

# Spatial statistics in environmental epidemiology

The planet earth is rich in carbon, oxygen, and hydrogen; it is third from the sun in an orbit that delivers just the right amount of heat and light; and it travels through space with a thin shell of protective ozone that has been crucial to the evolution of terrestrial-based organisms. Our environment has in recent times been seen as delicate and not to be taken for granted. The study of the interaction of one or many organisms with its environment is known as ecology. In this entry, we shall concentrate on one organism, the species *Homo sapiens*, which is unique in its ability to modify the environment of all organisms on this planet.

Changes in the environment, anthropogenic or natural (*see* **Global environmental change**), can have deleterious effects on humans. The study of these effects can range in scale from the molecular to the demographic. This entry is concerned with the demographic scale, where the study of disease and its progression through the population of *H. sapiens* is known as epidemiology.

Environmental epidemiology is a discipline that relies to a large extent on observational studies; even the case–control studies that are a standard tool of epidemiologists [3] (*see* **Deviance function**) do not approach the rigor of the design of agricultural field trials [14]. There is a very real tension in environmental epidemiology between the desire to establish a network of causation and the lack of experimental control available in observational studies. However, when the ‘signal’ is strong enough, sometimes it is possible to establish causative relationships.

One of the earliest success stories in environmental epidemiology is the study by Snow [26] that related the street addresses of cholera victims to putative pollution sources, water supply pumps. Through the use of maps and the careful study of each case of a cholera epidemic in London in 1848, John Snow discovered that cholera is a waterborne infectious disease (*see* **Disease mapping**). This is all the more amazing because it was done some 20 years before Koch and Pasteur established the beginnings of microbiology as we know it today.

To make progress with noisy data and vaguely formulated hypotheses, statistical models have been developed that allow inference even down to individual levels, such as in frailty models [4]. Many of these models are hierarchical and some are spatial, and it is on these we shall concentrate in this entry. Methods of longitudinal data analysis [18], which emphasize changes in disease state over time, are not addressed here; see instead the entry on **longitudinal studies**.

## Spatial Statistical Methods

Spatial aspects of environmental epidemiology can be extremely important; see [16] for an extensive review of spatial statistical methods used in this field. Disease incidences that cluster geographically can sound alarm bells among the general public and government health agencies. Are such disease clusters associated with pollution sources? Raised case intensities near a source of ionizing radiation (*see* **Radioactivity**), or directional preference related to a dominant wind direction near a waste-product incinerator, may provide evidence of a link between environmental insult and human health. The types of data observed can vary from case locations (usually residence addresses) to case counts within small areas (that are often administratively defined, such as counties, zip codes, etc.). Spatial point process models (*see* **Point processes, spatial**) are appropriate for analyzing case-location data [1, 8], and **Markov random field models** have often been used for small-area counts [2, 24] (*see* **Small area estimation**). A review of statistical methods for detecting spatial patterns of disease around putative sources of hazard is given by Lawson and Waller [17], and a review of spatial smoothing of small-area disease maps is given by Cressie et al. [7].

Environmental epidemiologists search for a causal link between a spatial variable  $S$  measuring pollution and the variable  $Y$  giving case locations or small-area case counts. Typically, both  $S$  and  $Y$  vary spatially and are observed with error. Since the statistical model for  $Y$  usually requires knowledge of  $S$ , some preliminary spatial (usually geostatistical – see below) analysis is done to obtain  $\hat{S}$ , which is then substituted into the model in place of  $S$  as if it were known perfectly. In the example that follows, it is seen that there is a better way to deal with the various sources of spatial variability.

**Example of a Spatial Hierarchical Statistical Model**

This example is taken from Cressie [6]. Suppose that a spatial domain  $D$  is made up of disjoint regions  $A_1, \dots, A_k$ , and that the number of nonaccidental deaths of people aged 65 years or older in a one-month period is counted in each region. For illustrative purposes, we assume that this population at risk is homogeneous. Should it be heterogeneous, stratification (e.g. male/female) is needed; however, the concepts given below remain unchanged. Let  $Y_1, \dots, Y_k$  denote the mortality counts in regions  $A_1, \dots, A_k$ , respectively. Suppose also that ambient air pollution measures (e.g. particulate matter with mean diameter less than or equal to  $10\mu\text{m}$ , often abbreviated to  $\text{PM}_{10}$ ) are available, and that we wish to relate the mortality counts with the severity of the pollution. This is a common problem environmental regulators face when trying to set pollution standards in the interest of public health (*see **Standards, environmental***). However, many times the associated epidemiological analyses are not spatial, the investigators having aggregated across regions. Unfortunately, this could incur a severe ecological bias, as is discussed in a subsequent section. More and more, regulations are being written for smaller regions, invoking the principle of **environmental justice**; see [27]. In these situations, a proper statistical analysis requires spatial statistics.

Let  $S(\mathbf{s})$  denote the logarithm of the  $\text{PM}_{10}$  value at location  $\mathbf{s} \in D$ . Although  $S(\cdot)$  is the exposure variable with which we wish to work, we have incomplete knowledge of it. We only observe it at monitoring stations located at  $\mathbf{s}_1, \dots, \mathbf{s}_n$  and, importantly, when we do take observations, it is hidden from us by a measurement-error process  $\varepsilon(\cdot)$ . That is, the observation process is

$$Z(\cdot) = S(\cdot) + \varepsilon(\cdot) \tag{1}$$

where  $\varepsilon(\cdot)$  is a zero-mean, white-noise process with variance  $\sigma^2$ . To make inference on  $S(\cdot)$ , we can use a geostatistical methodology known as **kriging** [5, Chapter 3; 19]. Specifically, let

$$\widehat{S}(\mathbf{s}) = \sum_{i=1}^n \lambda_i(\mathbf{s})Z(\mathbf{s}_i), \quad \mathbf{s} \in D \tag{2}$$

where the coefficients are chosen to minimize mean squared prediction error, subject to the unbiasedness condition,  $E[\widehat{S}(\mathbf{s})] = E[S(\mathbf{s})]$ ,  $\mathbf{s} \in D$ .

Let us now turn to the mortality counts. The spatial model for  $Y_1, \dots, Y_k$  is rather simple. Consider a generic random vector  $\mathbf{X}$  and let  $[\mathbf{X}]$  denote its (joint) density. We assume that, conditional on the hidden mortality process  $\lambda(\cdot)$ ,

$$\begin{aligned} [Y_1, \dots, Y_k] &= [Y_1] \dots [Y_k] \\ [Y_i] &= \text{Poi} \left[ \int_{A_i} \lambda(\mathbf{u})\delta(\mathbf{u}) \, \mathbf{d}\mathbf{u} \right], \\ & \quad i = 1, \dots, k \end{aligned} \tag{3}$$

where ‘Poi’ denotes the Poisson distribution,  $\lambda(\cdot)$  is the intensity function of the underlying point process of mortality events, and  $\delta(\cdot)$  is the density of the (homogeneous) population at risk. Then the mortality rate for region  $i$  is  $Y_i / \int_{A_i} \delta(\mathbf{u}) \, \mathbf{d}\mathbf{u}$ ,  $i = 1, \dots, k$ . We condition on knowing  $\delta(\cdot)$ , although in reality it has probably been estimated from demographic studies.

Conditional on  $\delta(\cdot)$ , we can characterize mortality through  $\lambda(\cdot)$ ; it is common to link mortality and exposure through the loglinear relationship

$$\lambda(\mathbf{s}) = \exp[\alpha + \beta S(\mathbf{s})], \quad \mathbf{s} \in D \tag{4}$$

Now,  $S(\cdot)$  is known imperfectly, and so an ad hoc solution is often obtained by substituting  $\widehat{S}(\cdot)$  into (4), in place of  $S(\cdot)$ :

$$\widehat{\lambda}(\mathbf{s}) \equiv \exp[\alpha + \beta \widehat{S}(\mathbf{s})], \quad \mathbf{s} \in D \tag{5}$$

While this might seem to be a reasonable solution based on technology available in the 1980s and 1990s, it is unsatisfactory from several more modern points of view.

First, the spatial variability in  $\widehat{S}(\cdot)$  is inherently different from that of the original process  $S(\cdot)$ ; by construction, kriging is much smoother than the process  $S(\cdot)$ , although it is calibrated to be unbiased. Second, even though kriging is unbiased,  $E[\widehat{\lambda}(\mathbf{s})] \neq E[\lambda(\mathbf{s})]$ . The first and second points can be addressed somewhat satisfactorily by developing constrained kriging [6, p. 289], which takes the same linear form as (2) but now is calibrated so that  $E[\widehat{S}(\mathbf{s})] = E[S(\mathbf{s})]$  and  $\text{var}[\widehat{S}(\mathbf{s})] = \text{var}[S(\mathbf{s})]$ . This last constraint ensures that  $E[\widehat{\lambda}(\mathbf{s})] = E[\lambda(\mathbf{s})]$ .

The third point is the difficulty of doing valid inference on  $\alpha$  and  $\beta$  given in (4). The negative log-likelihood of  $\alpha$  and  $\beta$  is, up to an additive constant,

$$l(\alpha, \beta) = \int_D \lambda(\mathbf{u})\delta(\mathbf{u}) \, \mathbf{d}\mathbf{u} - \sum_{i=1}^k Y_i \log \left[ \int_{A_i} \lambda(\mathbf{u})\delta(\mathbf{u}) \, \mathbf{d}\mathbf{u} \right] \quad (6)$$

where recall that  $\lambda(\mathbf{u}) = \exp[\alpha + \beta S(\mathbf{u})]$ . Minimizing (6) to obtain **maximum likelihood estimates** of  $\alpha$  and  $\beta$  is straightforward if  $S(\cdot)$  were known. If we substitute  $\hat{\lambda}(\cdot)$  given by (5), into (6), we obtain a **pseudo-likelihood function**  $\hat{l}(\alpha, \beta)$  (in the sense of Gong and Samaniego [12]). This could then be minimized with respect to  $\alpha$  and  $\beta$ .

However, proper statistical inference must recognize that (5) is used instead of (4) and hence it must account for the covariational properties of the small-area counts  $\mathbf{Y}_{k \times 1} \equiv (Y_1, \dots, Y_k)'$ , and the geostatistical data  $\mathbf{Z}_{n \times 1} \equiv [Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)]'$ . This is a formidable task and will almost certainly rely on the **delta method** and/or other asymptotic approximations. Fortunately, there is another way, namely a general approach to solving scientific questions through hierarchical statistical modeling (*see* **Hierarchical model**).

In the context of relating mortality with particulate-matter exposure, the two substudies, one geostatistical and one small-area counts, can be linked naturally at a higher level of the hierarchy. Replace (4) with:

$$[\log \lambda(\cdot)] = \text{Gau}[\alpha + \beta S(\cdot), C(\cdot, \cdot; \theta)] \quad (7)$$

where ‘Gau’ denotes a **Gaussian process**, here with covariance function

$$C(\mathbf{u}, \mathbf{v}; \theta) \equiv \text{cov}[\log \lambda(\mathbf{u}), \log \lambda(\mathbf{v})] \quad (8)$$

that depends on parameters  $\theta$ . Explanatory models of the sort given by (7) are motivated by the descriptive geostatistical models found in [9]. A fully Bayesian approach would create a higher level, where a **prior distribution**  $[\alpha, \beta, \theta, \sigma^2]$  is specified for the unknown parameters (*see* **Bayesian methods and modeling**). An empirical Bayesian approach would look for frequentist-based estimating equations for  $\alpha$ ,  $\beta$ ,  $\theta$ , and  $\sigma^2$  that depend on the count data  $\mathbf{Y}$  and the exposure monitoring data  $\mathbf{Z}$ ; for the purpose of

process prediction, these estimates would then be treated as the true values of the parameters.

In subsequent discussion of this example, we consider the fully Bayesian approach. However, it is important to emphasize that there is a place for **empirical Bayes methods** in the preliminary stages of a scientific investigation when little is known about the parameters. Frequentist estimation of  $\alpha$ ,  $\beta$ ,  $\theta$ , and  $\sigma^2$ , together with good model diagnostics, can lead to better model specification at *each* level of the hierarchy (e.g. perhaps the relation (4) should not be loglinear). And familiarity with ranges of  $\hat{\alpha}$ ,  $\hat{\beta}$ ,  $\hat{\theta}$ , and  $\hat{\sigma}^2$  from previous, like studies will lead to better prior distributional assumptions.

Inference from the fully hierarchical Bayesian model (*see* **Hierarchical Bayes**) proceeds from the joint posterior distribution,

$$[\lambda(\cdot), S(\cdot), \alpha, \beta, \theta, \sigma^2 | \mathbf{Y}, \mathbf{Z}] \quad (9)$$

which is obtained via Bayes’ Theorem. While it is typically not possible to obtain this distribution analytically, we can usually simulate from it using **Markov chain Monte Carlo** (MCMC) methods. Each hierarchical model has its own special features, which means that care is required in constructing the MCMC, but the principles of the construction are straightforward. A **Markov chain** simulation is constructed from conditional distributions that are generally immediately available from the hierarchical specification [11, 25].

This example illustrates what we believe the future will hold for spatial statistics in environmental applications. Rather than a piecemeal analysis, a hierarchical model consisting of (1), (3), (7), and perhaps a prior, enables exact finite-sample inference (up to the approximation afforded by MCMC). One of the attractive things about such a simulation-based inference is that the **posterior distribution** of the mortality intensity  $[\lambda(\cdot) | \mathbf{Y}, \mathbf{Z}]$  is immediately available by looking only at realizations of  $\lambda(\cdot)$  from the MCMC samples. Or, for that matter, inference on any function of the parameters is also immediately straightforward upon computing that function for each of the MCMC samples.

### Ecological Bias

There is one feature about the behavior of spatial statistical models that deserves special mention. Spatial

## 4 Spatial statistics in environmental epidemiology

---

aggregation from neighborhoods into regions, from regions into states, and from states into continents, is natural. Health data may be collected at one level of aggregation, environmental covariates at another, and policy decisions might be made at a third. However, changing the level of spatial aggregation can lead to completely different conclusions about, for example, the effect of particulate matter on respiratory illness (see **Respiratory epidemiology**). This phenomenon goes under the rubric of *ecological bias* or *ecological regression* in the epidemiology literature [13, 24], and under *modifiable areal unit problem* in the geography literature [10]. Statisticians know it as Simpson's paradox, often seen with **categorical data**. If one is designing a health-effect study a priori, it is possible to use multilevel designs that also collect covariate information at the individual level [21]. Otherwise, one could try to resolve this problem by building models at the point (or individual) level and then aggregating up to the desired level for inference. This is often difficult to do properly because covariates do not enter linearly, nor are they always available at the desired level of aggregation; typically, approximations based on Taylor-series expansions are used (e.g. [22]).

If one is capable of specifying all joint distributions and probabilities (up to a finite set of unknown parameters), fully or empirical Bayesian predictors of individual-level quantities are, in principle, available from MCMC (e.g. [20]).

### Spatio-temporal Analysis

Spatial analyses are more interesting and more consequential if a temporal component can be incorporated. Causation is much easier to investigate when the data can be modeled dynamically in time. Using a hierarchical model, manifestly complex space-time interactions can be handled straightforwardly at higher levels of the hierarchy (e.g. [15]).

In the example given above, where a model was proposed for case counts and  $PM_{10}$  exposure data, time was averaged out over a one-month period. There is, however, strong evidence to suggest that **meteorology** has an important effect on  $PM_{10}$  readings. Because meteorology is inherently dynamic,  $PM_{10}$  should be too, and hence one could modify the spatial model to incorporate both time and meteorological conditions

at the second level by modeling spatio-temporal processes that are hidden behind incomplete and noisy observations.

### References

- [1] Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A* **154**, 143–155.
- [2] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**, 9–25.
- [3] Breslow, N.E. & Day, N. (1980). *Statistical Methods in Cancer Research*, Vol. 1, *The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon.
- [4] Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models, *Biometrics* **47**, 467–485.
- [5] Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition, Wiley, New York.
- [6] Cressie, N. (1999). Statistical analysis of data from a Geographic Information System, in *GIS en Waarachtig! Symposium Statistische Software 1999*, Geodan, Amsterdam, pp. 21–38.
- [7] Cressie, N., Stern, H.S. & Reber Wright, D. (2000). Mapping rates associated with polygons, *Journal of Geographical Systems* **2**, 61–69.
- [8] Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point, *Journal of the Royal Statistical Society, Series A* **153**, 349–362.
- [9] Diggle, P.J., Tawn, J.A. & Moyeed, R.A. (1998). Model-based geostatistics, *Applied Statistics* **47**, 299–326.
- [10] Fotheringham, A.S. & Wong, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis, *Environment and Planning, Series A* **23**, 1025–1044.
- [11] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.
- [12] Gong, G. & Samaniego, F.J. (1981). Pseudo maximum likelihood estimation: theory and applications, *The Annals of Statistics* **9**, 861–869.
- [13] Greenland, S. & Robins, J. (1994). Ecological studies – biases, misconceptions, and counterexamples, *American Journal of Epidemiology* **139**, 747–760.
- [14] Kempthorne, O. (1952). *Design and Analysis of Experiments*, Wiley, New York.
- [15] Knorr-Held, L. & Besag, J.E. (1998). Modeling risk from a disease in time and space, *Statistics in Medicine* **17**, 2045–2060.
- [16] Lawson, A. & Cressie, N. (2000). Spatial statistical methods for environmental epidemiology, in *Handbook*

- of *Statistics*, Vol. 18, P.K. Sen & C.R. Rao, eds, Elsevier, Amsterdam, pp. 357–396.
- [17] Lawson, A.B. & Waller, L. (1996). A review of point pattern methods for spatial modelling of events around sources of pollution, *Environmetrics* **7**, 471–488.
- [18] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.
- [19] Matheron, G. (1963). Principles of geostatistics, *Economic Geology* **58**, 1246–1266.
- [20] Mugglin, A. & Carlin, B.P. (1998). Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones, *Journal of Agricultural, Biological, and Environmental Statistics* **3**, 111–130.
- [21] Navidi, W., Thomas, D., Stram, D. & Peters, J. (1994). Design and analysis of multilevel analytic studies with applications to a study of air pollution, *Environmental Health Perspectives* **102**, 25–32.
- [22] Plummer, M. & Clayton, D. (1996). Estimation of population exposure in ecological studies, *Journal of the Royal Statistical Society, Series B* **58**, 113–126.
- [23] Richardson, S. (1992). Statistical methods for geographical correlation studies, in *Geographical and Environmental Epidemiology: Methods for Small Area Studies*, P. Elliot, J. Cuzick, D. English & R. Stern, eds, Oxford University Press, Oxford, pp. 181–204.
- [24] Richardson, S., Montfort, C., Green, M., Draper, G. & Muirhead, C. (1995). Spatial variation in natural radiation and childhood leukaemia incidence in Britain, *Statistics in Medicine* **14**, 2487–2501.
- [25] Smith, A.F.M. & Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, Series B* **55**, 3–23.
- [26] Snow, J. (1855). *On the Mode of Communication of Cholera*, Churchill, London. [Reprinted by Hafner, New York (1965).]
- [27] Waller, L.A., Louis, T.A. & Carlin, B.P. (1999). Environmental justice and statistical summaries of differences in exposure distributions, *Journal of Exposure Analysis and Environmental Epidemiology* **9**, 56–65.
- (See also **Distribution function; Morbidity and mortality; Population health surveillance; Space-time covariance models**)

NOEL CRESSIE