

# Impact of Cliff and Ord (1969, 1981) on Spatial Epidemiology

Sylvia Richardson<sup>1</sup> and Chantal Guihenneuc-Jouyau<sup>2</sup>

<sup>1</sup> Centre for Biostatistics and MRC-HPA Centre for Environment and Health, Imperial College London, UK.

<sup>2</sup> University Paris Descartes MAP5 UMR CNRS 8145, Paris, and INSERM U780, Paris, FR.

## 1 Introduction

As defined by Elliott and Wartenberg (2004), spatial epidemiology is the description and analysis of geographic variations in disease with respect to demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors. Spatial analyses abound in the epidemiological literature, with an early example found in the monograph edited by Doll (1984) on the geography of disease, in which large scale geographical variations of mortality for a number of chronic diseases were used to formulate hypotheses on the potential influence of life-style and environment. The paper on testing spatial autocorrelation by Cliff and Ord (1969) was highly instrumental in encouraging going beyond the simple display of disease maps and environmental atlases. It gave impetus to the development of a more formal statistical analysis framework, aimed at finely characterising the scale of spatial dependence and the type of geographical patterns, whether in the disease rates themselves or in the residuals of geographical correlation studies. Recent books on *Spatial Epidemiology* by Elliott et al (2000), Lawson (2006, 2008) and Pfeiffer et al (2008) all discuss the central concept of autocorrelation formalised in Cliff and Ord (1969) in order to lay the foundations for more sophisticated analyses. They also describe a series of recent examples, which show how the field has moved on from large scale descriptive studies towards more powerful small area studies that take advantage of the advances in geographical information systems.

## 2 Characterisation of spatial disease patterns

Before computing any autocorrelation index, it is important to pay attention to the geographical scale and resolution of the data, which will both influence autocorrelation measures as shown in Griffith et al. (2003). Moreover, spatial gradients, analogous to the trend observed in many time series, will influence the autocorrelation indices considered in Cliff and Ord (1969, 1981), as these will capture any non stationary patterns of variation. In spatial epidemiology, the influence of gradient-like contrasts on the epidemiological interpretation of association results was highlighted in the early work of Lazar (1982) and Pocock et al (1982), in which strong regional contrasts were shown for both disease and environmental variables, resulting in fewer degrees of freedom for interpretation than if substantial local heterogeneity is observed.

Following on from their 1969 paper, Cliff and Ord (1981) introduced a useful general definition of spatial autocorrelation of a variable  $Y_i$  as  $\sum_{i,j} w_{ij} A_{ij}$  with weights  $w_{ij}$  chosen to measure spatial closeness of areas  $i$  and  $j$ , and  $A_{ij}$  a function quantifying the dependence between the values of  $Y_i$  and  $Y_j$ . Their discussion highlighted two important interlinked aspects: first that

space had to be treated quite differently from time as there is no natural ordering, and second that there was a need to characterise *a priori* the form of spatial dependence of interest. Indeed, the chosen form of the weights  $w_{ij}$  is related to the scale and the type of dependence one is trying to measure, with binary weights based on adjacency being the most commonly used to study local dependence of disease outcomes, whereas distance based weights are more appropriate when a smoothly varying dependence is hypothesised, such as could be observed in an environmental pollutant.

In spatial epidemiology, spatial autocorrelation across a map is more commonly referred to as *clustering* (not to be confused with localised clusters). Walter (1993) describes the performance of autocorrelation statistics in the analysis of regional cancer incidence data, Wakefield et al (2000a) further point to some extensions of the basic definition and to connections with moving windows and related methods. One important aspect that was highlighted in Besag and Newell (1991) is that the significance of autocorrelation of disease rates, e.g when taking  $Y_i$  to be a Standardised Mortality Ratio (SMR), can be influenced by unequal variances of the SMRs, and that spurious autocorrelation might arise simply from the occurrence of high variable rates in small populations. In the same spirit, the performance of Moran's  $I$  under heteroscedasticity was studied by Waldhor (1996), who proposed an approximation of the moments of  $I$  that accounted for population size. Tango's statistic, which is based on proportions rather than rates and account for their inherent sampling variability, was proposed instead of Moran's  $I$  (Tango, 1995, Wakefield et al. 2000a). Here, a smooth dependence with distance is used to measure the closeness of areas. Recently, Kulldorff et al (2006) have performed a comprehensive investigation of a variety of clustering indices in the context of cancer mapping, which confirm that Moran's  $I$  is not adapted to the testing of autocorrelation of count data and that alternative indices have better performance.

Autocorrelation is nowadays used most commonly at a *latent level* than directly on epidemiological data, due to possible confusion between the Poisson noise and the true spatial dependence, particularly for small area noisy data, which has become increasingly the focus of interest. Hence, the emphasis has shifted from *testing* autocorrelation to *including* a spatial structure at a second level in hierarchical disease mapping models (see e.g. Wakefield et al, 2000b and Banerjee et al 2004). Weights  $w_{ij}$  are part of the formulation of a large class of spatial models, the conditional autoregressive (CAR) models (Besag et al. 1991). To be precise, after accounting for Poisson or binomial fluctuations of disease counts  $Y_i$  at the first level, the underlying (latent) (log)relative risks  $\theta_i$  are modelled using a conditional gaussian distribution with mean  $\mu_i$  and variance  $\sigma_i^2$ , where  $\mu_i$  is linked to the mean of the neighbouring  $\theta_j$  via an autoregressive formulation:  $\mu_i = \sum_j w_{ij}\mu_j$ , with additional constraints on the weights and the conditional variances in order to ensure symmetry of the variance covariance matrix of the  $\theta_i$ s. Thus, 'autocorrelation' has become an integral part of the hierarchical toolkit widely used in disease mapping.

In parallel to the frequent use of binary weights when computing autocorrelation coefficients, commonly used CAR models in epidemiology also rely on binary weights. In a simulation study of disease mapping models, Best et al. (1999) perform a comparison between binary and distance based weights and note more sensitivity to the choice of weights than to usual gaussian assumptions. Beyond the definition of weights, the type of spatial model included in hierarchical disease mapping models has also been extensively discussed, with conditional, convolution, mixture and partition based, and joint models based on specifying directly the variance-covariance matrix of the  $\theta_i$ , being proposed (see Besag et al., 1991, Wakefield et al, 2000b, Green and Richardson, 2002, Knorr-Held and Rasser, 2000, Lawson, 2008). A recent comparison of a large range of Bayesian spatial models for disease mapping shows that using

direct spatial modelling of variance-covariance matrix may lead to oversmoothing of the true disease rates, while convolution and mixture models adapt more readily to local changes (Best et al. 2005).

There is now wide spread use of hierarchical disease mapping models in chronic (e.g. prostate cancer in the UK, Jarup et al. 2001, thyroid cancer in New Caledonia, Truong et al. 2007), infectious diseases (e.g. cholera in Mexico, Borroto et al, 2000) and social epidemiology (e.g. to study neighborhood characteristics, Aucoin 2007), demonstrating eloquently the blooming legacy of Cliff and Ord seminal paper in this domain.

### **3 Importance of taking into account spatial autocorrelation in investigations of geographical association between disease and risk factors**

Analysis of geographic variations in incidence or mortality in relation to exposure to environmental variables is a subject of great interest in public health. The aim of such analyses is mainly to assess and test associations between health indicators and life-style or environmental exposure. But, as discussed in Cook and Pocock (1983), Richardson (1992) and Richardson and Monfort (2000), the potential spatial structure in each involved variables requires specific attention in order to be able to interpret associations meaningfully in a context of a regression model. The characterization of spatial patterns introduced by Cliff and Ord is then of great importance in the context of multivariate geographical regression. Different methodological aspects have been developed focussed on either modifying existing tests that assumed independence or alternatively, on directly modelling the autocorrelation structure of the regression residuals (see review by Marshall 1991).

First, authors have proposed to modify tests of association between health indicators and environmental variables to account for their spatial dependence. Clifford et al. (1989) propose a general method of testing such association based on computing the cross-correlation between two processes. The main idea is then to estimate an 'effective sample size' taking into account the spatial autocorrelation of each process, in order to adjust for the existing pattern of dependence in each data set. Applications of this procedure to study relationships between male lung cancer and different types of industry was discussed in Richardson et al (1992). Discussion and extension of this work can be found in Dutilleul et al. (1993, 2008) and in Fotheringham et al. 2009 (chapter 6).

Other approaches consider how to model the autocorrelation structure of regression residuals. A range of models for the covariance matrices of residuals have been suggested that are based on specifying different forms of spatial dependence (exponential, disc, Bessel ...) (see for instance Ripley 1981). Comparison of such models and evaluation of the impact of choosing one particular spatial model for the residuals on the estimation of the association coefficients of interest are studied by several authors (Richardson et al 1992, Wakefield and Morris 1999, Diggle et al 1998). Clear evidence of sensitivity is apparent. Many recent applications of such spatial regression modelling can be found in the epidemiology literature, as for example in Havard et al. (2009) where the relationship between air pollution and socioeconomic status is examined in France at a small area level.

Finally, as in disease mapping, current ecological regression approaches typically include the autocorrelation structure via hierarchical models. Poisson or Binomial generalized linear models are modified to include both a linear predictor containing fixed effects of a number of

ecological covariates and a random effects part. Clayton and Bernardinelli (1992) and Clayton et al. (1993) propose to include random effects that follow a CAR model together with an ecological regression part, and apply this to lung cancer in Sardinia and lip cancer in Scotland. Such models of extra spatial variability can be seen as natural extension of hierarchical models for mapping disease risk to ecological regression approaches.

Ecological correlation studies are now mostly done using such hierarchical formulation. For instance, Van Leeuwen (1999) studies association between stomach cancer and water contamination, Biggeri (2005) association between lung cancer and atmospheric pollutants. Hierarchical formulation allows also to consider latent geographical covariate as in Maheswaran et al (2006). The problem of simultaneously modelling data obtained on different geographical scales is a subject of current interest, see Best et al. (2000). Consequences of misspecification of the within-area exposure distribution in the hierarchical model are discussed in Fortunato et al (2007).

## 4 Discussion

Cliff and Ord introduced the crucial idea of assessing and testing autocorrelation in the processes under study as a way of characterising their spatial patterns. Autocorrelation structure is now currently modelled through hierarchical approaches in disease mapping risk and in regression analysis. The amount of published papers dealing with theoretical and methodological aspects of such analyses (estimation bias, spatial model formulation, impact of choosing a specific spatial model ...) and on applications in epidemiology underlines the importance of their work. Resulting improvement on risk estimates, association estimates, which carries to the interpretation of such associations is clearly a consequence of the importance of taking into account the autocorrelation of variables in principled statistical analyses in spatial epidemiology.

## Acknowledgement

Sylvia Richardson gratefully acknowledges support from the ESRC National Centre for Research Methods (grant number RES-576-25-0015).

## References

- Auchincloss, A.H., Diez Roux, A.V., Brown, D.G., Raghunathan, T.E. and C. A. Erdmann. (2007) Filling the Gaps: Spatial Interpolation of Residential Survey Data in the Estimation of Neighborhood Characteristics, *Epidemiology*, **180**, 469–478.
- Banerjee, S., Carlin, B.P., and A.E. Gelfand (2004) Hierarchical modeling and analysis for spatial data, Chapman & Hall/CRC Press.
- Besag, J. and Newell, J. (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, **154**, 143–55.
- Besag, J., York J., and Mollié, A. (1991). A Bayesian image restoration, with applications in spatial statistics. *Annal of the Institute of Mathematical Statistics*, **43**, 1–59.
- Best, N., Richardson, S., and A. Thomson. (2005) A comparison of Bayesian spatial models for disease mapping *Statistical Methods in Medical Research*, **14**, p35–59.

- Best, N., K. Ickstadt and R. Wolpert (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions, *Journal of the American Statistical Association*, **95**, 1076–1088.
- Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. and E.M. Conlon. (1999) Bayesian models for spatially correlated disease and exposure data, *Bayesian Statistics 6*, M. Bernardo, J. O. Berger, A.P. Dawid and A. F.M. Smith eds, Oxford University Press, 131–156.
- Biggeri, A., Bonannini, M., Catelan, D., Divino, F., Dreassi, E. and C. Lagazio. (2005) Bayesian Ecological Regression with Latent Factors: Atmospheric Pollutants Emissions and Mortality for Lung Cancer, *Environmental and Ecological Statistics*, **12**, 397–409.
- Borroto, R. J. and R. Martinez-Piedra. (2000) Geographical patterns of cholera in Mexico, 1991-1996, *Int. J. Epidemiol.*, **29**, 764–772.
- Clayton, D. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk, In *Geographical and environment epidemiology: methods for small areas studies* (P. Elliott, J. Cuzick, D. English, and R. Stern ed.), Oxford University Press, 205–220.
- Clayton, D., Bernardinelli, L and Montomoli, C. (1993) Spatial correlation in ecological analysis, *International Journal of Epidemiology*, **22**, 1193–1202.
- Cliff, A.D., and J.K. Ord. (1969) The problem of spatial autocorrelation. In London Papers in Regional Science, A. J. Scott (ed.), Pion, London, 25–55.
- Cliff, A.D., and J.K. Ord. (1981) Spatial processes: models and applications, Taylor & Francis.
- Clifford, P., Richardson, S., and D. Hemon. (1989) Assessing the Significance of the Correlation between Two Spatial Processes, *Biometrics* , **45**, 123–134.
- Cook, D.G., and S. J. Pocock. (1983) Multiple Regression in Geographical Mortality Studies, with Allowance for Spatially Correlated Errors, *Biometrics*, **39**, 361–371.
- Diggle, P. J., Tawn, J.A., and R. A. Moyeed. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society (Series C): Applied Statistics*, **47**, 299–350.
- Doll, R (editor) (1984) *The geography of disease*. British Medical Bulletin. Churchill Livingstone.
- Dutilleul, P. (1993) Modifying the t Test for Assessing the Correlation Between Two Spatial Processes, *Biometrics*, **49**, 305–314.
- Dutilleul, P., Pelletier, B. and G. Alpargu (2008) Modified F tests for assessing the multiple correlation between one spatial process and several others, *Journal of Statistical Planning and Inference*, reader reaction, **138**, 1402–1415.
- Elliott, P., Wakefield, J.C., Best, N.G. and D.J. Briggs.(2000) Spatial epidemiology: methods and applications, Oxford University Press.
- Elliott, P. and D. Wartenberg. (2004). Spatial epidemiology: current approaches and future challenges, *Environ Health Perspect.* **112**, 998–1006.

- Fortunato, L., Guihenneuc-Jouyaux, C., Tirmarche, M., Laurier D. and D. Hémon. (2007) Misspecification of within-area exposure distribution in ecological Poisson models, *Environmental and Ecological Statistics*, Published online, DOI 10.1007/s10651-007-0053-9.
- Fotheringham, S. and P. Rogerson. (2009) *The SAGE Handbook of Spatial Analysis*, SAGE Publications Ltd.
- Griffith, D. A., Wong, D.W.S., and T. Whitfield. (2003) Exploring relationships between the global and regional measures of spatial autocorrelation, *Journal of Regional Science*, **43**, pp. 683-710.
- Green P.J., Richardson S. (2002) Hidden markov models and disease mapping. *Journal of the American Statistical Association*, **97**, 1055-1070.
- Havard, S., Deguen, S., Zmirou-Navier, D., Schillinger, C., and D. Bard. (2009) Traffic-Related Air Pollution and Socioeconomic Status. A Spatial Autocorrelation Study to Assess Environmental Equity on a Small-Area Scale, *Epidemiology*, **20**, 223–230.
- Jarup, L., Best, N., Toledano, M.B., Wakefield, J., and P. Elliott. (2001) Geographical epidemiology of prostate cancer in Great Britain, *International Journal of Cancer*, **97**, 695 – 699.
- Knorr-Held L, Rasser G. (2000) Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**, 13-21.
- Kulldorff, M., Song, C., Gregorio, D., Samociuk, H. and L. DeChello. (2006) Cancer Map Patterns Are They Random or Not?, *Am J Prev Med*, **30**, S37-S49.
- Lawson, A. B. (2006) *Statistical methods in spatial epidemiology*, 2nd ed., In *Wiley Series in Probability and Statistics*, Chichester: Wiley & Sons.
- Lawson, A. B. (2008) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, CRC Press.
- Lazar, P. (1982). Problems of concurrent trends in etiological research. In: *Trends in cancer incidence*, ed. K Magnus, Hemisphere Publishing Corporation, Washington.
- Maheswaran, R., Haining, R.P., Brindley, P. and N. G. Best. (2006) Outdoor NO<sub>x</sub> and stroke mortality: adjusting for small area level smoking prevalence using a Bayesian approach, *Statistical Methods in Medical Research*, **15**, 499-516.
- Marshall, R. J. (1991) A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **154**, 421–441.
- Pfeiffer, D., Stevenson, M., Stevens, K.B., Robinson, T.P., Rogers, D.J., and A. C. A. Clements. (2008) *Spatial Analysis in Epidemiology*, Oxford University Press.
- Pocock, S. J., Cook, D. G., and Shaper, A.G. (1982) Analysing geographic variation in cardiovascular mortality: methods and results, (with discussion) *Journal of the Royal Statistical Society. Series A (General)*, **145**, 313–341.

- Richardson S. (1992) Statistical methods for geographical correlation studies In : *Geographical and Environmental Epidemiology : Methods for Small Area Studies*, Eds P. Elliott, J. Cuzick, D. English, R. Stern. Oxford University Press.
- Richardson, S., Guihenneuc, C. and V. Lasserre. (1992) Spatial linear models with autocorrelated error structure, *The Statistician*, **41**, 539–557.
- Richardson S. and Monfort, C. (2000) Ecological correlation studies. In: *Spatial epidemiology: methods and applications*, Oxford University Press.
- Ripley, B.D. (1981) *Spatial Statistics*, New York, Wiley.
- Tango. T. (1995) A class of tests for detecting ‘general’ and ‘focussed’ clustering of rare diseases. *Statistics in Medicine*, **14**, 2323–5.
- Truong T., Rougier Y., Dubourdieu D., Guihenneuc-Jouyaux C., Orsi L., Hémon D, and P. Guénel. (2007) Time trends and geographic variations for thyroid cancer in New Caledonia, a very high incidence area (1985-1999), *Eur J Cancer Prev.*, **16**, 62–70.
- Van Leeuwen, J.A., Waltner-Toews, D., Abernathy, T., Smit, B. and M. Shoukri. (1999) Associations between stomach cancer incidence and drinking water contamination with atrazine and nitrate in Ontario (Canada) agroecosystems, 1987-1991, *Int. J. Epidemiol.*, **28**, 836–840.
- Wakefield, J., Kelsall J.E. and Morris, S.E. (2000a) Clustering, cluster detection and spatial variation in risk. In: *Spatial epidemiology: methods and applications*, Oxford University Press.
- Wakefield, J., Best, N.G. and Waller, L. (2000b) Bayesian approaches to disease mapping. In: *Spatial epidemiology: methods and applications*, Oxford University Press.
- Wakefield, J., and S. Morris (1999) Spatial dependence and errors-in-variables in environmental epidemiology, *Bayesian Statistics 6*, M. Bernardo, J. O. Berger, A.P. Dawid and A. F.M. Smith eds, Oxford University Press, 657–684.
- Waldhor, T. (1996) the spatial autocorrelation coefficient moran’s  $i$  under heteroscedasticity, *Statistics in Medicine*, **15**, 887–89.
- Walter, S.D. (1993) Assessing spatial patterns in disease rates. *Statistics in Medicine*, **12**, 1885–94.