# Critical windows: A method for detecting lagged variables in ecological time series

Jean-Sébastien Pierre, Maurice Hullé, Jean-Pierre Gauthier, Claude Rispe

# Critical windows: A method for detecting lagged variables in ecological time series

Jean-Sébastien Pierre*, Maurice Hullé**, Jean-Pierre Gauthier** and Claude Rispe***

November 20, 2020

**Authors details:**

* CNRS INEE, Université de Rennes 1, OSUR, UMR 6553 Ecologie Biodiversité Evolution, campus scientifique de Beaulieu, 35042 Rennes - cedex (France)

** INRAE Agrocampus Ouest, Université de Rennes 1, IGEPP, F35650 GR. (France)

*** INRAE UMR BioEpAR Atlanpole-Chantrerie CS 40706 44307 Nantes Cedex 3, (France)

## Summary

Many developmental processes in the life sciences, ecology and even in economics depend strongly on the environmental conditions occurring in a bounded time interval, the results occurring often far later. Examples are as diverse as plant phenology, grapewine maturation, diapause induction and so on. The method proposed here, aims at detecting quickly such effects. The basic idea is to regress the recorded results of a series of replications of the process against a function of an independent time series. This variable is defined on a set of periods of time, systematically scanned by varying their lower and upper bounds. In simple cases when this function is the integral and the effect strictly limited in a time window, the response model, under the form of correlation coefficients, is tractable and its shape is predictable. It is the same when the window is a bell-shaped function and can be fitted with other weighting functions as the Beta and the polynomials. The null hypothesis of absence of influence of any past interval is tested by Monte-Carlo simulation. The most likely window of influence is determined by the maximum correlation coefficient, and the bivariate confidence interval is estimated by bootstrap. The period found with a rectangular shaped window can be used as a starting point for more specific windows. This technique has the advantage of avoiding to split the climatic series into arbitrary slices, thus multiplying the predictors and complicating the models selection. It is closely linked to continuous lag distributed models with the simplification that the variable of interest is not explicitly time

dependent. Examples are given for the prediction of aphids population dynamics, male morphs induction in aphids, and the phenology of the ash tree.

## Keywords

critical periods, correlation, time series, climate, lag analysis

## Introduction: What are critical periods ?

Some processes involving so called "critical periods" are fairly common in various fields of biology, ecology, behavioural sciences and psychology. They share the following features: a) the effect of a driving factor is lagged; b) the driving factor only has an effect during a limited period; c) the effect of the factor is cumulative during this period. The concept comes from Ethology and was coined by Lorenz [13]: the strength of imprinting in birds and mammals, depends on the presence of the mother-like stimuli during a short period after hatching, and has no effect thereafter. This concept is now commonly used in developmental psychology.

Many examples can be drawn also from Ecology. For example, one can quote the proportion of insects entering diapause and the moment where they do so and, similarly the proportion of endodormancy in plants which depends upon environmental cues like the decrease of photoperiod and temperature in a given period of time. The research on plant phenology constitutes a large frame for such investigations. Similar relationships can be found between the amount of cold received during winter and the proportion of insects emerging from diapause. The number of sexual forms induced in various species of aphids depends on the decrease of temperature , as well as of photoperiod during a short period of the fall. The quality of grapes in vineyards is known to require rainfall at fruit set and a sunny weather at maturation. Many other examples can be found emphasizing the interest of modeling critical periods, which were studied since several decades under different names and with different approaches.

The present article aims at detecting statistically such phenomenon and at finding the range of time in which environmental conditions influence them. Practically, the procedure is not new and arises from at least two statistical approaches. One is the systematic scanning of time periods, and another is the use of lag distribution models.

## Systematic scanning

The idea of scanning systematically the periods of influence of an external factor on a biological phenomenon was firstly applied to the phenology of vegetable crops, affected by weather, by Goldwin [9]. This author outlined the principles of the method, without studying in depth the underlying statistical model. The idea was then applied to the outbreaks of aphids by Thomas *et al.* [23], with a first proposition of graphical representation, and then, in a quite different manner by Pierre *et al* [15]. These latter authors made the first attempt to define the method as a bivariate correlogram, to propose an interpretable graphical representation of it, and to find its theoretical shape. Several cases of critical climatic periods were then found in various aphids: nine different species of winged aphids(A'brook,[1]) , *Phorodon humuli*,(Thomas *et al*[23]), *Sitobion avenae* F. (Pierre and Dedryver[16], Pierre *et al*[15], *Rhopalosiphum padi* L., Rispe *et al*[20]). The procedure suggested by Goldwin[9] was also developped further by Coakley *et al.*[5] under the name "Window" and later by Pietravalle *et al.*.[17] under the appellation "Window Pane". Both of them applied the method to the incidence of the wheat disease *Septoria nodorum*. Coakley suggested the boostrap method to find the variance and confidence interval of the best correlation coeffcient. Hullé and Gauthier rewrote the FORTRAN program written by Pierre *et al.*[15] in language C, adding bootstrapping and Monte-Carlo simulation to test correctly the correlation peaks. The software application, named Criticor was then distributed freely by the French National Institute of Agronomic Research (INRA, now INRAE). The program description and notice is available on the site Yumpu[2].

This very method was then used by Ennaïfar *et al*[8] for the take-all desease (*Gaeumannomyces graminis* var. *tritici*) incidence of wheat, and Atlan *et al*[3] for the flowering of the european gorse (*Ulex europaeus* L.). Recently, Pierre (2017) deposited in CRAN a version written in *R* [19] language named Rcriticor in reference to the INRA (now INRAE) version. Statistically, this method has some advantages, especially that of reducing the number of possible predictors only to one for each type of climatic series. A widespread alternative is to split the past climate into arbitrary slices (weeks, months, decades,etc.) and to achieve a multiple regression of any kind (linear model, neuron network, learning,..) between the variable of interest and the set of predictors so created (*e.g.* A'Brook 1981[1], Tsai *et al.*2016[24], Ji and Peters[11] among many others). Then comes the tedious work of selecting models, to disentangle the correlations between predictors. The research of critical periods does directly most of the work ahead, as we will show

on a phenologic example. The problem of model selection is reduced therefore to a small number of predictors, one or at most two by type of climatic series. The correlations between them are furthermore reduced providing that the critical periods for diverse climatic series do not overlap. As described above, the method that we propose is close to lag analysis, a statistical field which has recently attracted the interest of various scientists, especially in the field of plant phenology. There is however a fundamental difference: the variable of interest is not time-dependent and is only supposed to be observed later that its predictor. At the contrary, in the lag distributed models, the predictor does not cease its action and influences the whole time series of the variable of interest in its future (Sims[22]). This difference has some statistical consequences that will be described in the Material and Methods section of this article.

## Lag distributed models

The other approach, largely developped in the field of econometrics (*e.g.* Seddighi[21]) consists in the discrete or continuous[22] lag-distributed models, sometime called briely lag analysis. In these models, the dependent variable is itself a time series or at least a variable indexed by time. The phenomenon observed at time $t$ depends on several values of another time series, lagged in the past with each one its coefficient. This approach was recently used with success in Ecology.

Among the most elaborate approach using lag analysis, we can cite the work of Ogle *et al.*. (2015)[14], who propose a comprehensive modelling of lag analysis in ecology, with a bayesian estimation of all relevant parameters. Their work is more ambitious than our one which aims only at a fast research to detect "hot spots" in a climatic series during the year. Although quite close to the preceeding approach, there is a subtle but noticeable difference as outlined above. So, Ji and Peters (2005)[11], working on biweekly series, present in their equation (1) an expression which would be the strict discrete equivalent of the equation (1) of the present article, excepted that the dependent variable is estimated in $t$ while in our equation (1) $Y$ is not indexed by the time but supposed to be observed only once during a period of time (typically the year). They refer also to the lag distribution model of Seddighi (2000)[21] in a a book on econometrics comprising a comprehensive chapter on lag analysis. There again, the model of equation (1) is the continuous version of the one defined by this author p 118, equation 4.28., with the same noticeable difference. In the materials and methods section of this article, we will assess more precisely the correspondence of this model with the "continuous time lag distributed models" approach defined

by Sims[22] and present some results about this continuous version. A domain where lag distributed models have shown their usefulness is that of plant phenology,especially when recorded by remote sensing. Wang *et al.*.(2001)[25] studied so the evolution of the Normalized Difference Vegetation Index (NDVI),in relation to climatic factors, namely precipitations and temperature with long time lags (up to 2 months). Ji & Peters (2005)[11] built a similar model for precipitations, referring to a lag model based upon a biweekly cutting of the rainfall data. Their Figure 1 is conceptually the same as our Figure 1 shown further.

Recently, Hufkens *et al.*.(2018)[10] incorporated such models in a comprehensive phenology modelling framework, available in the repository CRAN (R development team 2019)[19], emphasising in their article on the importance of plant phenology for ecological processes. Keenan *et al*[12], study the temperature sensitivity of plant phenology. They define it quite simply, following various authors, as the ratio of the variation of the date of occurence of a phenological stage, by the variation of temperature in a given year. They draw from their work interesting conclusions on the effect of the observation time (denominator of the ratio) on the measure of this sensitivity. Incidentally, studying the correlation of the phenological date with the duration of the observation time, they draw exactly the same correlogram as we do in their Figure 1c. In 2013, Chuine[4] provided a comprehensive overview of phenology models that enlights the importance of lag distributed models.

## Aims of the article and examples provided

In this paper, we analyse some properties of the temporal bivariate correlogram resulting from the systematic scanning of many periods of time, and propose an underlying model in two simple cases (rectangular and bell-shaped window), as well as a technique of data analysis for detecting critical periods. We sketch also the properties of some other filtering windows, the Beta shape and the polynomial one. This technique is then applied, to various ecological problems, namely the outbreaks and biology of aphids and, to join the important field of plant phenology, the occurence of a given development stage of of the ash tree (*Fraxinus excelsior* L.) found in a public database.

# 1 Material and Methods

## 1.1 Modelling critical periods

### 1.1.1 The regressive process.

The idea of critical period can be expressed mathematically by a regressive process between a time series $U$ and a random variable $Y$: The transfer function between $U(t)$ and $Y$ is the a scalar product (product of functions in a Sobolev space) between $X(t)$ and a window function $\Phi(t)$. $Y$, the exogeneous variable, is therefore the sum of this transfer function and an error $\varepsilon$.

$$Y = k\int_{-\infty}^{T} \Phi(t) dU(t) + \varepsilon \quad (1)$$

where $k$ is a regression coefficient, $\varepsilon$ an random variable associated to each realisation of the process, $\Phi(.)$ a window function weighting the time series $U(t)$. When needed, $Y$, $dU(t)$ and $\varepsilon$ will be indexed by $i$ to infer the model from replicates.

Figure 1: Illustration of the transfer function and shape of the window function. a: rectangular window function b: resulting transfer function
c: bell-shaped window function d: resulting transfer function
The red bar and red dot indicate the result of the process, and the red arrows the application from the integral to the final observed value of $Y$. With the rectangular window (a and b) the limits of the critical period is sharp, with a bell-shaped window, (c and d) they are fuzzy.

The model is illustrated by Figure 1, with two kinds of weighting functionss $\Phi(.)$. For statistical analysis, it is convenient to consider $U(t)$ as a pathwise solution of some stochastic differential equation such as

$$dU(t) = X(t)dt + \sigma dW(t) \quad (2)$$

where $X(t)$ is the derivative of a trend or drift, and $W(t)$ the standard brownian motion amplified by a standard deviation or volatility $\sigma$.

For example, the daily record of mean temperatures in a year can be considered as the sum of a seasonal sinusoidal function and of a random noise. This noise is more realistically modeled as

a "pink" noise than by a white noise, taking in consideration some autocorrelation, but at the moment we do not enter into such accurate considerations.

By substituting (1) in (2), the model splits easily in the sum of a deterministic and of two stochastic parts:

$$Y = k\int_{-\infty}^{T} \Phi(t)X(t)dt + k\int_{-\infty}^{T} \Phi(t)\sigma dW(t) + \varepsilon \tag{3}$$

in which the first term is a determinist function (drift), the second term, a stochastic integral which results in a random effect $A_i$ attached to each year in most applications. The third term, $\varepsilon$ is an error, a random variable associated to each replicate of (1) on which we do not do any particular hypothesis, excepted $E(\varepsilon) = 0$.

If necessary, in the context of the general linear model the most habitual hypothesis are normality, constant variance and null autocorrelation. This model is somewhat similar to a continuous ARIMA model [6], except that we focuse on the "far" past and not at the close past of the process.

Furthermore, the $Y$ response is completely dissociated temporally from the time series from which it originates. Actually, it is enough to consider that $Y$ is observed at any time $t_{obs} \geq T$. $Y$ is unique, and only linked to a realisation of the series, what we shall explain later. We focus our interest on two simple cases of windows defined on $R$:

i) a rectangular window

$$\Phi(t) = \lambda\left[H(t-\alpha) - H(t-\beta)\right] \tag{4}$$

where $H(t-u)$ is the Heaviside unit step function switching from 0 to 1 on u. For proper scaling, we will set $\lambda = \frac{1}{\beta - \alpha}$ to ensure that the weights sum to one over $R$. We assume also that $T > \beta > \alpha$. This last assumption models the fact that $Y$ is observed after that the effect of $U(t)$ vanished.

In summary, this hypothesis is a simple cutoff of the effect onto time $\alpha$ and after time $\beta$. This is the crudest model.

ii) a bell shaped window centered on given date $t^*$. For instance a gaussian density function of pseudo-standard deviation $c$:

$$\Phi(t) = \frac{1}{c\sqrt{2\pi}} e^{-\frac{(t-t^*)^2}{c^2}} \tag{5}$$

We qualify $c$ as a "pseudo" standard deviation because $\Phi(t)$ is used as a mere weighting function, not as a probability density function. Then $c$ governs only its spreading. By respecting the scaling by $\sqrt{2\pi}$ of the gaussian function, the weights over $R$ sum also to one. In this continuous case, we cannot set $T$ as strictly posterior to the effect of $U(t)$, and choose, as observation time, a $T$ arbitrary large enough to ensure that

$$\int_T^\infty \Phi(t)dt < h \tag{6}$$

$h$ being arbitrarily small. For practical purposes, $T > t^* + 4c$ can be convenient in analogy to the fact that the densities of the gaussian function are negligible under -4 and above +4 standard deviations. In summary, the effect of the times series increases progressively from a time about $t^* - 4c$ to the time $t^*$, then decreases and vanishes about a time $t^* + 4c$. The effect is smoother and more fuzzy, thus is a little more realistic.

Fig 1 illustrates its shapes in both cases.

We note that these two shapes of windows are not at all the only possible, an infinitely many others can be imagined. With those two shapes of kernel, some useful results can be mathematically obtained, which are fully developed in appendix A where two other types of windows are briefly discussed. In the first case, and when replicates are available the process reduces obviously to the simple form:

$$Y_i = k\int_\alpha^\beta X_i(t)dt + k\int_\alpha^\beta dW_i(t) + \varepsilon_i \tag{7}$$

Where $dW_i(t)$ denotes the pathwise $i$th realisation of the brownian motion. To distinguish clearly the two sorts of error, we can set:

$$A_i = k\sigma\int_\alpha^\beta dW_i(t) \tag{8}$$

as representing the process error due to the random noise in the case of a simple rectangular window. The properties of the brownian motion ensures that $E(A) = 0$ and

$$\mathrm{var}(A) = k^2\sigma^2(\beta - \alpha) \tag{9}$$

The pathwise independence of the brownian motion ensures also the pathwise

independence of the errors $A_i$. Equation (7) is therefore an ordinary regression model, although implying both a stochastic integral, and an observation error $\varepsilon_i$ on which we can make any kind of hypothesis, conditioning only the type of regression we can carry out (normal, Poisson, binomial, beta, etc.)

If the replicates consist in a set of years (or of any period of interest), that later process may be illustrated as in Figure 1b, in the case of a rectangular window: Each year a given variable is sampled one time, and is regressed on the past. The past is taken into account through the window $\Phi(t)$, and is therefore proportionnal to the area between $X(t)$ and the time axis in an interval $[\alpha, \beta]$. In the second case, the integral is weighted by all the past, but its contributions are concentrated around the middle time of influence $t^*$. The boundaries of the influential period are fuzzy.

The major point of interest in such processes is to determine wether the shape and boundaries of the window could be estimated. For that purpose, we shall devise an empirical correlogram, and search its properties in the two cases described above, (rectangular and bell-shaped function). We hypothesize that the Bravais-Pearson correlation coefficient can be used as a criterion function to evaluate the window of influence of $U(t)$. This point is proven in appendix A1.

### 1.1.2 Exploring other shapes of the window

An infinity of functions $\Phi(.)$ can be used in equation(1) to filter the time series $U(t)$. Basically we use the simpler, the rectangular window, to find critical periods. After that a period is found, its bounds can be used as initial boundary guess to find a different window. Seddighi[21] and Sims[22] insist on the general difficulty to find a good shape for the equivalent function in lag distribution models, without setting some strong constrains on them. In the discrete case, trying to find an unconstrained set of coefficients leads quickly to an overparametrisation of the model assorted with a weak gradient, making the convergence of most algorithms unsure and unstable. Some types of functions are classical: the exponential decreasing from present to past, and polynomial structure. We did not keep the decreasing exponential because, as we set it in § 1.1, our model is not fully of type "lag distribution", the dependent variable having no temporal link with the predictor. We considered three shapes (plus the rectangular one) either for their ecological

relevance or for their versatility:

1. Rectangular (the default window).
2. Gaussian (bell-shaped, no precise limits).
3. Generalized Beta (strictly bounded, very versatile).
4. Polynomial (different shapes, bounded and normalised by its integral for summing to 1 in the boundaries of the window). This window can thus be strictly bounded and is also very versatile in shape. We chose to limit the degree of the polynomial to 5.

In appendix A, the rectangular and gaussian cases are discussed in depth. The beta and polynomial cases are just outlined as providing few analytic results but entering in the general frame of the equation (A23) in Appendix A. In the package Rcriticor, three novel functions were added to optimize the choices 2,3 and 4. We'll give the results of their use in the "Results" part.

## 1.2 Precision of estimations, tests

Testing the significance of the best correlation coefficient as found in § 2 is not straightforward as all the calculated coefficients are themselves correlated. The task seems mathematically untractable and that led us to adopt some simulation approaches. We chose two of them: Monte Carlo permutations for testing the significance of the higher peak and a bootstrap estimation for assessing significance limits. Those two strategies are used in both implementations described lower in § 1.3.

### 1.2.1 Monte Carlo permutation test for peaks significance

The choice of a maximum correlation coefficient in absolute value leads to a high risk of spurious correlation. The classical test of the correlation coefficient, using the $t$ approximation

$$r = \frac{t}{\sqrt{n-2+t^2}} \tag{10}$$

is flawed in this case. Its significance value is however given in both softwares as an indication. Alternatively, a p-value is calculated by random permutations of the dependent values with replacement. Given a threshold risk $\alpha$, the observed peak (or sink, respectively) is considered as significant if it occurs in les then $100\alpha$ percent of the permutations.

This procedure allows also to draw contour maps of the areas including significant correlations in the plane (a,d) defined above ($d = b - a$).

An example is given in Figure **??**c showing the area where the correlation between the sums of temperatures above $3^{o}$ C and the peak of outbreak for the grain aphid *Sitobion avenae* F. are significant. The observed peak, figured as a black dot, is inside a significant area at risk $\alpha = 0.05$. Criticor and Rcriticor give also an histogram of the expected distribution of the correlation coefficients under the null hypothesis that there is no correlation between the independent series and the dependent set. Examples are given in Figure 2.

Figure 2: Histogram of the absolute values of the maximum correlation coefficients obtained through 1000 random permutations of the dependent variable.
a: in the software CRITICOR (INRAE).The observed coefficient (-0.85) is thus declared significant at risk $\alpha = .05$.
b: in the package Rcriticor of CRAN. Case of the male flights of Ropalosiphum padi see also Figure **??**. Observed coefficient: 0.668 (vertical black line), p-value=0.00019

### 1.2.2 Bootstrap tests for the confidence limits of critical periods and peaks

Another important question is to assess confidence areas for the best correlation coefficient and for the asociated bounds of the corresponding critical period. In both softwares, a bootstrap procedure is proposed to achieve this task. At each bootstrap run, a sample of the available cases of the dependent variable is drawn with replacement and each case is conveniently associated to its corresponding time series (typically a year). Each run produces a pseudovalue and a given number of runs (1000 or more) permits an estimation of the variance-covariance matrix of the maximum (respectively minimum) correlation coefficient, and of $a$ and $d$, respectively the beginning and duration of the associated critical period. This is done by computing the Efron's bootstrap estimator [7]. Let us note that another estimator of the variance-covariance matrix of those coefficients may be obtained via the Fisher's information matrix (see appendix B).

## 1.3 Computer implementation

Two implementations of the method were achieved: one, maintained by the french National Research Institute for Agriculture food and Environment (previously INRA, now INRAE) under the acronym Criticor, and the other maintained by J.S. Pierre is available in CRAN as a package and named Rcriticor. Criticor is written in C and Rcriticor in the R programming language[19]. As

all time series available are discrete, the softwares implement discrete versions of the process described above, where integrals are replaced by discrete sums of trapezes. Some example of results are given here below.

## 1.4 Set of data analysed

**Grain aphids peaks of population**

*Sitobion avenae F.*, the grain aphid,is a direct pest for cereal crops, especially for winter wheat. Forcasting models were achieved by different method by our laboratory since 1975, and population in the field were recorded during 15 years among which 7 were intensive. The outbreaks of *Sitobion avenae* F. were firstly correlated to the sum of day degrees above $3^oC$ [16] in the month of february. This was obtained by linear multiple regression in which the independent variables were the sum of temperatures in each month preceding the outbreak. This statistical procedure was then followed by the research of critical periods [15], the example given here. Aphids were sampled weekly, and the variable studied here is the recorded peak of population. This example is provided, along with the data, in the CRAN package Rcriticor. Let us note that we switched after that toward mechanistic models of population dynamics[18] themselves incorporated in the commercial package Colibri® belonging now to the firm Bayer Cropscience.

**Induction of male morphs of the cherry-oat tree aphid**

This example comes from our common work with Rispe *et al* [20]) trying to find the factors influencing the production of sexuate morphs in other cereal aphid, *Rhopalosiphum Padi* L, the bird cherry-oat aphid. This species spends most of its time as parthenogenetic morphs, but autumnal conditions induce the production of sexuate morphs. Among them the males are the easiest to record, as they are currently caught in the English and European suction traps network. The photoperiod is well known as being the main factor influencing this cycle switch

**Occurence of the stage 11 (first leaf unroll) in the ash tree**

This series was taken from the public database PEP725, devoted to phenological data. In this database, we chose the ash tree in the village of Cardedeu (Catalunia, Spain) because it provided a long series of observations spanning from 1953 to 2000, that is 47 years. Unfortunately somme yearly observations lack and, ass a result, only 23 years are available:

1953,1955,1969,1973,1980-1984,1987-2000. We chose as phenological data the date of leafs unfolding, a good marker of spring revival of the trees. The climatic data were obtained from the meteorologic service of Catalunya which give access freely to the data. Unfortunately, for the site Cardedeu, only monthly means were available and only for temperature and rainfall. To recover one part of the climatic variablity at a scale smaller then the month, we achieved a daily interpolation of these data by cubic splines before applying the research of critical periods. Obviously, this procedure leads to work on a smoothed version of the climatic series and the stochasticity of climate is known less accurately than for the other examples. The R scripts of reconstitution of the data are provided as supplementary material.

# 2 Results

## 2.1 The sensitivity to late winter temperatures in the grain aphid populations

Aphids outbreaks occur generally in june, and thus february appeared as a critical period for temperature, preceding the event of interest (outbreak or population peak) of five to six months. The present method was then used to define more accurately this period. Figure 3a shows the correlogram obtained in this case. The figure is much more noisy than the theoretical figures 6 and Figure 7 (in appendix A), but looks more like Figure 7, evoking a bell-shaped influence. The best correlation was found as $\rho = 0.981$ for a period beginning at day 37 after january the first, i.e. february 6, and lasting 44 days, that is until march 22. This new period, longer than the month, was used further for forcasting [15]. Figure 3b shows the confidence area of the maximum correlation position by bootstrap, while Figure 3c (significance map) and Figure 3d (histogram of permutations) ensure that the maximum correlation observed is significant at the risk $\alpha = 0.05$.

Figure 3: Intensity of outbreaks of the grain aphid *Sitobion avenae* F.

a: Correlogram crossing the intensity of outbreaks of the grain aphid *Sitobion avenae* F. and the sum of temperatures above $3^o$C during 7 years in Britanny (France). The observed maximum corresponds to a period beginning at day 37 (february 6th) and spending 44 days (until march 22th).

b: The same correlogram with bootstraping (500 subsamplings). The red diamond indicates the bootstrap estimator of the maximum of correlation. Red dots figure individual pseudovalues. The

ellipse is the bivariate confidence interval (95%) around the bootstrap estimator.

c: Significance map obtained by Monte-Carlo permutations (5000 replicates). read area: correlation coefficients significant at $\alpha = 0.05$.Two other areas are figured corresponding to the thresholds 0.01 (black) and 0.1 (yellow)

d: histogram of extreme correlation coefficients obtained among 10000 random permutations.

**Other shapes of windows.**

On this example, with a gaussian window we get, at most, $\rho = 0.788$, a lower value than with the rectangular window. Trying the Beta function, we find $\rho = 0.981$, exactly like with the rectangular window, resulting from the fact that the optimization algorithm converges toward $\alpha = 1$ and $\beta = 1$, the values where the Beta distribution is identical to the uniform one. The best correlations found with a polynomial window were obtained at .9709 for the second degree and .9755 for the fifth degree. To summarize, there is no reason here to challenge the rectangular window.

## 2.2 The sensitivity to induction factors of sexuality in the bird cherry-oat tree aphid

The Criticor method shows that, surprisingly, high temperatures in late summer and autumn favor the proportion of males in the suction traps. The critical period for temperature, the span of which is estimated from August 24th and October 17th. Figure 4a shows the correlogram and Figure 4b the scatter plot corresponding to the estimated critical period. This shows that, if photoperiod drives the induction of sexuals, the importance of their production is strongly determined by autumn temperature. The observation of a positive correlation is the inverse of our initial hypothesis.

Figure 4: a - Correlogram obtained between the proportions of males of R. padi in the autumn flight and mean summer temperatures at Rothamsted. Periods are from July 19th to October 17th, with duration from 1 to 60 days. The arrow indicates the highest correlation peak. b- Linear relationship between the corresponding sum of temperature and the rate of males production.

This fact has a strong importance in aphids population dynamics: sexuals go back to the winter host, the cherry-oat tree where females lay diapausing winter eggs which are cold resistant. The part of the population which remains parthenogenetic, at the opposite, stays on winter and volunteer cereals. Those latter are susceptible to be killed by harsh frosts. The knowledge of this fact has applications on risk forecasting: this species transmits viruses on winter cereals in autumn, and that as more as the population is constituted of more parthenogenetic morphs and less sexual morphs.

**Other shapes of window.**

All our attempts to fit other shapes of window than rectangular resulted in failures. In all cases, the absolute values of the correlation coefficients were lower than in the case of rectangular window. In some cases (Beta, polynomials of degrees 2,3,4,5) the correlation coefficient turned to be negative. We do not present these results in details.

## 2.3 Ash tree phenology in Spain

The date of occurence of the stage 11 (leaf unfolding) of the ash tree was correlated with past temperature and rainfall.

### 2.3.1 Temperature

We found a negative relation (Figure 5a) between temperature (threshold $0^0$C) with a correlation coefficient of -0.667 (maximum in absolute value) for a period beginning at Julian day 43 (February 10) and during 74 days (till Julian day 117 i.e. April 26). The bootstrap estimation (1000 resamplings, Figure 5b) corrects somewhat this crude estimation, obtaining a slightly better correlation of $\rho = -0704$ for a slightly shorter period beginning at day 55 (55.196 estimated) and lasting 59 days (exact estimate : 59.147).That is a period of two months between February 22 and April 23. Figure 5d shows the scatter plot obtained for this last period, a fairly linear relationship, rather surprising as, theoretically, following the theory of day-degrees, an inverse hyperbolic relation is expected. Figure 5c shows the significance map obtained, showing a very large area in which the influence of temperature is highly significant (black area, $p<0.01$). This result is obviously banal but shows that day-degrees have little influence on the leaf unfolding of the ash tree before late february. This is probably because before that, winter coldness is necessary for the

breaking of endodormancy. The effect of day-degrees accumulation acts therefore only after that dormancy breaking.

### 2.3.2 Rainfall

Rainfall has a longer influence than temperature (Figure 5e), the period detected covering three months, from Julian day 61 (March 2) to Julian day 184 (July 4). The absolute value of the correlation coefficient is lower than for temperature, ( $\rho = .465$ ). On the correlogram (Figure 5a) it appears as the merging of a short period beginning at mid March and lasting a few days and another much later not visible on the figure. The bootstrap then corrects strongly this estimation by setting a new one for a period ongoing from day 65 (bootstrap: 65.456 - March 4) to day 139 (bootstrap: 139.095 - May 18, Figure 5g). This estimation has the interest of being more predictive although with a lower correlation coefficient ( $\rho = 0.334$ ). In effect, this period ends just at the earliest beginning of leaf unfolding. We therefore kept this estimation for further multivariate regression. The scatterplot show a positive relation but with a noticeable deviation from linearity (smoothing spline, red curve,Figure 5h). This encouraged us to use a General Additive Model (GAM) to combine temperature and rainfall in an explanative model. The positive effect of rainfall suggests that it delays the leaf unfolding of the ash tree. This is a little surprising as rain is supposed to be beneficial for the plants development. The inverse correlation of rainfall and temperature may be suspected here: the rainfall results often in a loss of temperature.

### 2.3.3 Bivariate model

To combine the effects of temperature and rainfall we used a General Additive Model. The choice of GAM was driven by a detectable nonlinearity of the residuals (Figure 5h). We compared two families of errors, the classic normal one and the Cox Proportional Hazard model as the variable of interest is of the survival time type. Table 1, a and b summarises the results. Although theoretically more correct, the Cox Proportional Hazard model does not do better than the standard normal one. The standard regression model is more powerful, although its Akaike Information Criterion is higher than that of the Cox model. Furthermore, the residuals pass the Shapiro-Wilks test of normality. The lack of significance of the rainfall effect would incite to take only into account the effect of temperature. We kept it, however, because of the improvement of the AIC it gives. At end, we get a parsimonious model, with only two predictors, directly issued from the analysis of

correlograms.

| parameters | p-value | sign. level | AIC | adj. $R^2$ |
|---|---|---|---|---|
| temperature alone | 0.000741 | *** | 173.512 | 0.426 |
| temperature | 0.00266 | ** | | |
| + rainfall | 0.0926 | . | 172.028 | 0.458 |

a

| parameters | p-value | sign. level | AIC | adj. $R^2$ |
|---|---|---|---|---|
| temperature alone | 0.031 | * | 100.556 | 0.185 |
| temperature | 0.047 | * | | |
| + rainfall | 0.115 | | 102.03 | 0.235 |

b

Table 1: Results for the univariate (temperature alone) and bivariate (temperature and rainfall) models, as predictors for the stage 11 of the ash tree in Cardedeu (Spain).

a: GAM model, gaussian family,

b: GAM model, Cox Proportional Hazard model.( $R^2$ : pseudo $R^2$ of Cox and Snell)

Figure 5: Research of predictors for the stage 11 of the ash tree in Cardedeu (Spain) - a to d: temperature - e to h: rainfall.From left to right: correlogram, bootstrap estimation with pseudovalues and confidence ellipse, significance map, scatterplot corresponding to the bootstrap estimator. In the frame $h$ the red curve is a smoothing spline indicating the non linearity of the relation which led to use a GAM model.

| model | components | BIC |
|---|---|---|
| complete | 12 predictors | 184.76 |
| | mean temperatures: January to June | |
| | rainfall: January to June | |
| best predictor alone | mean temperature in March | 177.26 |
| stepwise regression | temperatures: February,March; rainfall: June | 173.60 |
| GAM | temperatures:February, March; rainfall: June (smoothed) | 171.74 |

Table 2: Summary of model selection with monthly predictors

### 2.3.4 Comparison with a monthly multivariate approach

An alternative frequently used is to introduce in a multiple regression a series of predictors obtained by cutting the time series into equal segments, on which the climatic variable is summed or averaged. Those segments can be months, half months, weeks, etc. The inconvenience is to introduce an artificial splitting of the climatic series of interest, and to manage a large set of predictors leading to the difficulty of finding a "good" subset of predictors. To illustrate this point, we present the results obtained in the ash tree case. The stepwise procedure was based on the Bayesian Information Criterion (BIC), more penalizing than AIC and leading to a slightly more parsimonious model (only three predictors, instead of four with AIC). We can then compare the characteristics of the two approaches. The comparison (Table 3) shows firstly that the correlogram based approach is basically more parsimonious than that by multiple regression, although, in the present case, the stepwise procedure does the job well. It shows also that there is a gain of accuracy in the definition of the periods, because the correlogram based approach allows to find bounds inside one month, at any place. In our example, we can see that the influence of rainfall in March, April and May are discarded by the process of predictors selection, at the advantage of June alone, while the scanning process detects a critical period covering march to may. This advantage of parsimony is shared by all the methods which research precisely the lags of influence of the climatic series.

| model | initial | final | final predictors |
|---|---|---|---|
| | predictors | predictors | (nature) |
| | (number) | (number) | |
| Rcriticor | 2 | 2 | -temperature: February 22th - April 26th |
| | | | -rainfall: March 4th - May 18th |
| multiple regression | 12 | 3 | -temperature: February |
| on monthly variables | | | -temperature:March |
| | | | -rainfall: June |

Table 3: Comparison of the structures of the models obtained by Rcriticor (correlogram based research) and by multiple regression.

**Searching other shapes of the window.**

We give here shortly the results of our trials to fit other shapes then rectangular for the filtering window. We tried, as announced in Materials and Methods, three cases besides the rectangular one: gaussian (without precise limits), Generalized Beta (strictly bounded) and polynomial (bounded and normalised). Table 4 summarizes the results obtained. As in the preceding cases, the results are somewhat deceiving. The gaussian windows does not improve the fit, the Beta case challenges very few the rectangular one (only giving less importance to the upper bound of the window) regarding the temperature, and converges to the rectangular one (uniform) regarding the rainfall. The polynomial shape gives the best results for degrees 3 and 5 for temperature and for degrees 3 and 4 for rainfall. The correlation coefficients get higher results than that of the rectangular shape only in the two polynomial degrees for rainfall. In the case of temperature, degrees 5 and 3 give contradictory results, the degree 5 leading to a positive correlation, and an artefactual scatterplot: one of the value gets the huge value of 1 billion, while all other points are flattened around zero. It is clear that the simulated annealing procedure found a spurious maximum which must be discarded. Concerning rainfall, the polynomial approaches are more interesting as giving more importance to the end of the period than to its beginning. At degree 4, especially, we get a shape of the window monotonous and convex, leading to reconsider the interest of an exponential decreasing model.

| window type | predictor | degree | coefficients | $\rho$ |
|---|---|---|---|---|
| rectangular | temperature | - | - | -0.478 |
| gaussian | temperature | - | $\mu = 100$ | |
| | | - | $\sigma = 80$ | -0.442 |
| Beta | temperature | - | $\alpha = 1$ | |
| | | - | $\beta = 1.1$ | -0.479 |
| polynomial | temperature | 3 | $a_0 = 2.383$ | |
| | | | $a_1 = 0.453$ | |
| | | | $a_2 = 3.195$ | |
| | | | $a_3 = -0.051$ | -0.380 |

| | | | | |
|---|---|---|---|---|
| polynomial | temperature | 5 | $a_0 = 57.830$ | |
| | | | $a_1 = 26.617$ | |
| | | | $a_2 = -203.800;$ | |
| | | | $a_3 = 308.400$ | |
| | | | $a_4 = -229.345$ | |
| | | | $a_5 = 2.621$ | 0.345 |
| rectangular | rainfall | - | - | 0.465 |
| gaussian | rainfall | - | $\mu = 100$ | |
| | | - | $\sigma = 40$ | 0.361 |
| Beta | rainfall | - | $\alpha = 1$ | |
| | | - | $\beta = 1$ | 0.465 |
| polynomial | rainfall | 3 | $a_0 = 122.448$ | |
| | | | $a_1 = 136.019$ | |
| | | | $a_2 = -142.186$ | |
| | | | $a_3 = 0.491$ | 0.562 |
| polynomial | rainfall | 4 | $a_0 = 122.001$ | |
| | | | $a_1 = 245.457;$ | |
| | | | $a_2 = -111.124;$ | |
| | | | $a_3 = -673.024$ | |
| | | | $a_4 = 3.176$ | 0.541 |

Table 4: Fitting different critical windows on the ash tree phenology (stage 11) in Cardedeu (Spain)

# 3 Discussion

We provide here a method for exploring the specific question of cumulative and delayed effects of a time series onto a measurable event. This case is relatively frequent in biology, ecology and psychology. We show here that our method is efficient when such effects are suspected. This

article has no other purpose than to contribute to statistical methods for exploring the past effects of some factors onto measurable events in biology. We bring some mathematical results about its statistical background. Under two sorts of transfer functions the maximum exists, and is theoretically obtainable in a quasi closed form (see appendix B for details). The maximum of correlation coincides with the maximum likelihood estimate (proven in appendix B). The visual inspection of the correlogram is useful to detect eventually a local maximum and/or a combination of separate critical periods. A R package is associated to achieve this exploration. We restricted our investigations to some simple cases. Some refinements are provided later: how to manage non-stationarity in the independent series such as regular trend, seasonality, truncation. Multiplicative trends or other functional forms would require the Ito or Stratonovitch calculus which were not required in our simple models where the white noise can be considered separately and additively.We provide here a method for exploring the specific question of cumulative and delayed effects of a time series onto a measurable event. The modelling of critical periods that we propose here is purely phenomenological. It says actually nothing about the underlying mechanisms which determine these influences of the past. It has however the advantage to focus the study of these eventual mechanisms in precise periods of time. This can be of great interest for building relevant mechanistic models in the first step of an ecological research. We show the kinship with the lag distribution models, emphasising however on a difference which results in a simplification: the lag itself is not modelled, neither estimated. The interest is focused on the size of the response. Even when the response is the date of realisation of an event such as the occurence of a phenological stage, this date is considered as a quantitative response, not as a time lag between a cause and an effect. Although it can be useful for building models in which the effects are delayed, it is not conceived as an "all purpose" toolbox susceptible to solve all type of related problems.

All the examples given here are purely local. As more and more data become available along with the corresponding meteorological records, it would be interesting to consider multisite strategies to study critical periods. A first idea, the simplest, is to use our method (or another) separately in each place, obtain local predictors, and build a multisite variance-covariance model where the sites play the role of a random factor. This could be valuable for distant sites few correlated together. Other more sophisticated ideas can also be considered but risk to ruin the simplicity and exploratory character of the method. For multisite phenological data, the approach

of Hufkens *et al.* seems more recommandable for such a purpose.

The method leads to parsimonious models of statistical explanation and prediction. This advantage is shared by all the published methods aiming at finding delayed effects in ecology and other sciences. It adds a tool for such investigations. It can take place in a toolbox aside with the proposals of Ogle *et al.et al.*[14] and of Hufkens *et al.*[10]. The system proposed by Ogle*et al.* [14] is a comprehensive modelling framework for the explanation of an ecological phenomenon subject to several influences of the climate components. It requires a strong corpus of hypothesis about the event studied. When the ecological subject is relatively well known, the work of these authors can lead to a robust modelling of the system. This is generally not the case in animal population dynamics where statistical relationships between populations and climate are rarely explained by mechanical models, for diverse reasons. In the case of cereal aphids, for instance, a statistical relation emerges from the late winter temperature and the outbreaks in June[15][16]. A mechanistic model of population dynamics proved to be useful when an sampling of population is available in March-April (Plantegenest *et al.* 2001[18]). Both models have could never could be connected simply because cereal aphids in winter are so rare that no accurate sampling of their density is available. Such difficulties are frequent in the case of insects, which cross periods of very low density in the year. Ogle *et al.*[14] insist also on the concept of ecological memory, a concept highly relevant to the case of critical periods as all goes as if the accumulation of the climatic series is retained during the critical period and forgotten after that.

The question arises also to extend the method to a direct multivariate approach by combining all meteorological data together. This can be considered by two ways: Firstly by eigenvector techniques such as PCA and try to maximize the effect of a linear combination of several weather records on the biological observations, secondly by searching in a $2^p$ space ( $p$ being the number of meteorological series) the most influential set of periods caracterised by their boundaries. The inconvenience would be the difficulty of visualising such a space of responses. Nevertheless, we are currently trying to achieve both of these approaches. The tool presented here can find its place in the armoury of statistical methods which allow to study the relationships between climate and populations. With few modifications it could also serve in other scientific fields where delayed effects are suspected whatever could be the time and space scale where they occur.

# 4 Deposit for the data

The data corresponding to *Sitobion avenae L.* are deposited in CRAN, along with the package Rcriticor. Those corresponding to *Ulex europaeus L.* and to *Rhopalosiphum padi L.* are deposited in Mendeley. The ash tree data are available in the PEP725 phenologic database (http://www.PEP725.eu), and the corresponding climatic data on the site of the .eu) Meteorological Service of Catalunya (https://es.meteocat.gencat.cat/?lang=es). the R scripts necessary to reproduce our calculations of these data are provided as supplementary material.

# 5 Aknowledgements

We thank warmly Jean-Pierre Masson, emeritus professor of statistics and probabilities at Agrocampus Rennes, for his carefull reading of the mathematical developments of this article. We thank Anne Atlan for allowing us to use her european gorse data, and Nicolas Parisey for his reading of the manuscript. We thank James Bell and Lynda Alderson of the Rothamsted Insect Survey (RIS, UK) for sending us the original data of aphids males flights. We thank also Dr Sarah Perryman e-RA curator at Rothamsted who made the English weather data available for us.

# References

[1] A'Brook J., (1981) Forecasting the incidence of aphids using weather data . *EPPO Bull.*, 13,2,229-233

[2] Anonymous, (2017) *CRITICOR, notice d'utilisation*. https://www.yumpu.com/fr/document/view/35920185/Criticor-notice-dutilisation-0-sommaire-inra

[3] Atlan, A., Hornoy, B., Delerue, F., Gonzalez, M., Pierre, J.S. and Tarayre, M. l. (2015). Phenotypic Plasticity in Reproductive Traits of the Perennial Shrub *Ulex europaeus* in Response to Shading: A Multi-Year Monitoring of Cultivated Clones. - *Plos One* 10: e0137500

[4] Chuine, I., Garcia de Cortazar-Atauri I., Kramer K., Hänninen H. (2013). *Plant Development Models. Phenology: An Integrative Environmental Science.* M. D. Schwartz. Dordrecht, Springer Netherlands: 275-293.

[5] Coakley S. M., McDaniel L. R., Shaner, G. (1985). Model for predicting severity of

*Septoria tritici* blotch on winter-wheat. *Phytopathology* 75 (11): 1245-1251.

[6] Cochrane, J. H. (2012). *Continuous-Time Linear Models.* Chicago, Chicago Booth University: 33.

[7] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7, 1-26.

[8] Ennaïfar, S., Makowski, D., Meynard, J. M. and Lucas, P. (2007) Evaluation of models to predict take-all incidence in winter wheat as a function of cropping practices, soil, and climate. - *European Journal of Plant Pathology* 118: 127-143.

[9] Goldwin G.K., 1982. A technique for studying the association between components of the weather and horticultural parameters. *Sciences Horticultur*e,16, 101-107

[10] Hufkens, K., Basler D., Milliman T., Melaas E.K., Richardson, A.D.. (2018). An integrated phenology modelling framework in R. *Methods in Ecology and Evolution* 9(5): 1276-1285.

[11] Ji, L. and A. J. Peters (2005). Lag and Seasonality Considerations in Evaluating AVHRR NDVI Response to Precipitation. *Photogrammetric Engineering & Remote Sensing* 71(9): 1053-1061.

[12] Keenan T.F., Richardson A.D., Hufkens, K. (2020). On quantifying the apparent temperature sensitivity of plant phenology. *New Phytologist* 225(2): 1033-1040.

[13] Lorenz, K. (1970). *Studies in animal and human behaviour,* Vol 1 & 2 , Methuen & Co LTD, Methuen. 1 & 2: 250-252.

[14] Ogle, K., Barber, J.J.,Baron-Gafford, G.A., Bentley, L.P., Young,J.M., Huxman, T.E., Loik, M.E., Tissue, D.T. (2015). Quantifying ecological memory in plant and ecosystem processes. *Ecology Letters* 18(3): 221-235.

[15] Pierre J.S., Guillome M., Querrien M.T., (1986) Une methode statistique et graphique de recherche des periodes de l'annee ou les populations animales sont particulierement sensibles a une composante donnee du climat (periodes critiques). Application au cas des pucerons des cereales. *Acta Oecologica, Oecol. Gener.*, 7, 365-380.

[16] Pierre, J. S. and C. A. Dedryver (1985). Un modèle de prévision des pullulations du puceron *Sitobion avenae* F. sur blé d'hiver. *Acta Oecologica, Oecol. Gener.*, 13-17

[17] Pietravalle S., Shaw M.W., Parker S.R.,van den Bosch F. (2003). Modeling of Relationships Between Weather and *Septoria tritici* Epidemics on Winter Wheat: A Critical Approach. *Phytopathology* 93(10): 1329-1339.

[18] Plantegenest, M., Pierre, J.S., Dedryver, C.A., Kindlmann, P. (2001). Assessment of the relative impact of different natural enemies on population dynamics of the grain aphid Sitobion avenae in the field. *Ecological Entomology* 26(4): 404-410.

[19] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[20] Rispe C., Hulle M., Gauthier J.P., Pierre J.S., Harrington R. (1998) Effect of climate on the proportion of males in the autumn flight of the aphid *Rhopalosiphum padi* L. (Hom., Aphididae). *J. appl. Entomol.*, 122, 129-136.

[21] Seddighi H.R., (2012) *Introductory Econometrics: A Practical Approach* (2nd edn), Routledge, New York, 385p.

[22] Sims C.A. (1971) Discrete approximations to continuous time distributed lags in econometrics. Econometrica 39,3,545-563.

[23] Thomas, G. G., Goldwin, G. K.,Tatchell, G. M. (1983) Associations between Weather Factors and the Spring Migration of the Damson-Hop Aphid, *Phorodon humuli*. *Annals of Applied Biology* 102(1): 7-17.

[24] Tsai, C.-W., et al. (2016). Phenological responses of ash (Fraxinus excelsior) and sycamore (Acer pseudoplatanus) to riparian thermal conditions. *Urban Forestry & Urban Greening* 16: 95-102.

[25] Wang J., Price K.P., Rich, P.M. (2001). "Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains." *International Journal of Remote Sensing* 22(18): 3827-3844.

# Appendix A: Shape and extrema of the empirical correlogram

## A.1 The case of a rectangular window.

Let us consider the following correlation function:

$$\rho(a,b) = \frac{Cov(S_{a,b}, Y)}{\sqrt{Var\left(S_{a,b}\right)\sigma^2{}_Y}} \tag{A1}$$

where $\sigma_Y^2$ is the variance of Y and $S_{a,b} = \int_a^b dU(t)$ , $dU\left(t\right)$ being defined as in equation (2), § 1.1.1 . This interval will be called further the trial period. $\rho(a,b)$ is the correlation

coefficient between $Y$ and $S_{a,b}$.

This correlogram can be used empirically to search for the period which influences most the observed variable $Y$. If such a period exists, then a maximum or minimum should be observed on a map plotting $\rho(.,.)$ against $a$ and $d = b - a$. Such empirical map can be seen in Figure 2a. We eliminate the trend or drift considered as fixed and keep only the stochastic parts. Then,

$$\forall b > a, Var(S_{a,b}) = \sigma^2 (b-a) = d\sigma^2 \tag{A2}$$

where $\sigma$ is the scale of $X(t)$, and $d$ the length of the interval $[a,b]$. To go further, we have to choose a shape for $\Phi(t)$. Let us treat the case of the rectangular window defined in equation (4) §1.1.1. We have to correlate a vector whose components are

$$S_{i,a,b} = \sigma \int_a^b dW_i(t) \tag{A3}$$

and one of which they are

$$Y_i = k\sigma \int_\alpha^\beta dW_i(t) + \varepsilon \tag{A4}$$

The covariance implies to get the expectation of their scalar product, the value of which is proposed below:

**Proposition:**

$$E(S_{a,b}.Y) = k\sigma^2\delta \tag{A5}$$

Where $\delta$ is the length of the intersection between intervals $[a,b]$ and $[\alpha,\beta]$.

**proof:**

Let us consider a single replication:

$$E(S_{a,b,i}.Y_i) = k\sigma^2 \int_\alpha^\beta \int_a^b dW_i(s)\, dW_i(u) + \varepsilon_i \sigma \int_a^b dW_i(t) \tag{A6}$$

and for any replication

$$E(S_{a,b}.Y) = k^2\sigma^2 E\left[\int_\alpha^\beta \int_a^b dW_i(s)\, dW_i(u)\right] + E\left[\int_a^b dW_i(t)\right] E(\varepsilon)$$

$$= k\sigma^2\delta \tag{A7}$$

As the increments of $W_i(t)$ are independent,

$$E\left[dW_i(t)\,dW_i(u)\right] \neq 0 \tag{A8}$$

if and only if $t = u$. in this case:

$$E\left[dW_i(t)\,dW_i(u)\right] = E\left[dW_i(t)\right]^2 = dt \tag{A9}$$

This condition is verified only in the common part of the intervals $[\alpha,\beta]$ and $[a,b]$, the interval $\Delta = [a,b] \cap [\alpha,\beta]$ of length $\delta$. The properties of the brownian motion imply then that $k^2\sigma^2 \int_\delta E\left(dW(t)\right)^2 = k^2\sigma^2\delta$. In the second term of the sum

$$E\left(\int_a^b dW(t)\right) = 0 \tag{A10}$$

what completes the proof.

As $dW(t)$ is centered, this expectation is the covariance of $S_{ab}$ and $y$, and the correlation coefficient is

$$\rho_{a,d} = \frac{k^2\sigma^2\delta}{\sqrt{d\sigma^2\sigma_Y^2}} = k^2\,\frac{\sigma}{\sigma_Y}\,\frac{\delta}{\sqrt{d}} \tag{A11}$$

So, the correlation coefficient is proportional to the length of the intersection of the window interval and of the trial interval and inversely proportional to the square root of the length of the trial interval.

As $\delta \leq d$, this quantity is maximum when $d = \delta$, i.e. when

$$\begin{aligned} a &= \alpha \\ b &= \beta \end{aligned} \tag{A12}$$

The interval $[a,b]$ for which $\rho$ is maximum is an unbiassed estimator of the influence interval $[\alpha,\beta]$. From the habitual properties of correlation coefficients it results also that $a$ and $b$ are least square estimates of $\alpha$ and $\beta$. Figure 5 shows , in perspective and contour plot the shape of the function $\rho(a,d)$, with its definite peak on the point $(\alpha,\delta)$.

We chose the representation in the plan $(a,d)$ rather than $(a,b)$ not to leave the graph empty under the bissector.

The shape of the surface (Figure 6) is straightforward from formula (A11). All is governed by the relative sizes of $d$ and $\delta$. The interval $[a,b]$, the length of which is $d$, the abscissa of

figure 5, grows monotonously from left to right. Two lines, the horizontal $a = \alpha$ and the negative bissector $a = \beta - d$ share the plane in four quadrants. In the lower left quadrant, a is lower then $\alpha$ and $\delta$ is null until $d + a \geq \alpha$. Then, $\delta$ grows faster then $\sqrt{d}$. In the lower right quadrant, $a + d > \beta$ and thus, $\delta = \beta - \alpha$, a fixed value, and the ratio decreases from left to right with $\sqrt{d}$. In the upper left quadrant, $a > \alpha$ and $a + d \leq \beta$. Thus $\delta = d$ and $\frac{\delta}{\sqrt{d}} = \sqrt{d}$, growing from left to right until $a + d > \beta$, which occurs in the upper right quadrant. The whole set results in a pyramidal shape with an acute peak. Of course, this is only a theoretical result, altered by uncontrolled sources of noise in the real world.

Figure 6: The shape of the expected correlogram under the hypothesis of a rectangular window:
a: perspective view;
b: contour plot.

## A.2 The case of a bell-shaped window.

In this case, the window has no definite bounds. The interesting things are to estimate $t^*$ and $c$ the pseudo standard deviation of the gaussian window. Let us return to expression (A6). It takes now the form:

$$E\left(S_{a,b}.Y\right) = k^2\sigma^2 E\left[\int_{-\infty}^{T}\Phi(t)\,dW(t)\int_{a}^{b}dW(u)\right] + E\left[\int_{-\infty}^{T}\Phi(t)\,dW(t)\right]E\left(\varepsilon_i\right) \qquad \text{(A14)}$$

Figure 7: The shape of the expected correlogram under the hypothesis of a bell shaped window:
a: perspective view;
b: contour plot.

And, for the same reasons of independence as in the previous case, and because

$$E\left(\varepsilon_i\right) = 0 \qquad \text{(A15)}$$

this reduces in:

$$E\left(S_{a,b}.Y\right) = k^2\sigma^2 E\left(\int_a^b \Phi(t)\left[dW(s)\right]^2\right) \tag{A16}$$

As $\Phi(t)$ is fixed for all $t$, we can factor the expectation in:

$$E\left(S_{a,b}.Y\right) = k^2\sigma^2\left(\int_a^b \Phi(t)E\left[dW(s)\right]^2\right) \tag{A17}$$

and using the properties of the Wiener process $E\left(dW^2(s)\right)$ is its variance at time $s$ thus:

$$E\left(dW^2(s)\right) = dt \tag{A18}$$

and (A16) reduces to

$$E\left(S_{a,b}.Y\right) = k^2\sigma^2\left(\int_a^b \Phi(t)\,dt\right) \tag{A19}$$

that is:

$$E\left(S_{a,b}.Y\right) = k^2\sigma^2\left(F(b,c) - F(a,c)\right) \tag{A20}$$

Therefore, $F(.,c)$ being the normal cumulative distribution function of standard deviation $c$. The expression is the covariance as both $S_{ab}$ and $Y$ are supposed centered. The value of $S_{ab}$ is:

$$S_{ab} = \int_a^b dW(t) \tag{A21}$$

and thus its variance is:

$$\text{var}\left(S_{ab}\right) = b - a = d \tag{A22}$$

The correlation coefficient is thus:

$$\rho_{a,b} = \frac{k^2\sigma^2\left(F(b,c) - F(a,c)\right)}{\sigma_Y\sqrt{d}} \tag{A23}$$

Does this function have a maximum in the plane $(a,d)$ with $d = b - a$ ? The following conjecture will help:

**Conjecture:**

As the Gauss function is symmetric, let us conjecture that the interval $\left[a^*,b^*\right]$, if any, that maximizes $\rho_{a,b}$ is symmetric, left and right of $t^*$. The mean of $F(.,c)$ is supposed to be $t^*$ meaning that $F(.,c)$ must be understood as $F(.,t^*,c)$

**Proof:**

The extrema of the coefficient $\rho - a,b$, regarding to the borders of the interval is given by the couple of equations :

$$\begin{cases} \dfrac{\partial \rho_{a,b}}{\partial a} = 0 \Leftrightarrow \\ \dfrac{\partial \rho_{a,b}}{\partial b} = 0 \Leftrightarrow \end{cases} \tag{A24}$$

Excluding $b = a$, this leads to the following simultaneous equations:

$$\begin{cases} F(b,c) - F(a,c) = 2f(a,c)(b-a) \\ F(b,c) - F(a,c) = 2f(b,c)(b-a) \end{cases} \tag{A25}$$

This is only possible if $f(a,c) = f(b,c)$, implying that $b$ is symmetric to $a$ with regards to $t^*$ QED

**Determination of** $h$

Equations (A25) however leave $a$ and $b$ indeterminate. So let us put $b - a = 2h$ and, knowing that $a$ and $b$ are symmetric, let us try to optimize the half span $h$. Without loss of generality, let us put $t^* = 0$ in equation (5,?2.2) (a simple change of origin), and considering that $k, \sigma$ and $\sigma_Y$ are positive constants, we reduce the problem to maximize:

$$\varphi_{a,b} = \frac{F(h,c) - F(-h,c)}{\sqrt{2h}} = \frac{2F(h,c) - 1}{\sqrt{2h}} \tag{A26}$$

the derivative of which is:

$$\varphi'_h = \sqrt{\frac{2}{h}} f(h,c) - \sqrt{2}\,\frac{2F(h,c) - 1}{4h^{\frac{3}{2}}} \tag{A27}$$

Equating to zero we find:

$$f(h,c) = \frac{2F(h,c) - 1}{4h} \tag{A28}$$

or with the change of variable $\dfrac{h}{c} = s$

$$z(s^*) = \frac{2F(s^*) - 1}{4cs^*} \tag{A29}$$

This equation has a solution in $s$ which can be found numerically. For $c = 20$, $s \simeq 3.024$ and

$h \simeq 60.479$. In the case of a gaussian bell-shaped window with the deviation parameter $c$, the procedure detects the span of influence of the time series as an interval $[t^* - h, t^* + h]$, where $h$ is solution of equation (A28).

The shape of the surface (Figure 7a) looks like a smoothed version of the one obtained at §A.1 and shown Figure **??**a. It is organised by the same crossed lines as in the rectangular window case. This case is rather more realistic than that described in 2.1 as nobody expects that the influence of a time series on a phenomenon could begin precisely at a given date and cease at another precise date. The interest of the bell shaped window is to show that, with a somwhat simple research procedure, a peak is detectable, bordering a window of influence whose boundaries are functions of the deviation parameter $c$.

## A.3 Generalized Beta case

The Generalized Beta density distribution can be used as weighting function, just as the gaussian distribution. It has one advantage: that of being precisely restricted between two boundaries. Its other feature is an extreme versatility, with a shape varying from exactly rectangular window to humped one, symmetrically or not, L or J shape, and even in U shape depending on the values of its two parameters, $\alpha$ and $\beta$. The scheme of reasoning is essentially the same as for the gaussian case, as we can start at equation (A23), just replacing the function $F$ which is the integral of the gaussian function by the integral of the generalized Beta distribution. This leads to :

$$\rho_{(a,b,\alpha,\beta)} = \frac{k^2\sigma^2\left[B(b,\alpha,\beta) - B(a,\alpha,\beta)\right]}{\sigma_Y\sqrt{(d)}} \qquad \text{(A30)}$$

where $B(.,\alpha,\beta)$ designs the integral of the Generalised Beta density of parameters $\alpha$ and $\beta$ and of support $[a,b]$ but, by definition, this difference is equal to 1 and thus:

$$\rho_{(a,b,\alpha,\beta)} = \frac{k^2\sigma^2}{\sigma_Y\sqrt{d}} \qquad \text{(A31)}$$

from this formula, the correlation coefficient seems to depend only on the length of the intersection of the trial window and of the "true" range of action of the Beta function but, this optimal range depends itself on the coefficients of the Beta distribution. There is thus no useful analytical solution but the range and the coefficients $\alpha$ and $\beta$ of the Beta distribution can be found by an optimization procedure. We used for that the function *optim* of $R$, with as method either

Nelder-Mead (the default), or SANN (simulated annealing) or L-BFGS-B. This last method allows to specify restrictions (box constraints) on the parameters and is necessary to maintain the two shape parameters of the Beta distribution greater than zero or sometimes greater or equal to 1 to obtain a humped shape.

### A.4 The polynomial case

Let us choose for $\Phi(.)$ a polynomial of degree $p$ scaled to unity on $[a,b]$:

$$P(t) = a_0 + a_1 t + a_2 t^2 + ... + a_p t^p$$
$$\Phi(t) = \frac{P(t)}{\int_a^b P(s)ds} \quad \text{(A33)}$$

Let us take an example for $p = 2$. We have:

$$P(t) = a_0 + a_1 t + a_2 t^2 \quad \text{(A34)}$$

$$\int_a^b P(s)ds = a_0(b-a) + \frac{1}{2}(b^2 - a^2) + \frac{1}{3}(b^3 - a^3) \quad \text{(A35)}$$

$$\Phi(t) = \frac{P(t)}{a_0(b-a) + \frac{1}{2}(b^2 - a^2) + \frac{1}{3}(b^3 - a^3)} \quad \text{(A36)}$$

or, putting $I = a_0(b-a) + \frac{1}{2}(b^2 - a^2) + \frac{1}{3}(b^3 - a^3)$ and $A_i = \frac{a_i}{I}$:

$$\Phi(t) = A_0 + A_1 t + A_2 t^2 \quad \text{(A37)}$$

and so on for higher degrees. Reasonably, to avoid overparametrisation and keep a simple shape for the window, we limited our investigations to the 5th degree. Under this form (A36), the function $\Phi(t)$ sums to 1 in the interval $[a,b]$ and is the normalised form of the polynomial. So as that for the Beta shape, there is no interesting analytical solution and the same procedure of optimisation was used.

## Appendix B: Maximum correlation and maximum likelihood

### B.1: Identity of both (normal errors case)

Let us go back to formula (3) adapting it to a rectangular window and supposing an intercept $\mu$:

$$Y_i = \mu + k\int_\alpha^\beta X_i(t)\,dt + k\sigma\int_\alpha^\beta dW_i(t) + \varepsilon_i \tag{B1}$$

$\varepsilon$ being a random variable that we'll suppose distributed distributed normally as $N\left(0,\sigma_r^2\right)$ with the standard hypothesis of independence of variance and no autocorrelation. This can be rewrote as:

$$Y_i = \mu + k\xi_i + \varepsilon_i \tag{B2}$$

Where

$$\xi_i(\alpha,\beta) = \int_\alpha^\beta X_i(t)\,dt + \sigma\int_\alpha^\beta dW_i(t) \tag{B3}$$

$\xi_i(\alpha,\beta)$ is a predictor of $Y_i$. It involves the result of a stochastic integral, but, as the different replicates of the time series $U_i(t)$ are realized, $\xi_i(\alpha,\beta)$ as predictor, must be considered as fixed. Supposing that $\xi_i(\alpha,\beta)$ is known ($\alpha$ and $\beta$ are known), the maximum likelihood estimators of $\mu$ and $k$ are well known: Let us $A$ be the two columns and $n$ rows matrix :

$$A = \begin{pmatrix} \mu & \xi_1 \\ \mu & \xi_2 \\ \cdots & \cdots \\ \mu & \xi_n \end{pmatrix} \tag{B4}$$

the estimated vector of parameters is:

$$\begin{pmatrix} \hat{\mu} \\ \hat{k} \end{pmatrix} = \left(A^T A\right)^{-1} A^T Y \tag{B5}$$

$\alpha$ and $\beta$ are unknown, and will be replaced, in what follows, by the trial parameters $a$ and $b$. The preceding equation, however, means that once a maximum likelihood is found for those two parameters, the solution for $\mu$ and $k$ is straightforward, being given by equation (B5).

From (B2), we know that conditionnaly to $\xi_i$, $Y_i$ is distributed as : $Y_i \sim N\left(\mu + k\xi_i, \sigma_r^2\right)$ Now, let us indicate the values of $\xi_i$ as $\xi_i(\alpha,\beta)$ for the unknown "true" bounds $\alpha$ and $\beta$ and $\xi_i(a,b)$ for any other period $[a,b]$ such as $b > a$. We can wright the likelihood of the observed vector $Y$ as follows:

$$L\left(Y|a,b\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_r^2}}e^{\frac{\left[Y_i-\mu-k\xi_i\left(a,b\right)\right]^2}{2\sigma_r^2}} \tag{B5.1}$$

And the log-likelihood:

$$LL\left(Y|a,b\right)=-\frac{n}{2}\ln\left(2\pi\sigma_r^2\right)+\sum_{i=1}^{n}\frac{\left[Y_i-\mu-k\xi_i\left(a,b\right)\right]^2}{2\sigma_r^2} \tag{B5.2}$$

We have then to equate to zero the partial derivative of this expression with respect to the bounds $a$ and $b$. Let us begin by $a$:

$$\frac{\partial LL\left(Y|a,b\right)}{\partial a}=0\Leftrightarrow-\frac{1}{\sigma_r^2}\sum_{i=1}^{n}k\left[Y_i-\mu-k\xi_i\left(a,b\right)\right]\frac{\partial\xi_i\left(a,b\right)}{\partial a}=0 \tag{B6}$$

And then we have to solve:

$$\frac{1}{\sigma_r^2}\sum_{i=1}^{n}k\left[Y_i-\mu-k\xi_i\left(a,b\right)\right]\frac{\partial\xi_i\left(a,b\right)}{\partial a}=0 \tag{B7}$$

Let us now replace $Y_i$ by its theoretical value:

$$\frac{1}{\sigma_r^2}\sum_{i=1}^{n}k\left[\mu+k\xi_i\left(\alpha,\beta\right)+\varepsilon_i-\mu-k\xi_i\left(a,b\right)\right]\frac{\partial\xi_i\left(a,b\right)}{\partial a}=0 \tag{B8}$$

A sufficient solution occurs for

$$\xi_i\left(\alpha,\beta\right)=\xi_i\left(a,b\right) \tag{B9}$$

For we get:

$$\sum_{i=1}^{n}\frac{\partial\xi_i\left(a,b\right)}{\partial a}\varepsilon_i=0 \tag{B10}$$

$$\frac{\partial\xi_i\left(a,b\right)}{\partial a}=X_i\left(a\right)+\sigma dW_i\left(a\right) \tag{B11}$$

The function $X$ being deterministic,

$$X_i\left(a\right)=X\left(a\right),\forall i \tag{B12}$$

equation (B10) becomes:

$$X\left(a\right)\sum_{i=1}^{n}\varepsilon_i+\sigma\sum_{i=1}^{n}dW_i\left(a\right)\varepsilon_i=0 \tag{B13}$$

The firs term equals to $0$ by construction of residuals, and the second is null in average, by independence of the increments $dW_i\left(a\right)$ and of the residuals of the regression $\varepsilon_i$. So, $a=\alpha$

ensures the nullity of the derivative, in expectancy. Let us now derive with respect to $b$

$$\frac{\partial LL\left(Y_i \middle| a,b\right)}{\partial b} = 0 \Leftrightarrow -\frac{1}{\sigma_r^2}\sum_{i=1}^{n} k\left[Y_i - \mu - k\xi_i\left(a,b\right)\right]\frac{\partial \xi_i\left(a,b\right)}{\partial b} = 0 \tag{B14}$$

After the same series of calculations as for $a$, we get:

$$\sum_{i=1}^{n} dW_i\left(\beta\right)\varepsilon_i = 0 \tag{B15}$$

Which is realized in expectation. This has as consequences that the estimates of $\alpha$ and $\beta$ are slightly biased and asymptotically unbiassed for large $n$s. The important point is that the determination of $\alpha$ and $\beta$ by maximum likelihood coincides with the maximisation of the correlation coefficient. After that, it is straightforward to find $\hat{\mu}$ and $\hat{k}$ from equation (B5).

## B.2: Consequences

From $\hat{k}$ we get $\hat{\rho}$ the maximum correlation coefficient from the common formula:

$$\rho = k\sqrt{\frac{\mathrm{var}\left(\xi\right)}{\mathrm{var}(Y)}} \tag{B16}$$

And we have estimates of the variances of $\hat{\alpha}, \hat{\beta}, \hat{\mu}$, and $\hat{k}$ by calculating the Fisher's information matrix from the log-likelihood value in the neighbourhood of these parameters. The difficulty comes from that the partial differentials of the log-likelihood with respect to $\hat{\alpha}$ and $\hat{\beta}$ include the realization of a stochastic process ( $\xi(\alpha,\beta)$ ) and thus are not the derivatives of continuous functions. We chose to treat the differentials in mean, one way to stabilise the Fisher information matrix. These estimates can be compared with those obtained by bootstrapping. Those are in development for the next version of Rcriticor.

# Highlights

- The study of relationships between climate and animal or plants populations is of prominent importance especially in the context of climate change.
- We focuse here explicitely on the hypothesis that one series may have an influence on a further event by a bounded window of time, long time before the effect. This window is called “critical period”.
- We propose a bivariate correlation function to detect such periods. Some close form estimators are proposed for it, bootstrapping and Monte-Carlo methods are proposed to estimate its variability, a CRAN package is available for that.
- We give three examples of ecological applications of the method.