Stat 406 – Spring 2020

Exam 2, due 4 pm (or so) Friday, 24 April.

Reminders:

1) Put your answers into a Word document, pdf document or R Markdown output file and e-mail it to me. Please put Stat 406 Exam 2 in the subject line.

2) Put your name on the last page (the honesty statement) and only there. If you put your name in the filename, I will remove it when I save the file.

3) Please work individually (major change from HW problems).

4) If you don’t understand what I’m asking, please ask. If R is not cooperating, definitely ask. I am very willing to answer questions like ‘how do I get R to <fill in the blank>?’. If something doesn’t seem right, please ask.

5) Note the honesty statement on the last page and include it with your answers.

**Short answer questions:** A few sentences should be sufficient to answer each question.

1) You are studying road damage across central Iowa. You have divided central Iowa into a 12 x 4 grid of regions (48 in all). Each region is approximately 4km x 4km. You have created an average measure of road damage for each region. The data are plotted here:



a) 5 pts When you do a Moran’s test, the estimated I = -0.066 and the one-sided p-value for the test of no spatial correlation is 0.68. Based on this information, what can you conclude about the spatial correlation of your response?

b) 5 pts You decide to do a local Moran’s I analysis. The plot of the Z scores for each region is:



Based on the local Moran’s I analysis, what can you say about the spatial correlation of your response?

Remember:

1) a Z score is computed as $Z\_{i}= \frac{I\_{i}-expected value}{sd}$, so positive Z values indicate a local I statistic larger than the expected value, and negative Z values indicate a local I statistic below the expected value.

2) when a Z score is larger than 1.96 or less then -1.96, the two-sided p-value (using a normal approximation) is < 0.05.

2) Imagine you have an X variable measured for each of 48 regions in problem 1. You want to chose an areal data model for these data.

a) 5 pts. You fit 4 models using maximum likelihood (ML, not REML) to these data. Each of these potential models has a “trend” model of $y\_{i}= β\_{0}+β\_{1} X\_{i}$, where X is a predictor variable (not the spatial x coordinate). The four models have different specifications of neighbor effects. Log-likelihood values and AIC statistics for each model are:

|  |  |  |
| --- | --- | --- |
| Model | Log likelihood | AIC |
| (1) SAR using binary weights | -57.6 | 121.2 |
| (2) SAR using row std. weights | -60.1 | 126.2 |
| (3) CAR | -54.3 | 114.6 |
| (4) Independence | -69.8 | 143.6 |

Of these four models, which is the most appropriate for these data? Briefly explain your choice.

b) 5 pts. Models (1), (2), and (3) have the same fixed effect = “trend” model but different specifications of neighbor effects. All three of these models have positive spatial correlation. What is it about these three models that results in different log likelihood and AIC values? Note: The answer is not “different neighbor specifications”.

c) 5 pts. You fit a 5th model, (5) with a quadratic relationship with X: $y\_{i}= β\_{0}+β\_{1} X\_{i}+ β\_{2} X\_{i}^{2}$ and a CAR model for neighbor effects. That fit has a log likelihood of -54.0 and an AIC of 116.0. When we fit spatial linear models using REML, I cautioned against using AIC to compare models with different fixed effects (e.g. a quadratic trend vs. a linear trend). Is it appropriate to use AIC to compare model (5) = quadratic to model (3) = linear? Briefly explain why or why not.

3) Here are plots of K(x) and g(x) and associated CSR envelopes for a spatial point pattern. There are a total of 88 points in a 3 x 3 square sampling window.



a) 5 pts. Based on the plot of K(x), what can you say about the characteristics of this point pattern? Briefly explain what features of the plot support your answer.

b) 5 pts. Based on the plot of log g(x), what can you say about the characteristics of this point pattern?

Briefly explain what features of the plot support your answer.

c) 5 pts. I haven’t given you a plot of G(x), the distribution of distances to the nearest neighbor. Based on your answers to a) and/or b), would you expect the G(x) plot to show evidence of clustering, inhibition, or be similar to CSR? Briefly explain your choice.

d) 5 pts. The confidence bands shown in the above plots are pointwise intervals. We have discussed overall tests based on combining information across a range of distances. These include the maximum absolute deviation and Loosmore-Ford tests. Do you see any concerns applying one of these to the K(x) or g(x) plots shown above? Briefly explain your answer.

4) We now turn to mapping the intensity of the point pattern used in question 3.

a) 5 pts. Here is a plot of the Diggle-Berman criterion for various values of the smoothing parameter. The Y axis is the mean-squared error using leave-one-out crossvalidation.



Based on this plot, what is the estimated smoothing parameter? An approximate answer is fine. Briefly explain your choice.

b) 5 pts. Here are three possible results for smoothing the point pattern. Do note the different scales for each plot and remember that the sampling window is 3 x 3. Which plot (left, middle, or right) is the most likely result when using the Diggle-Berman smoothing parameter that you chose above. Briefly explain your choice.



**Data analysis problem.** 45 pts.

I have written two versions of the data analysis problem. One is an areal data analysis; the other is a point pattern analysis. **Choose one**. The format of your answer will be similar to that in Exam 1.

Please write up your answer in three parts:

a) a concise but complete description of what you did, i.e., the statistical methods paragraph (or paragraphs) of a scientific paper.

b) a bulleted list for each choice you made during your analysis. State what you needed to decide, summarize the options you considered, identify the choice you made, and give a short justification for that choice. An example for one choice from Exam 1 is below.

c) answers for the investigators “want to see” items (slightly different lists for the two versions of the problem), and any supplemental material.

Here is an example of how I want you to organize your answer to part b.

* Choose a reasonable variogram model
	+ considered Spherical, Exponential, Matern, k=1, …., with and without a nugget
	+ chose Spherical with a nugget
	+ because it <provide reason or reasons>

If you aren’t familiar with a statistical methods paragraph, I have provided (lightly) edited versions of two good answers from Exam 1 in <https://pdixon.stat.iastate.edu/stat406/exams/Stat%20Methods%20paragraphs.docx>

Please follow the requested format for your answer. I graded that lightly on Exam 1; I will pay more attention to that on Exam 2.

Please include your R code as an Appendix. I will only look at it when I don’t understand one of your answers. In that case, it helps understand what you were trying to do.

**Areal data problem**: iowa.Rdata contains **iowa,** a SpatialPolygonsDataFrame with information about the number of confirmed COVID-19 cases in Iowa by county. This is time-wide data, with one row for each of the 99 IA counties, and information about the number of confirmed cases reported by IA Dept. Public Health on 15 March, 22 March, and 13 April 2020.

The variables of most interest are:

 log10pop: log transformed population for each county

 logmar15, logmar22, logapr13: log(N+1) transformed case counts on 15 Mar, 22 Mar, and 13 Apr.

The data frame also includes the original untransformed variables, in case you want to look at them:

 pop: population of the county

 mar15, mar22, and apr13: case counts on those dates

The other variables are quantities from the Census file with county population. The most relevant is:

 COUNTY: the name of the county.

The analysis focuses on logapr13, the log N+1 transformed count on 13 Apr.

The investigators want to see:

 A map of log(N+1) counts on 13 April. You do not need to smooth this.

 Whether log transformed counts are spatially correlated. (see neighbor note below)

 Whether there is an association between the Mar 22 count and Apr 13 count (both log transformed),

 after accounting for spatial correlation with neighbors.

 If so, what is the regression slope and its confidence interval?

 Whether adding log10pop to the previous model improves the predictions.

 Whether the data, at least approximately, satisfy the assumptions of equal variance and no lack of fit.

 Whether any counties have unusually large case counts on Apr 13 given their Mar 22 count

 and population. Unusually large is an observed case count more than 5 times the predicted value.

 If so, which counties are these?

 Hint: think about the residuals. FYI, log(5) approx. = 1.61.

Note about neighbors: By default, poly2nb() considers “queen’s” neighbors. That is two polygons are neighbors even if they touch only at a corner. There are a lot of these pairs in Iowa. For COVID spread, it makes more sense to define neighbors as sharing more than a corner. That is done by poly2nb( … , queen=F). As supplemental material, please include a map showing the neighbor connections.

**Point pattern problem:** iowaPPP.Rdata contains **covid.ppp**, a spatial point pattern object with locations of cases, reconstructed from the county-level case counts. Because analysis of all 1710 locations is very, very slow, I have sampled 10% of the locations. The sampling window is the Iowa state border. Locations are UTM coordinates so distances are in m. These locations are not exact, but they are good enough for this problem. Log10pop.Rdata and logmar22.Rdata contain image files (**log10pop.im** and **logmar22.im**) with the population (log 10 transformed) and the log N+1 transformed case counts on 22 Mar, respectively.

The investigators want to see:

 A map of locations of cases on 13 April.

 A map of smoothed intensity on 13 April.

 Whether locations of COVID cases are clustered? If so, what is the spatial scale of that clustering?

 An estimate of the number of excess cases within ca. 20km of a case.

 Whether there is an association between the log transformed Mar 22 count and the intensity of cases

 on Apr 13?

 If so, what is the regression slope and its confidence interval?

 Whether adding log10pop to the previous model improves the predictions of Apr 13 intensity?

Two plots showing how the predicted intensity on Apr 13 depends on the log transformed number of cases on Mar 22, based on a model with both logmar22 and log10pop. One plot is for an area with 10,000 people (log10pop=4); the other is for an area with 100,000 people (log10pop=5). Include 95% confidence intervals for the predicted intensity on each plot.

Notes about computing:

1) Because the Iowa border is relatively complex, edge corrections are slow. I suggest just using 199 random simulations for any envelope computations. Remember to adjust nrank to the appropriate value to get 95% pointwise confidence bounds when you have 199 simulations.

2) If you use RMarkdown, you can add cache=TRUE to the code block header to save the results for subsequent Markdown runs. E.g., ```{r envelope, cache=TRUE} That way you don’t need to wait a long time each time you make a revision to your document.

3) By default Kest and related functions choose a “good” set of distances, r, to evaluate K(r) at. You can find the closest r to 20km by: which.min(abs( K$r – 20000)), where K is the name of the Kest result. Alternatively, you explicitly specify r, e.g. Kest(…. , r = 20000).

**Honesty statement:** You are allowed to use notes, books, and static online resources. You are not allowed to ask anyone besides me for help. That includes not using online chat and not posting questions to online forums. You are encouraged to ask me questions. In particular, please ask if you and R are not cooperating.

Check the appropriate statement and sign where indicated.

 This exam is my work and only my work. I received no assistance except (perhaps) from Dr. Dixon.

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

If you did receive assistance from someone other than me, provide a short summary of who helped and what help you received.