Week 4.

Stat 222, Spatial Statistics. Lecture MWF 9am, Math-Sci 5203.
Professor: Rick Paik Schoenberg, frederic@ucla.edu, www.stat.ucla.edu/∼frederic

DAY EIGHT. Monday, 4/23/01.

1) Examination of plots for coal-ash data, Cressie chapter 2.

- p32: Grid of observation locations.

- p34: Observations, on transformed locations.

North, South, East, West from now on mean transformed coordinates. The space has been rotated clockwise about 150 degrees. Easier to see rotated coordinates in top-left of p37. Compare with p32.

- p37: row and column means and medians.

Shows trends. Places where mean > median (like row 6 or column 12) are because of a few large outliers. Right-skew. Notice not much NS trend, but there is WE trend. Values get LOWER as you go East (the y-axis on bottom-left of p37 is weird). This cannot simply be explained by the fact that as you go East, there are more Northern observations, because there's not much a North-South trend.

- p39: $Z(s+e)$ vs $Z(s)$.

For the top, e is a unit vector to the East. For the bottom, e is a unit vector going North. Outliers are obvious. For both directions, for the most part large values are near large values. One expects the points to fall near the line $Y = X$. More scatter in the vertical direction.

- p41: Boxplot of variogram cloud.

The variogram cloud can be defined as the set of triplets $\{s_1, s_2, [Z(s_1) - Z(s_2)]^2\}$. You can then imagine indexing the locations and making a 3-d plot, where you put s_1 on the x-axis, s_2 on the y-axis, and the squared difference on the Z-axis. Such a plot may be called a variogram cloud. Alternatively, if you're interested in the variogram along a certain direction only, you can just fix some vector e. (Here e is a unit vector pointing East.) Then you could look at triplets $\{h, s, [Z(s+he) - Z(s)]^2\}$. Again, you can imagine a variogram cloud: a 3-d plot where h is on the x-axis, s is on the y-axis, and the squared difference is on the z-axis.

Notice that for each h, you have a bunch of values of the squared difference $[Z(s+he) - Z(s)]^2\}$: practically one for each observation, s. So for each h, you could take all those squared differences and make a boxplot of them. That is what is shown on p41.

DAY NINE. Wednesday, 4/25/01.

1) Continuing with the plots for coal-ash data.

• p41-42: Boxplot of variogram cloud and sqrt differences cloud.

p42 shows boxplots of squared differences $[Z(s+he)-Z(s)]^2\}$ versus h.

p43 shows boxplots of *square-root* differences $[Z(s+he)-Z(s)]^{1/2}\}$ versus h. Taking square roots seems to have a stabilizing effect; the distributions are less skewed to the right, since the few very large values are now not so prominent, since they are getting square-rooted, rather than squared.

Usually such a boxplot of the variogram cloud will be steadily increasing. It is important to look for departures from this: for example if the variogram seems constant everywhere, that suggests white noise (WN), and if the variogram suddenly increases or tapers off at some points, those are usually worth noticing.

Note how we're adding levels of complexity with each plot: first we just looked at the data on p37, then looked at one-unit differences on p39, and now h-unit differences on p41-42. Next, we'll look at how the variogram can be broken down by row or column, with the pocket plot.

• p44: Pocket plot.

Fix e = a unit vector pointing North. While the variogram cloud shows all contributions to the estimate of $\gamma(h)$ as h varies, now we'll see which rows contribute more to $\hat{\gamma}(h)$ and which contribute less. (If e were a unit vector pointing East or West, we'd be breaking it down by columns instead of rows.)

How does the pocket-plot work? Fix h. For the purpose of explanation, say $h=3$. So for instance row 1 and row 4 are h units apart. Same with row 2 and row 5. Take any such pair of rows and call them j and k. So for instance, j might be 1 and k might be 4.

Let $\bar{Y}_{j,k}$ = the mean of the square root of the absolute values of the differences between observations in row j and row k that are in the same column. For instance, on p34 we see that for $j=1$ and $k=4$ we have the differences $\sqrt{|10.59-10.94|}$, $\sqrt{|10.43-9.53|}$, and $\sqrt{|9.32-10.61|}$. Average these three square roots, and you get $\bar{Y}_{1,4}$.

Generally, one expects $\bar{Y}_{1,4}$ to be greater than $\bar{Y}_{1,3}$ and less than $\bar{Y}_{1,5}$. We want to investigate not just how big $\bar{Y}_{1,4}$ is, but how it compares to, say, $\bar{Y}_{2,5}$, $\bar{Y}_{3,6}$, $\bar{Y}_{4,7}$, etc. So let $\bar{\bar{Y}}_3$ denote the weighted average of all of these $\bar{Y}$'s, weighted by the number of pairs of differences in each $\bar{Y}$. (i.e. we'd give $\bar{Y}_{1,4}$ a weight of 3.) Then subtract $\bar{\bar{Y}}_3$ from each of the $\bar{Y}$'s. That is, take $\bar{Y}_{1,4}-\bar{\bar{Y}}_3$.

Let $p_{1,4}$ denote $\bar{Y}_{1,4}-\bar{\bar{Y}}_3$. $p_{1,4}$ indicates how much more (or less) rows 1 and 4 contribute to $\hat{\gamma}(3)$ than the typical pair of rows that are 3 units apart.

Now, consider $p_{1,2}$, $p_{1,3}$, $p_{1,4}$, $p_{1,5}$, ...

If there is something unusual about row 1, i.e. if row 1 is strangely different from all its neighboring rows, then these values $p_{1,2}$, $p_{1,3}$,... will tend to be positive. A pocket plot shows

a boxplot of the $p_{j,k}$'s, for each j. That is, on the x-axis of p44 is the "row number", which is j. On the y-axis is a box-plot. Corresponding to row 1, we have a boxplot of $p_{1,2}$, $p_{1,3}$, $p_{1,4}$, ...

[Question: What does it mean if for some j, the $p_{j,k}$'s are very *negative*? That means that row j is unusually *similar* to its nearby rows.]

From the figure on p44, we see that rows 2, 6, and 8 have mostly positive $p_{j,k}$'s. For rows 6 and 8 this is probably because of individual outliers (see p34). For row 2, there are lots of strangely low values in this row. So in the case of row 2, the pocket plot is not indicating an outlier, but rather a whole unusual row, which we might call a "pocket" of nonstationarity.

- p47: Median polished data.

Median polishing is a simple way of fitting the model:

$$Z(j,k) = f_0 + f_1(j) + f_2(k) + WN(j,k),$$

where j is the row, k is the column, and by f_0 we just mean some number indicating an overall center of the data, $f_1(j)$ is a row effect, $f_2(k)$ is a column effect, and $WN(j,k)$ is white noise, i.e. values that are independent and identically distributed with mean 0.

How does it work? a) First you take each row of the data, compute its median, and subtract the row median from each observation. Do this for each row. Then b) you take each column of the resulting differences and find its median, and subtract it from each observation. Do this for each column. c) Repeat, until convergence (i.e. until each row and column has median 0).

(I expanded on this a bit in class, showing some of the initial computations.)

After median polishing, one may look at the residuals (these are the entries in the middle of the table). See if they look like WN, and look for strangely large or small terms. Some jump out on p47, such as 6.66 at column 5, row 6. Also note that some rows and columns have unusually large or small median effects: for instance row 2 has a value of -1.52 and column 15 has -1.69.

Why median polish instead of mean polishing? Median polishing is resistant to outliers, in the sense that one outlier can do weird things to the mean polish, and also, it can be shown that estimates of the covariogram $C(h)$ are less biased when estimating it based on the residuals of median polishing than the residuals of mean polishing.

Also, note that for each observation on p34, its value corresponds to the overall effect (9.82, seen in the bottom-right corner of the table on p47) + row effect + column effect + residual. For instance, take column 5, row 2. The value is 8.75 on p34. From p47, we decompose this as

$(9.82) + (0.35) + (-1.52) + (0.10)$,

which is indeed equal to 8.75.

- p51: Variogram estimates of Z and of residuals.

On p51, robust means the mean-sqrt-difference to the fourth power. We can see how the classical estimator compares to the robust one. Top of p51 is the East-West variogram;

the North-South one is on p79. For the East-West direction, there is a row trend (see p37), causing the variogram to steadily increase.

The bottom of p51 shows the variogram of the residuals after median polishing. The variogram looks pretty flat, suggesting that the residuals don't look too different from white noise. This means that the model

$$Z(j,k) = f_0 + f_1(j) + f_2(k) + WN(j,k),$$

might not be unreasonable, in which case the only main features of the data are the row and column trends. This is why Cressie says on the bottom of page p50: "most of the correlation present in Figure 2.11 seems to be due to the trend".