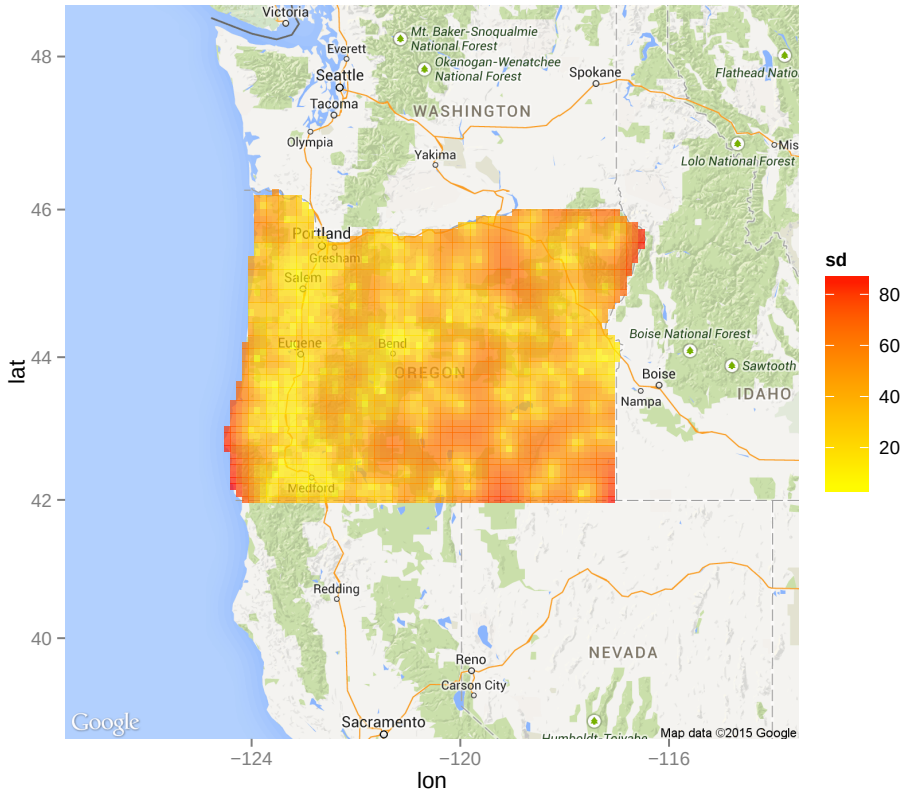

Figure 2.11: Standard deviation grid

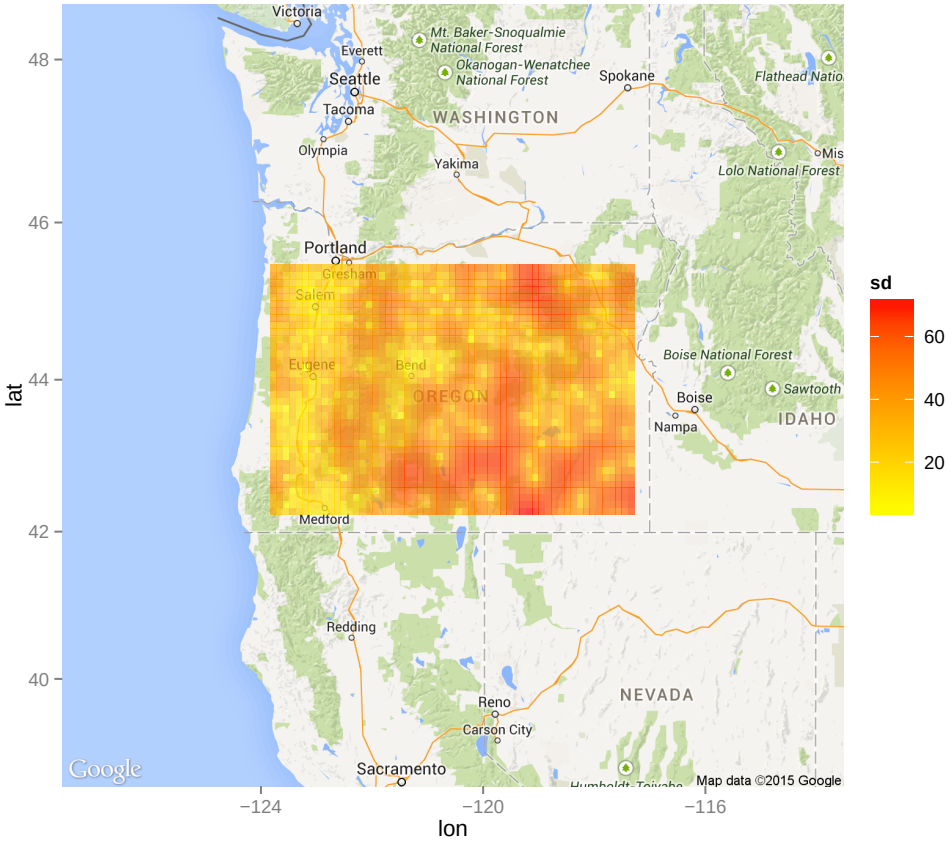

Figure 2.12: Cropped standard deviation plot

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.

Chen, C., Kim, A. Y., Ross, M., & Wakefield, J. (2014). *SpatialEpi: Methods and Data for Spatial Epidemiology*. R package version 1.2.1. `http://CRAN.R-project.org/package=SpatialEpi`

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.

Jr, P. J. R., & Diggle, P. J. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, *1*(2), 14–18. ISSN 1609-3631. `http://CRAN.R-project.org/doc/Rnews/`

Kahle, D., & Wickham, H. (2013). *ggmap: A package for spatial visualization with Google Maps and OpenStreetMap*. R package version 2.3. `http://CRAN.R-project.org/package=ggmap`

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

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. `http://had.co.nz/ggplot2/book`

Wickham, H., & Francois, R. (2014). *dplyr: a grammar of data manipulation*. R package version 0.2. `http://CRAN.R-project.org/package=dplyr`

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Computational Research*. Chapman and Hall/CRC. ISBN 978-1466561595. `http://www.crcpress.com/product/isbn/9781466561595`