ST433/533 Applied Spatial Statistics

Lab activity for 10/28/2020

A. CLARIFICATION QUESTIONS

(1) When looking at Ripley's K, it appears as though inhomogeneous and clustered data are the same. Is there a reason to distinguish the two?

Yeah, they are difficult to distinguish. If you have repeated looks at the data (say data each year) and the clusters stayed in the same location each replication, then you could say the process was inhomogeneous and not clustered. Or if the locations of a cluster can be explained by known covariates (see that last example of coffee shops below) then you can say the process is inhomogeneous not clustered.

(2) What are the importance and difference between all the K_hat(r). Which one is calculated as in the slides?

They are all different edge correction methods. Assuming a really large spatial window this isn't needed and you can apply the method in the slides. If the window is large and samples are iid then the expected value of the curve is πr^2. However, there are always edges and so circles around some points will include areas outside the window (and thus with no samples) which will bias the simple K estimator. The four curves spit out by the `Kest` function all have different edge corrections. The formulas are complicated so for your purposes we don't need to go into the details. You can get the curve from the notes using the `correction = "none"` option in the `Kest` function, but this is not the best option.

(3) How can we add the intensity (different level) of events to the spatial point pattern model? (eg. a crime like robbery vs a crime like gun shut)

You are prescient, this is a topic in next week's notes.

(4) In Ripley's K function, why is the denominator of P(t) n^2?

Because the sum if over of n^2 pairs of points so then p(t) represents an average.

B. STUDENT DISCUSSION QUESTIONS

(1) A potential group discussion topic is examples of each type of spatial point pattern:

(a) completely random sample: Locations of meteor strikes/maybe locations of stores?

(b) clustered: Locations of a species trees in a forest (seeds spread locally)/crime/cancer cases

(c) regular: Locations of trees in a forest competing for sun/crops on an ag field

(d) inhomogeneous: Locations of super markets in high population areas

(2) Say a group of territorial animals, i.e. hyenas, hunt together and form a cluster within their territory. This cluster of hyenas maintains a distance from other groups of predators in the area. How do you measure this type of two-level spatial dependence? Can you distinguish this type of spatial dependence from other one-level types by using the K functions or other methods?

For small r, K(r) will be higher than the random K (pi*r^2); for medium r, it will be lower than random K.

(3) What would be a way to randomly generate a sample that has clustering/inhibition?

(a) Set the cluster locations, generate Gaussian data points around each cluster.

(b) First make a regular grid, then perturb the points a bit.

(4) How can we determine if the difference shown by Ripley's K is significant?

Bootstrapping procedure where you resample the points. (BR: we will discuss this in the test of a completely random sample notes)

(5) How would you treat a discrete model differently than a continuous one?

Say we just have county counts, Y_i for county i. BR: could model county counts as Poisson and use a CAR model for these data.

(6) How is point pattern data different than point-referenced data? BR: Another student asked: Is point pattern data the same as point-referenced data with binary response variable (in terms of data type. not calculation and processing)?

(a) Point pattern data are locations, while in point referenced data the response is something measured, like temp or rainfall (Gaussian).

(b) Maybe Poisson is a better connection? BR: Spatstat often sets Y(s)=1 for observed points and Y(s)=0 for randomly selected non-observed points, and then these binary data are fit using spatial logistic regression. This is a good approach for answering some specific questions, but not all questions.

(7) Marked point pattern is briefly mentioned in the videos. Why don't we use methods for the point referenced data to handle it?

Marked point patterns different and so the same methods wouldn’t apply. BR: could pretend s’s were fixed and just model the $Y(s_1),\ldots,Y(s_n)$ using geostats, but this would ignore dependence between locations and the responses measure at the locations.

(8) Do you think that it is possible to combine SPP with time dependence? If so, how?

BR: Sure, there is a big literature on this. One common question is "If an event occurred at location s at time t, does this increase or decrease the probability of an event at location s at time t+1?"

(9) There was a question a couple lectures ago about what is the minimum range that clusters should take when deciding to do a "cluster" analysis. Can you use Ripley's K to determine what type of spatial analysis to use in this context? Does it make sense to do this?

BR: If the spatial locations are in fact clustered, this seems like a good way to set the number of clusters. This could be useful for modeling nonstationarity (different correlation in each cluster) or large datasets (divide and conquer).

(10) Once we figure out what the basic distribution of the point pattern data is, i.e. completely random / clustered / with inhibition, what would be the next step? For prediction, will we first estimate the pdf of the locations, based on which we can then predict the probability of the event happening at a new location? Is the inherent location effect a latent pattern or would we actually be modeling a probability density based on what we observe?

BR: This depends on your objective. You might estimate the density as you suggest and this could be the final output of the analysis. You could also test for interactions and this could be the final output. Or you could test for covariate effects. We'll do several of these analyses in future lectures.

(11) How would you go about addressing the assumptions for point pattern data?

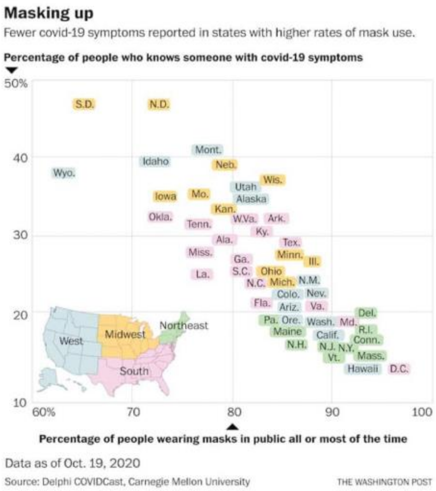
BR: The main assumptions of clustering/independence/inhibition can be checked using Ripley's K. We haven't really discussed model yet so we don't really have any assumptions other than these interactions. We should keep this issue in mind as we move forward.

(12) Since there is no response variable in point pattern data, it reminds me of the unsupervised learning where the goal is to learn about patterns without any response variable. Are (and could) those methods, like k-means clustering or neural networks, be applied to point pattern data meaningfully?

BR: Well, technically the spatial locations are the responses. But they are atypical because they are vectors, not scalars, and we don't usually assign the responses, say Gaussian distributions. But k-means is certainly a useful estimation procedure for the clustering models we'll discuss. I haven't seen neural networks applied here, but probably they have been, they seem to be applied to everything these days.

C. BRIAN’S DISCUSSION QUESTIONS

(1) I recently case across this plot of a Facebook survey from the website https://covidcast.cmu.edu/,


Masking up

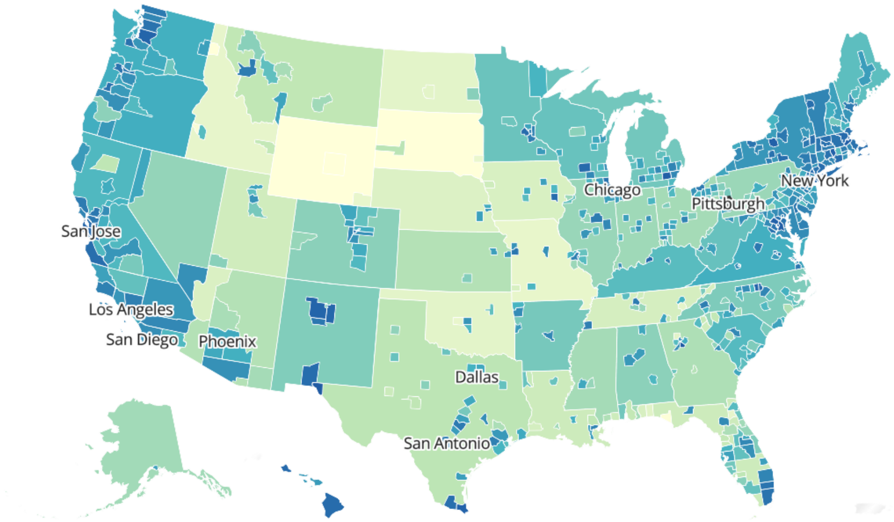
Fewer covid-19 symptoms reported in states with higher rates of mask use.

Data as of Oct. 19, 2020

Source: Delphi COVIDCast, Carnegie Mellon University

THE WASHINGTON POST

This is neat, and they also provide county level data (below is % wear a mask),



They call this their “smooth” rates, which appears to be imputing the state average for counties with a small sample size. For county i, let N_i, Y_i and X_i be the number respondents, number of people who report wearing a mask and covariates, respectively.

(a) What covariates might you add to the model? X = population density, # hospitals, political variables, adherence to other rules, etc.

(b) Write a spatial model for these data. Be specific!

Y_i ~ Poisson(N_iZ_i) where log(Z_1,…,Z_n) ~ CAR with mean that depends on X

-- or --

BR: Y_i ~ Binomial(N_i,Z_i) where logit(Z_1,…,Z_n) ~ CAR with mean that depends on X

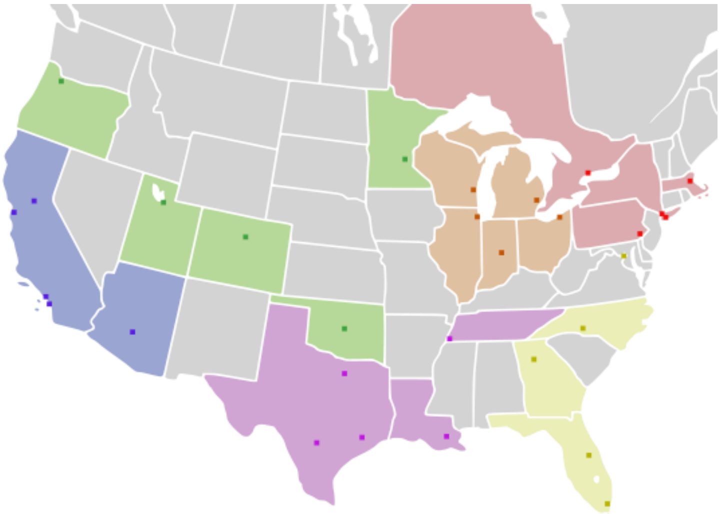
(c) How would you summarize your results for public consumption?

Plot posterior mean and SD of Z_i by census track or county.

(d) What are some advantages and disadvantages of the spatial analysis compared to their approach?

The spatial model gives finer-scale output and is more accurate if we use right covariates, and we get to learn which covariates are significant. On the other hand, the spatial model is more technical (difficult to explain).

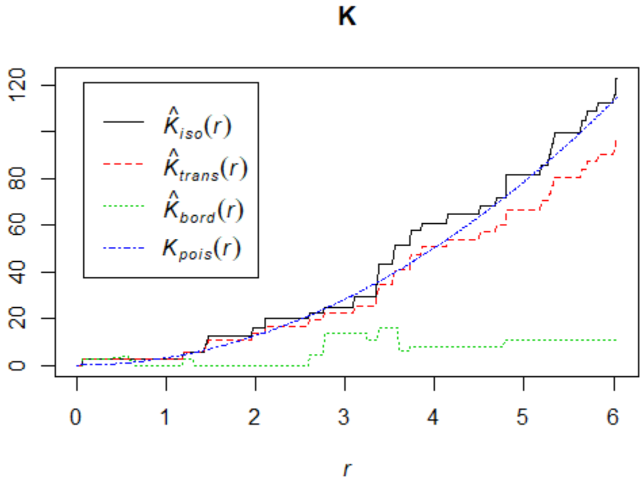
(2) Here are the locations (thanks, Wikipedia!) of the NBA franchises (let's ignore Canada for now, sorry, Canada!).



(a) Based on a visual inspection of the points, which of the adjectives apply to this point pattern?

~~Homogeneous~~, Inhomogeneous, ~~clustered~~, ~~repulsive~~, ~~completely random sample~~

(b) Below is the Ripley's K function for these data. What does this tell us about the point pattern?

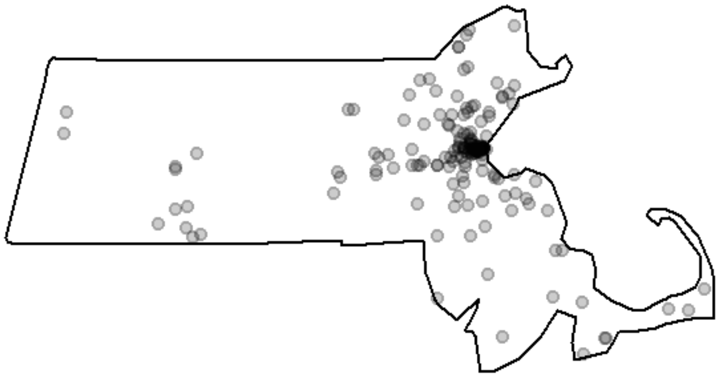


BR: For the most part there doesn't need appear to be strong interaction between locations.

(c) Now let's say we have the location of the NFL franchises and we conduct a similar analysis. How might you test the hypothesis that the degree of clustering/repulsion is the same for both leagues?

BR: Perhaps we could devise some sort of resampling method to get 95% intervals for the k-functions and then we could compare the intervals for the two leagues to see if they overlap. Another approach would be to fit a model (next lecture) and compare 95% intervals for the model parameters.

(3) Here are the locations of the Starbucks Franchises in Massachusetts (thanks, https://mgimond.github.io/Spatial/point-pattern-analysis-in-r.html).



(a) Based on a visual inspection of the points, which of the adjectives apply to this point pattern?

~~Homogeneous~~, Inhomogeneous, clustered, ~~repulsive~~, ~~completely random sample~~

(b) The big cluster is Boston. If you could zoom in just around Boston, would you expect the locations to be clustered or repulsive?

BR: Like the hyenas example above, at a very small scale I would expect repulsion because you wouldn't put two stores right next to each other. But at a medium scale there is probably clustering that could be explained by covariates such as ...

(c) What covariates might explain potential clustering?

BR: Zoning (e.g., no Starbucks in residential or industrial areas), median income, etc.

(d) How would you test for spatial clustering/repulsion after adjusting for these covariates?

BR: In the next lecture we will add covariates to the intensity and then model the dependence of the residuals.