Problem Set 1: Geostatistics

Spatial Analysis III

Biostatistics 140.698, 2025

Due Date: Monday February 10 10:00pm

Additional Information

- The following files accompanying this problem set have all been posted: *R4ProbSet1.R*, *Q2 Data.csv*, and *Attribute_Ozone2007.csv*, and two zipped data folders *Ozone Monitors 2007* and *USA States*. You will need to unzip these folders, and we recommend storing all the files together in the same place.
- Use the R file *R4ProbSet1.R* for questions 1-3.
- Questions 1 & 2 can be completed after Geostatistics lecture 1.
- Question 3 can be completed after Geoistatistics lectures 1-4 (but started sooner)

You are allowed to work in groups of no more than three. For group work only one assignment should be handed in with group members listed. You'll need to convert all electronic submissions into one pdf file and submit the file in the drop box on CoursePlus prior to the due date/time. Late assignments will be penalized 10% and only accepted within 2 days of the original due date and time. If you are using Rmarkdown to complete your assignments, please make sure all plots fit inside the document margins. Please also do not include unnecessary code/output to your submission. Include any additional R programming that you might have used outside of what as been provided.

1. Running Through an R Demonstration (20 pts)

Use code provided in the *R4ProbSet1.R* file. Open up Rstudio and then open up this file as a script and run each command. This question loads the R package *geoR*, a common package for performing various geostatistical analysis, using the Swiss Rainfall and Wolfcamp Aquifer data sets presented in class. You will reproduce some of the analysis and graphics shown in class in addition to some further analysis. When complete or as you are going through answer the following questions.

Swiss Rainfall Data

a. Do you think the rainfall measurements exhibit spatial dependence? Briefly explain and provide a graphic to support your answer.

b. Do you think the residuals from the linear model regressing the variable altitude (called elevation in class) on rainfall exhibit spatial dependence? Briefly explain and provide a graphic to support your answer.

c. Do you think altitude accounted for much of the spatial variation in the rainfall measurements? Briefly explain.

d. Using the set of commands provided suggest values for the range, partial sill, and nugget for a spherical semivariogram model drawn to the rainfall semivariogram estimates. Provide the graph of the rainfall semivariogram estimates with the drawn in spherical function line as well.

Wolfcamp Aquifer Data

e. Do you think the pressure measurements exhibit spatial dependence? Briefly explain and provide a graphic to support your answer.

f. Do you think there is any evidence for a large scale spatial trend as a function of either or both coordinates? Briefly explain and provide a graphic to support your answer.

g. Look at and run the code under the section “Trend Analysis and Residual Spatial Variation” and answer the following

   - Write out the linear regression model used to account for a trend the pressure measurements (with estimated parameters). What are the assumptions on this linear regression model?
   - What is your assessment of residual spatial variation, that is after accounting for the trend do the regression residuals still exhibit spatial dependence? This is asking the exact same question as in (c) above, but comment on what else might have changed.

2. Large Scale vs. Small Scale Spatial Variation (30 pts)

The data file *Q2 Data.csv* has 6 variables with 200 records. Variables y_1 and y_2 are two simulated spatial data sets (outcome variables) with coordinates *coordx* and *coordy*. The variables x_1 and x_2 are potential explanatory variables or covariates for each outcome respectively.

(a) Read the data file *Q2 Data.csv* into your current R session using the *read.csv* command. Nothing is to be handed in for this.

For parts (b)-(d) do not consider the information provided in (e) when interpreting your results.

(b) Perform some exploratory spatial data analysis on the variables y_1 and y_2 with focus on large scale spatial variation and interpret your findings. Recall in the spatial literature, the terms large scale and small scale spatial variation are often taken to mean first order and second order variation.

(c) Continue from (b) above and explore spatial variation in the outcomes y_1 and y_2. Estimate and plot semivariograms for y_1 and y_2. Interpret results from both (b) and (c).

(d) Perform simple linear regressions for y_1 using x_1 as a covariate and for y_2 using x_2 as the covariate. Estimate and plot the residual semivariograms for each. Interpret.

(e) Below are the statistical models used to simulate these data,

$$\begin{aligned} Y_1(\mathbf{s}) &= 20 + 3X_1(\mathbf{s}) + \epsilon_1(\mathbf{s}), \qquad \epsilon_1(\mathbf{s}) \sim N(0, \sigma_1^2) \;\; corr(\epsilon_1(\mathbf{s}_i), \epsilon_1(\mathbf{s}_j)) \neq 0 \\ Y_2(\mathbf{s}) &= 20 + 3X_2(\mathbf{s}) + \epsilon_2(\mathbf{s}), \qquad \epsilon_2(\mathbf{s}) \sim N(0, \sigma_2^2) \;\; corr(\epsilon_2(\mathbf{s}_i), \epsilon_2(\mathbf{s}_j)) = 0, \end{aligned}$$

where coordinates $\mathbf{s} = (coordx, coordy)$, $X_1(\mathbf{s})$ are uniform random numbers generated independent of Y_1, $X_2(\mathbf{s})$ are actually the *coordx* values scaled up by a factor of 10, $\epsilon_1(\mathbf{s})$ is a spatially dependent Normal random variable, and $\epsilon_2(\mathbf{s})$ is a Normal random variable spatially independent. Thus $Y_1(\mathbf{s})$ is generated from adding a spatially unstructured variable $X_1(\mathbf{s})$ to spatially structured errors $\epsilon_1(\mathbf{s})$. In contrast, $Y_2(\mathbf{s})$ is generated from adding a spatially structured variable $X_2(\mathbf{s})$ to spatially unstructured errors $\epsilon_2(\mathbf{s})$. With this knowledge, comment on the behavior of the respective semivariograms of the outcome variables compared to their respective residual semivariograms. That is compare the spatial dependence structure of Y_1 with the spatial dependence of the residuals from the regression of Y_1 on X_1, and the spatial dependence of Y_2 with the spatial dependence of the residuals from the regression of Y_2 on X_2

3. Kriging the EPA's AQS Ozone Data (50 pts)

This data set corresponds to the EPA's 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 and USA States 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 2x2 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.

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 select a semivariogram function and set of parameter estimates that appear to fit the semivariogram well. Select the semivariogram function based on the shape of the estimated semivariogram and "eyeball" an initial set of parameter estimates.

e. Following the results from (b) go ahead and fit a large-scale spatial trend based on the easting (or x) coordinate, ignoring the 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). Select the semivariogram function based on the shape of the estimated semivariogram and "eyeball" an initial set of parameter estimates as you did in (d).

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?

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

   (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.

(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.

(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?

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.

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

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.

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 doesn't exist just say so. Level of detail for the written models should be commensurate with that found in the lecture notes.

   (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.

   (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.