Week 1.

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

DAY ONE. Monday, 4/2/01.

1) Administrative.

a) Stat 100 not required. See Dean Dacumos in Math-Sci 8142a to enroll.
b) Office hours: Wed 10am - 11:30am, Math-Sci 6167.
c) Course homepage: www.stat.ucla.edu/courses/stat222_1.php
d) Grading. Midterm (40%) Friday May 11 in class, Written project (50%) due Monday June 4, oral presentation (10%) last week of class. Midterm will be extremely easy. Written project is basically just find some interesting spatial data and analyze it (see course webpage). No extensions or makeups. Students with disabilities see me by 2nd week.
e) Texts. Cressie, "Statistics for Spatial Data", and Ripley, "Spatial Statistics". We will mostly use Cressie. Expensive ($130?), and not yet in the bookstore. Will be on reserve at SEL/EMS library.
f) No auditors.

If less than 8 students enroll....

* Mandatory attendance. Absences and lateness will severely hurt your grade. 2 absences or 3 latenesses is one grade down.
* Longer oral presentations. (5-10 min if 8 or more, otherwise 10-15 min).

2) Overview: geostatistical, lattice, and spatial point process data.

Geostatistical: $Z(s)$ at locations s in R^d. Maybe a continuum of s's.
Lattice: $Z(s)$ on a fixed collection of *connected or organized* points in R^d.
PP: $Z(s)$ at pts of a random point process on R^d.

3) Examples

Geostatistical: Pennsylvania coal-ash data.
Lattice: North Carolina SIDS data.
Point process: Florida pine trees data.
Numerous others (environmental, biomedical, meteorological, etc.)

DAY TWO. Wednesday, 4/4/01

1) Administrative and introductory remarks.

a) Ignore outline on web.

b) Attendance during lectures will not be required (except for last week of classes).

c) I will be instructing how to implement techniques in the computer program R. Manual: see Venables and Ripley.

d) Date of final projects and oral presentations may be changed. To be announced.

e) If book does not arrive by Friday, I will start putting notes on web.

(Go to http://www.stat.ucla.edu/courses/stat222_1.php and click on "handouts".)

We will begin by considering the geo-statistical case. Process Y takes values everywhere in some space like $\mathbf{R}^2$. For asymptotics, think of the region of observation extending to the entire space. For example, the square region $[-n, n] \times [-n, n]$ as $n \to \infty$.

2) Stationarity.

Complete: Dist of $Z(s)$ does not depend on s, and further, joint dist. of $\{Z(s); s \text{ in some set } \mathcal{S}\}$ is the same as the dist. of $\{Z(s); s \text{ in } \mathcal{S} + h\}$.

Second-order: $E[Z(s)] = E[Z(s+h)] < \infty$ for any h, and $Cov[Z(s), Z(s+h)] < \infty$ and only depends on h, not s.

Isotropic if 2nd-order stationary and $Cov[Z(s), Z(s+h)]$ only depends on $|h|$.

Intrinsic (Cressie, p40): $E[Z(s)] = E[Z(s+h)] < \infty$ and $V[Z(s+h) - Z(s)]$ doesn't depend on s and is $< \infty$.

Complete implies 2nd order, and intrinsic implies 2nd order.

Question: Does 2nd order imply intrinsic?

3) Variogram.

Main idea: values at locations near each other are NOT independent. But values at far-away locations are nearly independent.

For intrinsic stationary processes, $2\gamma(h) = V[Z(s+h) - Z(s)]$.

Typically $\gamma(h)$ (called the semivariogram) increases as $|h| \to \infty$.

Covariogram is estimated using

$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum\limits_{N(h)} [Z(s_i) - Z(s_j)]^2$, summed over $s_i - s_j = h$,

i.e. the mean square of these differences.

$N(h)$ is the set of all pairs of locations (s_i, s_j) that are h apart. By $|N(h)|$ we mean the number of pairs in this set.

Properties of the variogram:

$\gamma(0) = 0$.

$\gamma(-h) = \gamma(h)$.

Conditionally negative definite: $\sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0$.

Generally, $\frac{\gamma(h)}{|h|^2} \to 0$, as $|h| \to \infty$.

Two properties of the estimated variogram $2\hat{\gamma}(h)$:
The estimate is unbiased, but not robust. (More on this later in the course.)
Nugget effect: often $\hat{\gamma}(h) \not\to 0$ as $|h| \to 0$. If $\hat{\gamma}(h) \to C_0 > 0$ as $|h| \to 0$, then C_0 is called the nugget effect. This implies the process Y is not L^2-continuous, but instead must vary discontinuously at the smallest scales, in little nuggets of variation. With this variation (as with any variation), it is often hard to dsicriminate between measurement variability and intrinsic variability in the underlying process being observed (see Cressie, p. 89).

3) Some very basic isotropic models.

a) White Noise (WN).
$Y(s)$ are all iid random variables. $\gamma(h) =$ constant.

b) Constant/Deterministic.
$Y(s) = \mu$ for all s, where μ is some constant. $\gamma(h) = 0$. (Same, if $Y(s) = \mu(s)$, where μ is any deterministic function of s.)

c) 100%-Correlation.
$Y(s) = A$ for all s, where A is a random variable. e.g. precipitation at stations around UCLA campus, or exam scores where all students cheat off one student.
Classic example of a non-ergodic process. The mean $E[Y(s)]$ is not increasingly well estimated by the sample mean $\sum_i Y(s_i)$ as the range of observations stretches (i.e. as $n \to \infty$).

d) Linear variogram.
Suppose $Y(s)$ is instrinsically stationary with variogram
$\gamma(h) = C_0 + b_l|h|$, for $0 < |h|$.
Estimate scalar parameters C_0 and b_l. (l in b_l just stands for linear.)

DAY 3. Friday, 4/6/01

1) More isotropic models.

(For all of these, $Y(s)$ is assumed intrinsically stationary, with $\gamma(0) = 0$. Also notice that all these models are isotropic, and that $\gamma(h) \to C_0$ as $|h| \searrow 0$.)

e) Spherical variogram.

$$\begin{aligned}\gamma(h) &= C_0 + C_s[\frac{3|h|}{2a_s} - \frac{|h|^3}{2a_s^3}], for 0 < |h| < a_s.\\ &= C_0 + C_s, for a_s \le |h|.\end{aligned}$$

Estimate C_0, C_s, a_s.

f) Exponential variogram.

$$\gamma(h) = C_0 + C_e[1 - exp\{\frac{-|h|}{a_e}\}], for 0 < |h|.$$

Estimate C_0, C_e, a_e.

g) Rational quadratic.

$$\begin{aligned}\gamma(h) &= C_0 + C_r\frac{|h|^2}{1 + |h|^2/a_r}\\ &= C_0 + \frac{C_r a_r |h|^2}{a_r + |h|^2}, for 0 < |h|.\end{aligned}$$

Estimate C_0, C_r, a_r.

h) Wave.

$$\gamma(h) = C_0 + C_w[1 - \frac{a_w}{|h|} sin(\frac{|h|}{a_w})], for 0 < |h|.$$

Estimate C_0, C_w, a_w.

i) Power.

$$\gamma(h) = C_0 + b_p|h|^\lambda, for 0 < |h|, 0 \le \lambda < 2.$$

Estimate C_0, b_p, λ.

See Cressie, p63 for figures. Since isotropic, each can be drawn purely as a function of $|h|$.

2) Geometric anisotropy.

Suppose $\gamma(h) = \gamma(|Ah|)$, where A is some 2×2 matrix (for $\mathbf{R}^2$), or A is $d \times d$ for $\mathbf{R}^d$.

That is, can linearly transform (e.g. stretch out or rescale) the coordinates according to the transformation A, so that after transforming the axes, the process is isotropic, i.e. γ only depends on the *size* of Ah.

3) 2nd-order stationarity implies intrinsic stationarity.

Suppose $Y(s)$ is 2nd-order stationary, so $cov[Z(s), Z(s+h)] = C(h)$. Need to show that $V[Z(s+h) - Z(s)]$ is a function of h only, not s.

$$\begin{aligned} V[Z(s+h) - Z(s)] &= V[Z(s+h)] + V[Z(s)] - 2cov[Z(s+h), Z(s)] \\ &= cov[Z(s+h), Z(s+h)] + cov[Z(s), Z(s)] - 2cov[Z(s+h), Z(s)] \\ &= C(0) + C(0) - 2C(h). \end{aligned}$$

Now, does intrinsic imply 2nd-order? Suppose $Y(s)$ is intrinsically stationary, so $V[Z(s+h) - Z(s)] = 2\gamma(h)$. Need to show that $cov[Z(s), Z(s+h)]$ is a function of h only.

From the first equation above,

$$\begin{aligned} 2cov[Z(s+h), Z(s)] &= V[Z(s+h)] + V[Z(s)] - V[Z(s+h) - Z(s)] \\ &= V[Z(s+h)] + V[Z(s)] - 2\gamma(h). \end{aligned}$$

This highlights the distinction: $V[Z(s)]$ might depend on s.

Simple example of an intrinsically stationary process that is not 2nd-order stationary: Let the space $\mathcal{S}$ consist of just three locations, 1, 2, and 3. Then construct $Z(1), Z(2)$, and $Z(3)$ as follows. Flip a coin to see if $Z(1)$ is -1 or 1. Then flip a coin to see if $Z(2)$ is $Z(1) - 1$ or $Z(1) + 1$. Then flip a coin to see if $Z(3)$ is $Z(2) - 1$ or $Z(2) + 1$.

Verify that $E[Z(s)] = 0$, $V[Z(s+1) - Z(s)] = 1$, $V[Z(s+1) - Z(s)] = 2$, so $\gamma(h) = h$. But $cov[Z(1), Z(2)] = 1$, and $cov[Z(2), Z(3)] = 2$.

4) Projects and oral presentations.

Oral presentations will be on Wed June 6 and Friday June 8, in class. ATTENDANCE ON THESE DAYS IS MANDATORY FOR ALL STUDENTS. Presentations will be 5-7 minutes each (plus questions). Written projects will all be due on Friday, June 8, in class.

5) READER. Tuesday 4/10/01 or so, at Course Reader Materials, 1141 Westwood Blvd.