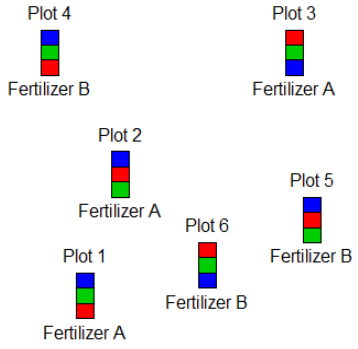
ST433/533 Applied Spatial Statistics

Lab activity for 9/30/2020

A. CLARIFICATION QUESTIONS

Can you show a graphical example of the split plot design? I'm most interested in this one for my research.

Split plot designs have two treatments (the simplification for one treatment is obvious). Say treatment 1 is fertilizer type (A or B) and treatment 2 is corn variety (red, green or blue). You first randomly assign plots fertilizer and then randomly assign subplots within each main plot to a corn variety. Here is my lame graphic. Data from these types of experiments are often analyzed using non-spatial mixed models with main effects for both treatments and random effects for the main plot. If the plots are spread out like this then that's OK, but if they were close (meaning within the effective spatial range) together then the random effects could be replaced by a spatial term (we've been denoting this Z).



(2) What is the difference between stratified sampling and cluster sampling?

In cluster sampling the cluster centers are random, so there is no guarantee they cover the whole domain (i.e., you might be unlucky and miss a part of the domain), whereas stratified sampling is guaranteed to have a certain number of observations in each stratum.

(3) The idea behind " Space-filling design/regular grid" wasn't fully clear.

The idea is just to spread out the points as much as possible. It turns out that if the region is a square and the number of points is m^2 then the most spread out you can be is a m x m grid.

(4) About stratified random sample by X (covariate), I'm not sure how to make the partition. In the example plot (slide 7), low x/medium x/high x points are mixed together and what should we do to divide these points into three clusters?

In this design there was stratification on X but not s, so you get the same number of observations in each X category, but there are no restrictions placed on the spatial locations of the sample so the low-X category might cluster or might be spread out.

(5) Why is there no nugget effect/minimal nugget effect in the regular grid spatial design vs. nugget effect in the cluster design or random design? Shouldn't there be some measurement errors present in the regular grid design or no matter what design it is?

The design is separate from the model parameters. You can use a regular or cluster sample for any spatial process (nugget or no nugget). The best design might depend on the true parameters, but the design doesn't determine the models you have to fit.

(6) I understand how much time we can save by choosing the DnC methods but wouldn't one disadvantage of this method be that you are not pooling information from the entire sample? Some information at the borders of the batches are lost? For example in the midterm, if we divide the sample into 5 states, some observations at the border of NC and SC are more correlated to each other than the other observations in either NC or SC. How much information will be lost in the DnC methods though? Or due to very large data, this will not matter?

Yes, just analyzing states separately would be inefficient. Maybe if there are thousands of observations in each state this isn't a problem, but generally this is a naïve approach. Typically results are pooled across regions somehow. For example, the simplest (not the best) approach would be to take the final estimate of the range to be the average of the estimated ranges from the state-level fits.

(7) How to estimate or how could be a good distance (from roads, river, etc) to take samples in order to avoid bias? (i.e. for taking soil samples)

I guess this depends on the nature of the bias and the application. Sorry I don't have a definitive answer.

(8) What could be the criteria to decide between Stratified sampling and Latin Hypercube sampling.

LHS is an example of a stratified sample. If the spatial domain in partitioned into a grid, then LHS would ensure that there are some samples in each row and each column to estimate row and column effects. Usually in spatial data analysis row and column effects are not of interest, but if they are, LHS would be a good design.

(9) For the DnC methods, simply average the results of mle will cause lots of bias? Is there some way to make the estimate better?

I don't know that it will cause bias, but it might not be the most efficient use of the data. The group assigned spatial partitioning will tell us all about it.

(10) Why it says low rank method will finally goes to a linear mixed model? The interpretations are not clear enough to me.

Well, the model is $Y = X*beta + B*a + e$ where X are covariates (fixed effects), B the spatial basis functions (random effects), $e \sim N(0,I)$ and $a \sim N(0,S)$, so it fits the definition of a linear mixed model.

(11) Is it common to combine sampling methods and if so, would this affect model building in any way, in terms of sample assumptions?

Yes, for example, you can do a cluster sampling with cluster centers stratified by X, or anything else. As long as the selection of the design points doesn't depend on Y, then model building doesn't need to consider the sampling design.

(12) In the NNGP method, it seems like there is a subjective choice for number of neighbors (m). In the Heaton paper, they used 25, but didn't really explain why they chose that number. What would be a good method to make this decision based on the data? One way could be to randomly thin the full dataset to a reasonable size and then use standard Kriging methods to estimate the spatial range parameter, and the choose a reasonable m based on the average number of points within that distance.

Yes! The group assigned NNGP will tell us all about it.

(13) For the NNGP method used in the Heaton paper, they decided to depart from fully Bayesian and use a grid-search for two of the parameters for this dataset. Does this imply that while fully Bayesian estimation is possible with this method, it's too time-consuming to be practical for very large datasets?

Right, they did the grid search rather than MCMC to make it faster. Grid search is generally better than MCMC when there are only a few parameters. Bayes will be slower for large datasets, but not impossible.

(14) Going back to GLMs, why wouldn't you want to add the nugget effect in our model? Would it be hard to capture for binary responses?

You can add a nugget effect. However, even without the nugget the variance of the response given the spatial term Z (and thus the probability p = exp(Z)/[1+exp(Z)]) is p*(1-p). You might have trouble separating an additional Gaussian nugget variance from this source of variance, but you can give it a shot if you want.

(15) For DnC methods, how to deal with the correlation between groups?

The group assigned NNGP will tell us all about it.

(16) For low-rank methods, It seems that the covariates are based on the coordinates. Why do you say there are many choices for the covariates?

You could have cos(4*pi*s) or s^10 or exp(-(s-4)^2), etc.

(17) Is it possible to combine methods? For example, Low Rank with Sparse Matrix.

Yes, there are definitely combinations of these methods. For example, LatticeKrig using sparse matrices and low-rank methods. We need to throw the kitchen sink at these hard problems!

(18) For the computation of the inverse of the covariance matrix $\Sigma(\theta)-1$, we just need to calculate $\Sigma(\theta)-1Y$, which is the same as solve sparse linear system $\Sigma(\theta)X=Y$. However, the computation of the inverse of a matrix is so hard even though it is sparse. The equivalent way is easy to calculate. Thus, why should we still focus on the calculation of the inverse of the covariance matrix $\Sigma(\theta)-1$?

Ah, excellent question. You don't have to compute the inverse if you are clever. I don't know of a way around computing the determinant though and so the calculation of the likelihood remains cubic in n.

(19) In low-rank methods p must be very close to n. Is it feasible to use this method with a really big data set, i.e. data set with over a million observations? Then we have to come up with a really big number of p...

This is a limitation for sure.

How do you stay abreast of new big data methods since the field is progressing so quickly? Also, how do you know a new method is good if just one person wrote a paper on it? (for us non-stats folks)

Well, journals and conferences I guess. I probably wouldn't use a method for a data analysis I cared about until it has gained some traction (citations, R package, etc).

On slide 19 of "Dealing with Large datasets", you say that its tricky to form groups and deal with the inter-group correlation. I was wondering, whether the best way to form groups is such that the intra-group correlation is high and inter-group correlation is low? Also, will the variance parameter estimates be less-reliable (underestimated standard errors) in this case, since we might be ignoring the overall variation?

This sounds like a job for the spatial portioning group ☺

B. STUDENT DISCUSSION QUESTIONS

(1) What is a real-life example of a treatment that would warrant a split plot design?

BR: The bee example at the end of this document.

(2) What are some real-world examples of when you would NOT want to use uniform data sampling or random data sampling?

BR: Where logistical constraints make this difficult, such as sampling a dense rain forest.

(3) When running Divide and conquer for something like the United States, would you rather use preset markers like states, or do a little gerrymandering and run lines that cut through states and maybe avoid cities in some of your regions?

BR: I would gerrymander to get homogeneous and equally-sized regions.

(4) A potential group discussion topic is how the data used in the first midterm (either for the response variable or covariates, or for both) could have benefited from any of the spatial design types presented in lecture.

BR: Well, if we could move the monitors then we might place them carefully to isolate treatment difference. For example, if neighboring states had different polices, say one closed everything while the other did nothing, then monitors close to each other but in different states might highlight the treatment effect.

(5) Do we care about spatial designs when analyzing large data sets?

BR: Design is less relevant when it's overpowered by data. One interesting application of design for big data is when the full dataset is too large to handle and you need to carefully select a subset for your analysis.

(6) If one were to design an experiment with cluster sampling, would the closeness of those clusters be affected by different spatial covariates at different spatial scales? Is there any way to measure the optimal distance from the inner-most point of the cluster to the outer-most point based on the spatial covariates present?

BR: Yes, you would probably want to ensure the cluster centers aren't right on top of each other. Some knowledge of the effective range would be needed to ensure there is not signification overlap, though if they are correlated it's not the end of the world, you have spatial tools at your disposal after all.

(7) Regardless of the type of sampling, how to decide the best/more convenient number of samples per area/cluster to be representative?

Well, more is always better. To pick the sample size it is common to first say how precise you want to be (say what standard error for the parameter of interest you can live with, or the power of the test you will conduct) and then collect sufficient data to achieve this precision.

C. BRIAN'S DISCUSSION QUESTIONS

(1) What covariate did you use? Why? Did you find it was significant?

- Population density within 60 km around each station, it was significant in GA not others.

- 2016 Presidential election results were significant for all states.

- Number of COVID cases, sort of significant.

(2) Some of you had a really large (~30) response value near Asheville, some discarded this as an outlier and some never mentioned it. Did you see this outlier? Do you know what explains it? What's the "right" thing to do with an outlier like this?

There were ~ 4 really big observations, but it seems they were corrected by the EPA in the course of the exam! The "right" thing might be to compare results with and without the outlier.

(3) Many used the notation $Y(s) = b_0 + X(s)b_1$, what's wrong with this?

Missing the Z and E terms. E(Y) = b0 + Xb1 is correct.

(4) Why do you think there is no nugget in this application?

The estimated nugget was small, maybe because the response was the difference of averages.

(5) Is it OK to use another pollutant (or change in the pollutant) as a covariate for change in PM?

It's OK, especially it is from the same emission source so they are correlated and this covariate is measure in places where the response is not so it helps with prediction.

(6) Is it better to use population and/or COVID-19 counts or rates as the covariate? Why?

Maybe better to use rates here, since this accounts for population size.

(7) How would you convince yourself that the COVID restrictions **caused** a decrease in PM?

- Make COVID a treatment, i.e., randomize restrictions (not feasible here).

- Add all other possible variables (maybe temperature, other policies, long-term trends, or some other differences between years) that might explain the change as covariates.

- Difference in differences

- Matching

(8) You have been tasked with designing a monitoring network for the bee population in Wake County. You can only collect data at 100 sites per year. The goal is to produce a map (i.e., of the entire county) that shows where the bee population is increasing and/or decreasing. Where would you conduct the sampling?

Since the goal is prediction and we don't know the parameters then a regular grid with a few clusters (not so close that you count the same bees twice) would be ideal. Up-sample locations known to have the host plants? Maybe stratify by habit?

(9) You have been tasked with designing a study to determine the effect of pesticides on the bee population in Wake County. You can only collect data at 100 sites per year. For each one of these locations you can decide whether or not to apply pesticide there. Where would you conduct the sampling?

Split plot where 50 sites are randomly selected and then each has a treated and untreated plot.