Discussion Papers
Department of Economics
University of Copenhagen

No. 07-25

Correlation, Regression, and Cointegration of Nonstationary Economic Time Series

Søren Johansen

Studiestræde 6, DK-1455 Copenhagen K., Denmark
Tel.: +45 35 32 30 82 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (online)

Correlation, regression, and cointegration of nonstationary economic time series

Søren Johansen*
University of Copenhagen† and CREATES

November 6, 2007

Abstract

Yule (1926) introduced the concept of spurious or nonsense correlation, and showed by simulation that for some nonstationary processes, that the empirical correlations seem not to converge in probability even if the processes were independent. This was later discussed by Granger and Newbold (1974), and Phillips (1986) found the limit distributions.

We propose to distinguish between *empirical* and *population* correlation coefficients and show in a bivariate autoregressive model for nonstationary variables that the empirical correlation and regression coefficients do not converge to the relevant population values, due to the trending nature of the data.

We conclude by giving a simple cointegration analysis of two interests. The analysis illustrates that much more insight can be gained about the dynamic behavior of the nonstationary variables then simply by calculating a correlation coefficient.

JEL Classification: C22

*The author acknowledges the support of the Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation.

†Address: Department of Economics, University of Copenhagen, Studiestræde 6, DK-1455 Copenhagen K, Denmark. Email: sjo@math.ku.dk

1. INTRODUCTION

In his presidential address at the meeting in the Royal Statistical Society November 17, 1925 Udne Yule stated

> "It is fairly familiar knowledge that we sometimes obtain between quantities varying with the time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly "significant"."

He goes on to show a plot of the proportion of Church of England marriages to all marriages for the years 1866-1911 inclusive, and in the same diagram, the mortality per 1.000 persons for the same years, see Figure 1. He comments

> "Evidently there is a very high correlation between the two figures for the same year: The correlation coefficient actually works out at +0.9512."

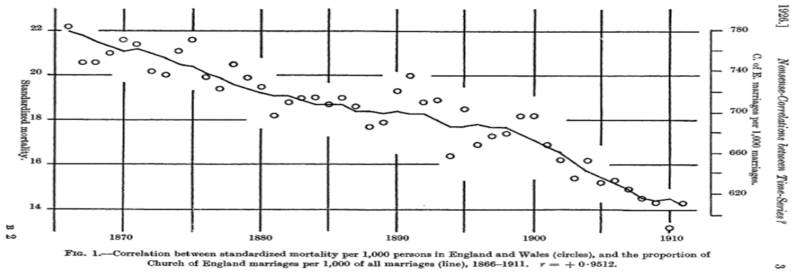

FIG. 1.—Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0{\cdot}9512$.

Figure 1: The proportion of Church of England marriages to all marriages for the years 1866-1911 (line), and the mortality per 1.000 persons for the same years (circles)

He then points out that

> "When we find that a theoretical formula applied to a particular case gives results which common sense judges to be incorrect, it is a generally as well to examine the particular assumptions from which it was deduced and see which of them are inapplicable to the case in point."

In order to describe the probability assumptions behind the "ordinary test" he invents an experiment which consists of writing corresponding numbers of (X_t, Y_t) on cards and defines the distribution of the correlation coefficient as what you get when you draw the cards at random and calculate the correlation coefficient. He then simulated the distribution of the empirical correlation coefficient calculated from

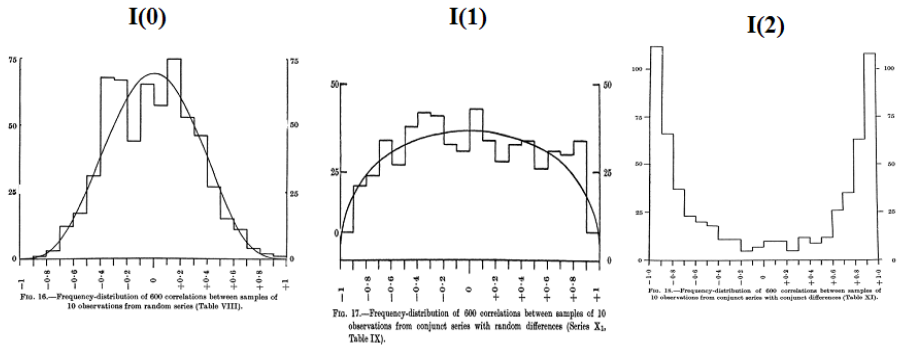

Figure 2: Simulation for $T = 10$ of the distribution of the empirical correlation coefficient for independent i.i.d. processes, $I(0)$, independent random walks, $I(1)$, and independent cumulated random walks, $I(2)$, Yule (1926).

two independent i.i.d. processes, from two independent random walks, and from two independent cumulated random walks, the latter having a U-shaped distribution, see Figure 2.

Thus, 80 years ago Yule pointed to what is wrong with just calculating correlation coefficients without checking the stationarity assumption behind the interpretation, and he suggested calling such correlations "nonsense correlations".

Granger and Newbold (1974) take up the point and note that

> "It is very common to see reported in applied econometric literature, time series regression equations with an apparently high degree of fit, as measured by the coefficient of multiple correlation R^2 but with an extremely low value the Durbin-Watson statistic. We find it very curious that whereas virtually every textbook on econometric methodology contains explicit warnings of the dangers of autocorrelated errors this phenomenon crops up so frequently in well-respected applied work."

They show by simulation of ARIMA models that regressions can be quite misleading. The important paper by Phillips (1986) solved the problem of finding the asymptotic distribution of correlation and regression coefficients, when calculated from a class of nonstationary time series. Thus the problem and its solution has been known for a long time but we still find numerous examples of misunderstandings in applied and theoretical work.

The paper by Hoover (2003) discusses Reichenbach's principle of the common cause, that is, "if event X and Y are correlated, then either X causes Y, Y causes X, or X and Y are joint effects of a common cause (one that renders X and Y conditionally probabilistically independent)", see Sober (2001). A counter example to this principle, according to Sober (2001), consists in considering Venetian sea levels and British bread

prices. Sober claims they are truly correlated but not causally connected by construction, therefore neither causes the other and there can be no common cause. Hoover points out that the statement "truly correlated" is based on calculating the empirical correlation coefficient, which is clearly a case of a spurious or nonsense correlation, as both series trend with time.

Another example is the analysis of the (trending) time series of sea level and temperature of the earth. Rahmstorf (2007) reports a correlation coefficient between the rate of change of sea level and temperature: $R = 0.88$ with a p value 1.6×10^{-8}.

Thus the problem pointed out and analyzed by simulation by Yule in 1926, followed up by Granger and Newbold (1974), and finally solved by Phillips (1986) is still present in applied and theoretical work.

2. DISTINCTION BETWEEN EMPIRICAL AND POPULATION VALUES

It is important in applications to distinguish between the sample, $(X_t, Y_t)_{t=1}^T$, and the population as expressed by the density of $(X_t, Y_t)_{t=1}^T$. We have different words for sample average, $\bar{X} = T^{-1}\sum_t X_t$ and population expectation, $E(X_t)$, and it is well know that for stationary (ergodic) processes it holds that $\bar{X} \xrightarrow{P} E(X_1)$, but for trending variables this need not not hold. For correlation and regression coefficients we use the same words for the sample and population concepts. We therefore suggest to use the terms *empirical* correlation coefficient and *population* correlation coefficient and similarly for regression coefficients in order to distinguish between the calculated values and their interpretation in the population.

The *empirical* correlation coefficient and *empirical* regression coefficient are calculated from a sample

$$R = \frac{\sum_t (X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_t (X_t - \bar{X})^2 \sum_t (Y_t - \bar{Y})^2}}, \ \hat{\beta}_{ols} = \frac{\sum_t (X_t - \bar{X})(Y_t - \bar{Y})}{\sum_t (X_t - \bar{X})^2} \tag{1}$$

and the *population* correlation correlation coefficient and *population* regression coefficient of the pair (Y_t, X_t) are defined as

$$\rho_t = \frac{Cov(X_t, Y_t)}{\sqrt{Var(Y_t)Var(X_t)}}, \ \beta_t = \frac{Cov(X_t, Y_t)}{Var(X_t)}, \tag{2}$$

provided the moments have a meaning. For stationary processes $\rho_t = \rho$ and $\beta_t = \beta$, but for nonstationary processes we may have to condition on initial values.

For ergodic processes with finite variance the law of large numbers shows that

$$R \xrightarrow{P} \rho, \ \hat{\beta}_{ols} \xrightarrow{P} \beta, \ T \to \infty.$$

Obviously such a result need not hold if the processes are nonstationary either with a deterministic trend or a stochastic trend. When using the qualifications *empirical* and *population* for the concepts one has to argue that one can estimate the *population* values

(ρ_t, β_t) by the *empirical* values $(R, \hat{\beta}_{ols})$, and that requires knowledge of the properties of the processes. In order to find the properties of the processes, however, we need a model for the data, which describes the variation of the data in a satisfactory way. Only then can we find out if the assumptions for interpreting an empirical correlation coefficient as a population correlation coefficient are satisfied.

3. INTERPRETATION OF CORRELATION AND REGRESSION IN NONSTATIONARY TIME SERIES

In order to understand the relation between empirical and population values for nonstationary processes, we assume that data is generated by an autoregressive model written in terms of differences and lagged levels

$$\begin{aligned} \Delta Y_t &= \pi_{yy} Y_{t-1} + \pi_{yx} X_{t-1} + \mu_y + \delta_y t + \varepsilon_{yt}, \\ \Delta X_t &= \pi_{xy} Y_{t-1} + \pi_{xx} X_{t-1} + \mu_x + \delta_x t + \varepsilon_{xt}, \end{aligned}$$

where $\varepsilon_t = (\varepsilon_{yt}, \varepsilon_{xt})'$ is i.i.d. $(0, \Omega)$. This is a dynamic stochastic model for the evolution of the processes Y_t and X_t, expressed as a model for how the changes depend on lagged levels, deterministic terms, and noise terms. The stochastic properties of the solution depend on the coefficients in the model in terms of the roots of the characteristic polynomial:

$$\det \Psi(z) = \det \begin{pmatrix} 1 - z - \pi_{yy} z & -\pi_{yx} z \\ -\pi_{xy} z & 1 - z - \pi_{xx} z \end{pmatrix} = 0.$$

We assume that the roots of $\det \Psi(z) = 0$ satisfy that either $|z| > 1$ or $z = 1$, thereby ruling out explosive roots and seasonal roots on the unit circle. Because $\Pi = -\Psi(1)$ we distinguish three cases:

3.1. Rank(Π) = 2.

Because there are no unit roots, the processes are trend stationary with a linear trend

$$Y_t = U_{yt} + \gamma_y + \xi_y t, \;\; X_t = U_{xt} + \gamma_x + \xi_x t,$$

where (U_{yt}, U_{xt}) is stationary with mean zero and variance $\Sigma = Var(Y_t, X_t)$. The population correlation and regression coefficients are

$$\rho = \frac{\Sigma_{yx}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}}, \;\; \beta = \frac{\Sigma_{yx}}{\Sigma_{xx}}. \tag{3}$$

The slopes are given by

$$\xi_y = \frac{\pi_{yx}\delta_x - \pi_{xx}\delta_y}{\pi_{yy}\pi_{xx} - \pi_{yx}\pi_{xy}}, \;\; \xi_x = \frac{\pi_{xy}\delta_y - \pi_{yy}\delta_x}{\pi_{yy}\pi_{xx} - \pi_{yx}\pi_{xy}}.$$

In this case, the average of X_t, say, $\bar{X} = \bar{U}_x + \gamma_x + \xi_x \bar{t}$ diverges and does not correspond to the population values $E(X_t) = \gamma_x + \xi_x t$, when $\xi_x \neq 0$, because of the deterministic linear trend. The dominating term of the product moment

$$\sum_t (X_t - \bar{X})(Y_t - \bar{Y}) = \sum_t (U_{xt} - \bar{U}_x + \xi_x(t - \bar{t}))(U_{yt} - \bar{U}_y + \xi_y(t - \bar{t}))$$

is $\xi_x\xi_y \sum_t (t-\bar{t})^2$, so that the empirical correlation and regression coefficients converge:

$$R \xrightarrow{P} \frac{\xi_x\xi_y}{|\xi_x\xi_y|} = \pm 1,\ \hat{\beta}_{ols} \xrightarrow{P} \frac{\xi_y}{\xi_x},\ T \to \infty. \tag{4}$$

These values are not the population correlation and regression coefficients (ρ, β), see (3). Thus, when calculating empirical correlation coefficients from trend stationary data, we cannot interpret them as approximations to the true population quantities, and the empirical results are entirely spurious, in the sense that the conclusions drawn from these cannot be considered conclusions about the correlation in the population.

3.2. Rank($\mathbf{\Pi}$) = 0.

In this case $\Pi = 0$, and the equations are

$$\Delta Y_t = \mu_y + \delta_y t + \varepsilon_{yt}, \quad \Delta X_t = \mu_x + \delta_x t + \varepsilon_{xt}.$$

These determine two random walks with quadratic trends. We assume for simplicity that $\delta_x = \delta_y = 0$, so that the trends are linear, and find

$$Y_t = Y_0 + \sum_{i=1}^{t} \varepsilon_{yt} + \mu_y t = Y_0 + S_{yt} + \mu_y t, \quad X_t = X_0 + \sum_{i=1}^{t} \varepsilon_{xt} + \mu_x t = X_0 + S_{xt} + \mu_x t.$$

The population values for means, correlation, and regression coefficients have to be defined by conditioning on (Y_0, X_0):

$$E(Y_t|Y_0, X_0) = Y_0 + \mu_y t, \quad E(X_t|Y_0, X_0) = X_0 + \mu_x t,$$

$$\rho = \frac{Cov(Y_t, X_t|Y_0, X_0)}{\sqrt{Var(Y_t|Y_0, X_0)Var(X_t|Y_0, X_0)}} = \frac{\Omega_{yx}}{\sqrt{\Omega_{yy}\Omega_{xx}}},\ \beta = \frac{Cov(Y_t, X_t|Y_0, X_0)}{Var(X_t|Y_0, X_0)} = \frac{\Omega_{yx}}{\Omega_{xx}}.$$

If $\mu_x\mu_y \neq 0$, the averages

$$\bar{Y} = Y_0 + \bar{S}_y + \mu_y \bar{t}, \quad \bar{X} = X_0 + \bar{S}_x + \mu_x \bar{t}$$

diverge and do estimate $E(Y_t)$ and $E(X_t)$. Because the linear trend dominates the processes, we get the result (4) for the empirical correlation and regression coefficient with (ξ_y, ξ_x) replaced by (μ_y, μ_x).

If, however, $\mu_x = \mu_y = 0$ the linear trends vanish and the random walks dominate the behavior of the processes. The limit of the random walk is the Brownian motion:

$$T^{-1/2} \sum_{i=1}^{[Tu]} \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} B_y(u) \\ B_x(u) \end{pmatrix},$$

so that the averages, normalized by $T^{-1/2}$, give the limits

$$T^{-1/2}(\bar{Y}, \bar{X}) \xrightarrow{d} (\bar{B}_y, \bar{B}_x) = (\int_0^1 B_y(u)du, \int_0^1 B_x(u)du.$$

The limits of R and $\hat{\beta}_{ols}$ are

$$R \xrightarrow{d} \frac{\int_0^1 (B_y(u) - \bar{B}_y)(B_x(u) - \bar{B}_x)du}{\sqrt{\int_0^1 (B_y(u) - \bar{B}_y)^2 du \int_0^1 (B_x(u) - \bar{B}_x)^2 du}}, \; \hat{\beta}_{ols} \xrightarrow{d} \frac{\int_0^1 (B_y(u) - \bar{B}_y)(B_x(u) - \bar{B}_x)du}{\int_0^1 (B_y(u) - \bar{B}_y)^2 du}.$$

These distributions were derived by Phillips (1986); see also the simulations of the distribution of R by Yule (1926) in Figure 2.

Notice that for random walks without deterministic trend, the empirical correlation does not converge in probability to anything. In fact the empirical correlation between two random walks can give any number between -1 and $+1$, even if the random walks are completely independent of each other and T is arbitrarily large. In this case an empirical correlation is completely spurious and has no interpretation as a population parameter.

3.3. Rank($\mathbf{\Pi}$) = 1.

When the matrix Π has rank 1, it has the representation

$$\Pi = \begin{pmatrix} \alpha_y \\ \alpha_x \end{pmatrix} \begin{pmatrix} \beta_y \\ \beta_x \end{pmatrix}' = \begin{pmatrix} \alpha_y\beta_y & \alpha_y\beta_x \\ \alpha_x\beta_y & \alpha_x\beta_x \end{pmatrix}$$

and the equations become (with $\delta_x = \delta_y = 0$ as in Case 2)

$$\begin{aligned} \Delta Y_t &= \alpha_y(\beta_y Y_{t-1} + \beta_x X_{t-1}) + \mu_y + \varepsilon_{yt}, \\ \Delta X_t &= \alpha_x(\beta_y Y_{t-1} + \beta_x X_{t-1}) + \mu_x + \varepsilon_{xt}. \end{aligned}$$

It is seen that the same linear combination $\beta_y Y_{t-1} + \beta_x X_{t-1}$ of the lagged levels enter both equations and that α_y and α_x describe the dynamic adjustment of the variables to deviations from the relation $\beta_y Y_{t-1} + \beta_x X_{t-1} = 0$. The model thus describes a feedback mechanism and is called the error correction or equilibrium correction model. Under the condition that $\alpha_y\beta_y + \alpha_x\beta_x \neq 0$, one can solve these equations and find the representation

$$Y_t = \beta_x S_t + Z_{yt} + A_y + \gamma_y, \quad X_t = -\beta_y S_t + Z_{xt} + A_x + \gamma_x,$$

where S_t is a random walk with a trend:

$$S_t = \sum_{i=1}^{t} \eta_i + \tau t, \; \eta_t = \frac{\alpha_x\varepsilon_{yt} - \alpha_y\varepsilon_{xt}}{\alpha_y\beta_y + \alpha_x\beta_x}, \; \tau = \frac{\alpha_x\mu_y - \alpha_y\mu_x}{\alpha_y\beta_y + \alpha_x\beta_x}. \tag{5}$$

The processes Z_{yt} and Z_{xt} are stationary and A_y and A_x depend on initial values and satisfy $\beta_y A_y + \beta_x A_x = 0$, see Johansen (1996, Theorem 4.2).

Note that the process (Y_t, X_t) is nonstationary due to the common trend S_t, but that the linear combination

$$\beta_y Y_t + \beta_x X_t = \beta_y Z_{yt} + \beta_x Z_{xt}$$

is a stationary process because the common trend and initial values are eliminated.

We say that Y_t and X_t are integrated of order one, $I(1)$, because they are nonstationary and ΔY_t and ΔX_t are stationary, and they are cointegrated because a linear combination is stationary, Granger (1981).

The dominating term in the product moment

$$\sum_t (X_t - \bar{X})(Y_t - \bar{Y}) = \sum_t (\beta_x(S_t - \bar{S}) + Z_{xt} - \bar{Z}_x)(\beta_y(S_t - \bar{S}) + Z_{yt} - \bar{Z}_y)$$

is $\beta_x\beta_y \sum_t (S_t - \bar{S})^2$. This implies that for *population* correlation and regression coefficients, the limits are

$$\rho_t \to -\frac{\beta_y\beta_x}{|\beta_y\beta_x|} = \pm 1,\ \beta_t \to -\frac{\beta_x}{\beta_y},\ t \to \infty,$$

and similarly we find the result (4) for the *empirical* correlation and regression coefficients, with (ξ_y, ξ_x) replaced by (β_y, β_x).

Hence in this cointegrated case, the empirical correlation and regression coefficients give consistent estimates of the limits of the population values. Thus correlation coefficients are not spurious even though a proper analysis of the data reveals more structure than can be captured by the asymptotic population correlation coefficient.

4. AN EXAMPLE OF A COINTEGRATION ANALYSIS

Consider quarterly data for the interest rates i_t^{aus} and i_t^{us} in Australia and United States in the period 1972:01 to 1991:01; Johansen (1996). The processes are fitted with a bivariate autoregressive model with three lags. By a statistical analysis it is found that the rank of Π can be taken to one, see Case 3 above, so that there is a unit root and i_t^{aus} and i_t^{us} can be considered nonstationary and cointegrated $I(1)$ variables. The cointegrating relation is found to be $i_{t-1}^{aus} - i_{t-1}^{us}$ which is stationary around the value 0.03, so that the estimated model is

$$\begin{aligned} \Delta i_t^{aus} &= \underset{[t=-4.33]}{-0.17}\,(i_{t-1}^{aus} - i_{t-1}^{us} \underset{[t=-6.30]}{-0.03}) + \ldots + \varepsilon_t^{aus}, \\ \Delta i_t^{us} &= \underset{[t=-0.58]}{-0.03}\,(i_{t-1}^{aus} - i_{t-1}^{us} \underset{[t=-6.30]}{-0.03}) + \ldots + \varepsilon_t^{us}. \end{aligned} \tag{6}$$

Because the adjustment coefficient of the equation for Δi_t^{us} is not significantly different from zero, $t = -0.58$, see (6), only the Australian interest rates adjust to a disequilibrium between the interest rates. We say that i_t^{us} is weakly exogenous; see Engle, Hendry, and Richard (1983). This implies that in (5) we take $\alpha_x = 0$, $\beta_x = -\beta_y = -1$, so that $\eta_t = -\varepsilon_t^{us}$. The common stochastic trend is therefore the cumulated disturbances $\sum_{i=1}^t \varepsilon_i^{us}$. In this sense the US interest rate is driving the interest rate in Australia.

The empirical correlation coefficient between the series i_t^{aus} and i_t^{us} is $R = 0.60$, and by the analysis above in Case 2, this is a consistent estimator of the limiting value of 1, which is hardly a useful result.

5. CONCLUSION

It is argued that one should distinguish between the *empirical* and the *population* correlation and regression coefficients. An empirical correlation coefficient is calculated from a sample, whereas a population correlation coefficient requires a population which we can estimate from a model.

It is shown in a bivariate autoregressive time series model for nonstationary processes, that the presence of a nonstationary deterministic or stochastic trend implies that the empirical correlation and regression coefficients cannot be interpreted as their population counterparts in any useful way. Hence they are entirely spurious, in the sense that the conclusions drawn from these cannot be considered conclusions about the correlation or regression in the population.

Calculating an empirical correlation coefficient between the changes of sea level and temperature gives 0.88, but it is not obvious how that can be interpreted in view of the trending nature of the data, nor what is the corresponding distribution. Similarly the correlation coefficient of 0.60 between i_t^{us} and i_t^{aus} is not a useful coefficient to calculate in view of the detailed statistical analysis of nonstationary processes that is now available.

The solution to the spurious correlation problem in practice is to model the data, and only when one is convinced that the model gives a good description of the data one can calculate the population counterpart of the empirical correlations and thereby avoid spurious or nonsense correlations.

6. REFERENCES

Engle R.F., Hendry D.F., and Richard J.-F. 1983. Exogeneity. *Econometrica* 51, 277–304.

Granger, C.W.J. 1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–130.

Granger, C.W.J., and Newbold, P. 1974. Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–120.

Hoover, K. 2003. Nonstationary time series, cointegration, and the principle of the common cause. *British Journal for the Philosophy of Science* 54, 527–551.

Johansen, S. 1996. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press, Oxford.

Phillips, P.C.B. 1986. Understanding spurious regression in econometrics. *Journal of Econometrics* 33, 311–340.

Rahmstorf, S. 2007. A semi-empirical approach to projecting future sea-level rise. *Science* 315, 368–370.

Sober, E. 2001. Venetian sea levels, British bread prices and the principle of the common cause. *British Journal for the Philosophy of Sciences* 52, 331–346.

Yule, U. 1926. Why do we sometimes get nonsense-correlations between time series? – A study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89, 1–63.