Stat 605 – Spatial Statistics – Spring 2022
Homework 2. Due: Friday, January 21, start of class.

Please answer the following questions and submit your answers using Gradescope.

1. Categorize the following examples of spatial data as to their data type: geostatistical data, lattice data, point pattern data, or none-of-the-above. Explain your answers. Also, if what's being described is point pattern data, explain whether it's marked or unmarked point pattern data.
   (a) Distribution of oaks and pines in a forest stand;
   (b) Number of squirrel nests in the trees in (a);
   (c) Percentage of Republican voters in each state in the continental U.S.;
   (d) Concentration of a mineral in soil;
   (e) Amount of snowfall during January at 50 locations in Alaska;
   (f) Elevations in the foothills of the Allegheny mountains;
   (g) Locations of an animal across time, using values reported by an electronic tag;
2. Plots using R.
   (a) Plot the data set `ca20`, using the `plot` function as illustrated on page 17 of the lecture notes. (Note that when I tell you to create a plot, this means you need to include it in what you turn in.)
   (b) What is this data set? You can find documentation on it using the R command `help(parana)`. (If you choose to copy-and-paste description from R, you need to make sure you state that's what you've done. You must give credit to your sources.)
   (c) Describe the general features of this data set, e.g. N-S or E-W trend, if any; where are the values higher? Lower? What is the range of values in this data set? Describe the shape of the distribution. (In particular, does it appear to be symmetric? Skewed? Do you see any need to perhaps log-transform the data?)
   (d) If the data needs to be log-transformed, include the 4-panel plot you get after transforming, and discuss the resulting plots; does it look as though this is an appropriate transformation? If the data doesn't need to be log-transformed, just write "Does not apply." (It is entirely possible that your answer is, "It's not clear whether a log-transform is necessary or appropriate.")
   (e) Create `ggplot` versions of the four panel plots for the data on the original scale (not log-transformed). Try changing some of colors, font sizes, dot sizes, etc., just to see what happens. I have posted my R code for the `Ksat` data sets on Canvas under Files.
3. Repeat parts (a)-(d) for the `parana` data set.
4. Repeat parts (a)-(d) for the `wolfcamp` data set.

Continued on next page.

5. We sometimes use the function $f(d) = \exp(-d/\lambda)$ to express the correlation between observations made at locations separated by the distance d, where $d \geq 0$. Here, λ is a positive constant, which we treat as a known constant.

   (a) Show that $f(d) \geq 0$ and that $f(d) \leq 1$. (We know that correlations must lie between -1 and +1; so you've just shown that $f(d)$ might correspond to the value of a correlation between random variables.)

   (b) Let $\lambda = 1$. Find d so that $f(d) = 0.05$; in other words, find the (minimum) distance between observations in order for them to be 'almost' uncorrelated.

   (c) Let $\lambda = 5$. Find d so that $f(d) = 0.05$.

   (d) True / false; explain briefly. The larger λ is, the quicker correlations die out with increased distance between observations.

   (e) If d is measured in km, what are the units for λ?

6. We sometimes use the function $f(d) = \exp(-d\lambda)$ to express the correlation between observations made at locations separated by the distance d, where $d \geq 0$. Here, λ is a positive constant, which we treat as a known constant.

   (a) Let $\lambda = 5$. Find d so that $f(d) = 0.05$. Compare this with your answer to 5(c).

   (b) If d is measured in km, what are the units for λ?

7. Logarithms using different bases.

   (a) Find $\log_{10}(x)$ where $x = 2, 3, 4, 5$. Using R: `> signif(log10(2:5), 4)`

   (b) Find $\ln(x) = \log_e(x)$ where $x = 2, 3, 4, 5$. Using R: `> signif(log(2:5), 4)`

   (Note that when you use the function called `log` in R, it assumes you want log base e, i.e. natural logarithm.)

   (c) Find the ratios, $\log_{10}(x)/\ln(x)$, $x = 2, 3, 4, 5$, again rounded to 4 digits.

   `> signif( log10(2:5) / log(2:5), 4 )`

   (d) Find $\log_{10}(e)$, e.g. by typing `> log10( exp(1) )`

   (e) Find the ratios, $\ln(x)/\log_{10}(x)$, $x = 2, 3, 4, 5$

   (f) Find $\ln(10)$, e.g. `> log(10)`

   This essentially explains why I say that if you decide to log-transform (the response variable), the base doesn't really matter. Selecting a different base simply means all the y-values are multiplied by the same constant value.

   The relationship is this: if you're considering bases a and b, then

   $$\log_b(x) = k \log_a(x), \text{ where } k = \log_b(a)$$

   $$\log_a(x) = k' \log_b(x), \text{ where } k' = \log_a(b)$$