James K. Pringle
140.663 Geostatistics
Dr. Curriero
Midterm Exam
April 14, 2014

Midterm Exam

1. (10 Points) Traditional geostatistics often considers all variation as small scale for performing spatial prediction.

(a) Explain what this means. Start by writing out a model corresponding to this.

According to the course slides, "little to no regression model formulation is apparent in the traditional geostatistical approach" (slide 41). Traditional geostatistics is "applied without the explicit specifications of any underlying probability distribution."

However, this question seems to be misleading (maybe misworded). If all the spatial variation is modeled on the small scale, then the model would look like

$$Y(\mathbf{s}) = b_0 + \epsilon(\mathbf{s})$$

where

$$\epsilon(\mathbf{s}) \sim N(0, \mathbf{\Sigma})$$

This means that the outcome has some constant expected value (b_0) everywhere (all points $\mathbf{s}$ in the area of interest), and that the randomness comes from an error term. Having all small scale variation means no large scale variation. Therefore, this approach does not model variation in the outcome by differing covariate values.

(b) With a focus on spatial prediction provide one potential negative consequence of such an approach, for example when dealing with spatial dependence, and

There are potential negative consequences as compared to allowing large scale spatial dependence. For example, if there is a significant linear relationship between the x coordinate and the outcome, the model above will not capture that. Therefore, its standard errors will be higher than in a model that does include the x coordinate large scale term. Thus the predictions using the model above will be less accurate.

(c) Provide an alternative but still optimal approach that can be used for spatial prediction. Start by writing out the model corresponding to your alternative approach.

An alternative is a model that allows for large scale spatial varation, i.e.

$$Y(\mathbf{s}) = b_0 + b_1 X_1(\mathbf{s}) + \cdots + b_k X_k(\mathbf{s}) + \epsilon(\mathbf{s})$$

where

$$\epsilon(\mathbf{s}) \sim N(0, \mathbf{\Sigma})$$

This model includes the model discussed above in (a) (by setting $X_1 = \cdots = X_k = 0$). It allows for the large scale spatial trends by including the X_j covariates.

2. (10 Points) In class covering cluster detection I showed an application of SaTScan applied to Baltimore City STI (sexually transmitted infections) at the block group level. Two resulting cluster detection maps were shown and reproduced here on the last pages of the exam; one adjusting for population (map on the left) and one adjusting for population and race/ethnicity (map on the right). Provide a brief interpretation of these results.

One reason some areas may have higher counts of STI cases than others is that they have higher populations. The map that shows clustering after adjusting for population shows four clusters that cannot be explained by population alone.

Likewise, one reason some areas may have higher counts of STI cases is that they have a higher proportion of a race/ethnicity that has a higher rate of STIs. The other map of clustering shows eight clusters of elevated STI counts that cannot be explained by the population level and the race/ethnicity breakdown alone.

Going from adjusting for population to adjusting for population and race/ethnicity, some clusters diminish (or disappear) and others appear. The clusters that diminish have race/ethnicity breakdowns that could explain the STI counts in that cluster. The clusters that appear have rates that are not unusually high for their population, but are when considering the population and race/ethnicity of that area.

3. (10 Points) You are a researcher interested in characterizing the spatial distribution of lead levels in the soil in a particular region using geostatistical methods of spatial prediction. Without any knowledge of the variation in the process you are studying, you propose a 10×10 regular grid (100 sample locations) with 1km spacing that covers the entire region of interest where predictions will be generated. Prior to sampling and analysis you realize money is left over to add an addition set of 40 sample locations. Describe at least two different approaches as to where you might put these additional locations in reference to the already established 10×10 grid and why. An example 10×10 regular grid is provided on one of the last pages of the exam if you find it useful for your descriptions, although it is NOT required that you use this or label those extra 40 sample locations.

Since all the sample points are within the area of interest, then the standard error of the predictions will be relatively large around the perimeter. Therefore, one approach would be to place 10 samples along each side (on the outside) of the 10×10 grid. This is done because it is important

to borrow imformation from the surrounding regions. This will bring down the standard error for predictions at points along the perimeter.

Another approach would be to sample data at points really close to points that are already being sampled (or to resample). This will help to pin down (better estimate) the semivariogram at the smaller distances range. This will help get a correct estimate of the nugget. The better the semivariogram is estimated, the better the spatial dependence can be predicted.

4. (15 Points) In class we discussed a geostatistical application regarding calcium content in surface soil which included 178 spatial locations and their corresponding measured calcium. On the last page of this exam two estimated semivariograms are shown: the one on the left shows the estimated residual semivariogram from an ordinary kriging model and the one on the right shows the estimated residual semivariogram from a considering a universal kriging model including the North coordinate as a covariate (note: dont get hung up on the titles on these plots, my description here is sufficient).

Explain what these two semivariogram estimates are measuring in terms of this problem.

The semivariogram for the ordinary kriging model (OK) shows the spatial variance of the calcium data. That is because the residuals in OK are just the geostatistical data minus their mean (or a centered version of the geostatistical data). Thus the variance of the residuals is the same as the variance of the original data, and hence the semivariogram shows the spatial variability of the calcium data.

The semivariogram for the universal kriging model (UK) shows the spatial variability of the the residuals after regressing the calcium onto the northing coordinate. This semivariogram shows the variability in the data after accounting for the variation due to the northing coordinate.

Also include the following: whether or not you think each shows evidence of spatial dependence and why they might look different.

Both semivariograms show evidence of spatial dependence (i.e. that the variance is a function of distance, that the variance increases as the distance between points increases). Since the UK semivariogram levels off at a lower variance (y) value than the OK semivariogram, it is the case that the northing coordinate explains some of the spatial variation of the data. In other words, the more different two points are in terms of the northing coordinate, the more different their calcium values will be.

Finally provide an estimate of the reduction in residual variance between the ordinary and universal kriging models proposed here. If it helps you may assume the residual semivariogram for the ordinary kriging model levels off at 200.

The UK semivariogram levels off at approximately 90. Hence the reduction in the residual variance is $200 - 90 = 110$ (assuming the OK semivariogram levels off at 200).

5. (20 Points) Consider the traditional regression setting where the identification and quan-

tification of risk factors (the X's and their $\hat{\beta}$'s) are of primary interest. We referred to this as the inference objective in class.

(a) Describe what is meant by the term residual spatial variation.

Residual spatial variation means that the residuals from some model vary spatially or are correlated/spatially dependent. An example is like UK from the problem 4 above, where the residuals of regressing the outcome onto some covariates (with an intercept) have a semivariogram that shows spatial variation.

(b) What are potential consequences for failing to recognize and account for residual spatial variation.

First of all, if it is assumed that the residuals do not vary spatially, or that the residuals are independent, then the standard errors will be smaller than they should be. This means that false postives for including covariates are more likely (if building a model), or at the least, it will give a false sense of precision in predictions.

(c) Provide a method (a set of statistical analysis steps) for diagnosing residual spatial variation.

A way to diagnose residual spatial variation is to create a semivariogram. It is a fast tool for assessing if there is residual spatial variation. First, create tolerance bins by breaking the distances into intervals. Then for each bin, take the difference of all data values that are separated by distances that fall in that bin, square the differences, sum up all those squares, and divide by the number of summands. That value is the semivariogram value at that point. See slide 22 for the mathematical details.

If the dots show an increasing function over the bins (over increasing distances), then it is possible to say that there is residual spatial variation. Furthermore, it is possible to fit different conditionally negative definite functions to these patterns using likelihood based methods or weighted least squares methods. Examples of these functions are the familiar exponential and spherical functions. They have three parameters: the nugget, the sill, and the range.

(d) In the usual multivariate setting where you have a pool of potential covariates to consider, explain why recognizing and accounting for residual spatial variation could wait and be applied to the more final model after evaluating all potential covariates. To begin it might be helpful to first write down a model describing such a scenario.

Checking the residual spatial variation can wait until the very end because this leads into the universal kriging model. It is definitely possible to choose what covariates predict the outcome well and then treat the rest of the spatial variation as small scale variation and put it into a universal kriging model. It isn't necessary to check the semivariogram for spatial variation at every step, because it is best to get covariates first for the large scale trend, because the expected value of the outcome is whatever the large scale trend would say it is. The residual spatial variation

comes in to explain the random noise around that expected large scale trend value.

6. (15 Points) We know kriging, whether ordinary or universal, produces predictions at locations where no data was observed. The kriging equations also produce a prediction variance, which is an estimate of the minimized prediction error at the unsampled location. Now consider predicting at unsampled locations that are far from the area encompassing the sampled data.

(a) Briefly describe what happens to kriged predictions at these locations from using an ordinary kriging model.

An unsampled location far from the area encompassing the sampled data is similar to having $\|\mathbf{h}\|$ increase very large (perhaps to infinity). According to slide 104, as $\|\mathbf{h}\| \to \infty$ the $Cov(\|\mathbf{h}\|) \to 0$. Therefore, the error term becomes uncorrelated with all other error terms, and the expected value (predicted value) of this unsampled point far away from all other points converges to the expected value of the large scale trend at that point. Since there technically is no large scale trend, the large scale trend is just b_0, and by the principles of regression, b_0 is the average of all the data points.

(b) Briefly describe what happens to kriged predictions at these locations from using a universal kriging model.

By the same reasoning as before, the unsampled point far away from all other points converges to the expected value of the large scale trend at that point.

(c) For either model, briefly describe what happens to the kriged prediction variances at these locations. For (c) also provide an intuitive justification for such behavior.

The kriged prediction variances become the sample variance. This is because the covariance of this faraway point with the points in the sample approaches 0. Therefore, $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}'$ approaches 0, so that the kriged variance (from slide 61)

$$\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{12}^{-1}\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{11}$$

which is the variance of the data.

One way to explain this is as follows. The variance value where semivariogram levels off is the variance of the data. Since the sample point to predict at is really far away from all other data points, its variance with the other points is going to be on the semivariogram where the variance levels off (it will be far enough down on the x axis).

Shorter answer questions

7. Are the models referred to by the terms landuse regression and trend surface synonymous with multivariate linear regression applied to spatial data? Explain. (4 Points)

Yes they are. They all assume independent errors (no spatial dependence). All these models have an outcome Y and a set of covariates $X_1, \cdots X_p$, and they use the least squares fit to predict Y given the covariates.

8. Will a universal kriging model ALWAYS outperform an ordinary kriging model for spatial prediction (yes or no)? Use an example to justify your answer. (4 Points)

   No. If there truly is no large scale variation, then any covariates included in UK will not improve prediction. Thus UK and OK will predict the same. Hence UK does not outperform OK in prediction in this case.

9. You want to identify and quantify risk factors in a regression model. The semivariogram of the outcome shows no evidence of spatial dependence. Do you need to check for spatial dependence in the residuals after including covariates in the model? Why or why not? (4 Points)

   Yes, you do need to check spatial dependence in the residuals after including covariates. This is because it is possible that even though the data do not exhibit spatial dependence (through the semivariogram), the residuals after regressing onto covariates may exhibit spatial dependence. See Problem Set 1, question 2 for an example.

10. Circle (box) only the TRUE statements. (4 Points)

    a. Directional variograms are commonly used as a diagnostic to check the assumption of isotropy.

    b. Kriging produces best linear unbiased predictions.

    c. The traditional variogram estimator (aka method of moments, classical estimator) is a reliable objectively based estimator of spatial dependence.

    d. Likelihood based methods for fitting variogram models uses information similar to that contained in the variogram cloud (all the squared pairwise differences).

    e. Spatial cluster detection algorithms are commonly used to assess global spatial dependence.

    Clustering is a behavior describing the entire map of data, a global property. (Cluster Detection, slide 5)

11. Consider exponential semivariogram with parameters τ^2 (nugget) of 10, σ^2 (partial sill) of 50, and ϕ (range) of 75. (4 Points)

    a. What is an estimate of the variance for data represented by this semivariogram?

    An estimate of the variance for the data is the total sill, i.e.

    $$\sigma^2 + \tau^2 = 60$$

b. What does the value $3\phi = 225$ represent?

For an exponential fit to a semivariogram, 3ϕ represents the "effective range." The effective range is the approximate distance where the exponential fit has reached its sill.