

# MAPPING AND MEASURING MULTIRESPONSE AIR POLLUTION FIELDS.

James V Zidek,  
University of British Columbia

January 16, 2003

## Abstract

**1. The simplest case.** A very active current area of environmental statistics involves the spatial prediction and modelling of air pollution fields. At the simplest level, one might imagine just two sites, one ("G" for "gauged") having a monitor and one ("U" for ungauged) not. Given say an hourly measurement at G, call it  $x$ , interest obtains in predicting  $y$  from it.

Since both  $x$  and  $y$  arise from natural processes, one would think that ideally, one could construct a physical model,  $\hat{y}_P = g_P(x)$  that would enable the unmeasured  $y$  to be predicted from  $x$ .

### QUESTION 1(a). DO SUCH PHYSICAL MODELS EXIST?

Alternatively, an enormous literature exists in statistics and geostatistics about data-based models  $\hat{y}_S = g_S(x)$  that accomplish the same thing (largely without knowledge of the environmental processes involved). As an example, we commonly see models of the form  $g_S(x) = a + bx$ . Many of the hierarchical Bayes models formally involve such things as Markov random fields or Kalman filters. Specifying these often complex models may involve subjective choices on the part of the statistician. Ideally these choices should be informed by subject area knowledge. Yet little has been done to incorporate such knowledge. (Exceptions: Alvo Meyer, some years ago, and Christopher Wikle and his co-authors very recently.)

A conceptually straightforward approach would input  $\hat{y}_P$  as the mean of the hierarchical model and then allow the data to amend it through Bayesian updating. However, this frontier in the development of prediction models deserves closer study and leads to the following question.

### QUESTION 1(b). HOW CAN $\hat{y}_P$ AND $\hat{y}_S$ BE SENSIBLY COMBINED?

[NOTE ADDED AFTER THE PRESENTATION: Discussion at the meeting gave an affirmative answer to Question 1(a) and that in fact, a number of those present were already working on this problem.]

The slope  $b = \sigma_{GU} / \sigma_{GG}$  of the statistical model involves the covariance  $\sigma_{GU}$  between  $x$  and  $y$ . It measures the degree of association between them.

If  $\sigma_{GU} = 0$ ,  $x$  and  $y$  do not co-vary and hence  $x$  proves a poor predictor of  $y$ . On the other hand, if it is large (and of necessity approaches its maximum possible value of 1, at least when  $\sigma_{UU} = \sigma_{GG}$ , while  $\sigma_{GU}$  approaches 0), then  $x$  will predict  $y$  almost perfectly. Clearly for (linear) statistical modelling, the spatial covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{UU} & \sigma_{UG} \\ \sigma_{GU} & \sigma_{GG} \end{pmatrix}$$

has an important role to play. However, one would expect this matrix to change in dynamic fashion with changes in wind direction, for example. This leads to a difficult question that has not been much studied if at all (Wikle and Dunsmuir have a related paper).

**QUESTION 1(c). HOW CAN PHYSICAL REASONING BE USED TO CONSTRUCT A DYNAMIC MODEL FOR THE SPATIAL CORRELATION MATRIX?**

**2. Multiple Sites.** In practice, urban areas will have a number of gauged (G) sites yielding say  $x_1, x_2, x_3$  and predictions will be needed for a number of ungauged (U) sites, say  $y_1, y_2$ . This leads to a more complex situation (where Question 1 still arises, of course.) However, the spatial covariance matrix will be of larger dimension:

$$\Sigma : 5 \times 5 = \begin{pmatrix} \sigma_{UU} : 2 \times 2 & \sigma_{UG} : 2 \times 3 \\ \sigma_{GU} : 3 \times 2 & \sigma_{GG} : 3 \times 3. \end{pmatrix}$$

**3. Multiple Responses.** Sites are almost never gauged to measure just one pollutant (partly because of the high start-up costs). Thus, each of the observations will be a vector, say:

$$x_1 : 3 \times 1 = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix}, \dots$$

The appropriate covariance is no longer spatial. In fact  $\Sigma : 15 \times 15$  must now represent both the association between sites as well as between responses within sites. One convenient representation of it would be a Kronecker product, justified under assumptions which, while plausible in certain situations, are highly implausible in others:

$$\Sigma = \Lambda : 5 \times 5 \otimes \Omega : 3 \times 3,$$

where  $\Lambda$  represents between site association,  $\Omega$  within site. This form is easily interpreted and moreover, greatly reduces the number of parameters to be estimated. Moreover, in unpublished work I have a generalized Kronecker structure that deserves some consideration. However, the importance of this covariance structure, particularly with the development of speciation networks (Brian Eder talked about at our last meeting) and response vectors potentially of dimension exceeding 100, we confront the following question:

**QUESTION 2. WHAT ARE APPROPRIATE, PRACTICAL COVARIANCE STRUCTURES FOR MULTI-RESPONSE ENVIRONMENTAL PROCESSES?**

**3. Time-Series.** Here the response vectors  $x_{t1}, \dots$  obtain at a sequence of times  $t = 1, \dots$ . The situation now is more complicated in that we have both “temporal association” (our *enemy*) and “spatial association” (our *friend*). Moreover for say, hourly data series they are inseparable. Any attempt to “factor out” our enemy so as to side-step its deleterious statistical consequences leads to the loss of our friend and any basis we might have had for successful spatial prediction. Naive approaches just do not work. Nhu Le and I along with our co-investigators have an approach to this problem that works in some applications. However, the topic needs further study since regulators and exposure modellers increasingly rely on small time averages of pollution fields.

(Hierarchical Bayes) statistical models (e.g. the Kalman filter of Mardia and Goodall, and the extended geostatistical models by Don Meyer and his collaborators) have recently been proposed for such processes. However, the subject deserves further study and I am not sure any of the approaches to date have provided a satisfactory answer to the following question in applications.

**QUESTION 3. HOW CAN THE INSEPARABILITY OF SPACE AND TIME BE HANDLED PRACTICALLY IN THE ANALYSIS OF DATA FROM SPACE-TIME PROCESSES.**

[The following was added after the presentation because it relates to two of the other presentations given the same day.]

**Design.** Where should new G's be placed and which if any of the current ones should be removed? The hierarchical Bayes, spatial prediction theory due to me and my co-investigators does provide one answer to that question, based on maximizing an entropy. However, our answer and others lead to the maximization of objective functions and in turn, (the np-hard problem of) combinatorial optimization along with the need for a practical optimization algorithms. Jon Lee and his co-investigators have produced an exact solution using branch and bound techniques, but these can only handle networks of up to 80 possible G sites from which as many as 40 are to be selected. This number is small compared to the numbers Dave Holland talked about at our previous meeting. In any case, we do not yet have a fully satisfactory practical solution to the following question.

**QUESTION 4. WHERE SHOULD THE GAUGED SITES BY PLACED?**