Stat 406 – Spring 2020

Exam 1, due 4 pm Wednesday, 11 March.

Reminders:

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

2) 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.

3) **Please put your name only on the last page** and make sure all pages are securely stapled.

4) 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 have been asked to estimate the total number of yellow perch larger than 2 inches long in the lake at Ada Hayden Park. You can collect data in different ways (sampling schemes) and estimate the total from your data using different estimators. You consider five different methods. A method is a combination of sampling scheme and estimator. Details of the different sampling schemes and estimators are not important. We’ll just call the 5 methods A, B, C, D, and E.

a) 4 pts. You go out to Ada Hayden and implement each of the five methods (A, B, C, D, and E). That gives you these estimates:

|  |  |
| --- | --- |
|  | Method: |
|  | A | B | C | D | E |
| Estimated number | 5,620 | 5,790 | 5,689 | 6,203 | 5,658 |

Based on other evidence you believe there are 5,700 yellow perch larger than 2 inches in Ada Hayden Lake. Assume that is the true number in the lake. Can you use this to determine whether or not a method (A, B, C, D, or E) is biased? If so, which is the least biased. Briefly explain both answers.

b) 4 pts. You program the computer to simulate each of the five sampling methods. The computer “lake” has 5700 yellow perch. Your program repeatedly draws samples using each method and estimates the number in the lake. Numeric summaries of the estimates for each method are:

|  |  |
| --- | --- |
|  | Method: |
|  | A | B | C | D | E |
| Number of simulations | 10,000 | 10,000 | 6,000 | 2,000 | 1,000 |
| Mean | 5,701 | 5,699 | 5,710 | 5,705 | 5,697 |
| Median | 4,963 | 5,706 | 5,652 | 5,149 | 5,690 |
| Standard deviation | 935 | 849 | 731 | 794 | 611 |
| Standard error | 9.35 | 8.49 | 9.44 | 17.75 | 19.32 |

There are 10,000 simulations of methods A and B and fewer simulations of methods C, D and

E (numbers in the above table).

Can you use these data to determine whether or not a method (A, B, C, D, or E) is biased? If so, which is the least biased. Briefly explain both answers.

c) 4 pts. Using some or all of the results in parts a and b, which method gives the most precise estimates? Explain your choice. If you can’t tell from these data, explain why not.

2) When you fit a variogram model to the binned empirical variogram, the fit.variogram() function sometimes gives you the error “Singular model fit”. This means that two different combinations of parameters have exactly the same fit to the binned empirical variogram.

Here are two situations where you will (or are likely to) get a “Singular model fit” error. For each, briefly describe why different combinations of parameters will have the same fit, and one reasonable thing you might do to obtain an appropriate model fit.

a) 4 pts. The empirical variogram is shown in panel a) below. You attempt to fit a Spherical model with a range, partial sill, nugget and measurement error. Because the measurement error variance is not known, you want to estimate both the measurement error variance and the nugget variance.

b) 4 pts. The empirical variogram is shown in panel b) below. You attempt to fit a Spherical model with a range, partial sill, and nugget.



3) Consider the small data set shown below. The observed values of Z(s) are shown next to their locations (blue dots).



Some estimates calculated from the data are in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Average Z | GLS mean | median Z | estimate of $σ^{2}$ | practical range |
| 5.6 | 10.0 | 2.8 | 34.6 | 0.5 |

You use ordinary kriging to predict the values at two prediction locations shown as red dots and marked by P1 and P2 on the plot.

Note: You should not try to compute the exact predicted values or prediction variances. You don’t have enough information to do that and you don’t need exact values to answer my questions. Use what you know (or have discovered by doing the HW) about the general behaviour of kriging predictions and kriging variances.

a) 4 pts. Which of the following four values is closest to the predicted value at P1: 2.0, 5.6, 10.0, or 34.6?

Briefly explain your choice.

b) 4 pts. Which of the following four values is closest to the predicted value at P2: 2.8, 5.6, 10.0, or 18.3?

Briefly explain your choice.

Ordinary kriging also estimates the prediction variance.

c) 4 pts. Which of the following three values is closest to the variance of the prediction at P1: 10.0, 34.6, 38.0? Briefly explain your choice.

d) 4 pts. Which of the following three values is closest to the variance of the prediction at P2: 10.0, 34.6, 38.0? Briefly explain your choice.

4) 4 pts. Residuals are often used to evaluate model assumptions. You use ordinary kriging to predict the values at the 8 locations used in problem 3. You calculate residuals, as the differences between the observed value and predicted value for each location. The plot of residuals against predicted values is below. What can you conclude from this plot? Briefly explain your answer.



5) 4 pts. A residual plot for a different data set is based on leave-one-out cross-validation predicted values and residuals. There are 100 observations in this data set. Based on this plot, do you have any concerns about assuming second order stationarity? Briefly explain your answer.



6) You want to compute distances between four locations (A, B, C, and D) scattered across the central US. Note: location C is in Louisiana. Each part of this question asks you to compare two ways to compute the distance. Your answer will be one of three choices: 1) you believe method a is more appropriate, 2) you believe method b is more appropriate, or 3) you believe the two distances will be very similar, e.g., within 1 km or less of each other, but don’t get hung up trying to define very similar. For each, briefly explain your choice. Your explanation is as important as your choice.



a) 4 pts. Consider locations A and B.

 Method a is Euclidean distance using UTM zone 15.

 Method b is great circle distance using long/lat.

b) 4 pts. Consider locations B and D.

 Method a is Euclidean distance using UTM zone 15 for both locations.

Method b is Euclidean distance using UTM coordinates in zone 15 for location B and UTM coordinates in zone 14 for location D.

c) 4 pts. Consider locations C and D. Long/lat for location C is measured using a GPS unit. For both methods, you will compute great circle distance.

 Method a gets long/lat for location D from a topo map using the NAD27 datum.

 Method b gets long/lat for location D from a topo map using the NAD83 datum.

**Data analysis problem (45 pts):**

7) The data in sstdata.csv are measurements of sea-surface temperature (SST) at 250 locations in the northern Gulf of Mexico (GOM) on one day in late February. Locations cover the area from Texas / NE Mexico on the west, to peninsular Florida on the east, Texas / LA / AL / MS / panhandle FL on the north, to 23.4° N on the south. The location is given in the lat (latitude) and long (longitude) variables, using a WGS84 datum. Sea surface temperature in °C is in the sst variable. The goals of the analysis are to estimate SST at some specific locations and to map SST.

The sstoutline.csv file has coordinates for the outline of the northern GOM.

The sstgrid.csv file has latitude and longitude coordinates for a fine grid across the northern GOM.

The sstgrid2.csv file has latitude and longitude coordinates for a coarse grid across the northern GOM.

I have done some preliminary analyses. You are welcome to repeat these, especially if you want to see more than I describe here, but your answer should not include any of the preliminary analyses.

Summary of Preliminary analyses:

1. There appears to be a trend in SST from N to S, and a weaker trend from W to E.
2. The empirical variogram, computed assuming a constant mean, looks linear.
3. You remember that a linear variogram often is the result of a trend in the mean.
4. Residuals after fitting a linear trend surface (lat and long) look approximately normal.
5. So, you decide that analyses should include a linear trend surface for the mean.
6. If there is evidence of spatial correlation in the residuals from the trend surface, predictions and your map should take advantage of that spatial correlation.

Three computing hints:

1. Problem size: Predicting SST on the fine grid may take a few tens of seconds (or more if a slower computer). Although the final product requires a map on the fine grid, I suggest you do preliminary computations / plots on the coarser grid. That’s why I give you sstgrid2.csv.
2. Examples of formulae that specify trend models are in prediction1.r. Look for swiss.ts2; the formula there is a linear trend surface with location stored in the x and y variables. The variable names in the trend model have to match names in the data set.
3. The range parameter for the Mat models is scaled differently. It is not the practical range, as it is for Sph, Exp, and Gau models. It’s usually much smaller; how much depends on the Matern k.

Analyze the data using what you believe is an appropriate method.

Provide:

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. (see below)

c) Predictions of SST at four locations: (-91.1, 28.6), (-86.2, 26.35), (-87.6, 24.6), and (-85.1, 24.0) (all long, lat), with the standard deviation of that prediction.

d) A map of SST using the fine GOM grid.

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.

An example of what I want for part b is:

* 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>

**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 no online chat and no 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.

I received the following assistance. Provide a short summary of who helped and what help you received.

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

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