BIOMETRICS 45, 123–134
March 1989

Assessing the Significance of the Correlation Between Two Spatial Processes

Peter Clifford,[1] Sylvia Richardson,[2] and Denis Hémon[3]

[1] Mathematical Institute, University of Oxford, 24–29 St Giles, Oxford OX1 3LB, England

[2] INSERM U170, 16 Avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France, and Laboratoire de Statistiques Médicales, Université de Paris V, 45, rue des Saints Pères, 75006 Paris, France

[3] INSERM U170, 16 Avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France

SUMMARY

Modified tests of association based on the correlation coefficient or the covariance between two spatially autocorrelated processes are presented. These tests can be used both for lattice and nonlattice data. They are based on the evaluation of an effective sample size that takes into account the spatial structure.

For positively autocorrelated processes, the effective sample size is reduced. A method for evaluating this reduction via an approximation of the variance of the correlation coefficient is developed. The performance of the tests is assessed by Monte Carlo simulations. The method is illustrated by examples from geographical epidemiology.

1. Introduction

The calculation of correlation coefficients for variables that are observed at a variety of different spatial locations has suggested intriguing hypotheses about the relationship between environmental factors and disease pathologies (Armstrong and Doll, 1975; Hoover and Fraumeni, 1975). It is well known that the null distribution of r, the product moment correlation coefficient, is influenced by spatial and temporal autocorrelation (Student, 1914; Bartlett, 1935; Richardson and Hémon, 1981).

In a preliminary report, Clifford and Richardson (1985) have suggested a method of approximating the critical values of r. Their method depends on an estimate of the variance of r. In this paper we investigate a general method of obtaining such an estimate and show that the test which results is related to a test based on the standardised covariance between the processes. Theoretical properties of the distribution of r are reviewed in Section 2. In Section 3 we report on an empirical simulation study of the performance of these procedures and in Section 4 we apply our methods to test the association between cigarette consumption, industrial risk factors, and deaths from lung cancer for French départements.

Some authors have devised measures of association that involve transformation of the data before standard techniques are applied. Student (1914) advocated a form of what would now be called prewhitening, and more recent work, such as that of Haugh (1976) with time series and Davies and Jowett (1958) can be seen as developments of this local approach, in that filters are applied to the data in order to reduce temporal or spatial

---

Requests for reprints should be addressed to the second author.

Key words: Correlation coefficient; Geographical epidemiology; Monte Carlo simulations; Significance tests; Spatial processes.

autocorrelation. However, the application of this approach to irregularly spaced data presents a number of problems that have not been thoroughly explored and information is lost by filtering techniques.

The problem of testing the association between autocorrelated variables can also be tackled by regression techniques (Ord, 1975) when a relationship between a dependent variable and a set of independent variables is postulated. Cook and Pocock (1983), in their study of the association between water hardness and heart disease, considered a regression model with stationary autocorrelated errors, the autocorrelation decreasing exponentially with distance. Mardia and Marshall (1984) have demonstrated the asymptotic properties of this approach. In general, these techniques involve a substantial amount of computing time and depend on an explicit parametric model for the autocovariance.

A nonparametric index of association has been proposed by Tjøstheim (1978) and generalised by Hubert and Golledge (1982). However, they assess the significance of the index with reference to a randomisation distribution that involves the spatial redistribution of one of the variables, the other staying fixed. This full set of permutations clearly does not preserve the autocorrelation of the permuted variable and hence the significance of the index is not correctly assessed.

2. Modified Tests of Association

2.1 *Basic Properties of r*

We are interested in data sets which consist of a set A of N locations numbered from 1 to N and a set of pairs of observations (X_α, Y_α), $\alpha \in A$, where each pair is indexed by its location. The correlation coefficient is then given by

$$r = \frac{s_{XY}}{s_X s_Y}, \tag{2.1}$$

where $s_{XY} = N^{-1} \sum_A (X_\alpha - \bar{X})(Y_\alpha - \bar{Y})$ is the sample covariance, $s_X^2 = N^{-1} \sum_A (X_\alpha - \bar{X})^2$, $s_Y^2 = N^{-1} \sum_A (Y_\alpha - \bar{Y})^2$, and where $\bar{X} = N^{-1} \sum_A X_\alpha$ and $\bar{Y} = N^{-1} \sum_A Y_\alpha$. If, conditional on X, the elements of Y are normal i.i.d. variables or conditional on Y, the elements of X are normal i.i.d., then r has the standard null distribution with p.d.f.

$$f_N(r) = \frac{(1 - r^2)^{(N-4)/2}}{B[\frac{1}{2}, \frac{1}{2}(N - 2)]}, \quad r \leqslant 1, \tag{2.2}$$

where B is the beta function.

Critical values of r are usually obtained from t-tables since $(N - 2)^{1/2} r/(1 - r^2)^{1/2}$ has a t-distribution with $N - 2$ degrees of freedom under these assumptions. This is also the t-statistic that is calculated in testing the significance of the linear regression either of Y on X or of X on Y.

2.2 *The Standardised Covariance*

For independent samples the correlation coefficient can be thought of as a standardised covariance. If the elements of Y are i.i.d., then conditional on $X = x$ the sample covariance s_{XY} has mean zero and variance $N^{-2} \sum_A (x_\alpha - \bar{x})^2 \sigma_Y^2$ where σ_Y^2 is the variance of the elements of Y. To standardise s_{XY}, the unknown quantity σ_Y^2 is replaced by an unbiased estimate $\sum_A (Y_\alpha - \bar{Y})^2/(N - 1)$ and s_{XY} is divided by the resulting estimate of its standard deviation. This leads to the expression $(N - 1)^{1/2} r$, whose significance would be approximately assessed with reference to tables of the normal distribution, relying on a central limit theorem to justify the approximation. Thus, a test based on the standardised covariance is equivalent to a test based on r.

In the general case, we suppose now that **X** and **Y** are independent but that both **X** and **Y** are multivariate normal vectors with constant means and variance–covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$, respectively. The conditional variance s_{XY} is given by

$$N^{-2} \sum_{\alpha,\beta} (X_\alpha - \bar{X})(X_\beta - \bar{X})\text{cov}(Y_\alpha, Y_\beta), \tag{2.3}$$

which is equal to $N^{-2} \sum_{\alpha,\beta} (X_\alpha - \bar{X})(X_\beta - \bar{X})\ \text{cov}(Y_\alpha - \bar{Y}, Y_\beta - \bar{Y})$ since $\sum_A (X_\alpha - \bar{X}) = 0$. Replacing $\text{cov}(Y_\alpha - \bar{Y}, Y_\beta - \bar{Y})$ by the unbiased estimate $(Y_\alpha - \bar{Y})(Y_\beta - \bar{Y})$ gives the expression s^2_{XY}, as a trivial estimate of the conditional variance of s_{XY}. It is clear that no progress can be made until some plausible restrictive structure is imposed on $\mathbf{\Sigma}_Y$.

2.3 *Structure for* $\mathbf{\Sigma}_X$ *and* $\mathbf{\Sigma}_Y$

We will assume that the set of locations A is a subset of some larger set Ω. For stationary processes Ω can, in principle, be arbitrarily large. For real spatial data, Ω is finite, perhaps equal to A itself. We assume that the set of all ordered pairs of elements of Ω can be divided into strata $S_0, S_1, S_2, \ldots$, so that the covariances within strata are constant, i.e., $\text{cov}(X_\alpha, X_\beta) = C_X(k)$ and $\text{cov}(Y_\alpha, Y_\beta) = C_Y(k)$, $(\alpha, \beta) \in S_k$, $k = 0, 1, \ldots$. Of course, if $(\alpha, \beta) \in S_k$, then $(\beta, \alpha) \in S_k$ for consistency. For stationary processes the stratification is indexed by directional lags. For isotropic processes the number of strata is reduced in the lattice case and when the data are irregularly spaced the strata can be indexed by a discrete distance function. The general formulation is flexible enough to permit spatially inhomogeneous variances and other aspects of nonstationarity. With this structure (2.3) becomes

$$N^{-2} \sum_k N_k \left[\frac{\sum_{A_k} (X_\alpha - \bar{X})(X_\beta - \bar{X})}{N_k} \right] C_Y(k), \tag{2.4}$$

where N_k is the cardinality of A_k, $A_k = (A \times A) \cap S_k$, and the summation over A_k is for all ordered pairs $(\alpha, \beta) \in A_k$.

For stationary processes, Clifford and Richardson (1985) have suggested using

$$\hat{C}_Y(k) = \sum_{A_k} (Y_\alpha - \bar{Y})(Y_\beta - \bar{Y})/N_k \tag{2.5}$$

as an estimate of $C_Y(k)$ for values of k corresponding to small spatial lags and shrinking $\hat{C}_Y(k)$ to zero otherwise. This was partially for computational convenience. Here we propose to consider the inclusion of all lags. The resulting estimate of the conditional variance of s_{XY} is therefore

$$N^{-2} \sum_k N_k \hat{C}_X(k) \hat{C}_Y(k). \tag{2.6}$$

It does not rely on an arbitrary notion of what constitutes a small lag, it is invariant to shifts in the mean, and it is unbiased for periodic processes. It has the additional advantage of being symmetric in X and Y, so that it is also the estimate of the conditional variance of s_{XY} given Y. Note that when $S_0 = A \times A$, (2.6) reduces to $N^{-1} s^2_X s^2_Y$.

2.4 *The Standardised Covariance and the Modified t-Test*

Using the estimate (2.6), the standardised covariance, W, becomes

$$W = N s_{XY} \left[\sum_k N_k \hat{C}_X(k) \hat{C}_Y(k) \right]^{-1/2}. \tag{2.7}$$

If we consider the correlation coefficient, in Appendix 1 it is shown that to the first order,

$$\sigma_r^2 \approx \frac{\text{var}(s_{XY})}{\text{E}(s_X^2)\text{E}(s_Y^2)}. \tag{2.8}$$

We therefore take as our estimate of σ_r^2

$$\hat{\sigma}_r^2 = \frac{\sum N_k \hat{C}_{\mathbf{X}}(k)\hat{C}_{\mathbf{Y}}(k)}{N^2 s_X^2 s_Y^2}. \tag{2.9}$$

We consider a modified t-test for the correlation coefficient by approximating the critical values of r by percentage points of the p.d.f. $f_{\hat{M}}(r)$, where $\hat{M} = 1 + \hat{\sigma}_r^{-2}$ and f is defined by (2.2). The quantity $(\hat{M})$ M can be thought of as an (estimated) "effective sample size" that takes into account the spatial autocorrelation in the variables **X** and **Y**. Note that when $S_0 = A \times A$, $\hat{M} = N + 1$. For positively autocorrelated processes, the estimated effective sample size is typically less than N. If one of the processes has negative autocorrelation it is, in principle, possible that the effective sample size will be larger than N.

Comparing (2.7) and (2.9), we see that

$$W = (\hat{M} - 1)^{1/2} r, \tag{2.10}$$

where $\hat{M} = 1 + \hat{\sigma}_r^{-2}$.

In Section 3 we consider the performance of the test obtained by assuming that W has a N(0, 1) distribution under the null hypothesis and compare this test with that obtained by assuming that r has p.d.f. $f_{\hat{M}}(r)$, that is, using a t-statistic with $\hat{M} - 2$ in place of $N - 2$.

3. Monte Carlo Simulations

The performance of the modified t-test or the test based on the standardised covariance has been assessed by Monte Carlo simulation in both a lattice and a nonlattice case.

3.1 *The Lattice Case*

Method We simulated stationary first-order isotropic simultaneous autoregressive processes defined by

$$X_{s,t} = a(X_{s-1,t} + X_{s+1,t} + X_{s,t-1} + X_{s,t+1}) + \varepsilon_{s,t}, \tag{3.1}$$

where $\{\varepsilon_{s,t}\}$ is a sequence of i.i.d. N(0, 1), $0 < |a| < \frac{1}{4}$. This class of processes has been widely discussed in the modelling of spatial patterns (Whittle, 1954; Besag, 1974; Cliff and Ord, 1975) and is relatively easy to simulate (see Appendix 2).

The simulation was performed on a DPS7 C.I.I. Honeywell Bull. A spectral decomposition similar to that given by Besag in the discussion of Bartlett's (1978) paper for autonormal processes was used to generate processes on a 26 × 26 lattice with zero on the boundary and then restricted to middle 12 × 12, 16 × 16, and 20 × 20 squares. This restriction was sufficient to render negligible the influence of the boundary condition.

For each simulation, a 26 × 26 field of i.i.d. N(0, 1) variables was first generated using a polar algorithm. Values of a equal to 0, .0945, .165, .2099, .2364 were chosen to give values of the nearest neighbour autocorrelation, $\rho_{\mathbf{X}}(1)$, equal to 0, .2, .4, .6, and .8, respectively.

For each value of a, 500 pairs (X_α, Y_α), $\alpha \in A$, were generated. Taking advantage of the lattice structure, each process **X** can also be rotated by 90° or reflected with respect to the diagonals, leading to eight copies (four rotations and/or two reflections) of the original process. These copies have the same set of estimated autocovariances, $\{\hat{C}_{\mathbf{X}}(k)\}$, which therefore is calculated only once. Thus, 4,000 trials, dependent in groups of eight, were obtained for each value of $\rho_{\mathbf{X}}(1)$ and $\rho_{\mathbf{Y}}(1)$.

The proposed statistics are denoted by W and $t_{\hat{M}-2}$, where $\hat{M}$ is the estimated effective sample size, $\hat{M} = 1 + \hat{\sigma}_r^{-2}$. If expression (2.6) for the estimated variance of s_{XY} gave an inadmissible negative estimate, it was replaced by the product $N^{-1}s_X^2 s_Y^2$, which is the estimate in the case of no autocorrelation. The standard t-statistic for r based on N observations is denoted by t_{N-2}. For the three statistics, W, $t_{\hat{M}-2}$, and t_{N-2}, a 5% nominal rejection level was chosen. For testing $t_{\hat{M}-2}$, the integer part of $\hat{M}$ was taken. The empirical variance of the rejection levels was estimated by first averaging the rejection indicator function over the eight dependent pairs (X_α, Y_α) (where X_α is obtained by rotations/reflections) and then calculating the empirical variance of this average over the 500 independent simulations.

Results Figure 1 illustrates the poor performance of the standard t-test in the presence of positive autocorrelation. The observed Type I error rates are represented for the three lattice sizes together with their 95% confidence interval. Note that the observed Type I error rates are significantly larger than 5%, even for low values of $\rho_X(1)$ and $\rho_Y(1)$.

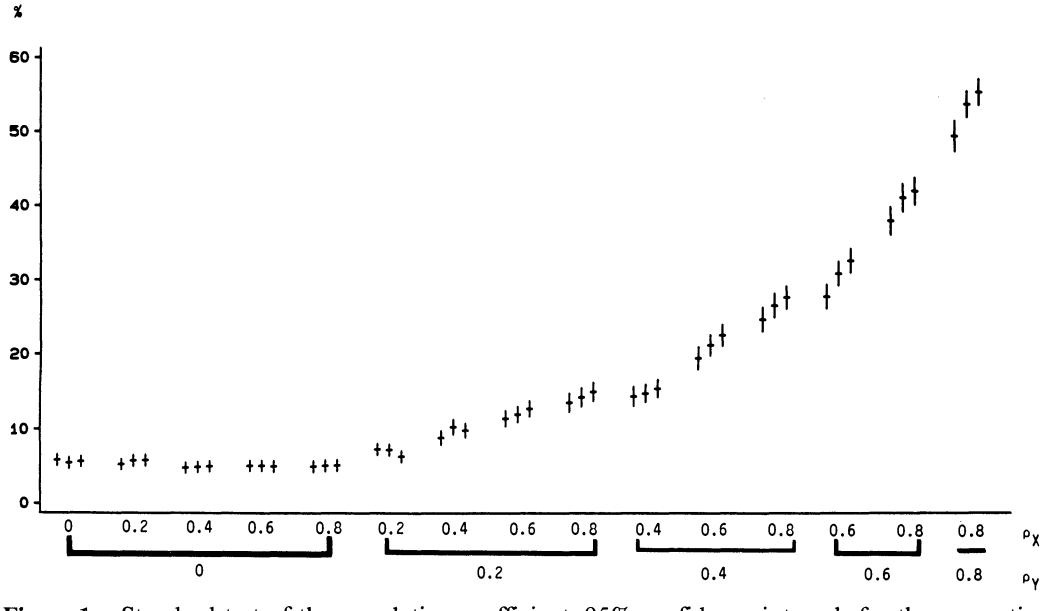

Figure 1. Standard test of the correlation coefficient: 95% confidence intervals for the proportion of Type I errors for a 5% nominal test in the case of two mutually independent simultaneous autoregressive processes on a lattice. For each value of the nearest neighbour autocorrelation for **X** or **Y**, $\rho_X(1)$ or $\rho_Y(1)$, the confidence intervals are plotted for three lattice sizes: 12 × 12, 16 × 16, 20 × 20.

When both **X** and **Y** are highly autocorrelated, observed Type I error rates vary between 25% and 55%, thus clearly showing that the testing procedure needs to be modified.

For the two proposed tests, the observed Type I error rates ranged from 4.2% to 5.85% for the W test (Figure 2) and from 4.1% to 5.9% for the $t_{\hat{M}-2}$ test, and are thus close to the nominal 5% level. Note that when one of the processes has no autocorrelation, W and $t_{\hat{M}-2}$ perform as well as the standard t-test. The confidence interval did not include the 5% nominal level in only two cases [20 × 20, $\rho_X(1) = .2$, $\rho_Y(1) = .2$ and 16 × 16, $\rho_X(1) = .6$, $\rho_Y(1) = .8$]. Inadmissible estimates were rare; they did not occur more than twice per lattice size in all the trials except in the case $\rho_X(1) = 0$, $\rho_Y(1) = .8$, where there were three negative estimates for the 12 × 12 lattice.

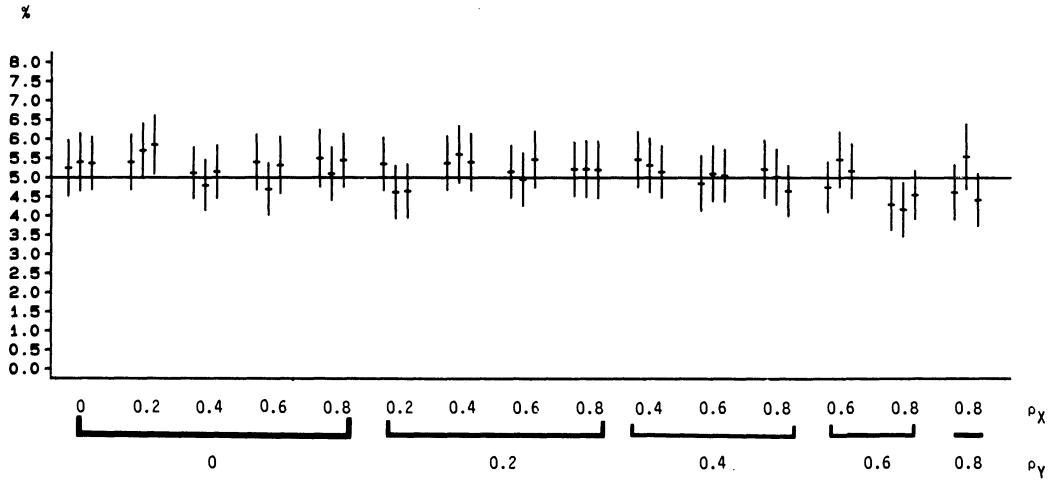

Figure 2. Test based on the standardised covariance W: 95% confidence intervals for the proportion of Type I errors for a 5% nominal test in the case of two mutually independent simultaneous autoregressive processes on a lattice.

A comparison of the empirical variance of r (averaged over the 4,000 trials) and of the average v_r of the estimated variance $\hat{\sigma}_r^2$ given by (2.9), was also made. For small to moderate autocorrelations, there is practically no difference between the empirical variance of r and v_r. As the autocorrelation increases, v_r is consistently too low. That (2.6) is negatively biased as an estimate of the variance of s_{XY} can be easily seen in the case $\mathbf{\Sigma}_X = \mathbf{\Sigma}_Y$.

Quantile plots of the distribution of the W statistic show that it has short tails compared with the normal distribution in the extreme case of $\rho_X(1) = \rho_Y(1) = .8$. The departure from normality is confirmed by a Kolmogorov–Smirnov test, which is significant at the 5% level but not at the 1% level. The tendency for short tails also occurs in other cases of $\rho_X(1)$ and $\rho_Y(1)$, but is more marked for higher autocorrelation. Nevertheless, the departure from normality of W and the order of the negative bias in v_r do not seem to be quantitatively altering the levels of the W and $t_{\hat{M}-2}$ tests even in the strongly autocorrelated cases.

3.2 *Nonlattice Case*

Choice of the model The structure of the network and the type of spatial dependence were both chosen in order to be similar to that of examples in geographical epidemiology which will be discussed in the next section.

The coordinates of the points of the network were identified with the geographical locations of the administrative centers ("préfectures") of French départements. The variables considered are the mortality rate for lung cancer (LC) for men, the cigarette sales (CS) per inhabitant, and the percentage of metal workers (MW) and the percentage of textile workers (TW) with respect to the male working population in each département.

The spatial structure of these variables was investigated by means of a variogram. In this analysis, $N = 82$ locations were retained after grouping the départements around Paris into one area. The distances between the centres of départements were partitioned into 15 classes of 50-km intervals each. This gives 15 strata $S_1, \ldots, S_{15}$; the stratum $S_0 = \{(\alpha, \alpha)\ \alpha \in A\}$. The number of strata chosen should take into account a balance between the sampling fluctuations of the estimated autocovariances when the number of strata is large and the introduction of bias when the number of strata is small. In making this judgement it is helpful to compute the observed semivariogram for several cases. For

the present data set we have found that the results were not sensitive to the choice of the number of strata.

The observed variogram of LC, i.e., the plot of

$$N_k^{-1} \sum_{(\alpha,\beta)\in S_k} (X_\alpha - X_\beta)^2, \qquad k = 1, \ldots, 15,$$

against the average distance, d_k, for départements in S_k, is shown in Figure 3. The variograms of CC and MW were similar and exhibited also an upward trend of fairly linear shape with increasing distance. Hence, a disc model for the covariance matrix (Ripley, 1981, p. 55) seemed appropriate and we chose it to simulate a spatially dependent process on this irregular network.

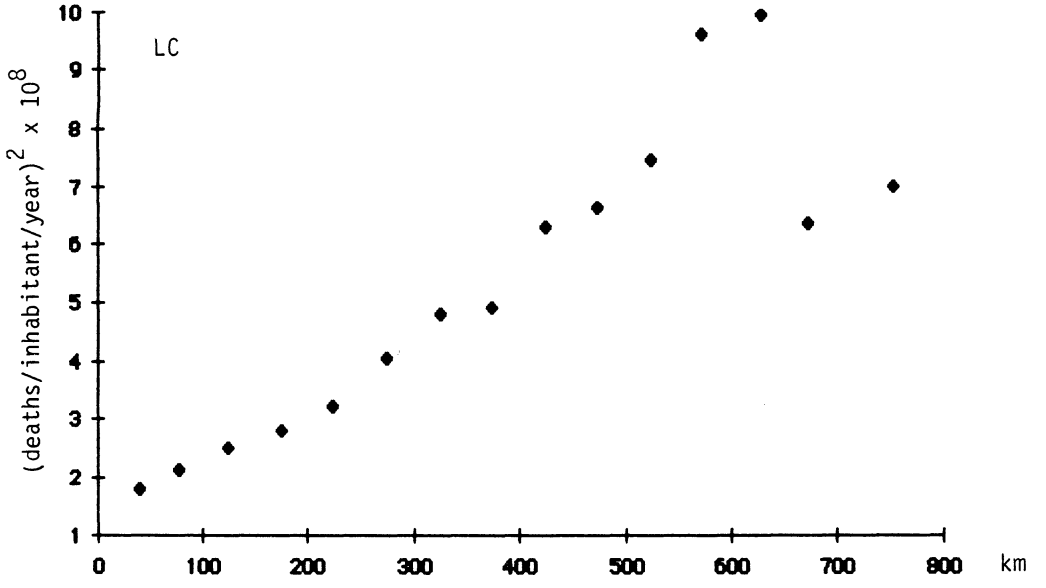

Figure 3. Variogram of the lung cancer mortality (LC). Fifteen classes of distance are considered; the number of ordered pairs in each class is (82; 400; 582; 674; 764; 822; 812; 726; 630; 476; 304; 172; 94; 58; 40). The abscissa corresponds to the average distance in kilometres within each class.

Method The average distance between départements in the first stratum S_1 is approximately 40 km. The parameters of the disc models for **X** and **Y** were chosen to be such that the autocorrelation at distance 40 km is equal to .2, . . . , .9. We also denote these by $\rho_X(1)$ or $\rho_Y(1)$.

For each chosen value of $\rho_X(1)$, the matrix $\boldsymbol{\Sigma}_X$ was triangularised ($\boldsymbol{\Sigma}_X = \mathbf{L}\mathbf{L}^T$) and then a realisation of **X** was obtained by first generating a vector of N i.i.d. N(0, 1) and then multiplying this vector by **L**. In each case 500 pairs of mutually independent processes (**X**, **Y**) were simulated and the statistics W and $t_{\hat{M}-2}$ calculated for each pair (**X**, **Y**) as in Sections 2.2 and 2.3, where the autovariances $\hat{C}_X(k)$ are defined as

$$\hat{C}_X(k) = N_k^{-1} \sum_{(\alpha,\beta)\in S_k} (X_\alpha - \bar{X})(X_\beta - \bar{X}).$$

The same procedure as in the lattice case was adopted for an inadmissible negative estimate for the variance of s_{XY}. A 5% nominal level was chosen to assess the performance of W, $t_{\hat{M}-2}$, and t_{N-2}.

Results Figure 4 demonstrates the increased proportion of Type I errors of the standard t-test in the case of positive autocorrelation of both processes. Figure 5 shows the performance of the W test, which is indistinguishable from the performance of the $t_{\hat{M}-2}$ test. All observed Type I error rates, even in the most highly autocorrelated case, were close to 5%. No inadmissible negative estimate, $\hat{\sigma}_r^2$, arose. Kolmogorov–Smirnov tests performed on the distribution of W were significant in two cases (.2 × .6 and .9 × .9).

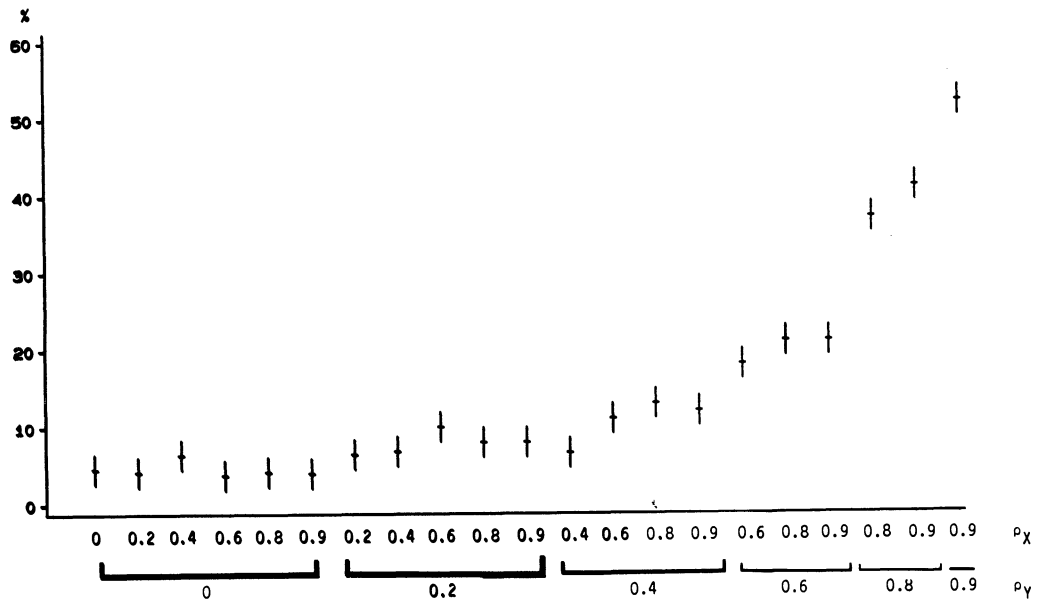

Figure 4. Standard test of the correlation coefficient: 95% confidence intervals for the proportion of Type I errors for a 5% nominal test in the case of two mutually independent processes generated by a disc model on a network of 82 points. (The parameters $\rho_X(1)$ and $\rho_Y(1)$ of the disc models for X and Y respectively are equal to the autocorrelation at 40 km.)

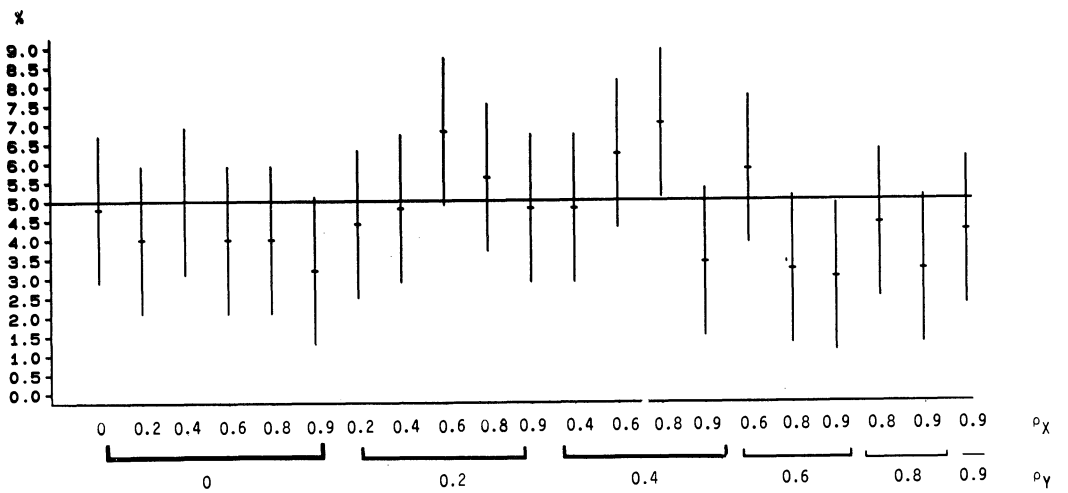

Figure 5. Test based on the standardised covariance W: 95% confidence intervals for the proportion of Type I errors for a 5% nominal test in the case of two mutually independent processes generated by a disc model on a network of 82 points. (The parameters $\rho_X(1)$ and $\rho_Y(1)$ of the disc models for X and Y respectively are equal to the autocorrelation at 40 km.)

4. Examples

Our examples concern the relationship between lung cancer, smoking and industrial factors. We calculate W and $t_{\hat{M}-2}$.

4.1 *The Data*

For 82 départements we considered male lung cancer mortality rate (LC) standardised over the age 35–74 and over a 2-year period, 1968–1969; cigarette sales per inhabitant (CS) in 1953 (a 15-year time lag was chosen to account for the delay between exposure and the onset of the pathology); and demographic data on the percentage of employed males in the metal industry (MW) and the textile industry (TW) recorded by census in 1962.

4.2 *Results*

Using the standard test based on r, there is a highly significant positive association both between LC and CS, and between LC and MW. The association between LC and TW, on the other hand, is less strong but still significant at the 1% level (Table 1). The W and $t_{\hat{M}-2}$ statistics for these three examples are shown in Table 1. One can see a substantial reduction of the degrees of freedom when the autocorrelation is taken into account.

Table 1
Comparison of the significance levels for tests of the association between lung cancer mortality rates and several risk factors given by standard test, W, and $t_{\hat{M}-2}$ tests

	r	t_{N-2}[a]	W	$\hat{M}$	$t_{\hat{M}-2}$
Cigarette sales per inhabitant (1953) (CC)	.76	10.48 $P \approx 10^{-21}$	2.94 $P = .0032$	15	4.22 $P = .001$
% male workers in metal industry (1962) (MW)	.63	7.16 $P \approx 10^{-11}$	2.48 $P = .0136$	16	3.00 $P = .01$
% male workers in textile industry (1962) (TW)	.28	2.57 $P = .01$	1.51 $P = .13$	30	1.52 $P = .15$

[a] Standard test (t-transformation with 80 d.f.).

For CS and MW the CS effective sample size is about 20% of the original sample size. Consequently, the significance levels are reduced but even after this "adjustment" these two factors are statistically significantly associated with lung cancer. For TW the significance disappears after adjustment.

These results can be considered in good agreement with current knowledge concerning life style and occupational risk factors for lung cancer (Schottenfeld and Fraumeni, 1982).

5. Discussion

Our study of the correlation coefficient has been motivated by its widespread use in epidemiology, where relatively small data sets of up to 100 irregularly spaced points are encountered. In this context, positively autocorrelated **X** and **Y** are most commonly observed.

We have confirmed both theoretically and empirically that the uncritical use of the correlation coefficient for testing association between positively autocorrelated processes leads to an inflated proportion of Type I errors, and have shown theoretically that the magnitude of this inflation is consistent with a reduction in the effective sample size. We have then investigated how the correlation coefficient behaves when it is adjusted to take account of this effect. We have done this in two related ways, the W and $t_{\hat{M}-2}$ tests, and

have shown that when adjustment is made the Type I error rate is much closer to the nominal level. These methods do not require the identification of a particular parametric model for the type of spatial autocorrelation and they cope equally well with regularly and irregularly spaced points. Isotropy need not be assumed. The strata can be defined by orientation as well as distance in the spirit of Granger (1969) because equations (2.7) and (2.9) can be easily adapted to any partition of the set of pairs of locations. Only simple calculations of autocorrelations, which can be done on small computers, are involved.

The detection of association between spatial data sets is not a simple problem. In this paper we have investigated one particular approach. As well as having reservations about the assumption of stationarity, in a detailed statistical analysis it should be recognised that association can exist simultaneously at a number of different geographical scales. The correlation coefficient is a single omnibus statistic that averages the scale-dependent association. Thus, for example, it is possible that negative association at small scales is swamped by positive association at large scales. We have not attempted to explore the ways in which these scale-dependent associations can be separated out. This is an important area of research.

ACKNOWLEDGEMENTS

The authors wish to thank Annie Mollié and Nicole Le Moual for computing assistance. This work was supported by Euratom Contract BI6-0126-F.

RÉSUMÉ

Des tests d'association modifiés, portant sur le coefficient de corrélation ou la covariance entre deux processus spatiaux autocorrelés, sont proposés. Ces tests peuvent être utilisés à la fois pour des données observées sur un lattice ou sur un domaine irrégulier. Il sont fondés sur l'évaluation d'un nombre d'observation corrigé qui prend en compte la structure spatiale de chaque processus.

Le nombre d'observations corrigé est inférieur à la taille de l'échantillon quand l'autocorrélation de chaque processus est positive. Une méthode pour évaluer cette réduction par l'intermédiaire d'une approximation de la variance du coefficient de corrélation est développée. La performance des tests est évaluée par des simulations de Monte-Carlo. La méthode est illustrée par des exemples d'épidémiologie géographique.

REFERENCES

Armstrong, B. and Doll, R. (1975). Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer* **15,** 617–631.

Bartlett, M. S. (1935). Some aspects of the time-correlation problem in regard to tests of significance. *Journal of the Royal Statistical Society* **98,** 536–543.

Bartlett, M. S. (1978). Nearest neighbour models in the analysis of field experiments. *Journal of the Royal Statistical Society, Series B* **40,** 147–174.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* **36,** 192–236.

Cliff, A. D. and Ord, J. K. (1975). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society, Series B* **37,** 297–348.

Clifford, P. and Richardson, S. (1985). Testing the association between two spatial processes. *Statistics and Decisions,* Supp. issue **2,** 155–160.

Cook, D. G. and Pocock, S. J. (1983). Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics* **39,** 361–371.

Davies, H. M. and Jowett, G. H. (1958). The fitting of Markoff serial variation curves. *Journal of the Royal Statistical Society, Series B* **20,** 120–142.

Granger, C. W. J. (1969). Spatial data and time series analysis. In *London Papers in Regional Science* 1. *Studies in Regional Science,* A. J. Scott (ed.), 1–24. London: Pion.

Haugh, L. D. (1976). Checking the independence of two covariance stationary time series: A univariate residual cross-correlation approach. *Journal of the American Statistical Association* **71,** 378–385.

Hoover, R. and Fraumeni, J. F. (1975). Cancer mortality in U.S. counties with chemical industries. *Environmental Research* **9,** 196–207.
Hubert, L. J. and Golledge, R. G. (1982). Measuring association between spatially defined variables: Tjostheim's index and some extensions. *Geographical Analysis* **14,** 273–278.
Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71,** 135–146.
Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* **70,** 120–126.
Richardson, S. and Hémon, D. (1981). On the variance of the sample correlation between two independent lattice processes. *Journal of Applied Probability* **18,** 943–948.
Ripley, B. D. (1981). *Spatial Statistics.* New York: Wiley.
Schottenfeld, D. and Fraumeni, J. F. (1982). *Cancer Epidemiology and Prevention.* Philadelphia: W. B. Saunders.
Student (W. S. Gosset) (1914). The elimination of spurious correlation due to position in time or space. *Biometrika* **10,** 179–181.
Tjøstheim, D. (1978). A measure of association for spatial variables. *Biometrika* **65,** 109–114.
Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41,** 434–449.

Received June 1987; *revised June* 1988.

Appendix 1

The Variance of r

Let $\mathbf{\Sigma}_\xi$ and $\mathbf{\Sigma}_\eta$ denote the variance–covariance matrices of the vectors with elements $\xi_\alpha = X_\alpha - \bar{X}$ and $\eta_\alpha = Y_\alpha - \bar{Y}$.

If we write

$$r = \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\eta}/(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}\,\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta})^{1/2}$$

then

$$\mathrm{var}(r) = \mathrm{E}\ \mathrm{tr}\left\{\frac{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}}\frac{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta}}\right\} = \mathrm{tr}\left(\mathrm{E}\,\frac{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}}\,\mathrm{E}\,\frac{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta}}\right) \tag{A.1}$$

by the independence of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$.

Expanding $\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta}$ about its expectation, $k_1 = \mathrm{tr}(\mathbf{\Sigma}_\eta)$, we have

$$\frac{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta}} = \frac{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}{k_1}\left[1 - \frac{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta} - k_1}{k_1} + \frac{(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta} - k_1)^2}{k_1^2} - \cdots\right]. \tag{A.2}$$

Note that in the case $\mathbf{\Sigma}_Y = \mathbf{I}$ we have $k_1 = N - 1$ so that we would be expanding in inverse powers of $N - 1$. To evaluate the expectation of (A.2) we calculate

$$\mathrm{E}(\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}})^2 = 2\mathbf{\Sigma}_\eta^2 + k_1\mathbf{\Sigma}_\eta$$

and

$$\mathrm{E}(\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}})^3 = 8\mathbf{\Sigma}_\eta^3 + 2k_2\mathbf{\Sigma}_\eta + 4k_1\mathbf{\Sigma}_\eta^2 + k_1^2\mathbf{\Sigma}_\eta,$$

where $k_2 = \mathrm{tr}(\mathbf{\Sigma}_\eta^2)$.

If we make the plausible assumption that the maximum eigenvalue of $\mathbf{\Sigma}_\eta$ is bounded above by λ_{B} as $N \to \infty$, then using the inequality $\mathrm{tr}(\mathbf{\Sigma}_\eta^2) \leqslant \lambda_{\mathrm{B}}\ \mathrm{tr}(\mathbf{\Sigma}_\eta)$, we have the second-order asymptotic expression

$$\mathrm{E}\left(\frac{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\eta}}\right) = \frac{\mathbf{\Sigma}_\eta}{k_1} - 2\,\frac{\mathbf{\Sigma}_\eta^2}{k_1^2} + 2\,\frac{\mathbf{\Sigma}_\eta k_2}{k_1^3}. \tag{A.3}$$

Finally, substituting (A.3) and a similar expression for $\mathrm{E}(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}/\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi})$ into (A.1), we have to the first order

$$\sigma_r^2 = \frac{\mathrm{tr}(\mathbf{\Sigma}_\eta\mathbf{\Sigma}_\xi)}{k_1 j_1}, \tag{A.4}$$

and to the second order

$$\sigma_r^2 = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\eta \boldsymbol{\Sigma}_\xi)}{k_1 j_1} - \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_\eta^2 \boldsymbol{\Sigma}_\xi)}{k_1^2 j_1} + \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_\eta \boldsymbol{\Sigma}_\xi)}{k_1 j_1}\left(\frac{k_2}{k_1^2} + \frac{j_2}{j_1^2}\right) - \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_\eta \boldsymbol{\Sigma}_\xi^2)}{j_1^2 k_1}, \tag{A.5}$$

where $j_n = \mathrm{tr}(\boldsymbol{\Sigma}_\xi^n)$.

In the case $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_\xi$ the second-order term in (A.5) becomes

$$-4\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_\eta^3)}{k_1^3} - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_\eta^2)\mathrm{tr}(\boldsymbol{\Sigma}_\xi^2)}{k_1^4}\right\} = -4\left\{\sum_{i=1}^{N-1} p_i^3 - \left(\sum_{i=1}^{N-1} p_i^2\right)^2\right\}$$

$$= -4\left\{\sum_{i=1}^{N-1}\left(p_i - \sum_{i=1}^{N-1} p_i^2\right)^2 p_i\right\} \leq 0,$$

where $p_i = \mu_i / \sum_{j=1}^{N-1} \mu_j$, $\{\mu_i\}_{i=1}^{N-1}$ being the eigenvalues of $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{\Sigma}_\xi$ (matrices of rank $\leq N-1$), i.e., the first-order approximation is no larger than the second-order approximation.

Note that in the case where (i) $\boldsymbol{\Sigma}_\xi$ and $\boldsymbol{\Sigma}_\eta$ commute, (ii) $\boldsymbol{\Sigma}_\eta$ is proportional to an idempotent matrix, and (iii) the eigenvalues of $\boldsymbol{\Sigma}_\xi$ are zero whenever the associated eigenvalues of $\boldsymbol{\Sigma}_\eta$ are zero, then the first-order term in (A.4) is exactly equal to σ_r^2. This follows from the remark that in this case r has p.d.f. $f_M(r)$ with effective sample size $M = 1 + \mathrm{rank}(\boldsymbol{\Sigma}_\eta)$, where $f_M(r)$ is the standard distribution of r defined in (2.2).

Appendix 2

Simulation of Simultaneous Autoregressive Process

For a process given by (3.1) with zero boundary, i.e., $X_{s,0} = X_{0,t} = X_{n+1,t} = X_{s,n+1} = 0$ ($s, t = 0, 1, \ldots, n+1$), the spectral decomposition of $\boldsymbol{\Sigma}_\mathbf{X}$ can be deduced from the expression given by Besag [discussion of a paper by Bartlett (1978)]. The process can be simulated by computing $\mathbf{P}^\mathrm{T}\mathrm{diag}(\delta_i^{1/2})\mathbf{Z}$, where $\mathbf{Z}$ is a vector of independent N(0, 1) variables and where $\mathbf{P}^\mathrm{T}\mathrm{diag}(\delta_i)\mathbf{P}$ is the spectral decomposition of $\boldsymbol{\Sigma}_\mathbf{X}$. If we denote the matrix $\{X_{s,t}\}_{s,t=1}^n$ by $\mathbf{X}$ and rearrange the vector of $\mathbf{Z}$ into an $n \times n$ matrix whose elements are Z_{ij} ($i, j = 1, 2, \ldots, n$), then, exploiting the properties of $\mathbf{P}$, we find that $\mathbf{X}$ is simulated by

$$\mathbf{X} = \mathbf{Q}\{Z_{ij}\lambda_{ij}\}\mathbf{Q},$$

where

$$\lambda_{ij} = 1 - 2a\{\cos[\pi i/(n+1)] + \cos[\pi j/(n+1)]\}$$

and

$$Q_{ij} = [2/(n+1)]^{1/2}\sin[\pi ij/(n+1)], \quad n \text{ even}.$$

The computational cost of a single realisation is therefore of the order of n^3 operations or $N^{3/2}$ since $N = n^2$. In contrast, a single triangulation $\mathbf{L}\mathbf{L}^\mathrm{T}$ of $\boldsymbol{\Sigma}_\mathbf{X}$ requires an order of N^3 operations and each realisation involves an order of N^2 operations.

References

Some Aspects of the Time-Correlation Problem in Regard to Tests of Significance
M. S. Bartlett
Journal of the Royal Statistical Society, Vol. 98, No. 3. (1935), pp. 536-543.
Stable URL:
http://links.jstor.org/sici?sici=0952-8385%281935%2998%3A3%3C536%3ASAOTTP%3E2.0.CO%3B2-2

Multiple Regression in Geographical Mortality Studies, with Allowance for Spatially Correlated Errors
D. G. Cook; S. J. Pocock
Biometrics, Vol. 39, No. 2. (Jun., 1983), pp. 361-371.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198306%2939%3A2%3C361%3AMRIGMS%3E2.0.CO%3B2-Q

Checking the Independence of Two Covariance-Stationary Time Series: A Univariate Residual Cross-Correlation Approach
Larry D. Haugh
Journal of the American Statistical Association, Vol. 71, No. 354. (Jun., 1976), pp. 378-385.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197606%2971%3A354%3C378%3ACTIOTC%3E2.0.CO%3B2-Q

LINKED CITATIONS

Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression
K. V. Mardia; R. J. Marshall
Biometrika, Vol. 71, No. 1. (Apr., 1984), pp. 135-146.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198404%2971%3A1%3C135%3AMLEOMF%3E2.0.CO%3B2-G

Estimation Methods for Models of Spatial Interaction
Keith Ord
Journal of the American Statistical Association, Vol. 70, No. 349. (Mar., 1975), pp. 120-126.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197503%2970%3A349%3C120%3AEMFMOS%3E2.0.CO%3B2-A

On the Variance of the Sample Correlation between Two Independent Lattice Processes
Sylvia Richardson; Denis Hemon
Journal of Applied Probability, Vol. 18, No. 4. (Dec., 1981), pp. 943-948.
Stable URL:
http://links.jstor.org/sici?sici=0021-9002%28198112%2918%3A4%3C943%3AOTVOTS%3E2.0.CO%3B2-R

On Stationary Processes in the Plane
P. Whittle
Biometrika, Vol. 41, No. 3/4. (Dec., 1954), pp. 434-449.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28195412%2941%3A3%2F4%3C434%3AOSPITP%3E2.0.CO%3B2-C