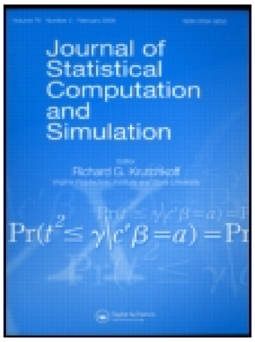

Journal of Statistical Computation and Simulation

Richard G. Krutchkoff

Journal of statistical computation and simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Lack of robustness in two tests of normality against autocorrelation in sample data

Pierre Durilleul & Pierre Legendre

To cite this article: Pierre Durilleul & Pierre Legendre (1992) Lack of robustness in two tests of normality against autocorrelation in sample data, Journal of statistical computation and simulation, 42:1-2, 79-91, DOI: 10.1080/00949659208811412

To link to this article: https://doi.org/10.1080/00949659208811412

Published online: 20 Mar 2007.

J. Statist. Comput. Simul., Vol. 42, pp. 79–91
Reprints available directly from the publisher
Photocopying permitted by license only

LACK OF ROBUSTNESS IN TWO TESTS OF NORMALITY AGAINST AUTOCORRELATION IN SAMPLE DATA

PIERRE DUTILLEUL and PIERRE LEGENDRE

Département de Sciences Biologiques, Université de Montréal,
C. P. 6128 Succursale A, Montréal, Québec, Canada H3C 3J7

(*Received 23 September 1991*)

Robustness against autocorrelation in time-series data is investigated for two tests of normality: the Kolmogorov-Smirnov test, in the class of normality tests using statistics based on the empirical cumulative distribution function, and the Shapiro-Wilk analysis-of-variance test, which regresses the ordered sample values on the corresponding expected normal order statistics. For a Gaussian first-order autoregressive process, it is shown by simulation that: 1. for short series, both tests are conservative for some range of negative values of first-order autocorrelation, and too liberal for medium-to-high positive and high negative values; 2. for moderate sample sizes, both tests are no longer conservative, but remain too liberal asymmetrically for high negative and positive values of first-order autocorrelation; 3. the Kolmogorov-Smirnov test, which traditionally suffers from lack of power in comparisons with the *W* test of Shapiro and Wilk, is more robust against autocorrelation in time-series data, whatever the sign of the first-order autocorrelation. We illustrate that these results also apply to spatially autocorrelated data along a transect.

KEY WORDS: Autocorrelation, Kolmogorov-Smirnov test, normality, robustness, sample data, Shapiro-Wilk test.

1. INTRODUCTION

Standard techniques of statistical inference are most often concerned with sets of observations drawn from independently and identically distributed random variables — that we will refer to as "random samples" in order to avoid lengthly repetitions, while autocorrelation in sample data may impair the ability to use these same tests in real-case studies and can alter the conclusions of statistical analyses performed without allowance for it. Autocorrelation is a nuisance in that it produces a bias when estimating variances and correlation coefficients, and does not provide minimum-variance unbiased linear estimators. This is well-known for such methods as the analysis of variance (Box, 1954; in time: Milliken and Johnson, 1984, Ch. 27; Crowder and Hand, 1990, Section 3.6; in space: Bartlett, 1978; Griffith, 1978; Legendre *et al.*, 1990), regression analysis (in time: Durbin and Watson, 1950, 1951; Walker, 1971; in space: Cliff and Ord, 1981, Section 8.2; Cook and Pocock, 1983), and correlation analysis (in time: Diggle, 1990, Section 8.3; in space: Bivand, 1980; Clifford *et al.*, 1989). Fortunately, valid statistical inference does not always require independence of the observations as a necessary condition; for instance, the necessary and sufficient condition for valid unmodified F testing in analysis of variance models allows for limited forms of dependence and heteroscedasticity (Huynh and Feldt, 1970; Rouanet and Lépine, 1970) although in most cases, unmodified F tests are not valid for autocorrelated sample data.

Goodness-of-fit tests for autocorrelated sample data are a problem of interest not fully explored by statisticians. Despite its practical importance, there is little literature about it. The distribution of Pearson (1900) chi-square statistic under the hypothesis of normality and its power were studied by simulation for stationary time-series processes by Gasser (1975). Testing the fit to a specified normal law through the chi-square statistic, Moore (1982) theoretically demonstrated that when large samples of time-series data come from a general Gaussian stationary process, positive correlation among the observations cannot be told apart from lack of normality. Gleser and Moore (1983) extended these results to stationary processes satisfying a positive dependence condition, for the Kolmogorov (1933) and Smirnov (1939) and the Cramér (1928) and von Mises tests, in the class of normality tests based on the empirical cumulative distribution function. More recently, Pierce (1985) demonstrated that any normality test performed on residuals in time-series autoregressive models of some well-specified order has the same limiting null distribution as the random-sample standard case when the parameters are estimated from the data.

In the present paper, we investigate by simulation the effect of positive and negative dependence among observations on two major normality tests, with respect to sample size, when the null hypothesis of a normal univariate marginal distribution is correct. The emphasis is on time, but a few illustrations of the effect of spatial autocorrelation will also be given. We selected a version of the Kolmogorov-Smirnov test adapted to account for the estimation of one or more parameters from the sample data (Stephens, 1974), and the analysis-of-variance omnibus test of Shapiro and Wilk (1965) which regresses the ordered sample values on the corresponding expected normal order-statistics. These tests are based on quite different approaches, using the empirical cumulative distribution function on the one hand and the least-squares estimation in the normal probability plot on the other; both were designed for the random-sample case. The question is: to what extent are these tests robust against autocorrelation, i.e. within what range of dependencies among observations do they remain valid? We performed our investigation of robustness through simulations, which are the procedure usually adopted in null percentage point and power studies, because of the hardly tractable mathematics underlying the complex effect of autocorrelation on both of these statistics. In Section 2, we briefly overview the historical background of the plethora of normality tests. We describe our simulation design in Section 3.1, and present and discuss our numerical results in Section 3.2. In Section 4, we propose a general solution that may help overcome the nuisance of autocorrelation, when present, on tests of normality.

2. TESTS OF NORMALITY: A BRIEF OVERVIEW

Karl Pearson may be regarded as having initiated the modern theory of testing for departures from normality. First, he showed how deviations from normality could be measured through the standard third and fourth moments of a distribution, respectively used to estimate skewness and kurtosis (or peakedness) (Pearson, 1895). Secondly, he developed the chi-square test, although not specifically to test for deviations from normality (Pearson, 1900). Pearson's chi-square test of normality contrasts the observed numbers of observations in each of the categories into which the distribution is discretised, with the numbers expected if the population from which the data are sampled is normal with known mean and variance. This test is well adapted for the cases where the distribution is discontinuous and where the parameters must themselves be estimated from the sample data; in all other cases, however, it is actually not very sensitive because it does not take into account the ordered nature of the data and should not be used routinely (see e.g. Stephens, 1974 and D'Agostino, 1982).

For a random sample $X_1, \ldots, X_n$ of size n, a general procedure using statistics based on the empirical cumulative distribution function $F_n(x)$ was developed to test for departures of fit from a specified parametric distribution, which can be used in particular for the normal law. Function $F_n(x)$ is defined as the proportion of sample values smaller than or equal to x, and is an estimate of the theoretical cumulative distribution function $F(x)$. Many statistics measuring the discrepancy between $F_n(x)$ and $F(x)$ have been proposed to test for fit. Three of the leading ones are: the Cramér (1928) and von Mises statistic

$$W^2 = \sum_{i=1}^{n} [F(X_i) - F_n(X_i) - 1/(2n)]^2 + 1/(12n),$$

the Kolmogorov (1933) and Smirnov (1939) statistic

$$D = \max(D^+, \ D^-)$$

with

$$D^+ = \sup_{1 \le i \le n}[F_n(X_i) - F(X_i)]$$

and

$$D^- = \sup_{1 \le i \le n} [F(X_i) - F_n(X_i)],$$

and the Anderson and Darling (1954) statistic

$$A^2 = - \left[\sum_{i=1}^{n} \{2F_n(X_i) - 1/\mathrm{n}\}\{\log F(X_i) + \log(1 - F(X_{n+1-i})\} \right] - n.$$

When $F(x)$ denotes the normal distribution function, large values of those statistics indicate nonnormality. All three of these discrepancy measures were primarily designed for the random-sample case where the hypothesised distribution function $F(x)$ is completely specified, i.e. for a simple null hypothesis. So they are not of much practical interest, despite a higher power in comparisons with the chi-square test. Using them when one or more parameters are not specified and must be estimated from the sample data, i.e. with a composite null hypothesis, produces very conservative tests. From earlier suggestions of Lilliefors (1967) and van Soest (1967), Stephens (1974) developed adjustments of the significance values that are valid for stated probabilities of type I error.

An alternative to formal statistical tests is provided by a graphical procedure called normal probability plotting, that was developed as an informal technique for judging deviations from normality. The objective is to plot the ordered sample values versus the inverse of a normal cumulative distribution function in such a way that if the underlying population is normally distributed, the graph will be a straight line. Some illustrations of how these deviations from linearity indicate the degree and type of nonnormality may be found in D'Agostino (1982), for instance.

Research into normality tests received some impetus with the introduction of the so-called analysis of variance W statistic by Shapiro and Wilk (1965). The innovation consisted of quantifying the information contained in the normal probability plot; they defined W as the F-ratio between the estimated variance obtained by weighted least-squares of the slope and the classical sample variance, to judge the adequacy of the linear fit

$$W = \frac{\left(\sum_{i=1}^{n} w_i X_i'\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

where $X_1' \leq \ldots \leq X_n'$ are the order statistics and the coefficients w_i $(i = 1, \ldots, n)$ are the optimal weights for a population assumed to be normally distributed. So W can be viewed as the square of the correlation coefficient obtained from the normal probability plot. Small values of W indicate nonnormality. Critical values for levels $\alpha = 0.01, 0.02, 0.05, 0.10$ and sample sizes between 3 and 50 are given in Shapiro and Wilk (1965). Royston (1982a) extended W for sample sizes up to 2000. Extensive simulation studies indicated that W has good power properties against a wide range of nonnormal distributions for a variety of sample sizes, and is therefore a very sensitive omnibus test statistic (Shapiro *et al.*, 1968). When carried out without adjustment for the composite null hypothesis, normality tests using statistics based on the empirical cumulative distribution function have suffered in comparisons with W (Shapiro and Wilk, 1965); with the appropriate adjustment, the powers of modified W^{2*} and $A2^*$ are comparable to that of W (Stephens, 1974) but no single test is optimum for all possible deviations from normality. For supplementary power results related to normality tests, good surveys are Stephens (1974, 1982), Pearson *et al.* (1977), D'Agostino (1982) and Royston (1982a), among others. From a practical point of view, the user may still prefer using statistics based on the empirical cumulative distribution function (W^{2*}, D^*, A^{2*}) because a different set of coefficients w_i $(i = 1, \ldots, n)$ is required for each n when computing W, although algorithms are available to compute the exact or approximate expected normal order statistics (Royston, 1982b) and the W significance values for any sample size between 3 and 2000 (Royston, 1982c).

3. LACK OF ROBUSTNESS AGAINST AUTOCORRELATION

3.1. *Simulation Design*

When the null hypothesis of a normal univariate marginal distribution is correct, to what extent are Kolmogorov-Smirnov D^* as modified by Stephens (1974) and Shapiro-Wilk W tests robust against autocorrelation, i.e. what is the range of their respective validity in the presence of dependencies among observations? With this aim, we adopted an approach often used in null percentage point and power studies, and performed our investigation of robustness through simulations, for a variety of autocorrelation values and sample sizes, because of the complex effect of autocorrelation which induces bias in variance estimations in both of these statistics. Autocorrelated sample data were generated from the following first-order autoregressive (AR(1)) processes.

Let us define the well-known temporal stationary first-order autoregressive process by

$$X_t = \rho X_{t-1} + \varepsilon_t \qquad \text{with} \qquad |\rho| < 1.0 \tag{1}$$

and the symmetric nearest-neighbour autoregression model along a spatial transect by

$$X_s = \beta(X_{s-1} + X_{s+1}) + \varepsilon_s \qquad \text{with} \qquad |\beta| < 0.5 \tag{2}$$

(Bartlett, 1978). To generate values of a normally distributed white noise ε with zero mean and unit variance, we used the SAS (Statistical Analysis System) random number function RANNOR which applies the Box-Muller transformation.

The procedure adopted for simulating n-size samples of temporally autocorrelated data from (1) obeys the following straightforward algorithm:

Initialisation: generate ε_1 (call RANNOR) and set $X_1 = \varepsilon_1$;

Step 2: generate ε_2 and compute $X_2 = \rho X_1 + \varepsilon_2$; . . .

Step $1000+n$: generate ε_{1000+n} and compute $X_{1000+n} = \rho X_{999+n} + \varepsilon_{1000+n}$;

End: retain as the n-size sample the set (X_i; $i = 1001, \ldots, 1000 + n$) of autocorrelated data in order to avoid influences of the origin of the simulated time series.

The procedure adopted for simulating spatially autocorrelated data along a transect, following equation (2), is somewhat different even if the basic idea remains to simulate more than n, say $50 + n$, autocorrelated data (X_i; $i = 1, \ldots, 50 + n$) and then, retain as n-size sample the set (X_i; $i = 25 + 1, \ldots, 25 + n$) in order to avoid edge effects. Let V denote the covariance matrix of $50 + n$ spatially autocorrelated data from (2). With slight modifications, we derived matrix V from the covariance matrix given in Cliff and Ord (1981, Section 6.2) for a two-dimensional simultaneous first-order autoregressive process. If $e = (\varepsilon_1, \ldots, \varepsilon_{50+n})$ denotes a vector of 50 + n generated values of ε, then the transformation

$$(X_1, \ldots, X_{n+50})' = V^{1/2} e$$

performed with the SAS matrix algebra procedure, provides the desired 50 + n spatially autocorrelated data.

The parameter values used when simulating autocorrelated sample data are: for process (1), $n = 20, 51, 100, 200$ and $\rho = -0.9$ to $+0.9$ by steps of 0.1; for process (2), $n = 20$ and $\beta = -0.45$, -0.4 to $+0.4$ by steps of 0.1, $+0.45$. For sample size $n = 20$ and for each ρ value in the time domain, 5000 samples were simulated (1000 for each β value in the space domain); for other sample sizes and each ρ value, 2500 samples were simulated. Different sets of pseudo-random numbers were generated for each simulation, to avoid dependence among results.

Kolmogorov-Smirnov tests were performed using the VERNORM procedure of the "R" package (Legendre and Vaudor, 1991) when the sample size was $n = 20$, and by the UNIVARIATE procedure of SAS Version 5 for higher sample sizes; SAS Version 5 does not allow to compute the Kolmogorov-Smirnov test for $n \leq 50$. For all sample sizes, Shapiro-Wilk tests were carried out through the UNIVARIATE procedure of SAS Version 6. Both tests were performed on the same simulated samples.

3.2. Results and Discussion

For a given sample size, the results are presented in terms of the observed significance level, that is, the proportion of times the normal null hypothesis is rejected among the (1000, 2500 or 5000) tests performed on the simulated data; the abscissa represents the first-order autocorrelation or autoregression coefficient values. When the null hypothesis of a normal univariate marginal distribution is correct, which is the case here because of the normality of the ε_i's and consequently of the X_i's, one would expect to observe significance levels equal to the nominal probability of type I error ($\alpha = 0.01, 0.05$, or 0.10), whatever the first-order autocorrelation or autoregression coefficient value; it should be kept in mind, however, that both tests were designed for the random sample case. We used Royston (1982c) algorithm to compute the significance values of W.

Figure 1 reveals interesting and, to some extent, somewhat surprising features of the behaviour of statistics D^* and W in the presence of autocorrelation in time-series sample data. First, both statistics are too liberal for medium-to-high positive autocorrelation values. This was expected from the theoretical asymptotic results of Gleser and Moore (1983); to our knowledge, quantitative results have never been published for the Kolmogorov-Smirnov and the Shapiro-Wilk tests, and in particular not for small samples. We have found this first feature to be present in small ($n = 20$) as well as moderate ($n = 51, 100, 200$) sample sizes. Secondly, the behaviour of both statistics for negative first-order autocorrelation values, i.e. in the presence of autocorrelation which is negative at odd time lags and positive at even lags, is asymmetrical with respect to positive values, for all the sample sizes that we considered. The justification has to be found in the reduction in induced bias when estimating variances, due to the alternation of negative and positive autocorrelation values at odd and even time lags. Finally, the surprising but most striking result comes from the conservative behaviour of both tests for short series, shown for $n = 20$ in Figure 1 (a) and (b) in the range of negative values of first-order autocorrelation between -0.4 and -0.8. This feature disappears with both tests for moderate sample sizes ($n = 100$ and more).

The comparison between D^* and W in terms of robustness against autocorrelation in time-series data suggests that the Kolmogorov-Smirnov test, which traditionally

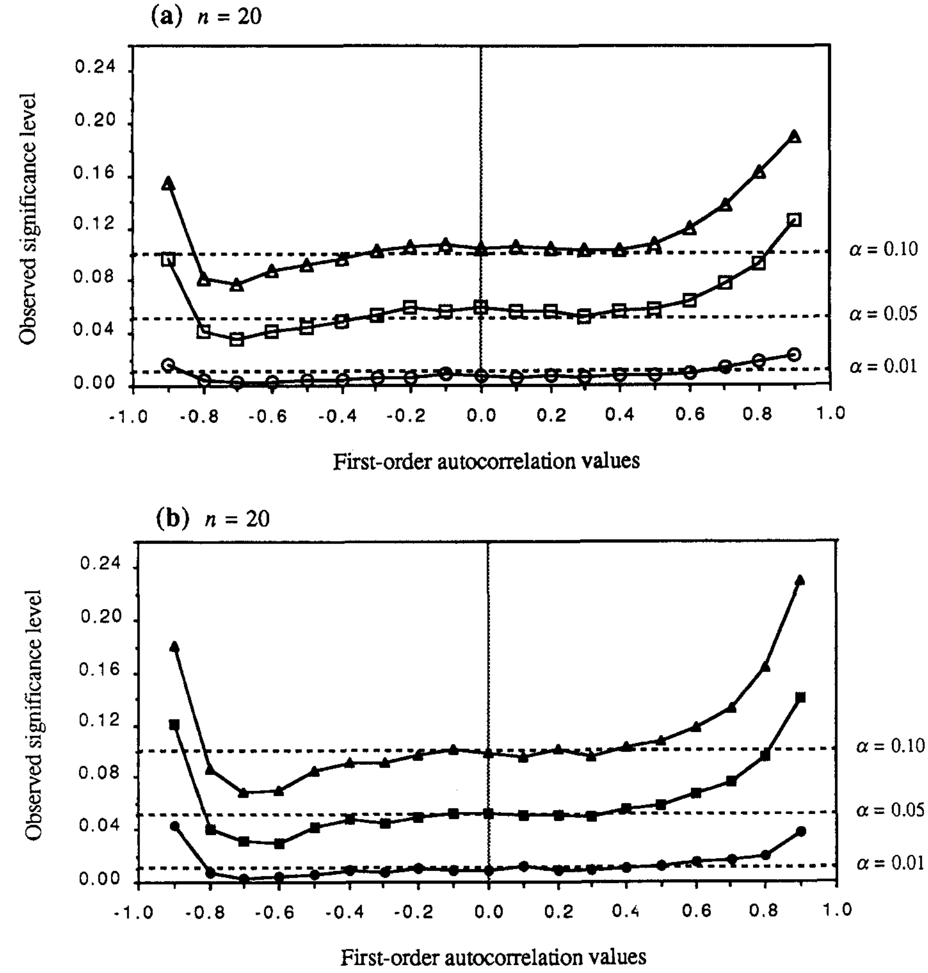
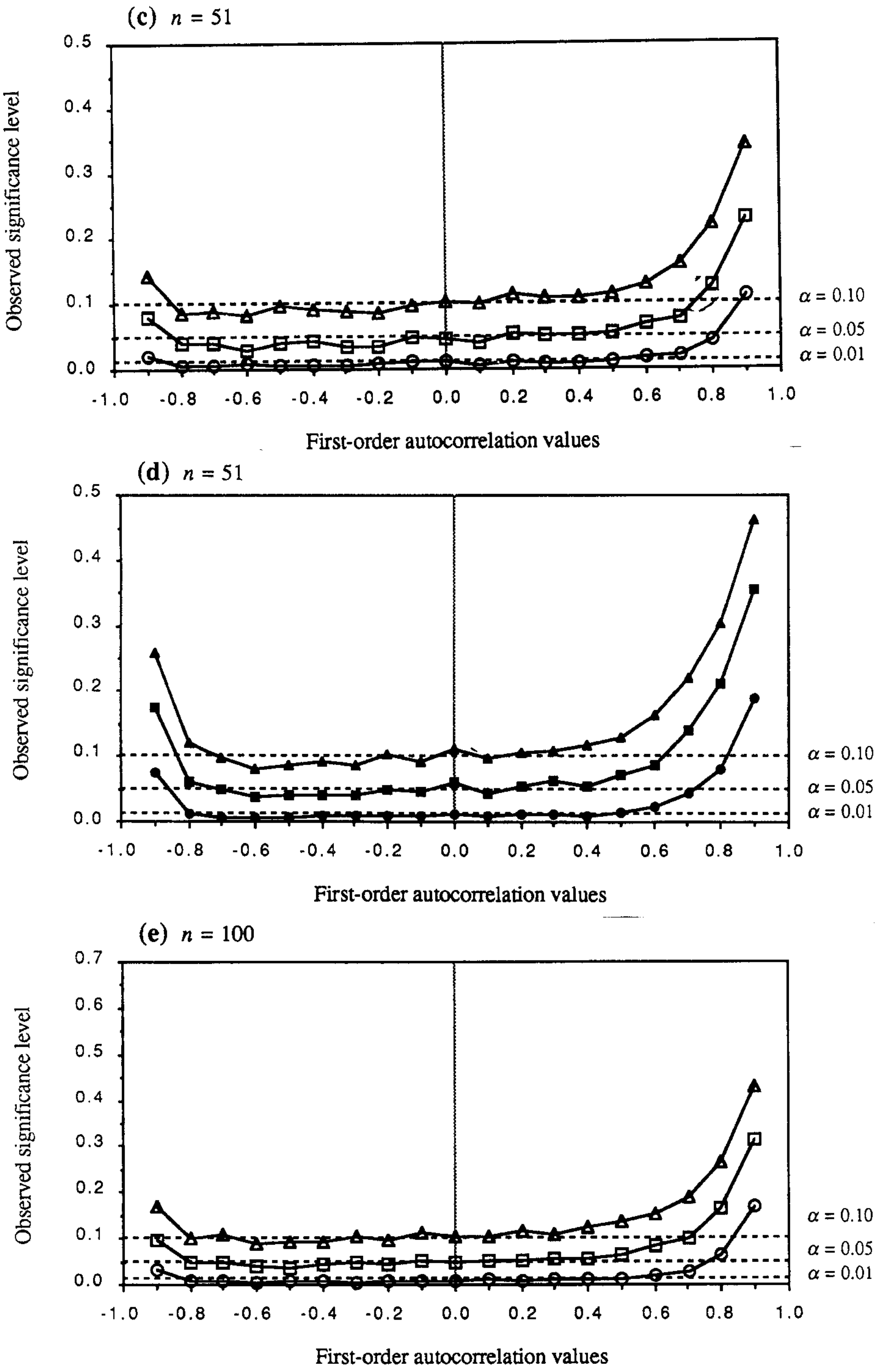


Figure 1 Kolmogorov-Smirnov (a)-(c)-(e)-(g) and Shapiro-Wilk (b)-(d)-(f)-(h) tests of normality, performed for simulated time series arising from a Gaussian AR(1) process: observed significance levels as a function of the first-order autocorrelation values. The nominal significance levels considered are: circles, $\alpha = 0.01$; squares, $\alpha = 0.05$; triangles $\alpha = 0.10$. The computed Kolmogorov-Smirnov tests account for the estimation of two parameters from the sample data, following Stephens (1974).

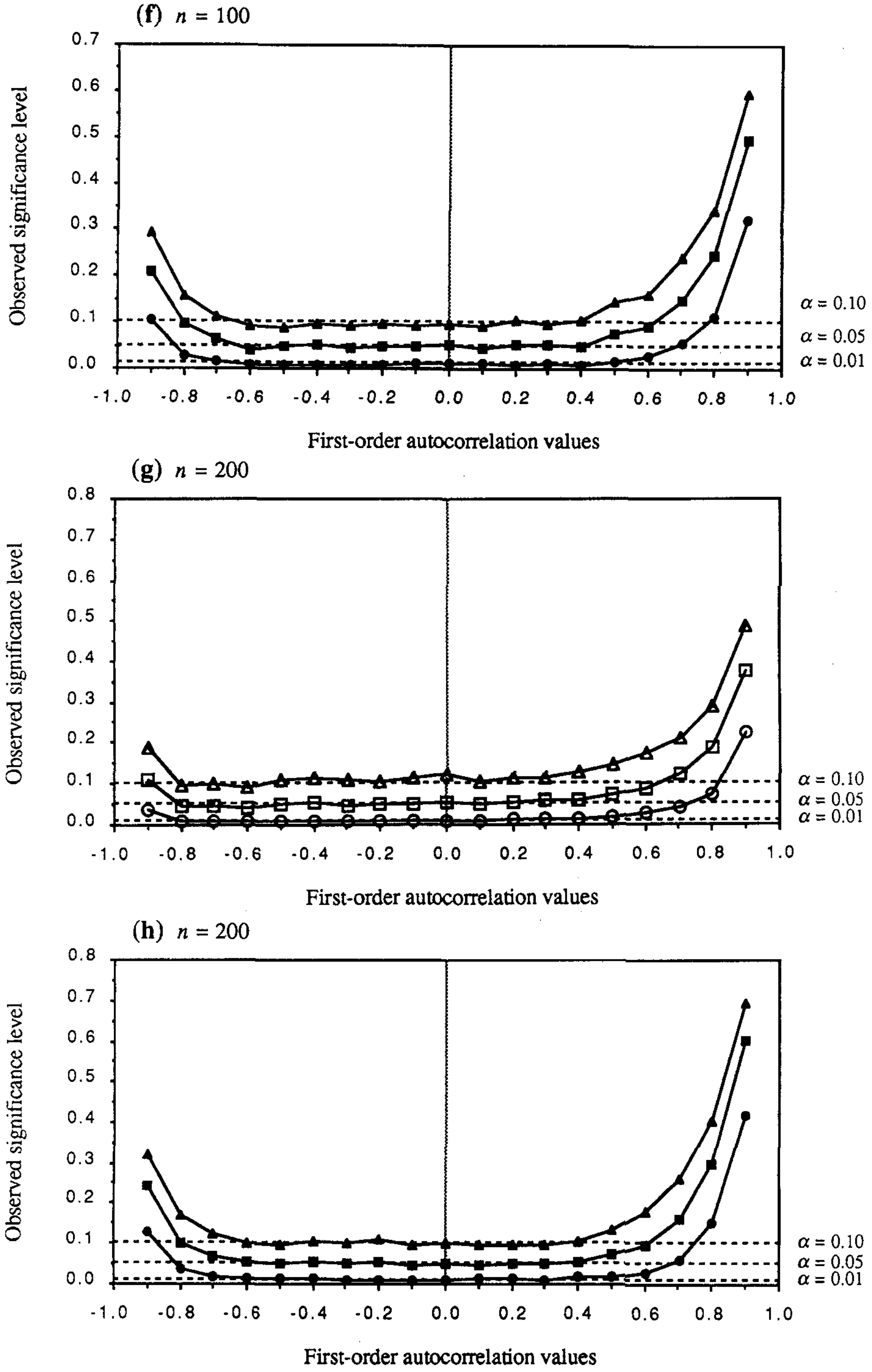


(c) $n = 51$

(d) $n = 51$

(e) $n = 100$

Figure 1 Continued


(f) $n = 100$

(g) $n = 200$

(h) $n = 200$

Figure 1 Continued

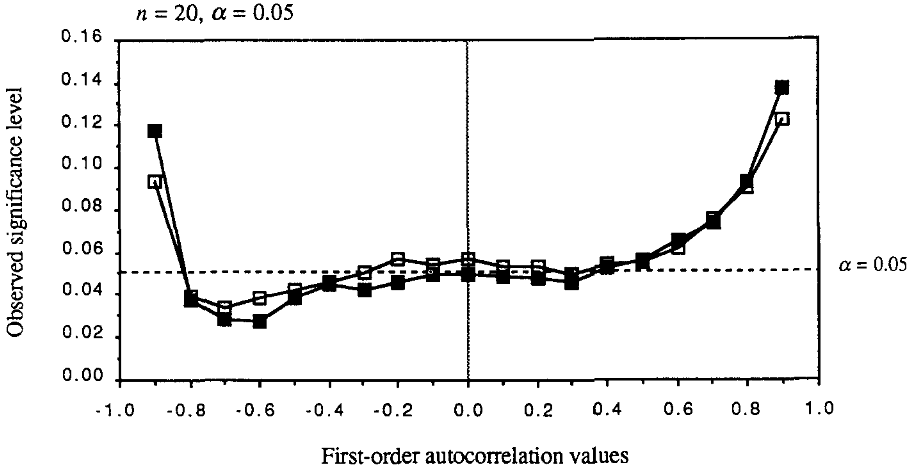


Figure 2 Kolmogorov-Smirnov (open squares) and Shapiro-Wilk (dark squares) tests of normality, performed for simulated time series arising from a Gaussian AR (1) process: observed significance levels as a function of the first-order autocorrelation value. The Kolmogorov-Smirnov tests account for the estimation of two parameters from the sample data, following Stephens (1974).

suffers from lack of power, is equally or more robust than Shapiro-Wilk's, depending on the sign or size of the first-order autocorrelation value. Figure 2, which presents the results for small sample size ($n = 20$) and nominal significance level $\alpha = 0.05$, shows that both statistics behave in the same way when the data display a small or moderate amount of positive autocorrelation, while the Kolmogorov-Smirnov statistic is less biased than Shapiro-Wilk's for negative and high positive first-order autocorrelation values.

Similar results are found with spatially autocorrelated data sampled along a transect. For $n = 20$ and $\alpha = 0.05$ in particular, the D^* and W tests remain too liberal for moderate-to-high positive autoregression coefficient values and too conservative for some range of negative values between -0.1 and -0.45, with a smaller bias for D^* (see Figure 3). The liberal behaviour of both statistics vanishes for high negative values of autoregression.

4. CONCLUDING REMARKS

We have explored the consequences of applying tests of normality designed for the random-sample case, to variables that are autocorrelated in time or in space along a transect. We have clearly shown that such applications may have misleading results: confusion between nonnormality and autocorrelation, conservative or liberal results. A way of overcoming the problem is to state the null hypothesis in terms of the normality of the generating process ε of the autocorrelated variable X under study, instead of the normality of the variable itself, because the former induces the latter. This suggests the following solution to the problem: one could

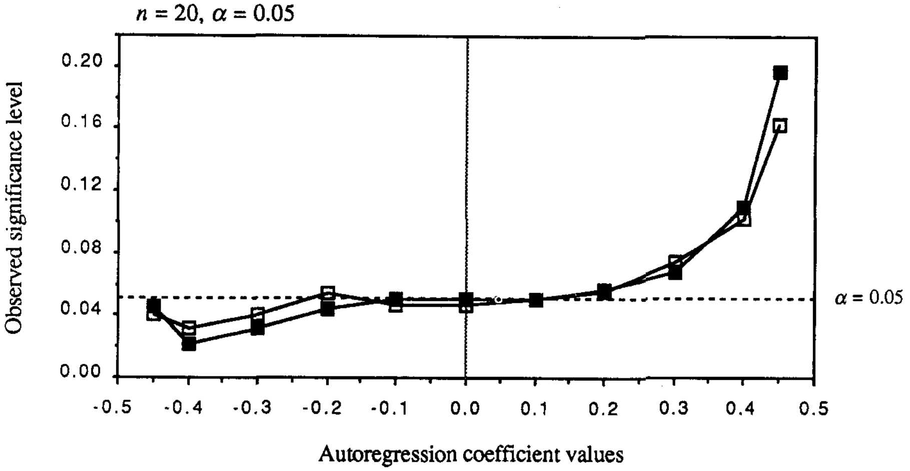


Figure 3. Kolmogorov-Smirnov (open squares) and Shapiro-Wilk (dark squares) tests of normality, performed for simulated one-dimensional spatial data arising from a Gaussian symmetric nearest-neighbour process: observed significance levels as a function of the autoregression coefficient value. The Kolmogorov-Smirnov tests account for the estimation of two parameters from the sample data, following Stephens (1974).

remove the effect of the autocorrelation and go back to the random-sample situation; this can be accomplished by identifying the autocorrelation structure in time or space, followed by a linear transformation based on the square root of the inverse of the estimated covariance matrix. We recently proposed a solution of this type to a specific, but not restricted, ecological problem in time-series analysis (Legendre and Dutilleul, 1991). In spatial analysis, the appropriate theory should be developed (Cliff and Ord, 1981, Ch. 7) in order to provide valid statistical models allowing for spatial autocorrelation; this may represent the long-term solution. For tests of normality in particular, the question of the effect of spatial autocorrelation in two-dimensional processes is too often ignored (as for instance in Griffith, 1987, Ch. 3); the results presented in this paper suggest that such processes should be investigated next.

The present paper was concerned with the comparison, in terms of robustness against autocorrelation, of the Kolmogorov-Smirnov D^* statistic as modified by Stephens (1974) on the one hand, and the Shapiro-Wilk W statistic on the other. From a practical point of view, our results may be summarised as follows: for some range of negative first-order autocorrelation or autoregression values, in small sample sizes, both statistics lead to rejecting the null hypothesis of normality in too few cases when it is true — which is not a major concern since the probability of committing a type I error is not increased; what is of concern, of course, is that the probability of a type II error is increased with nonnormal data. For small and moderate sample sizes, both tests incorrectly reject too often the normality hypothesis when the data are highly positively autocorrelated, and to a lesser extent also when they are highly negatively autocorrelated; this means that high positive or negative correlation among the observations cannot be told apart from lack of normality. The numerical results we presented for small and moderate sample sizes push the balance in favour of D^*, whose properties are as good as or better than those of W, and counterbalance the slightly poorer power reported for D^* in the random-sample case. This may give new impetus to the use of goodness-of-fit statistics based on the empirical cumulative distribution function when autocorrelation is present in the sample data.

Acknowledgements

This work has been supported by NSERC grant No. OGP0007738 to Pierre Legendre. This is contribution No. 381 of the Groupe d'Écologie des Eaux douces, Université de Montréal.

References

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association* **49,** 765–769.

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

Bivand, R. (1980). A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. *Quaestiones Geographica* **6,** 5–10.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and correlation between errors in the two-way classification. *Annals of Mathematical Statistics* **25,** 484–498.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models and applications*. Pion Limited, London.

Clifford, P., Richardson, S. and Hémon, D. (1989). Assessing the significance of the correlation between two spatial processes. *Biometrics* **45,** 123–134.

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

Cramér, H. (1928). On the composition of elementary errors. *Skandinavisk Aktuarietidskrift* **11,** 141–180.

Crowder, M. J. and Hand, D. J. (1990). *Analysis of repeated measures*. Chapman and Hall, London.

D'Agostino, R. B. (1982). Departures from and tests for normality. In: *Encyclopedia of Statistical Sciences* **2,** J. Kotz and N. L. Johnson, eds., Wiley, New York, 315–324.

Diggle, P. J. (1990). *Time series: a biostatistical introduction*. Clarendon Press, Oxford.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression, I. *Biometrika* **37,** 409–428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika* **38,** 159–178.

Gasser, T. (1975). Goodness-of-fit tests for correlated data. *Biometrika* **62,** 563–570.

Gleser, L. J. and Moore, D. S. (1983). The effect of dependence on chi-squared and empiric distribution tests of fit. *Annals of Statistics* **11,** 1100–1108.

Griffith, D. A. (1978). A spatially adjusted ANOVA model. *Geographical Analysis* **10,** 296–301.

Griffith, D. A. (1987). *Spatial autocorrelation: a primer*. Association of American Geographers, Washington.

Huynh, H. and Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Stastical Association* **65,** 1582–1589.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4,** 83–91.

Legendre, P. and Dutilleul, P. (1991). Comments on Boyle's Acidity and organic carbon in lake water: variability and estimation of means. *Journal of Paleolimnology* **6,** 94–101.

Legendre, P. and Vaudor, A. (1991). The R package: multidimensional analysis, spatial analysis. Département de sciences biologiques, Université de Montréal, Montréal.

Legendre, P., Oden, N. L., Sokal, R. R., Vaudor, A. and Kim, J. (1990). Approximate analysis of variance of spatially autocorrelated regional data. *Journal of Classification* **7,** 53–75.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* **62,** 399–402.

Milliken, G. A. and Johnson, D. E. (1984). *Analysis of messy data.* Van Nostrand Reinhold, New York.

Moore, D. S. (1982). The effect of dependence on chi squared tests of fit. *Annals of Statistics* **10,** 1163–1171.

Pearson, E. S., D'Agostino, R. B. and Bowman, K. O. (1977). Tests for departure from normality: comparison of powers. *Biometrika* **64,** 231–246.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II. Skew variation in homogeneous material. *Phil. Trans.* A **186,** 343–414.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* (5) **50,** 157–175.

Pierce, D. A. (1985). Testing normality in autoregressive models. *Biometrika* **72,** 293–297.

Rouanet, H. and Lépine, D. (1970). Comparison between treatment in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology* **23,** 147–163.

Royston, J. P. (1982a). An extension of Shapiro and Wilk's *W* test for normality to large samples. *Applied Statistics* **31,** 115–124.

Royston, J. P. (1982b). Algorithm AS 177. Expected normal order statistics (exact and approximate). *Applied Statistics* **31,** 161–165.

Royston, J. P. (1982c). Algorithm AS 181. The *W* test for normality. *Applied Statistics* **31,** 176–177.

SAS Institute Inc. (1985). *SAS® User's Guide: Basic, Version 5 Edition.* SAS Institute Inc., Cary (N.C.).

SAS Institute Inc. (1990). *SAS® Procedures Guide, Version 6 Edition.* SAS Institute Inc., Cary (N.C.).

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52,** 591–611.

Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests of normality. *Journal of the American Statistical Association* **63,** 1343–1372.

Smirnov, N. V. (1939). On the discrepancy of the empirical distribution curve (in Russian). *Matematicheskii Sbornik* **6,** 3–26.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* **69,** 730–737.

Stephens, M. A. (1982). EDF statistics. In: *Encyclopedia of Statistical Sciences* **2,** J. Kotz and N. L. Johnson, eds., Wiley, New York, 451–455.

van Soest, J. (1967). Some experimental results concerning tests of normality. *Statistica Neerlandica* **21,** 91–97.

Walker, A. M. (1973). On the estimation of a harmonic component in a time series with stationary dependent residuals. *Advances in Applied Probability* **58,** 21–36.