ST433/533 Applied Spatial Statistics

Lab activity for 8/19/2020

A. CLARIFICATION QUESTIONS

(1) I do not understand the properties of being stationary and being isotropic. I thought both assumptions are stating that for pairs of locations with the same distance, they will have the same spatial correlation.

Isotropic says that correlation depends only on distance, stationary allows the correlation to vary by distance and angle (anisotropic).

(2) By looking at the plot directly, we may just draw a rough conclusion about stationary/isotropy/anisotropy, what other methods can we use to test stationary/isotropy/anisotropy more systematically and draw a more reliable conclusion?

Yes, we'll do this in video 5b.

(3) Why did we start with an exponential relationship with distance? Can it be proportional to d? Is there any simple way to find this out before going into detail spatial modeling? Or do we have to model using different assumptions and see which one produces better outcomes?

Yes, there are many possible correlation functions: matern, squared expo, etc… I just picked exponential as a simple first example.

(4) Can you give tips in making choropleth maps? How can we assign point observations that include latitude and longitude coordinates to a region like a state?

In R the function to do this is map.where("state", x=, y=), as in the solution I'll post tonight.

(5) You mentioned that you can never know for sure if the data is stationary or not, unless you created the data like you did in the gridded examples. But that doesn't seem to be a big deal if stationary models work well on nonstationary data. Is there a case where a stationary model poorly predicts nonstationary data?

Sure, but for prediction the nonstationarity (NS) would have to be very extreme. NS matters more for uncertainty quantification of predictions and standard errors for beta.

(6) What if there are multiple spatial patterns in a dataset, i.e. for variable A, the spatial covariance looks like isotropic, but for variable B, the spatial covariance looks like nonstationary.

Yes if we do a multivariate analysis, but not for what we're doing now which is treating one variable as a covariate.

(7) It is hard to see whether we should add a nugget error or not by the spatial plot. Can you illustrate it more clearly?

You will do this on HWK 2 using variogram. You can also use AIC/BIC.

(8) Do we assume Z_i (spatial error) are independent of e_i (random error)?

Yes! Sorry if I didn't mention this.

(9) In the graphs at the end of Lecture 3, where we identify what the correlation structure is, we are still looking at a one-dimensional distance right? Since both s_1 and s_2 can be taken to be a one-dimensional axis?

These are plots in 2D, s1 is long, s2 is lat.

B. STUDENT DISCUSSION QUESTIONS

(1) Where in your previous experiences could you have applied, or did apply, spatial statistics? What is one thing you are looking forward to learning in the course?

Health care and apps to HC. Banking, how spending varies by state and how this relates to COVID.

(2) What are the features of the maps that enable you to differentiate between Isotropic, anisotropic or nonstationary?

Find something that disobeys isotropy, e.g., some points that are the same distance apart but with different correlations. If one area is crazy, it's nonstationary. If the data have the same statistical properties when the map is rotated by 90 degrees this is evidence of isotropy.

(3) In the lectures, Dr. Reich mentioned several potential covariants that could affect air quality information and then mentioned that even once all those were considered, there could still be error. How would you all go about picking covariants? Do you include all the thousands you might be able to think of [distance from every type of source pollution, day of week, meteorological data, altitude, location, etc.], or do you leave it at the most obvious spatial and time covariants? I know some methods of determining which factors have the most influence after they are chosen but I have less real-world practice at when to stop adding variables to consider. What advice do you all have?

It depends on what data you have and what people have done (literature). SAS does PROC GLM SELECT (BR: R does too ☺) or LASSO. Look at multicollinearity/PCA. BR: Covariates are more valuable for prediction when they are not spatially-correlated (because Z can't easily explain these trends).

(4) How would changing the assumed Gaussian distribution of the spatial covariance to a non-Gaussian one (say an inverse-gamma distribution) affect the assumptions of stationary vs. non-stationary correlations? Would the functional form just change, or would this have other implications? Could one theorize that the spatial covariance that seems non-stationary when assuming a Gaussian distribution become stationary when assuming a non-Gaussian distribution?

Definitely. Another angle is the transformation you apply to the data. So maybe Y has a certain covariance, say nonstationary, but then maybe log(Y) looks stationary.

(5) When you are looking at a model, which correlation value will you weight more Zi or Ei?

One measure is the percentage of variation explained by each component, so sigma2/(tau2+sigma2) versus tau2/(tau2+sigma2).

(6) Why is maximum likelihood estimation more general than least square estimation? When should you use maximum likelihood estimation and when should you use least square estimation?

Least squares is usually only used for estimating mean parameters for data that is approximately normal, MLE can be used for non-Gaussian data and covariance parameters.

(7) How to distinguish the mean trend from the spatial error?

Well, this is very tricky. One approach is to use AIC/BIC. Generally, it's easier to explain a model with known variables in the mean rather than spatial correlation. But that doesn't mean it's the "right" thing to do.

(8) Does a stationary spatial covariance model have to isotropic? If an anisotropy model depends only on the angle, is it a stationary model?

The three classes of covariance models we discussed are stationary and isotropic, stationary and anisotropic, and nonstationary, so the covariance can be stationary but not isotropic. If the covariance depends only on angle but not distance…well…I'm not sure about that.

C. BRIAN'S DISCUSSION QUESTIONS

(1) Your final project will be to analyze 2020 election data. Identify election-related data of each type:

(a) Point-reference: voter home locations (s), response is % GOP (Y)

(b) Areal: census tract data like demographics and density; data aggregated to the state level.

(c) Point pattern: pattern of where red states occur; political rally locations.

(2) Say on August 19, 2020 we observe one sea surface temperate (SST) measurement each for location $s_1,\ldots,s_n$. Explain:

(a) Why are mathematical assumptions needed to estimate the correlation between SST at location s_1 and s_2?

We need pseudo replications.

(b) What mathematical assumptions are sufficient to estimate this correlation?

Stationarity, isotropy, etc.

(c) How might you verify the assumptions you made in (c)?

Graphically, variogram by subregions.

(3) Let s_i be the location of home i and Y_i be its sale price. A model for these data is

$$Y_i = X_i b + Z_i + e_i$$

Give a factor that might contribute to

(a) The mean term, $X_i b$: size of house; age; county; dist to beach or city; house is stone/brick/etc; size of plot; comps ($ last five sales in neighborhood), inflation rate/year.

(b) The spatial term, Z_i: unobserved material prices; unobserved distance to library; unobserved zoning variable.

(c) The nugget term, e_i: Competition; ugly carpet; aggressive realtor.

(d) Do you think it's possible to add enough terms to X_i so that the spatial term is not needed? Probably.

(e) Do you think it's possible to make accurate predictions without covariates in the mean term? Maybe.

(f) Assuming both are possible, would you prefer (d) or (e)? (d)! Some say (e).

(4) Give one real-life example (a different one for each case) where each of these assumptions is likely violated

(a) Isotropy: House prices in Wilmington (angle from coast matters).

(b) Stationarity: House prices in New York State (spatial range is larger in upstate NY than NYC).

(c) Homoskedasticity (the same variance for all observations): House prices in New York State (larger variance in NYC than upstate NY).

(5) Would you say (and give a reason) that each of these datasets is

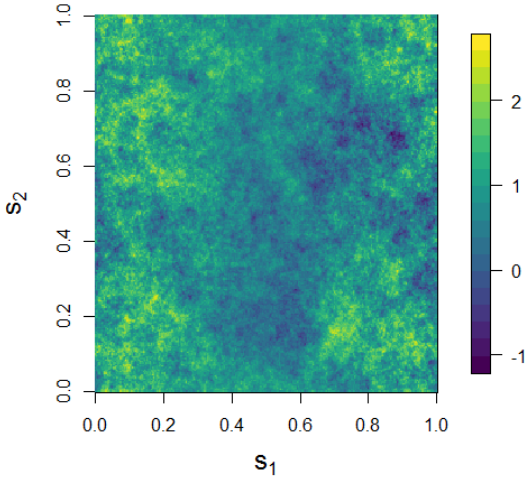
(i) stationary and isotropic

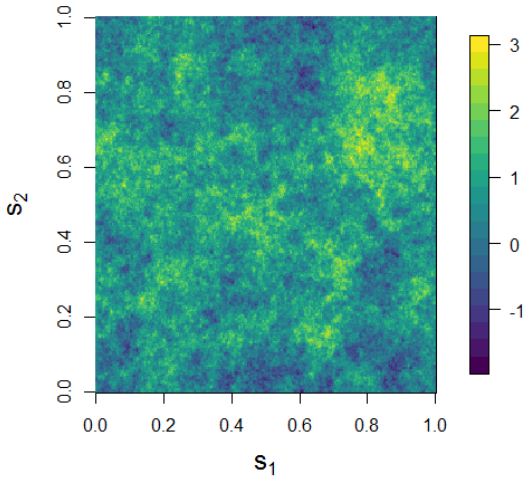
(ii) stationary and anisotropic

(iii) nonstationary

(a) Non-stationary because the variance is larger for s1=0 and s1=1 than s1=0.5

(b) Isotropic





(c) Isotropic

(d) Anisotropic because there are different patterns in the NW/SE direction than the SW/NE direction