James K. Pringle
140.663 Geostatistics
Dr. Curriero
Assignment 3
April 11, 2014

Assignment 3

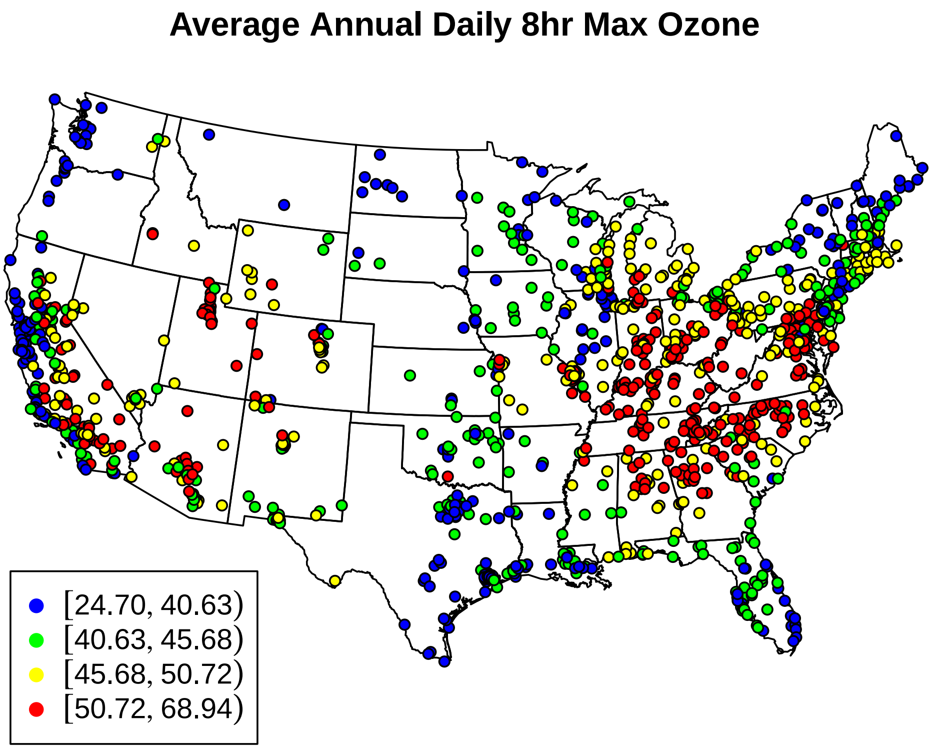
Questions 5

5 Kriging the EPAs AQS Ozone Data.

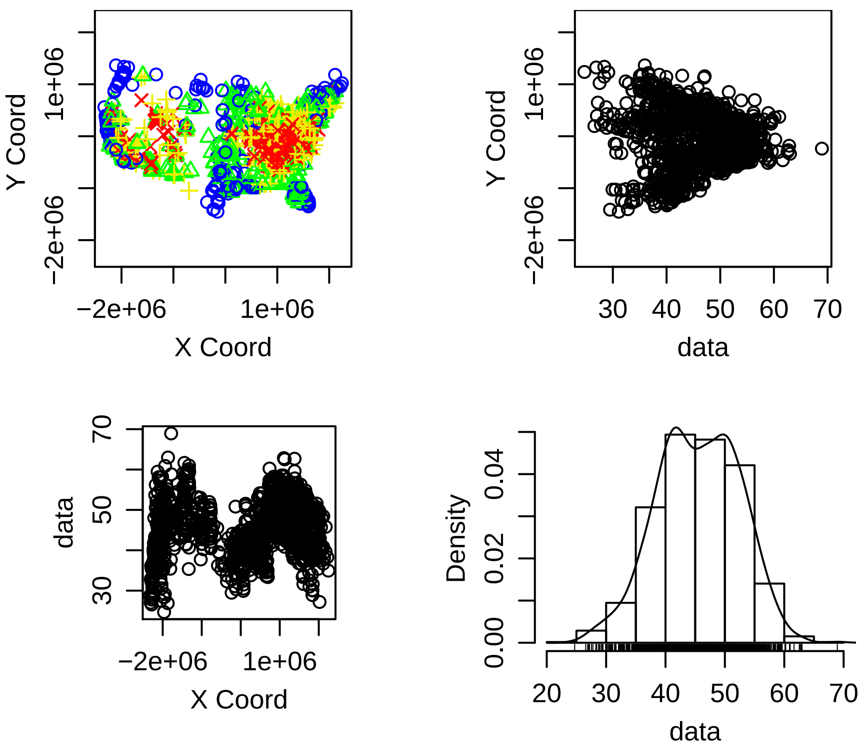
This data set corresponds to the EPAs AQS average annual daily 8 hour maximum ozone for 2007. The questions and code provided run through a kriging analysis of this data based on several different approaches. All the data you will need for this problem are in the two zipped folders `Ozone_Monitors_2007_reproj` and `States_reproj_lower48` posted on Courseplus. Unzip these and save them in your R working directory.

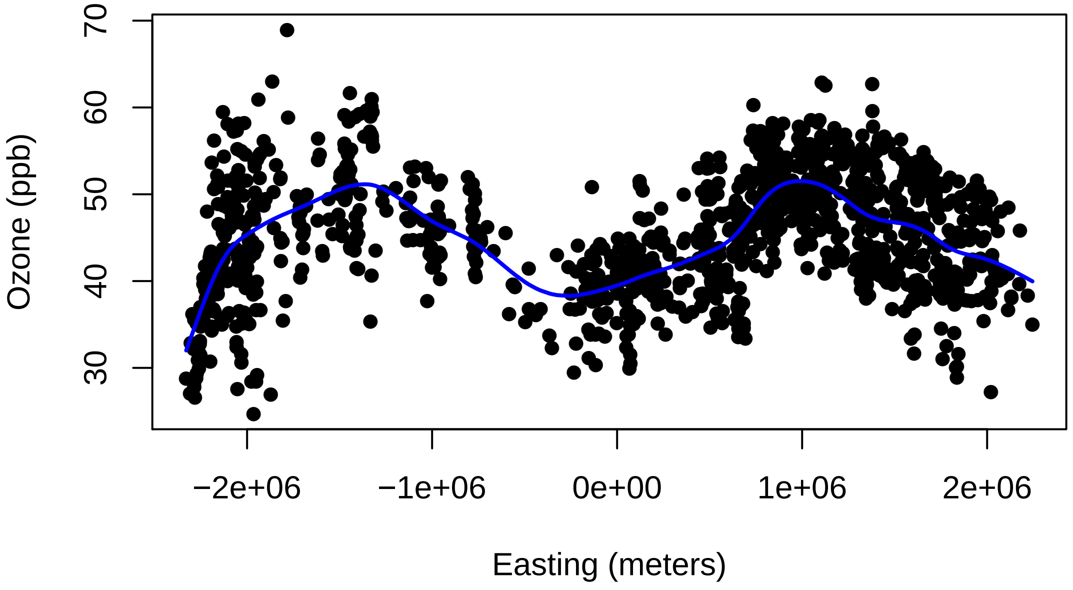
Exploratory Spatial Data Analysis of the Ozone Data

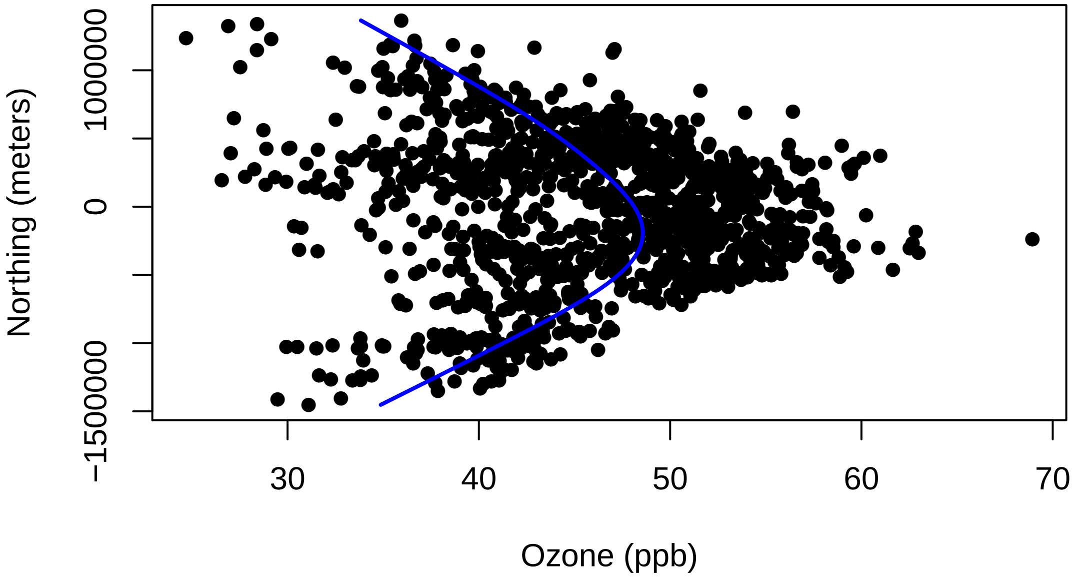
a. Produce a map of ozone with symbols signifying magnitude.



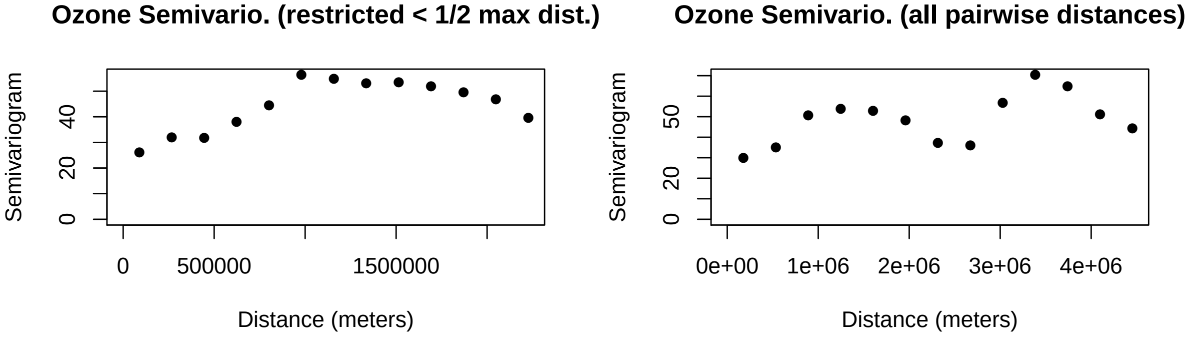
b. Produce a 2×2 display of 4 descriptive plots using the `plot(geodata object)` command for the ozone data. This is a large area to consider, so to better see possible spatial trends across the US, plot the data separately versus the x and y coordinates.


Y Coord
1e+06
−2e+06
X Coord
−2e+06
1e+06
Y Coord
1e+06
−2e+06
data
30
40
50
60
70
data
30
50
70
X Coord
−2e+06
1e+06
Density
0.00
0.02
0.04
data
20
30
40
50
60
70

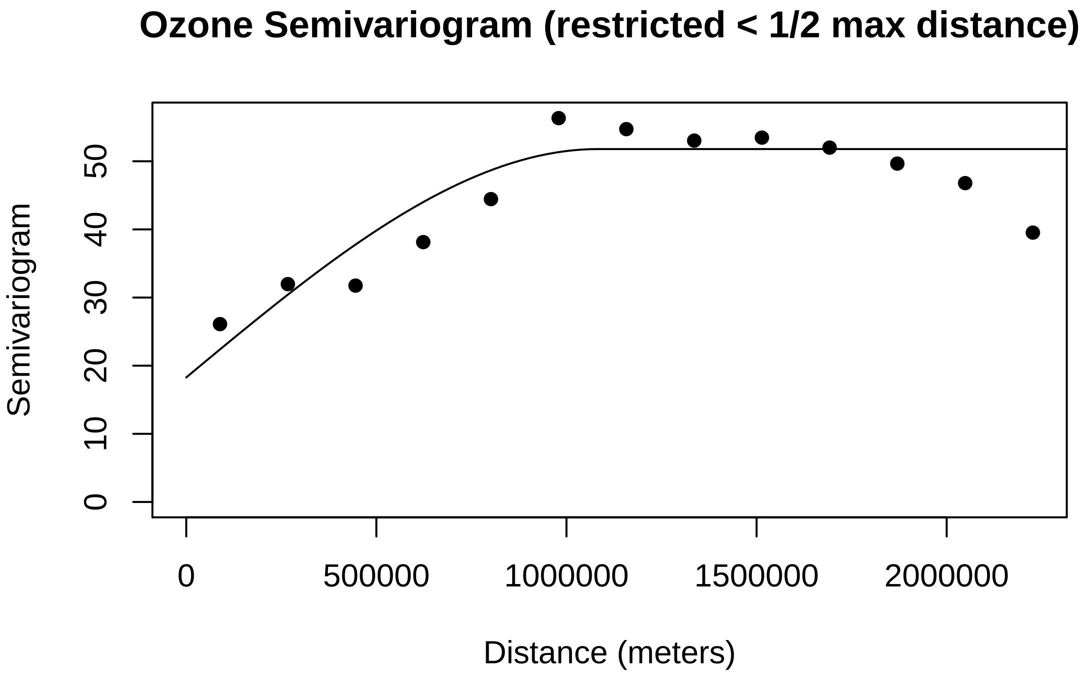

Ozone (ppb)
30
40
50
60
70
Easting (meters)
−2e+06
−1e+06
0e+00
1e+06
2e+06



c. Estimate and plot the semivariogram of the ozone data using the default binning in `variog`. Actually estimate and plot the semivariogram with and without restricting the distances to be within half the maximum inter-point distance (so estimate two semivariograms). In the future though whenever asked to estimate a semivariogram/variogram always restrict it to be within half the maximum inter-point distance.



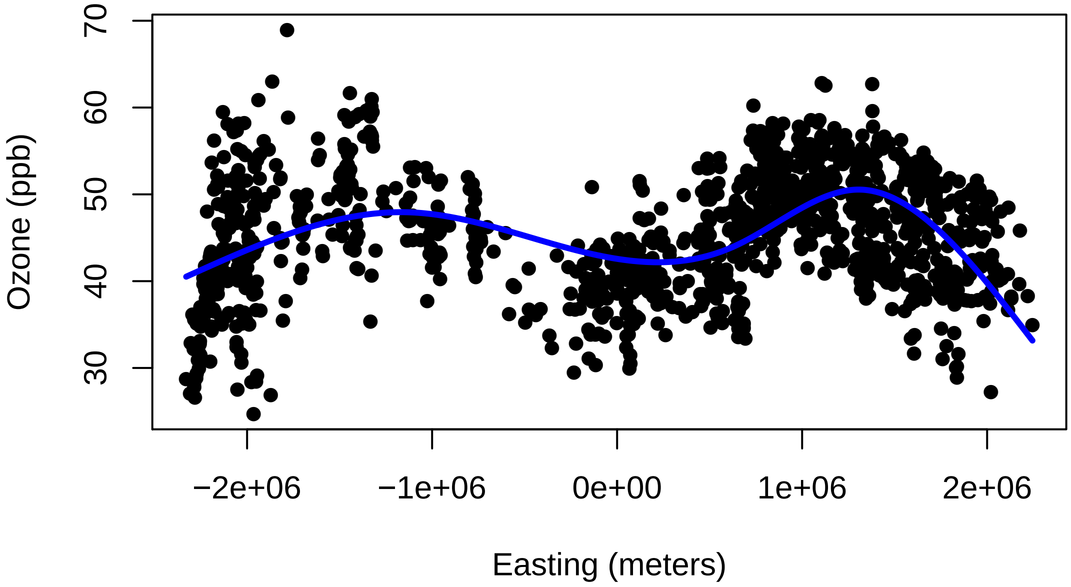
d. For the ozone semivariogram based on distances restricted to be within half the maximum inter-point distance use the `eyefit` command and select a semivariogram function and set of parameter estimates that appear to fit the semivariogram well. For Mac users who have problem getting the `eyefit` command to work just eyeball a set of parameter estimates.

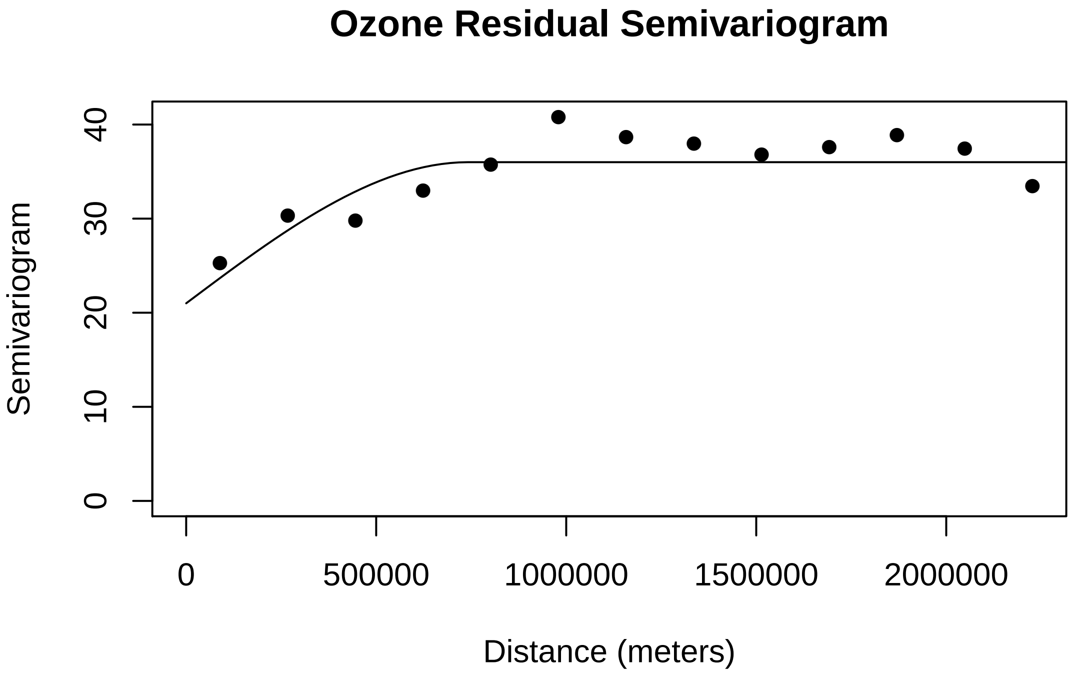


This is a spherical variogram function with parameters

- $\phi = 1.0833 \times 10^6$
- $\tau^2 = 18.28$
- $\sigma^2 = 33.51$

e. Following the results from (b) go ahead a fit a large scale spatial trend based on the easting (or x) coordinate, ignoring and trend in the northing (or y) coordinate. The code I provide fits a natural spline of the easting coordinate with 4 degrees of freedom to try and match the apparent trend. Now using the residuals from this model, estimate and plot the semivariogram (residual semivariogram) and use the `eyefit` command and select a semivariogram function and set of parameter estimates that appear to fit the semivariogram well. Select the same semivariogram function as you did in (d).





This is a spherical semivariogram function with parameters

- $\phi = 7.4228 \times 10^5$

- $\tau^2 = 21$
- $\sigma^2 = 15$

f. With the information generated (a) - (e) address the following and reference specific plots in your answers/interpretations.

(i) Does the ozone data appear to be Normally distributed?

The data do appear to be Normally distributed according to the density and histogram from the 2×2 plot above in part b.

(ii) Argue for the existence of a large scale spatial trend in the ozone data.

We have examined the relationship between the easting and northing coordinates and the outcome (ozone). Both the easting and the northing component seem to have a nonlinear relationship with ozone, as can be seen in the plots under part b. If there were no relationship, we would expect the scatter plots in b. to be more of a uniform blob of points. In contrast, the scatter plots in b. seem to have a pattern. Therefore, knowing the value of the easting and the northing coordinates gives at least some knowledge of the ozone amount.

(iii) Describe the difference in the two estimated semivariograms from (c) and what might be influencing the pattern seen in the semivariogram estimated based on all pairwise distances.

Obviously, the semivariogram which has distances up to the maximum distance contains the semivariogram restricted to half the distance. However, in the full semivariogram, the data has bigger bins than in the restricted semivariogram. The full semivariogram shows that beyond half the max distance, the variance dips (around 2.5 Mm, where Mm means megameter) then spikes (around 3.5 Mm) then settles down again (around 4.5 Mm) as the distance decreases.

It seems from the plot of Ozone vs. Easting that the two peaks in the spline are about 2.5 Mm apart. Since the fitted trend seems sinusoidal, with period about 2.5 Mm, then it would be reasonable that points that far apart would show a lower variance (values are more similar than different as shown by the spline). Furthermore, 2.5 Mm is about the maximum difference in the northing coordinates, so the points that are 2.5 Mm apart and more are separated more along an East-West line than a North-South line. The spike at 3.5 Mm might be related to the fitted trend (it is about there that we have a period and a half in the Ozone vs. Easting trend).

(iv) Specify the spatial regression model (its either ordinary or universal kriging) for what the semivariogram estimated in (c) is for and what the semivariogram estimated in (e) is for. So two models need to be specified. Also for each describe what data the semivariogram is estimating spatial dependence of.

In (c) the model is ordinary kriging. It is

$$Y(s) = b_0 + \epsilon(s)$$

and here $\epsilon(s) \sim N(0, \Sigma)$. The variance matrix Σ is estimated with the spherical model to approximate the semivariogram. The parameters for the spherical model are

- $\phi = 1.0833 \times 10^6$
- $\tau^2 = 18.28$
- $\sigma^2 = 33.51$

Here the semivariogram estimates the spatial dependence of the data $Y(s)$, the geotagged ozone quantities.

In (e) the model is universal kriging,

$$Y(s) = b_0 + b_1 X(s) + \epsilon(s)$$

Here, $X(s)$ represents the value of the natural spline estimated from Ozone vs. Easting. Thus $X(s)$ is only a function of the easting coordinate. The error structure is

$$\epsilon \sim N(0, \Sigma)$$

where the variance matrix Σ is estimated with the spherical model to apprximate the semivariogram. The parameters for this spherical model are

- $\phi = 7.4228 \times 10^5$
- $\tau^2 = 21$
- $\sigma^2 = 15$

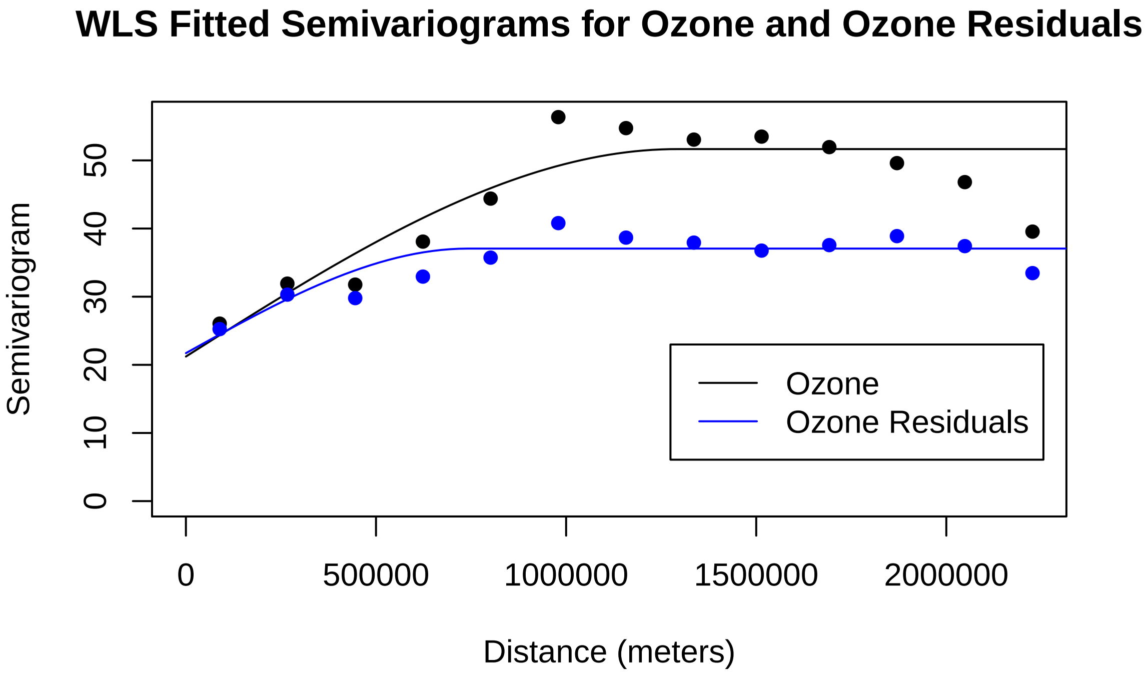
Here the semivariogram estimates the spatial dependence of the residuals of $Y(s)$, the ozone data, after accounting for the value of the natural spline at s.

(v) Describe any difference in the fitted semivariogram functions arrived at in (d) and (e). How have the total sills changed and provide an interpretation for this?

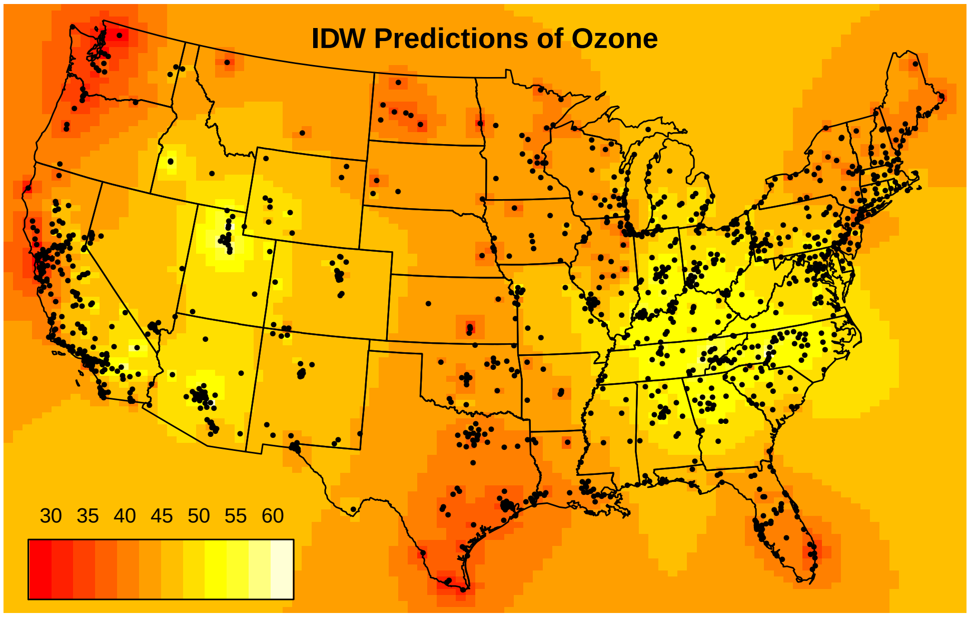
The total sill is different in the fitted semivariograms. The covariate (the natural spline value at the easting coordinate) accounted for some of the spatial variation in the outcome (ozone). Therefore the residuals have less spatial variation (less total sill) than the outcome data do. That is why the fitted semivariogram for the residuals levels off at a lower value than the fitted semivariogram for the outcome data.

Kriging the ozone data

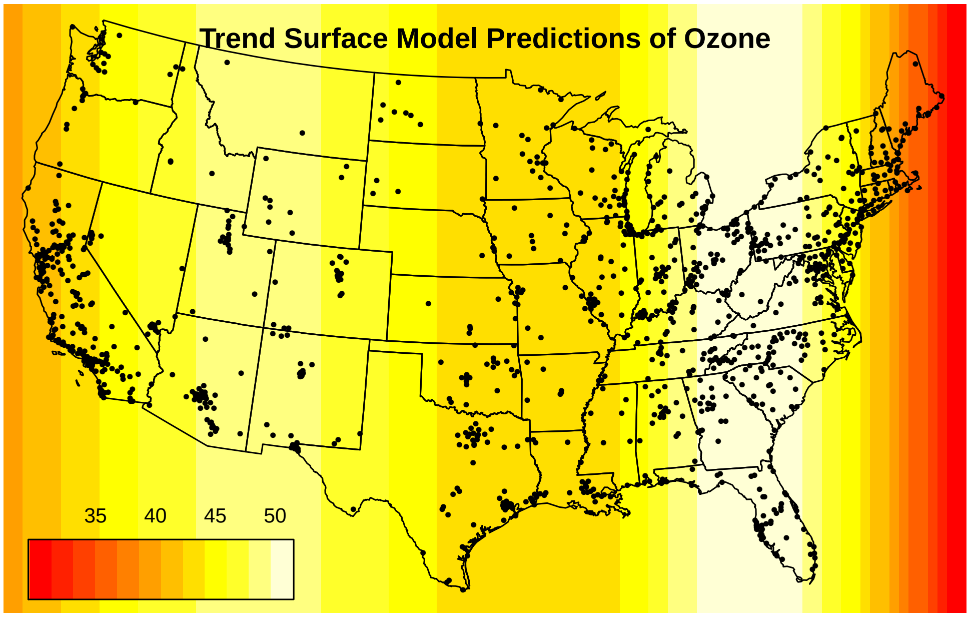
g. Using weighted least squares, fit the semivariogram function from (d) to the ozone data using the initial values selected in (d). Again using weighted least squares fit the semivariogram function from (e) to the residuals of the model used in (e) using the initial values selected.

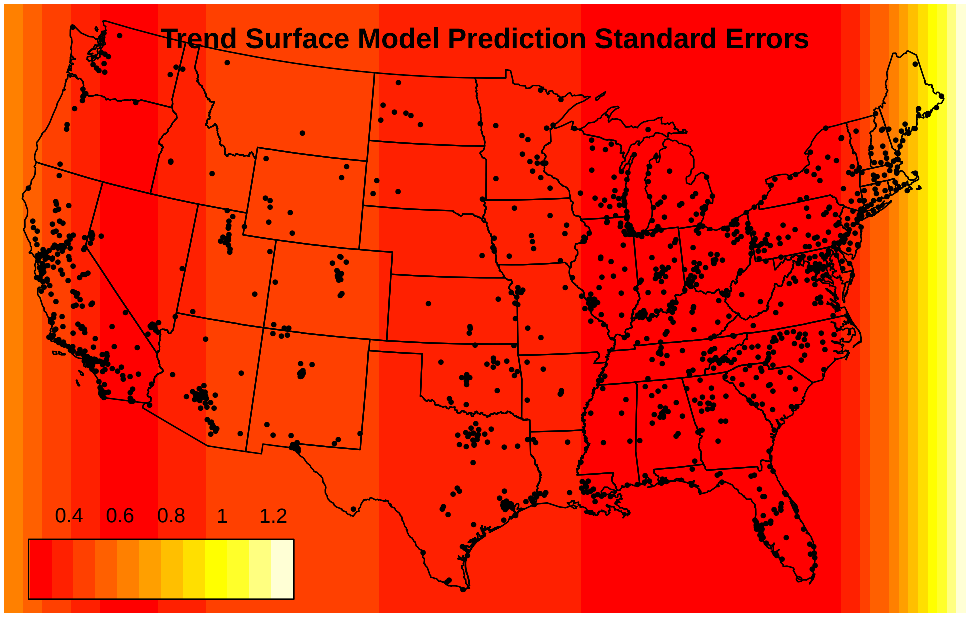


h. Produce a map of IDW predicted ozone.

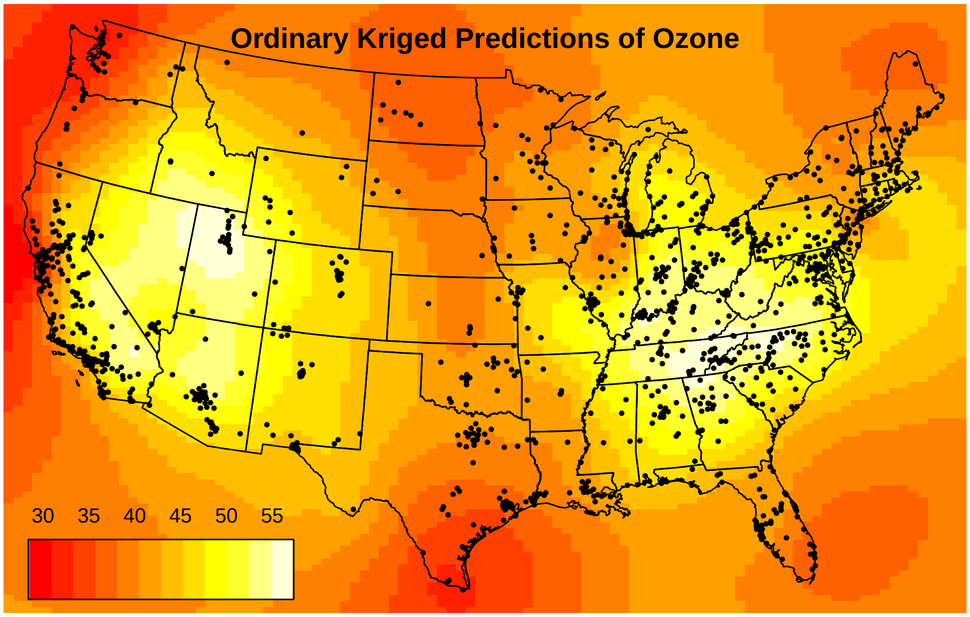


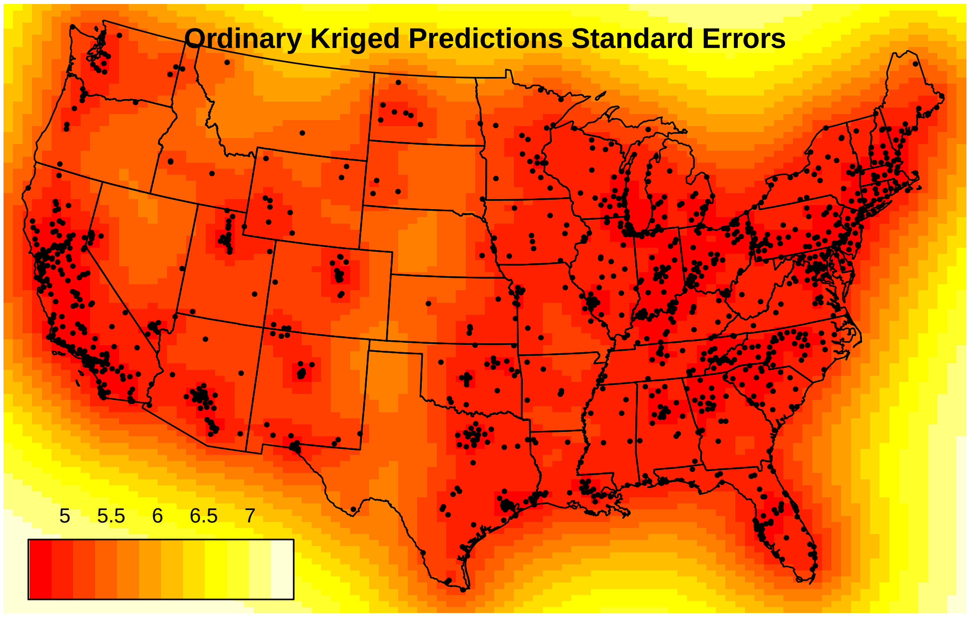
i. Produce a map of trend surface model ozone predictions and a map of predicted standard errors. Specify the trend using the natural spline (with 4 degrees of freedom) of the easting coordinate as utilized previously.


Trend Surface Model Predictions of Ozone
35
40
45
50

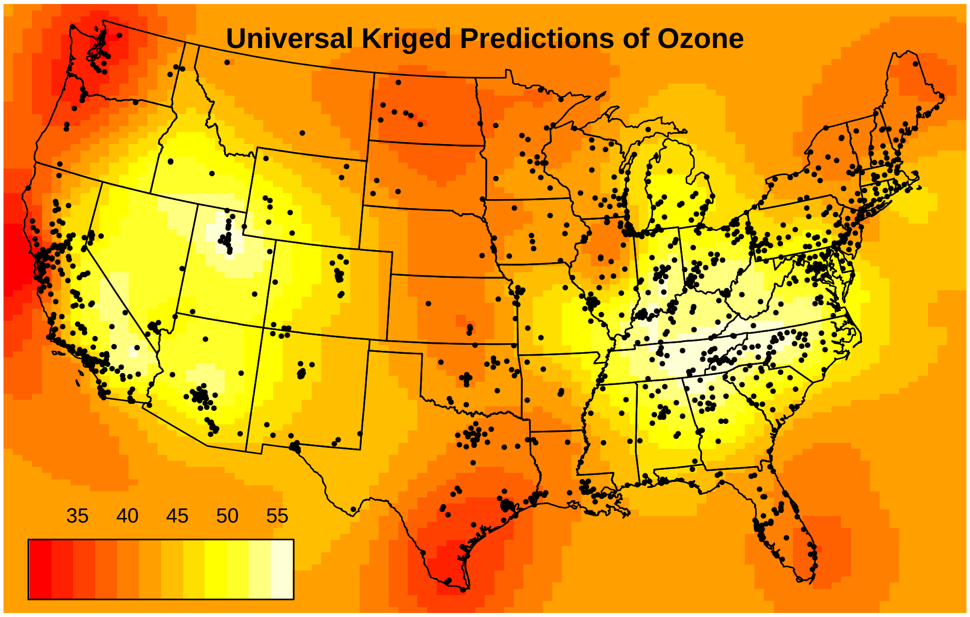


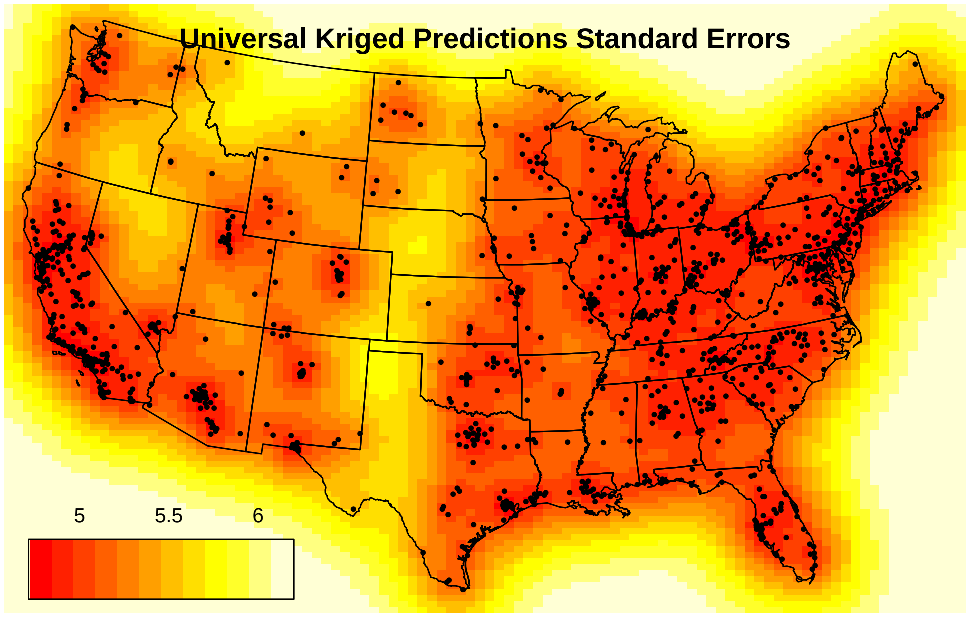
j. Produce a map of ordinary kriged ozone predictions and a map of corresponding prediction standard errors.


Ordinary Kriged Predictions of Ozone
30
35
40
45
50
55



k. Produce a map of universal kriged ozone predictions and a map of corresponding prediction standard errors. For the trend use the same natural spline on the easting coordinate as in the trend surface model predictions.


Universal Kriged Predictions of Ozone
35
40
45
50
55



l. With the information generated (g) - (k) address the following and reference specific plots in your answers/interpretations.

(i) Write out the statistical regression models used for generating the predictions in (h) through (k). If a statistical model doesnt exist just say so. Level of detail for the written models should be commensurate with that found in the lecture notes.

(h) Inverse distance weighting (IDW) is not a statistical model.

(i) The trend surface (TS) model is

$$Y(s) = b_0 + b_1 X(s) + \epsilon(s)$$

where $Y(s)$ is the value of the ozone at spatial location s. The covariate $X(s)$ is the value of the natural spline (applied to Ozone vs Easting coordinate data) with four knots evaluated at spatial location s. The error terms are assumed independence with common variance, $\epsilon(s) \sim N(0, \sigma^2)$.

(j) The ordinary kriging (OK) model is

$$Y(s) = b_0 + \epsilon(s)$$

where $Y(s)$ is the value of the ozone at spatial location s. The error term is distributed as $N(0, \Sigma)$ so that Σ can encapsulate spatial dependence. The variance matrix Σ is approximated using a variogram that is estimated from the data $Y(s)$ with a spherical model and weighted least squares fitting.

(k) The universal kriging model is a combination of OK and TS

$$Y(s) = b_0 + b_1 X(s) + \epsilon(s)$$

where $Y(s)$ is the value of the ozone at spatial location s. The covariate $X(s)$ is the value of the natural spline (applied to Ozone vs Easting coordinate data) with four knots evaluated at spatial location s. The variance matrix Σ is approximated using a variogram that is estimated from the residuals after regressing Y onto X with a spherical model and weighted least squares fitting.

(ii) For each of the spatial prediction approaches considered (IDW, trend surface, ordinary and universal kriging) describe the behavior of the predictions and prediction standard errors as prediction locations get further away from the sampled data.

IDW behaves very mechanically. It shows hot spots (higher predictions) where the original data do. It has no standard errors. Getting farther away from data points, IDW gets very smooth.

TS looks clunky because it has vertical bands. The predicted value depends only on the easting coordinate. Its standard errors are lower than the others because it assumes the errors are independent. Getting away from sampled data doesn't affect the predictions because it is only a function of the easting coordinate. The standard errors behave the same way since they are also only a function of the easting coordinate.

OK doesn't assume a large scale trend. Getting away from data points, OK predicts closer to the national average. That is a property of the model. The standard errors also grow.

For UK, as the estimated point is farther and farther away from data points, it approaches its expected value given the covariates. The standard errors are less than in OK because the including the covariate increases the R^2 value of the model.

(iii) Spend some time studying the difference between the spatial prediction approaches presented with this data. There is nothing to write down or hand in for this, but I'm hoping it might generate some questions.

Done