Supplementary online appendix to accompany the article "Characterizing bird migration phenology using data from standardized monitoring at bird observatories"

Endre Knudsen[1,*], Andreas Lindén[2], Torbjørn Ergon[1], Niclas Jonzén[3], Jon Olav Vik[1], Jonas Knape[3], Jan Erik Røer[4], and Nils Chr. Stenseth[1]

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, P.O.Box 1066 Blindern, N-0316 Oslo, Norway
[2]Department of Biological and Environmental Sciences, Integrative Ecology Unit, P.O.Box 65 (Viikinkaari 1), FIN-00014, Helsinki University, Finland
[3]Department of Theoretical Ecology, Ecology Building, Lund University, SE-22362 Lund, Sweden
[4]Lista Bird Observatory, N-4563 Borhaug, Norway

Only a limited amount of details could be given within the page limit of our printed review article. We here expand our presentation with a more thorough account of some methods less known and less covered. The following sections will present smoothing methods, options for fitting parametric seasonal distribution curves, quantile regression, and non-parametric linear trend analysis. Table 1 summarizes some basic mathematical notation followed.

Smoothing methods

Smoothing methods use observations of a response variable variable Y and predictor variable(s) $X_1 \dots X_p$ to estimate the trend in the data. This estimate, called the *smooth*, is less variable (in the sense "less wiggly") than Y, hence its name. Usually, no rigid form is assumed for the relationship between predictor(s) and response, so the smooth is of a non-parametric nature. We will only consider the case of a single predictor variable (also called *scatterplot smoothing*). The challenge is to determine the function $s = \mathcal{S}(\mathbf{y}|\mathbf{x})$ of bird migration phenology, using observations of the number of birds y_i observed at day x_i. The domain of s is the same as the domain of $\mathbf{x}$, and s may be defined for all x within the domain (splines), or only for the sample values of X, $x_1, x_2, \dots, x_n$ (all other smoothing methods treated in our main text). Various methods for determining the smooth are described below. Our exposition builds on those

*E-mail: endre.knudsen@bio.uio.no

Notation	Explanation
x_i (y_i)	observation of predictor (response) variable at time i
$\mathbf{x}$	vector of observations $x_i,\ i = 1, 2, \ldots, n$
X	predictor variable
$\mathbf{X}$	vector of predictor variables
$\|\mathbf{x}\|^2$	squared norm; $\|\mathbf{x}\|^2 = \sum_{i=0}^{n-1} \|x_i\|^2$
$\{i : i > 0\}$	set notation; set of i satisfying the condition $i > 0$
$f(x; \mu, \sigma)$	function of x, with parameters μ and σ
μ, σ $(\hat{\mu}, \hat{\sigma})$	population (sample) mean and standard deviation
ave(x)	the average function
sgn(x)	the sign function; sgn$(x)=\begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$

Table 1: Some mathematical notation followed in this appendix

of Silverman (1986), Hastie and Tibshirani (1990), Percival and Walden (2000), Venables and Ripley (2002), and Wood (2006).

Bin smoother

A bin smoother partitions the range of predictor values into a set of disjoint intervals, usually of equal length, and calculates a summary statistic (here, the average) for each of these intervals. Cutpoints $c_0 = -\infty < \cdots < c_K = \infty$ are chosen to define the K subsets of data points within each interval:

$$I_k = \{i : c_k \leq x_i \leq c_{k+1}\}, \qquad k = 0, \ldots, K-1 \tag{1}$$

The smooth s is then given by

$$s = \text{ave}_{i \in I_k}(y_i), \qquad \text{for all } x \in I_k. \tag{2}$$

The discontinuities at the cutpoints result in a jagged smooth and a high sensitivity to the location of cutpoints. Unless the cutpoints define a particularly meaningful partitioning reflected in the data, other smoothers should be preferred. The bin smoother is, however, widely used for summing up migration numbers over a predefined period of time, *e.g.*, three or five days.

Moving windows

A moving windows smoother calculates the summary statistic over a local neighborhood ("window") around the point of interest. In the present case, the data are a regularly spaced time series, and it is useful to choose a symmetric neighborhood; *i.e.*, the point-of-interest x_i and the k nearest points to the left and right. This is done for all n points of the data series; hence the 'moving' part of the name. Formally, we can define a symmetric neighborhood of width $2k + 1$ as

$$N_S(x_i) = \{\max(i-k, 1), \ldots, i-1,\ i\ , i+1, \ldots, \min(i+k, n)\}, \tag{3}$$

and the moving average as

$$s(x_i) = \text{ave}_{j \in N_S(x_i)}(y_j). \tag{4}$$

Missing data points can be allowed simply by omitting them from calculation of the summary statistic. The smoother can, however, be severely biased near the ends of the data series, where trends are poorly picked up.

Local weighted regression

The problem of bias near the ends of the data series can be alleviated by fitting a trend within the local moving window, rather than just a constant. This is the essence of local weighted regression, also called LOESS smoothing. For instance, we may fit a straight line:

$$s(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0, \tag{5}$$

where $\hat{\alpha}(x_0)$ and $\hat{\beta}(x_0)$ are the weighted least-squares estimates for the data points in the local neighborhood $N_S(x_0)$. Weights are traditionally calculated using the tricube weight function:

$$W\left(\frac{|x_0 - x_i|}{\max_{N_S(x_0)}|x_0 - x_i|}\right) = W(u) = \begin{cases} (1-u^3)^3, & 0 \le u < 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Kernel smoothing

Mathematically, kernel smoothing of a time series is a scaled convolution between the data and a *kernel* function $K(x)$ defining local weights and satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$. This can be conceptualized as sliding the kernel function along the series, multiplying its value with the value of the data for all non-zero data points, and then summing over these products. Hence, for each data point, the kernel smooth is a weighted average of this and the surrounding data points.

If the kernel function is non-negative for all x, the kernel smooth will also be so, which is suitable for our purpose and required for estimating a probability density. Popular choices for the kernel function include the Gaussian probability density function (equation (24)), the Epanechnikov,

$$K(t) = \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2), \quad |t| < \sqrt{5}, \tag{7}$$

the biweight,

$$K(t) = \frac{15}{16}(1 - t^2)^2, \quad |t| < 1, \tag{8}$$

cosine,

$$K(t) = \frac{\pi}{4}\cos\frac{\pi}{2}t, \quad |t| < 1, \tag{9}$$

rectangular (uniform),

$$K(t) = \frac{1}{2}, \quad |t| < 1, \tag{10}$$

and triangular,

$$K(t) = 1 - |t|, \quad |t| < 1 \tag{11}$$

kernel functions. (For all above kernels, $K(t) > 0$ for t as stated, and $K(t) = 0$ otherwise.)

Smoother kernel functions yield somewhat smoother kernel estimates; in particular, the discontinuities at the ends of the rectangular and triangular kernels result in somewhat jagged smooths.

Wavelet analysis

Wavelet analysis (Daubechies 1992, Torrence and Compo 1998, Percival and Walden 2000) can be regarded as kernel smoothing with kernel functions, *wavelets*, satisfying certain mathematical conditions that make their use particularly attractive. There are two main variants: the *continuous* wavelet transform (CWT), designed for continuous-time time series (or similar data) defined over the real axis, and the *discrete* wavelet transform (DWT), designed for discrete-time time series defined over a range of integers. While the former is the variant most frequently used in ecological studies, it yields redundant information when dealing with discrete-time time series.

A wavelet (literally, "small wave") is in the continuous case some function $\psi(\cdot)$ defined over the real axis and satisfying some basic conditions, first and foremost those of $\psi(\cdot)$ integrating to zero,

$$\int_{-\infty}^{\infty} \psi(t)dt = 0, \tag{12}$$

and $\psi^2(\cdot)$ integrating to unity,

$$\int_{-\infty}^{\infty} \psi^2(t)dt = 1. \tag{13}$$

The 'Mexican hat' wavelet has been one of the most popular choices for ecological applications:

$$\psi(t) = \frac{1}{\sqrt{2\pi}\sigma^3}\left(1 - \frac{t^2}{\sigma^2}\right)e^{-t^2/2\sigma^2}, \tag{14}$$

which is the normalized second derivative of a Gaussian function, and is named after its characteristic shape.

Wavelets can be scaled by its width λ ("scale" in wavelet terminology) and shifted t units in time:

$$\psi_{\lambda,t}(u) = \frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-t}{\lambda}\right). \tag{15}$$

The continuous wavelet transform of a signal $x(t)$ is the collection of time- and scale-dependent variables ("wavelet coefficients")
$\{(\lambda, t) : \lambda > 0\ ,\ -\infty < t < \infty\}$ calculated by convolving the signal with appropriately scaled and shifted wavelets:

$$\mathcal{W}(\lambda, t) = \int_{-\infty}^{\infty} \psi_{\lambda,t}(u)x(u)du. \tag{16}$$

Essentially being weighted local averages, wavelet coefficients reflect local features of the signal as observed at a particular observation scale. By varying λ, the wavelet transform is able to pick up features of the signal at a range of observation scales, while preserving the temporal information of the signal. This is a particularly attractive feature of the wavelet transform as compared to other forms of multi-scale analysis such as Fourier transform and autocorrelation analysis, leading to its popularity as an exploratory tool.

The continuous wavelet transform is however redundant in the sense that wavelet coefficients are not independent; they vary only slightly across scales λ and across time at larger scales. In contrast, the discrete wavelet transform is an orthonormal transform where wavelet coefficients are independent across scales and time. The DWT transforms the data into a mathematically equivalent representation. As for the CWT, there is an inverse transform allowing perfect reconstruction of the data. Drawbacks for practical use with data such as those describing the daily number of migrating birds, are that the data series needs to be of length n a power of two, wavelet coefficients are only calculated at scales $2^k, k = 0, 1, \ldots, K$, the number of wavelet coefficients at scale 2^k is only $n/2^k$, and the wavelet transform becomes sensitive to the choice of endpoints for the series.

These drawbacks are avoided by using a variant of the DWT, the MODWT (maximum overlap discrete wavelet transform; Percival and Walden 2000), also known as shift-invariant or non-decimated DWT. This yields an additive decomposition of the series $\mathbf{x}$ into a sum of 'details' $\mathbf{d}_j$ at scales $j = 1, 2, \ldots, J$ and a 'smooth' $\mathbf{s}_J$ at the largest scale J:

$$\mathbf{x} = \sum_{j=1}^{J} \mathbf{d}_j + \mathbf{s}_J, \tag{17}$$

where $\mathbf{d}_j$ and $\mathbf{s}_j$ are of length n equal to the time series, and are functions of, respectively, the wavelet coefficients $\mathcal{W}_j$ at scale j, and the so-called scaling coefficients $\mathcal{V}_J$ at the largest scale J. The latter are obtained by filtering (applying a circular convolution) the time series with a 'father wavelet', and the primer by filtering the series with an appropriately scaled 'mother wavelet'. This is a sequence of real numbers $\{h_l\}$, $l = 0, 1, \ldots, L-1$. The 'father wavelet' $\{g_l\}$ (also called scaling filter) is related to the 'mother wavelet' (also called wavelet filter) as follows:

$$g_l = (-1)^{l+1} h_{L-1-l}. \tag{18}$$

In our example analysis, we used the LA(8) wavelet, *i.e.*, the 'least asymmetric' wavelet of nominal length 8. As the length L increases, the mother wavelet approaches a shape that is roughly Gaussian, while the father wavelet approaches a shape resembling the 'Mexican hat' wavelet.

As for continuous wavelets, discrete wavelets need to satisfy $\sum_{l=0}^{L-1} h_l = 0$ (coefficients summing to zero) and $\sum_{l=0}^{L-1} h_l^2 = 1$ (squared coefficients summing to one). Also, the wavelet needs to be orthogonal to its even shifts:

$$\sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{l=-\infty}^{\infty} h_l h_{l+2n} = 0. \tag{19}$$

While the MODWT is not an orthonormal transform, it still has the important properties of energy preservation,

$$\|\mathbf{x}\|^2 = \sum_{j=1}^{J} \|\mathcal{W}_j\|^2 + \|\mathcal{V}_J\|^2, \tag{20}$$

i.e., the variance of the series (here, across time) is decomposed into a set of scale components. Hence, we can calculate the *wavelet empirical power spectrum*, which shows how variance is dispersed across scales rather than across time:

$$\hat{\sigma}_x = \sum_{j=1}^{J} P_W(\tau_j), \tag{21}$$

where $\tau_j = 2^{j-1}$ and

$$P_W(\tau_j) = \frac{1}{n}\|\mathcal{W}_j\|^2, \qquad j = 1, \ldots, J. \tag{22}$$

Due to circular convolution, boundary conditions will bias the wavelet and scaling coefficients close to the ends of the series, but an unbiased power spectrum can be calculated from the coefficients not affected by the boundaries.

Splines

A spline $\hat{g}(x)$ is a piecewise polynomial approximation of a smooth function. A smooth function is here understood as one that has derivatives of all finite orders, while a spline will typically consist of cubic polynomials joined together at so-called *knots* $\{x_i^* : i = 1, \ldots, k\}$ in a way that makes the entire spline continuous up to and including the second derivative. These *cubic splines* have several desirable attributes; in particular, they can in many cases be shown to be optimal or near-optimal interpolators. While *interpolation splines* assume that $g(x_i^*) = y_i$, *smoothing splines* treat the y_i as random variables and estimate the $g(x_i^*)$ in order to minimize a weighted sum of the squared approximation errors and some measure of the roughness. For cubic splines, the expression to be minimized is

$$\sum_{i=1}^{n} [y_i - g(x_i)]^2 + \lambda \int g''(x)^2 dx, \tag{23}$$

where the parameter λ determines the balance between producing a good fit and a smooth function. When used in a (multiple) linear regression, this is referred to as fitting a penalized *regression spline*.

An appropriate choice of the smoothing parameter λ can be determined by (generalized) cross-validation, and the model fitted using penalized least squares / penalized likelihood (Wood 2006) or the method of backfitting (Hastie and Tibshirani 1990). This is however not always successful, frequently due to overfitting. Using splines also involves a choice of basis functions, which determines the parameterization. The many possible parameterizations of cubic splines yields what is known as B-splines, P-splines, thin-plate splines etc. There is also a choice of the number and locations of knots, which is critical for interpolation splines, but usually less critical for smoothing splines, due to the influence of the smoothing parameter (provided the number of knots allows the flexibility needed).

Parametric seasonal models

Seasonal distribution curves

Suitable candidates for the seasonal curve to be fitted can be found among the density functions of continuous statistical distributions (Evans et al. 2000, Walpole et al. 2002). A well-known alternative would be the normal distribution,

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2}, \qquad -\infty < x < \infty. \tag{24}$$

The skew-normal distribution (Azzalini 1985) is most easily expressed for the special case

$$f(x;\alpha) = 2\phi(x)\Phi(\alpha x), \qquad -\infty < x < \infty, \tag{25}$$

where $\phi(x)$ is the standard normal density function, $\Phi(x)$ its cumulative density function, and α is the shape parameter determining the skew (no skew when $\alpha = 0$). For practical applications, we also need a location parameter ξ and a scale parameter ω, so we fit $g(x) = \xi + \omega f(x)$, where $f(x)$ is defined in equation (25).

The beta distribution can be expressed as

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1, \tag{26}$$

where α and β are real and positive, and Γ is the gamma function, which for a real argument can be defined as $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt = (x-1)\Gamma(x-1)$.

Both the skew-normal and the beta distributions are able to model asymmetric distribution. The beta distribution is particularly flexible. In Bayesian statistics, it is used as a conjugate prior for the Bernoulli, binomial and geometric distributions, and its multivariate generalization, the Dirichlet distribution, for the multinomial distribution.

Malo (2002) proposed a flexible distribution to model unimodal phenological events:

$$f(x;a,c,d,e) = a\{\sin[\pi \left(\frac{x}{c}\right)^d]\}^e, \tag{27}$$

where a is the maximum intensity, c is the duration of the event in the population, d determines the asymmetry (negative skew if $d < 1$ and positive skew if $d > 1$), and e determines the tails (no tails if $d = 1$ and $e \leq 1$, *i.e.*, an abrupt start and end). An advantage of using such non-standard distributions, is that the parameters may be more directly interpreted in terms of relevant features of the phenology.

Model fitting

The problem of fitting a non-linear curve to the data can be stated as the general regression model:

$$Y = E(Y) + \epsilon = f(\mathbf{X};\mathbf{\Theta}) + \epsilon, \tag{28}$$

where $\mathbf{X}$ is the vector of predictor variables, and $\mathbf{\Theta}$ is the vector of unknown parameters (Neter et al. 1996, Venables and Ripley 2002). As in linear regression, we seek the parameter values that maximize the fit with the data. However, there is generally no closed-form mathematical expressions for determining these parameter values, so they need to be determined by numeric optimization algorithms. Hence, parameter estimates are not guaranteed to be truly optimal, since optimization algorithms may find a local maximum / minimum rather than a global. There are three main approaches to model fitting: least squares (LS), maximum likelihood (ML), and Bayesian estimation (which will not be treated here).

Least squares fitting (Neter et al. 1996, Walpole et al. 2002) seeks to minimize the overall difference between the observations and the model. As for linear regression, this can be done by minimizing the ordinary least squares (OLS) criterion,

$$Q = \sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{29}$$

assuming random and uncorrelated error terms of zero expectation and equal variance. Unlike linear regression with normally distributed error terms, the parameter estimates are not normally distributed, are not unbiased, and do not have minimum variance. Exact inference is not possible, but inference can be made on basis of large-sample approximations or bootstrapping.

The assumption of equal variance is, however, not very realistic when analyzing data on the daily number of migrating birds. Hence, a better approach may be to minimize the weighted least squares (WLS) criterion (Neter et al. 1996):

$$Q_w = \sum_{i=1}^{n} w_i (y_i - f(x_i))^2, \tag{30}$$

where the weights w_i determine how much each observation influences the sum-of-squares and, hence, the parameter estimates. Intuitively, data points with large error variance should be assigned small weights, *e.g.*, by defining weights as the reciprocal of variance, $w_i = 1/\sigma_i^2$. However, error variances are usually not known, so weights will have to be assigned somewhat arbitrarily or determined numerically. The latter is in fact the basis for fitting generalized linear models (GLMs; McCullagh and Nelder 1989), since maximizing the likelihood requires a method called iteratively reweighted least squares (IRLS).

Maximum likelihood estimation (Neter et al. 1996, Morgan 2000) is appealing in the case of fitting seasonal distribution curves, where we assume that the intensity of bird migration follows a parametric density function, but have little knowledge of the error variance. ML estimation searches the parameter space for the combination of parameters values that maximizes a likelihood function,

$$\mathcal{L}(\mathbf{\Theta}) = f(\mathbf{x}|\mathbf{\Theta}), \tag{31}$$

where $\mathbf{\Theta}$ is the vector of parameters. The likelihood function is specified in terms of the joint probability for observing the data, given the parameters. This is most simply done if the data are assumed to be independently and identically distributed, when the likelihood function can be written as a product

of n probabilities,

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\Theta}) = f(x_1|\boldsymbol{\Theta}) \cdots f(x_n|\boldsymbol{\Theta}), \tag{32}$$

which typically simplifies the expression for the maximum likelihood estimator.

For linear regression, ML estimates will be similar to OLS estimates, and hence unbiased, if the error terms are independently normally distributed with equal variance. Otherwise, ML estimators are not unbiased, but are asymptotically unbiased and normally distributed under mild regularity conditions.

Statistical software may include specialized functions for fitting the more common parametric distribution functions. Alternatively, GLM is a useful framework for fitting seasonal distribution curves to bird migration data, as it builds on ML estimation and allows a flexible specification of the error distribution. A Poisson distribution is appropriate for modelling count data, but in the case of bird migration data, the typically large day-to-day variation frequently leads to overdispersion (Venables and Ripley 2002); *i.e.*, an observed variance which is larger than the variance modelled. The variance of a Poisson distribution is equal to its mean, but the extra-Poisson variance can be modelled by specifying a quasi-Poisson or negative binomial error distribution. The former acommodates overdispersion by estimating the dispersion parameter $\phi = \sigma^2/\mu$, while the latter has an extra parameter which can be used to adjust the variance relative to the mean.

Quantile regression

Quantile regression (Koenker and Bassett 1978) is an elegant alternative to performing ordinary regression on estimated sample quantiles. See Cade and Noon (2003) for a proper introduction aimed at ecologists. While ordinary linear regression assumes identically and independently normally distributed errors, the general idea behind quantile regression is to allow heterogeneity in the error distribution. Such heterogeneity in, *e.g.*, variance and skewness, will lead to different slopes for different levels of the response variable. Instead of fitting a mean to the data, quantile regression fits a specified regression quantile τ by minimizing a sum of weighted absolute errors:

$$\min_{b\in\mathbb{R}} \left[\sum_{t\in\{t:y_t \geq x_t b\}} \tau|y_t - x_t b| \quad + \sum_{t\in\{t:y_t < x_t b\}} (1-\tau)|y_t - x_t b| \right] \tag{33}$$

(Koenker and Bassett 1978). Hence, it does not assume a specific shape for the error distribution. This makes it attractive for modelling the timing of bird migration, where different quantiles may represent different populations or population segments, between-year variability has been shown to vary between quantiles, and models are likely to be misspecified due to a large number of unobserved covariates.

Non-parametric linear trend analysis

Any of the presented smoothing methods can of course also be used for modelling non-linear trends over time. A more popular approach is to fit a generalized additive model (GAM; Hastie and Tibshirani 1990,Wood 2006), where the smooth usually is estimated by a spline function. If a straight line fits within the confidence band of the smooth, a linear model for the trend is justified (Neter et al. 1996). In practice, however, a linear trend is usually assumed *a priori*. Hence, the important assumption of the error term following a normal distribution with zero mean and constant variance, is frequently broken. Transformations can be made in an attempt to meet this, but these may unintentionally alter the error structure, and make the biological interpretation of the model less clear. Standard regression techniques are also very sensitive to outliers. These should clearly not be removed if they are due to a "real" anomaly, which appears likely if the response is estimated from an adequate sample size.

There exists a range of robust regression techniques which are resistant to the presence of outliers. These include M-estimators, the least median of squares (LMS) estimator, and the least trimmed squares (LTS) estimator (Venables and Ripley 2002). Outliers may also be downweighted and the model fitted using the weighted least squares criterion.

Alternatively, trends can be estimated using a non-parametric method. As illustrated in our main text, these can also be used for identifying intervals of the predictor variable over which linear trends can be assumed. Typically, non-parametric methods replace the assumptions of linearity in trend and normally (or elsehow parametrically) distributed variables or error terms with the less restrictive assumptions of a monotonous trend and the appropriateness of the median for estimating the center of symmetry. Under assumptions of normality, they are usually less powerful (smaller asymptotic relative efficiency) than parametric methods, though often not substantially so (Gibbons 1997).

As noted in the main text, the slope from a simple linear regression and Pearson's correlation coefficient are closely related. A simple approach to non-parametric linear trend analysis is hence to make use of a non-parametric correlation coefficient such as Spearman's rho,

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}, \tag{34}$$

where $d_i = \text{rank}(x_i) - \text{rank}(y_i)$ for paired observations (x_i, y_i), or Kendall's Tau,

$$\tau = \frac{4U}{n(n-1)} - 1 = 1 - \frac{4V}{n(n-1)} = \frac{2S}{n(n-1)}, \tag{35}$$

where U is the number of concordant pairs of observations on (x, y) (satisfying $\text{sgn}(x_j - x_i) = \text{sgn}(y_j - y_i)$), V is the number of discordant pairs (*i.e.*, $\text{sgn}(x_j - x_i) = -\text{sgn}(y_j - y_i)$), and $S = U - V$ (Gibbons 1997).

The observed data are ranked by assigning integers according to their relative magnitude within the sample of observations; *i.e.*, $\text{rank}(x_i)$ is the order of appearance of x_i when the elements of $\mathbf{x}$ are listed from least to largest (or *vice versa*). Both Spearman's rho and Kendall's Tau assume there are no ties (equal ranks of x_i and x_j (or y_i and y_j) for some $i \neq j$) in the population sampled.

If there are tied observations, ρ can still be calculated as Pearson's correlation coefficient between ranked data where tied observations are assigned equal ranks (*e.g.*, using the average of the ranks the tied observations would have if they were not tied).

A test for for the null hypothesis of the slope β of the linear regression line being equal to a specified value β_0 can be performed by testing the significance of the correlation coefficient between the observed x_i and the residuals $y_i - \beta_0 x_i$ (reducing to y_i if $\beta_0 = 0$). A confidence interval for β can also be constructed; see Gibbons (1997).

As evident from the above and the formula given in our main text, the Mann-Kendall statistic S for time series data is proportional to Kendall's Tau where $\mathbf{y}$ is a time index ($y_j > y_i$ when $j > i$), thus, replacing the notion of 'concordant pair' with 'positive difference over time', and 'discordant pair' with 'negative difference'. More specifically, $S = \tau n(n-1)/2$, and is asymptotically normally distributed with $\mathrm{E}(S) = \mu = n(n-1)/4$ and $\mathrm{var}(S) = \sigma^2 = n(n-1)(2n+5)/72$.

References

Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Statist 12:171–178

Cade BS, Noon BR (2003) A gentle introduction to quantile regression for ecologists. Front Ecol Env 1:412–420

Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia, U.S.A.

Evans M, Hastings NAJ, Peacock JB (2000) Statistical distributions. Wiley, New York, U.S.A., 3rd edn.

Gibbons JD (1997) Nonparametric methods for quantitative analysis. American Sciences Press, Columbus, Ohio, U.S.A., 3rd edn.

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall / CRC, Boca Raton, Florida, U.S.A.

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Malo JE (2002) Modelling unimodal flowering phenology with exponential sine equations. Funct Ecol 16:413–418

McCullagh P, Nelder JA (1989) Generalized linear models. Monographs on statistics and applied probability. Chapman & Hall, London, U.K.

Morgan BJT (2000) Applied stochastic modelling. Arnold texts in statistics. Arnold, London, U.K.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear regression models. Irwin, Chicago, Illinois, U.S.A., 3rd edn.

Percival DP, Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge, U.K.

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall / CRC, Boca Raton, Florida, U.S.A.

Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bull Am Meteorol Soc 79:61–78

Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer Verlag, New York, U.S.A., 4th edn.

Walpole RE, Myers RH, Myers SL, Ye K (2002) Probability and statistics for engineers and scientists. Prentice-Hall, Upple Saddle River, New Jersey, U.S.A., 7th edn.

Wood SN (2006) Generalized additive models: an introduction with R. Chapman & Hall / CRC, Boca Raton, Florida, U.S.A.