Cahiers de Géostatistique, Fascicule 3
Compte–rendu des Journées de Géostatistique,
25–26 Mai 1993, Fontainebleau, pp. 159–165

BINOMIAL COKRIGING THE RISK OF A RARE DISEASE

M. A. OLIVER[1], C. LAJAUNIE[2], R. WEBSTER[3], K. R. MUIR[4] & J. R. MANN[5]

[1] *School of Geography, The University, Birmingham B15 2TT, England*
[2] *Centre de Géostatistique, 35 rue St-Honoré, 77305 Fontainebleau, France*
[3] *ETH Zürich, ITÖ, Grabenstrasse 3, 8952 Schlieren, Switzerland*
[4] *Queen's Medical Centre, The University, Nottingham NG7 2UH, England*
[5] *The Children's Hospital, Ladywood Middleway, Birmingham B16 8ET, England*

RESUME On résume deux modifications du krigeage ordinaire pour estimer le risque d'une maladie rare à partir de fréquences observées et on compare leurs performances. On suppose que les fréquences ont une répartition binomial le qui dépend du risque. L'adaptation la plus simple est assez stable vis-à-vis la taille du voisinage, mais elle lisse fortement. Si l'on ajoute la contrainte que l'espérance et la valeur vraie sont égales (cokrigeage conditionel sans biais) on augmente la sensibilité de l'estimation. Cependant on a besoin de plus de points (au moins 100 points) pour obtenir des estimations stables avec un intervalle de confiance raisonnable.

ABSTRACT Two novel modifications of ordinary kriging for estimating the underlying risk of a rare disease from observed frequencies are summarized and their performances compared. The frequencies are assumed to be binomial and dependent on the risk.

INTRODUCTION

In studying the geographical epidemiology of rare diseases we want to estimate the risk of people's developing them and to know whether this varies spatially. In general our only knowledge of the risk derives from observed frequencies of the disease within small neighbourhoods. These are poor local estimates of the risk because they embody errors arising from the the small risk and the limited population. Lajaunie (1991) and McNeill (1991) showed how data of this kind might be analysed geostatistically, and Oliver *et al.* (1992, 1993) applied the method to estimate the underlying risk of childhood cancer by binomial cokriging. Lajaunie (1991) also pointed out that for a constant overall risk the expected local risk at a place equals the actual risk there, and this information can be used to condition the cokriging. It results in conditionally unbiased estimates.

In this paper we summarize the methods and then compare our estimates from them of the risk of cancer among children in the West Midlands Health Authority Region (WMHAR) of England.

THE DATA

The Region covers about 25 000 km^2, and it includes rural, urban, industrial and suburban environments. During 1980 to 1984 inclusive there were 595 cases of cancer

among some 1.13 M children under 15 years of age living there. The coordinates of the home of each diagnosed child are known. The cases are distributed among 344 of the total 838 electoral wards (the smallest area for which population is recorded). So, the 595 cases, their spatial coordinates, and the childhood populations of 838 wards at risk and their locations constitute the data.

The overall estimate of the risk is the number of cases divided by the total population, and is 0.000528. The first step in estimating the local risk is to calculate the local frequency, $F(\mathbf{x}_i)$, i.e. the ratio of the number of cases within each ward, $L(\mathbf{x}_i)$, to the number of children living there, $n(\mathbf{x}_i)$: $F(\mathbf{x}_i) = L(\mathbf{x}_i)/n(\mathbf{x}_i)$, where the $\mathbf{x}_i, i = 1, 2, \ldots$, denote the centroids of the wards. As above, these are crude estimates. Their average is 0.000586 and the variance 1.3225×10^{-6}.

THEORY AND ANALYSIS

To proceed further we assume that there is an underlying risk, $R(\mathbf{x})$, of a child's developing cancer and to which all children are exposed. We also assume that the different cases occur independently, so that R is the only source of correlation among them. This is reasonable for a non-contagious disease. Hence the observed frequencies, $F(\mathbf{x}_i)$, are conditionally independent for a fixed risk:

$$F(\mathbf{x}_i) = \frac{1}{n(\mathbf{x}_i)} \mathrm{Bi}[R(\mathbf{x}_i), n(\mathbf{x}_i)] \; . \tag{1}$$

where $i = 1, 2, \ldots, n$ wards containing $n(\mathbf{x}_i)$ children

Estimating the risk variogram

First we compute an experimental variogram of the frequencies, γ_F, with the usual formula. It appears in Fig. 1a. This variogram embodies error arising from the binomial character of the frequencies and the uneven distribution of the children. Nevertheless, it is our starting point from which we develop the variogram of the risk.

The conditional independence of the frequencies and the standard expressions for their expectations and variances for given risk lead to the following expectation of semivariance:

$$\mathrm{E}[\{F(\mathbf{x}_i) - F(\mathbf{x}_i + \mathbf{h})\}^2] \; = \; 2\gamma_R(\mathbf{h}) + \{\mu(1-\mu) - \sigma_R^2\}\left\{\frac{n(\mathbf{x}_i) + n(\mathbf{x}_i + \mathbf{h})}{n(\mathbf{x}_i).n(\mathbf{x}_i + \mathbf{h})}\right\} , \tag{2}$$

where $\gamma_R(\mathbf{h}) = \frac{1}{2}\mathrm{E}[\{R(\mathbf{x}_i) - R(\mathbf{x}_i + \mathbf{h})\}^2]$, σ_R^2 is the variance of the risk, and μ is the mean estimated without bias by $\bar{F}$, the average of the data, and $n(\mathbf{x}_i)$ and $n(\mathbf{x}_i + \mathbf{h})$ are the numbers of children in the wards centred at $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{h}$, respectively,

This equation does not define a semivariance in the strict sense because the quantity depends on $\mathbf{x}$. However, the average over the whole Region is correct, and from it we obtain the relation

$$\hat{\gamma}_R(\mathbf{h}) \quad = \quad \hat{\gamma}_F(\mathbf{h}) - \frac{1}{2}\{\bar{F}(1-\bar{F}) - \hat{\sigma}_R^2\}\overline{\left\{\frac{n(\mathbf{x}_i) + n(\mathbf{x}_i + \mathbf{h})}{n(\mathbf{x}_i).n(\mathbf{x}_i + \mathbf{h})}\right\}} , \tag{3}$$

where the quantity beneath the bar is the average over all pairs of wards involved in calculating $\hat{\gamma}_F(\mathbf{h})$. The variance σ_R^2 is unknown, but it can be estimated iteratively as the sill of a fitted model (Oliver *et al.*, 1993).

Figure 1b shows the experimental variogram of the risk. The solid line is the fitted model, Whittle's (1954) *elementary correlation*. The model is bounded, and its effective range is about 50 km. This suggests that the risk is patchy in its distribution.

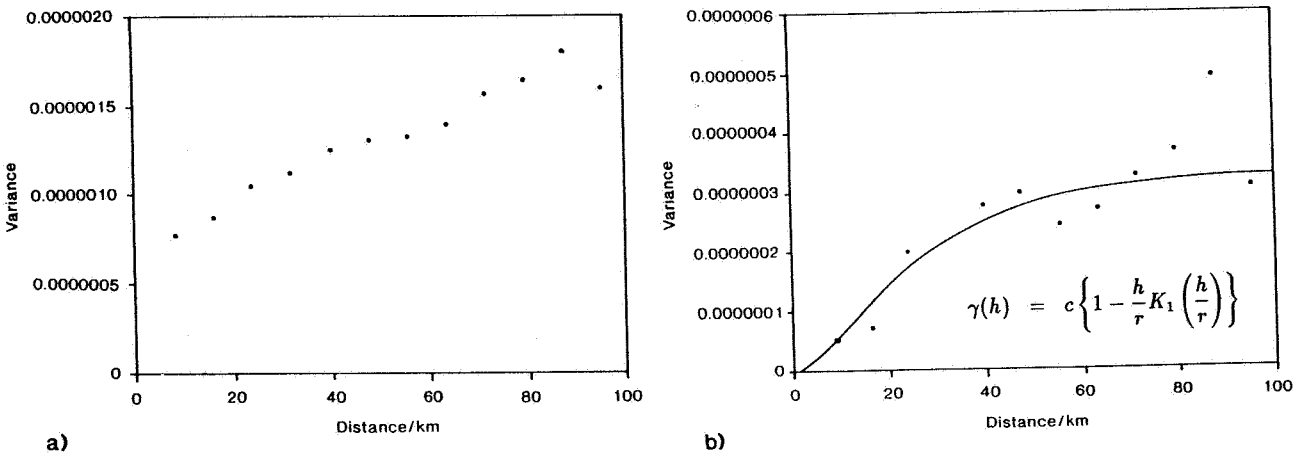

Figure 1) Experimental variograms of childhood cancer in the West Midlands: a) variogram of frequency, γ_F, b) variogram of risk, γ_R, the solid line is the fitted model.

Binomial cokriging

To estimate the risk at an unknown place we have two options. The first is an extension of ordinary cokriging. Assuming the mean to be unknown, the risk at a place for which we have no record, say $\mathbf{x}_0$ is

$$\hat{R}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i F(\mathbf{x}_i) , \tag{4}$$

where N is the number of data and λ_i are weights. The estimate is unbiased if $\sum_{i=1}^{N} \lambda_i = 1$, for $\mathrm{E}[R(\mathbf{x}_i)] = \mathrm{E}[F(\mathbf{x}_i)] = \mu$. Under this constraint it involves solving the kriging system:

$$\begin{aligned} \sum_{i=1}^{N} \lambda_i C^F(\mathbf{x}_i, \mathbf{x}_j) + \psi &= C^{FR}(\mathbf{x}_0, \mathbf{x}_j) \quad \forall\, j , \\ \sum_{i=1}^{N} \lambda_i &= 1 , \end{aligned} \tag{5}$$

where ψ is a Lagrange multiplier, the $C^F(\mathbf{x}_i, \mathbf{x}_j)$ are the covariances of the frequencies, and the $C^{FR}(\mathbf{x}_0, \mathbf{x}_j)$ are the covariances between the frequency and the risk. The covariances of the risk, C^R, are obtained from the variogram, and from these the other covariances are derived by

$$C^{RF}(\mathbf{x}_0, \mathbf{x}_i) = C^R(\mathbf{x}_0, \mathbf{x}_i) \quad \text{and} \quad C^F(\mathbf{x}_i, \mathbf{x}_j) = C^R(\mathbf{x}_i, \mathbf{x}_j)$$

except when $i = j$, for which

$$C^F(\mathbf{x}_i, \mathbf{x}_i) = \left\{1 - \frac{1}{n(\mathbf{x}_i)}\right\} C^R(\mathbf{x}_i, \mathbf{x}_i) + \frac{1}{n(\mathbf{x}_i)} \mu(1 - \mu) .$$

The second option is to krige without bias conditional on R. The condition is

$$\mathrm{E}[\hat{R}(\mathbf{x}_0) \mid R(\mathbf{x}_0)] = R(\mathbf{x}_0) . \tag{6}$$

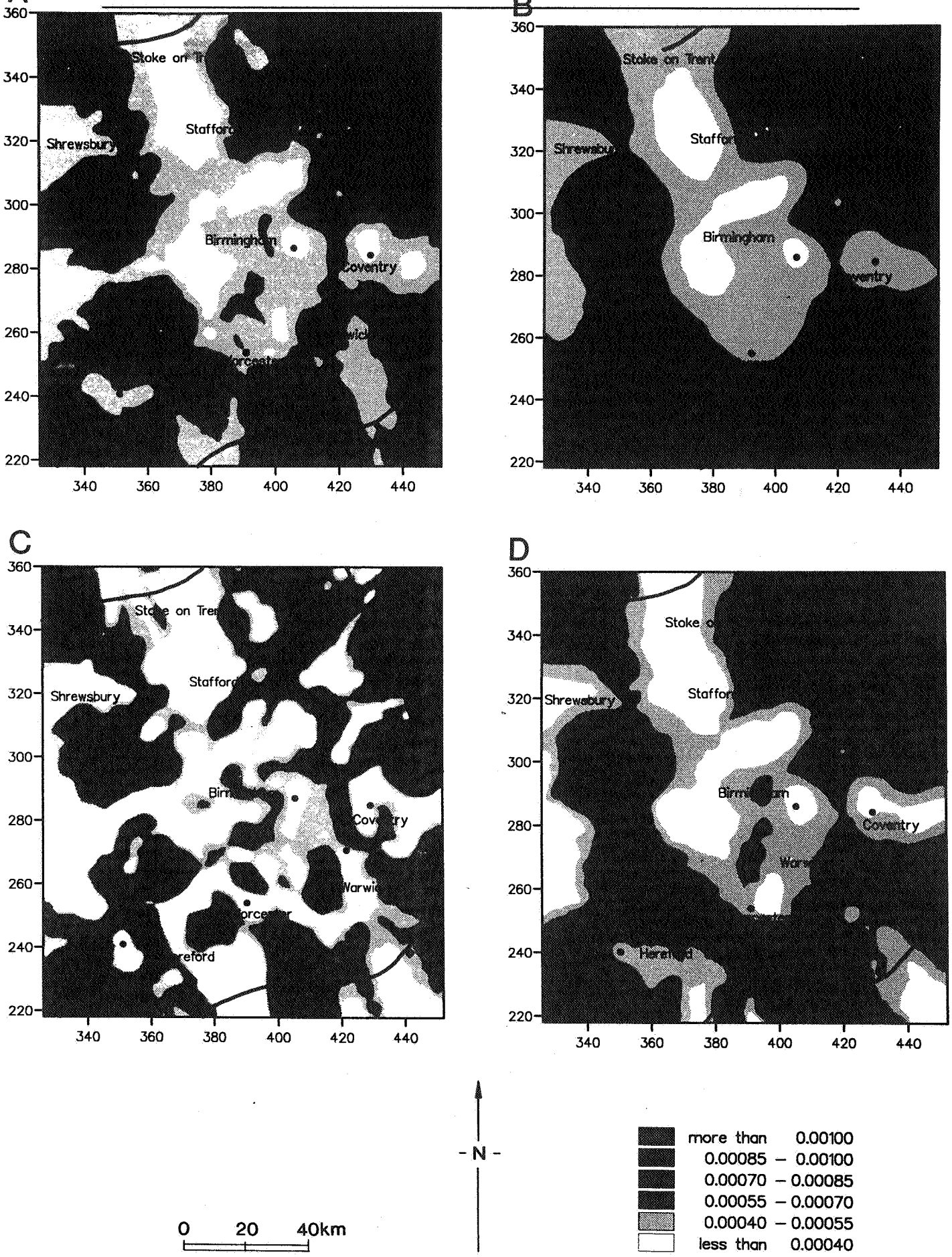

Figure 2. Maps of the estimated risk of childhood cancer: a) and b) are for ordinary binomial cokriging with 20 and 100 points in the neighbourhood, respectively; c) and d) are for conditionally unbiased binomial cokriging with 20 and 100 points, respectively.

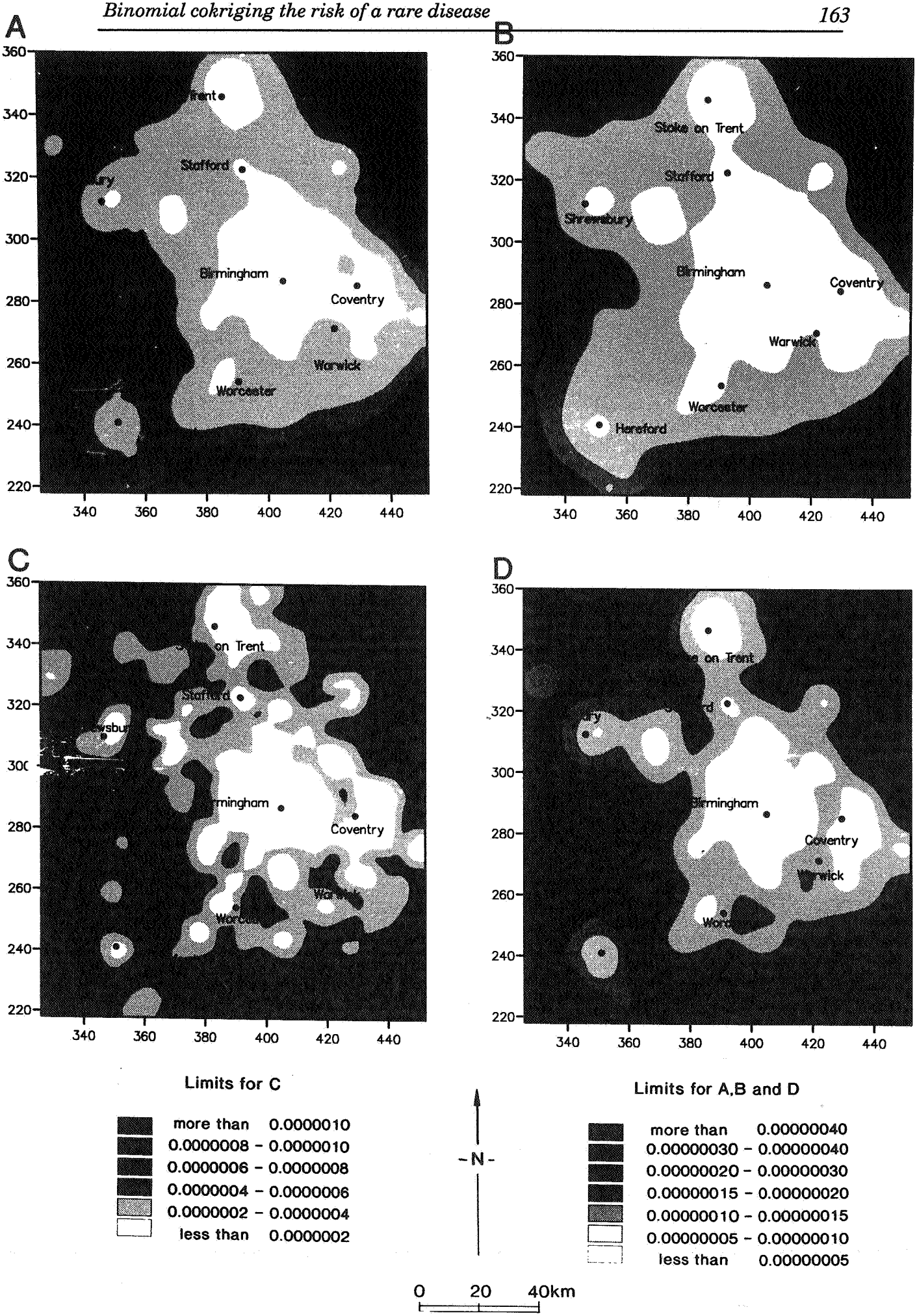

Figure 3. Maps of the estimation errors: a) and b) are from ordinary binomial cokriging with 20 and 100 points in the neighbourhood, respectively; c) and d) are from conditionally unbiased cokriging with 20 and 100 points, respectively.

Lajaunie (1991) has shown that

$$\mathrm{E}[R(\mathbf{x}_i) \mid R(\mathbf{x}_0)] = \mu + \frac{C^R(\mathbf{x}_0, \mathbf{x}_i)}{\sigma_R^2}\{R(\mathbf{x}_0) - \mu\} \ . \quad (7)$$

So in addition to the weights' summing to 1 to assure unbiasedness we have the extra condition:

$$\sum_{i=1}^{N} \lambda_i C^R(\mathbf{x}_0, \mathbf{x}_i) = \sigma_R^2 \ . \quad (8)$$

The full kriging system is therefore

$$\begin{aligned} \sum_{i=1}^{N} \lambda_i C^F(\mathbf{x}_i, \mathbf{x}_j) + \psi_1 C^R(\mathbf{x}_0, \mathbf{x}_i) + \psi_2 &= C^{FR}(\mathbf{x}_0, \mathbf{x}_j) \qquad \forall\, j \ , \\ \sum_{i=1}^{N} \lambda_i &= 1 \ , \\ \sum_{i=1}^{N} \lambda_i C^R(\mathbf{x}_0, \mathbf{x}_i) &= \sigma_R^2 \ . \end{aligned} \quad (9)$$

RESULTS

We solved the above kriging systems, and we first compared the performances of the two methods for different numbers of points in the neighbourhood and at locations in different environments. We estimated the risk at three points. We chose one point in each of rural, urban, and suburban areas with sparse, dense and intermediate populations and data, respectively. Table 1 summarizes the results.

Ordinary cokriging produces fairly stable estimates in all three environments for numbers of points ranging from 20 to 139. The variances decrease somewhat with increasing numbers of points. Conditionally unbiased cokriging, however, is less stable. Estimates based on only the 20 nearest points have much larger variances, and it seems as though about 100 points are needed for stability. The variances are greatest for the rural areas, where the data are most sparse, and least for the urban areas where they are most dense.

We estimated the risk over the whole Region at 2 km intervals by the two procedures. We set the kriging neighbourhood to 20 and 100 wards, respectively. The maps of the risk for ordinary binomial cokriging, Fig. 2a and b, are fairly similar. The one using 20 points is somewhat more 'noisy', however. The maps made using conditionally unbiased cokriging with 20 and 100 points, Fig. 2c and d contrast more: that using 20 points is very 'noisy'. The latter is more different from the equivalent one for ordinary kriging, Fig. 2a, than the maps using 100 points are, Fig. 2b and d. The estimation errors for ordinary binomial cokriging, Fig. 3a and b, are generally smaller than those for the conditionally unbiased estimation, Fig. 3c and d. The errors from either method are locally more variable using 20 points (Fig. 3a and c) than with 100 points (Fig. 3b and d). The estimation variances for conditionally unbiased cokriging are three times larger with 20 points than with 100, Fig. 3c and d. These results largely confirm the summary in Table 1.

CONCLUSIONS

It seems that ordinary binomial cokriging is somewhat more sensitive to the number of points used for estimation than ordinary kriging, presumably because of the implicit estimation of the mean, μ, within a given neighbourhood. Conditional unbiased binomial cokriging, however, is much more sensitive to the size of neighbourhood, as the estimation

Table 1. Estimates and kriging variances for urban, suburban and rural points.

	Estimates ($\times 10^{-3}$)					
	ordinary binomial			conditional non-biased		
Points	urban	suburban	rural	urban	suburban	rural
20	0.548	0.357	0.662	0.444	0.381	0.520
40	0.516	0.335	0.705	0.553	0.232	0.482
70	0.523	0.351	0.669	0.624	0.279	0.720
100	0.507	0.317	0.653	0.510	0.156	0.693
139	0.506	0.336	0.658	0.502	0.325	0.810
	Variances ($\times 10^{-6}$)					
	ordinary binomial			conditional non-biased		
Points	urban	suburban	rural	urban	suburban	rural
20	0.0145	0.0700	0.1842	0.0431	0.5541	0.8382
40	0.0126	0.0566	0.1382	0.0284	0.2477	0.4584
70	0.0124	0.0537	0.1325	0.0192	0.1171	0.3374
100	0.0123	0.0530	0.1319	0.0155	0.0910	0.2957
139	0.0122	0.0525	0.1317	0.0140	0.0795	0.2709

variances show. The difference between the estimation errors for the two methods decreases as the number of points increases, however, and the overall interpretation placed on the two maps made with 100 neighbouring points would be the same. Where the implications are significant it should be worth pursuing the more sensitive method.

REFERENCES

LAJAUNIE, C. 1991. *Local Risk Estimation for a Rare Noncontagious Disease based on Observed Frequencies.* Note N-36/91/G. Centre de Géostatistique, Ecole des Mines de Paris: Fontainebleau.

MCNEILL, L. 1991. Interpolation and smoothing of binomial data for the Southern Africa Bird Atlas project. *South African Statistical Journal* **25**, 129-146.

OLIVER, M. A., LAJAUNIE, C., WEBSTER, R., MUIR, K. R. & J. R. MANN. 1993. Estimating the risk of childhood cancer. In *Geostatistics Tróia '92*, Vol. 2, ed. A. Soares, Kluwer Academic Publishers: Dordrecht, pp 899-910.

OLIVER, M. A., MUIR, K. R., WEBSTER, R., PARKES, S. E., CAMERON, A. H., STEVENS, M. C. G. & MANN, J. R. 1992. A geostatistical approach to the analysis of pattern in rare disease. *Journal of Public Health Medicine,* **14**, 280-289.

WHITTLE, P. 1954. On stationary processes in the plane. *Biometrika* **41**, 434-449.