PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023

GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Sponsored by:

On time lag detection between time series sampled by eddy covariance systems

Domenico Vitale[1*], Dario Papale[2]

[1] *Department of Methods and Models for Economics, Territory and Finance (MEMOTEF), Sapienza University of Rome, via del Castro Laurenziano, 9, 00161, Roma, Italy, domenico.vitale@uniroma1.it*
[2] *Department for Innovation in Biological, Agro-food and Forest Systems (DIBAF), University of Tuscia, via San Camillo de Lellis, 01100, Viterbo, Italy, darpap@unitus.it*
[*] *Corresponding author*

Abstract. *This work introduces a new procedure for time lag detection between high-frequency time series sampled by means of the eddy covariance technique. The proposed methodology is based on the assessment of the cross-correlation function between variables subject to (i) a preliminarily filtering procedure based on pre-whitening, to avoid the risk of spurious correlations, and (ii) to a resampling technique based on block-bootstrapping, to enhance the accuracy of time lag detection between variables with correlation of low order of magnitude. We expect that the proposed procedure will become useful for the centralized data processing pipelines of research infrastructures (e.g. ICOS-RI) where the use of completely data-driven procedures constitutes an essential prerequisite.*

Keywords. *Pre-whitening; Block bootstrap; Time series alignment; Low signal-to-noise ratio.*

---

1 Introduction

Accurate quantification of greenhouse gases (GHGs) emitted to and removed from the atmosphere by natural ecosystems are nowadays calculated by using the eddy covariance (EC) technique [2]. This technique employs a sonic anemometer for wind velocity components and a gas analyzer for scalar atmospheric concentrations and involves high-frequency sampling (at least 10 observations per second). EC fluxes are derived from the covariance (normally within an averaging period of 30 minutes) between vertical wind speed (w) and the atmospheric scalar variable of interest (s), which can be temperature, water vapor (H_2O), carbon dioxide (CO_2), nitrous oxide (N_2O), methane (CH_4) or any other trace gas.

The calculation of EC fluxes requires the instantaneous quantities of w and s be simultaneously measured. Such a condition is not perfectly fulfilled during field measurements because, mainly to avoid possible wind flow distortions, there is no perfect co-location of the anemometer and the gas analyzer. Correcting mis-alignments between raw EC data is a key step in the calculation of derived fluxes. A failure to reach this target entails, in fact, a systematic error on flux estimates, whose sign (negative in case of underestimation or positive in case of overestimation) depends on the procedure adopted for the detection of the time lag existing between time series involved in the covariance estimation [4].

The prevalent solution for time lag detection in EC data processing pipelines is accomplished by assessing the cross-covariance function between w and s. In particular, the optimal time lag can be detected in correspondence of the lag that maximizes (in absolute terms) the cross-covariance function

between time series [5]. The effectiveness of the procedure depends on the shape of the cross-covariance function, which in turn depends on the stochastic properties of the time series involved. Generally, the procedure is effective under second order stationary conditions when the signal-to-noise ratio (SNR) is moderate/high. In these circumstances, in fact, the cross-covariance function exhibits a distinct and pronounced peak (either positive or negative) and the optimal time lag can be easily detected. In other circumstances, in particular when fluxes are of small magnitude, as occurs for trace gases or during dormant/senescence periods for instance, the cross-covariance function can be characterized by multiple local minima or maxima of similar magnitude. Consequently, the detection of the optimal time lag becomes problematic.

The aim of this work is to overcome the limitations of the procedures based on the maximization of the cross-covariance function (CovMax, hereinafter). To this end, we propose a completely data-driven procedure where time lag is detected by assessing the cross-correlation function (CCF) between raw EC data subject to (*i*) a preliminary filtering procedure based on pre-whitening and (*ii*) a resampling technique based on block-bootstrapping. By combining pre-whitening and resampling, the assessment of the CCF for time lag detection becomes more realistic, informative and suitable for variables having correlation of low order of magnitude, as in the case of EC fluxes characterized by low SNR.

2 Methods

2.1 Avoiding spurious correlations by pre-whitening

Let $Y = Y_t$ and $X = X_t$ be two time series indexed by time t, the correlation between X and Y at lag k can be estimated by the sample CCF defined by:

$$r_k(X,Y) = \frac{\sum(X_t - \overline{X})(Y_{t-k} - \overline{Y})}{\sqrt{\sum(X_t - \overline{X})^2}\sqrt{\sum(Y_t - \overline{Y})^2}}, \tag{1}$$

where $\overline{X}$ and $\overline{Y}$ are the sample mean of X and Y, respectively, and the summations are done over all data where the summands are available.

For white noise (WN) processes, $r_k(X,Y)$ is approximately normally distributed with zero mean and variance $1/n$, where n is the total number of paired data. This leads to the conventional 5% significance limits of the CCF estimates equal to $\pm 1.96/\sqrt{n}$. That is, any peak outside the interval $\pm 1.96/\sqrt{n}$ is judged to be statistically significant at 0.05 level. The approximate variance of $1/n$ applies only when data are independent and identically distributed (*iid*), a condition that is almost never met for real, observed time series. Under the assumption that both X and Y are stationary and that they are independent of each other, it turns out that the sample variance of $r_k(X,Y)$ is approximately

$$\frac{1}{n}\Big[1 + 2\sum_{k=1}^{n} \rho_k(X)\rho_k(Y)\Big], \tag{2}$$

where $\rho_k(X)$ and $\rho_k(Y)$ are the autocorrelation estimates at lag k of X and Y, respectively.
Suppose for simplicity that X and Y are both first-order autoregressive (AR) processes with coefficients ϕ_X and ϕ_Y, respectively, then $r_k(X,Y)$ is approximately normally distributed with zero mean and variance approximately equal to

$$\frac{1 + \phi_X\phi_Y}{n(1 - \phi_X\phi_Y)}. \tag{3}$$

From Eq. (3) it can be seen that when ϕ_X and ϕ_Y are close to 1, the ratio of the sampling variance of $r_k(X,Y)$ to the nominal value of $1/n$ approaches infinity. As a consequence, using the $\pm 1.96/\sqrt{n}$ rule in

deciding the significance of the sample CCF may lead to many more false positives than the nominal 5% error rate, even when time series are independent of each other.

The statistical significance of the CCF estimates is a typical representation of the so-called *spurious* correlations problem existing between time series [3, 6]. To avoid the risk of spurious correlations, a viable solution is to disentangle the linear association between X and Y from their autocorrelation. By examining Eq. (2), it can be seen that the approximate variance of $r_k(X,Y)$ is $1/n$ if at least one of X and Y is an *iid* sequence. Such a condition can be achieved by transforming one of the variables in a WN process, a procedure known as pre-whitening. The transformation of X in a WN process can be performed by means of AR models (with differencing possible to remove stochastic time trends):

$$\tilde{X}_t = (1 - \pi_1 B - \pi_2 B^2 - \ldots - \pi_p B^p) X_t = \pi(B) X_t, \tag{4}$$

where $\tilde{X}_t$ is a WN, π_i are the AR coefficients and B is the backshift operator such that $B^m X_t = X_{t-m}$. In this work, the AR order selection was performed by the Akaike Information Criterion (AIC), and parameters estimation via Yule-Walker method. After transforming the X-variable, the same filter was used to transform the Y-variable in $\tilde{Y}_t$, which does not need to be a WN. Since pre-whitening is a linear operation, any linear relationship between the original series will be preserved and can be retrieved by assessing the CCF between transformed $\tilde{X}_t$ and $\tilde{Y}_t$ variables [1]. As for the CovMax procedure, the optimal time lag can be retrieved in correspondence of the peak (in absolute terms) of the CCF between pre-whitened variables.

2.2 Time lag detection between time series with correlation of low order of magnitude

A time lag detection procedure based on the assessment of the CCF between pre-whitened variables is effective when the order of magnitude of the correlation is equal to -1, which holds for EC fluxes with moderate/high SNR. This is because the signal dominates over the noise and the estimate of the CCF in correspondence of the *true* time lag will be far greater than the conventional significance limits. When the correlation is low, as in the case of trace gas fluxes with low SNR, things becomes more complicated because the peak of the CCF in correspondence of the expected time lag will not be so pronounced as to dominate over the other estimates of the CCF at different lags. For example, in the case of a sample size of 36000 paired observations and an order of magnitude of the correlation between variables < -1, the peak of CCF is close to the 5% significance limits ($\pm 1.96/\sqrt{36000} \approx \pm 0.01$). Thereby, it can often happen that the peak of the CCF is detected in correspondence of an erroneous time lag.

If measurements from repeated sampling under the same conditions were available, it would be easier to distinguish between *true* and *false* peaks of the CCF, as the former would remain more stable than the latter, which instead would tend to cancel out after averaging. With this goal in mind, we mimic a repeated sampling by means of a block bootstrap resampling with the twofold aim of (*i*) increasing the accuracy of time lag detection and (*ii*) obtaining a quantification of the associated uncertainty. In particular, we built $R = 99$ bootstrap resamples of paired $\tilde{X}_t$ and $\tilde{Y}_t$ values of size n equal to the length of time series, and where each resample is formed by randomly choosing $k = n/L$ blocks (with replacement) of length $L = 400$ time steps (i.e. 20 sec for EC data sampled at 20 Hz), a temporal window large enough to include the *true* time lag and preserve the correlation structure between variables.

The CCF was then estimated for each of the R bootstrap resamples and, for each of them, the *candidate* time lag is detected in correspondence of the peak (in absolute terms) of the CCF. To further improve the accuracy and reduce the effect of noise, each peak is identified on a smoothed versions of the CCF obtained by means of a centered moving average of width 13 time steps. By analyzing the distribution of the 99 candidate time lags, regardless of the significance level, the most frequently observed value (mode) is selected as the *optimal* time lag, while the 50% highest density interval (HDI) provides a measure of the associated uncertainty.

3 Application

In the following, we show an application of the procedure outlined in Section 2 (hereinafter PWB) to real data sampled from an EC system for the measurements of N_2O fluxes operated at the Easter Bush grazed, managed grassland site (3.2065 W, 55.8655 N, 190 m a.s.l.) in Scotland, for two periods during the 2019 growing season. Results are compared with the widely used CovMax procedure. N_2O was selected as the X-variable, while either vertical wind speed (W) and sonic temperature (T) as the Y-variable. Notice that T and W are temporally aligned because sampled by the same instrument. At the same time, since air temperature and scalar concentration variables are co-related each other because parcels movement obeys to the same thermodynamics laws, T can be used in place of W for the detection of time lag existing for variables sampled by other sensors. For this reason, PWB were performed either with W and T. Among the two *optimal* time lag detected, the ones to which corresponds a higher correlation (in absolute value) is chosen as the reference. All the procedures were performed in a search temporal window of $\pm$10 sec.

Illustrative examples of time lag detected by PWB on real EC data are given in Figure 1. For the paired time series depicted in Fig. 1a, the use of PWB may not be necessary since the optimal time lag (1.6 sec) is in correspondence of the peak of the cross-covariance function. A different situation occurs for time series depicted in Figs. 1b and 1c. Here the time lag detected by means of the CovMax could lead to biased flux estimates since the optimal time lag is in correspondence of a local maxima (Fig. 1b) and of a minimum (in absolute terms, Fig. 1c) of the cross-covariance function. Time lags detected by PWB are instead more stable, exhibiting a dominant peak at 1.6 sec, even in presence of heteroscedasticity.

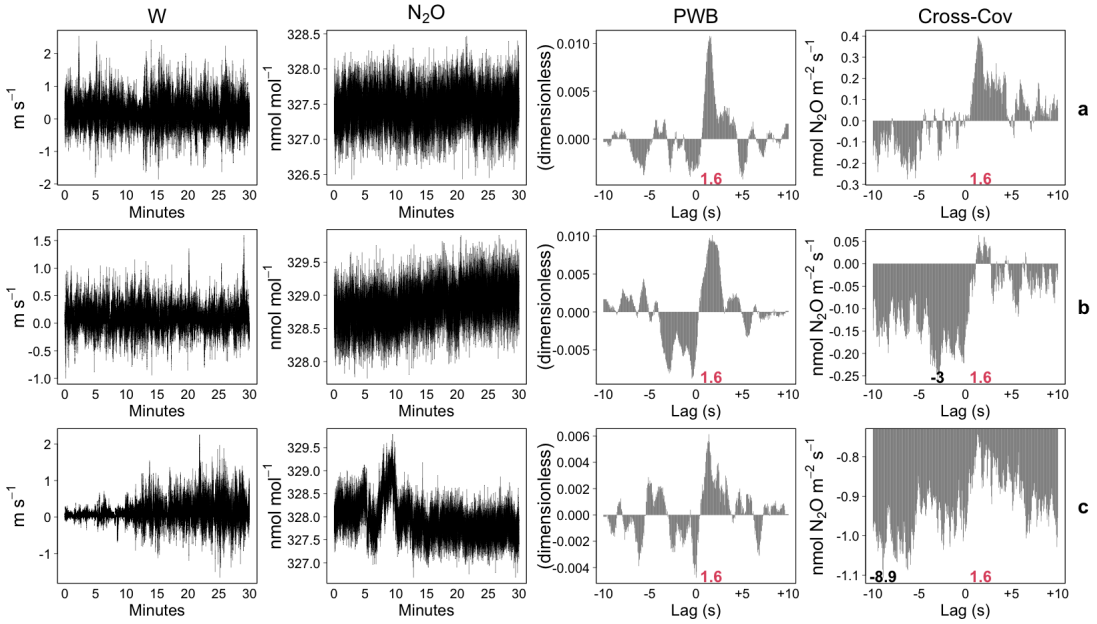

Figure 1: From left to right: vertical wind speed (W), nitrous dioxide (N_2O) atmospheric concentrations, cross-correlation function after pre-whitening with bootstrap (PWB), cross-covariance function. Numbers at the top of x-axis indicate the optimal time lag detected by PWB (in red) and the one in correspondence of the peak (in absolute terms) of the cross-covariance function (in black).

Method	10	20	30	40	50	60	70	80	90
CovMaxW	−9.10	−6.15	−3.10	0.00	+1.55	+1.80	+2.65	+4.15	+6.55
CovMaxT	−9.35	−6.50	+1.35	+1.55	+1.70	+1.85	+2.20	+2.80	+4.60
PWB	−0.85	+1.30	+1.50	+1.55	+1.65	+1.70	+1.85	+2.20	+2.90

Table 1: Deciles of the distribution of time lags detected by several approaches. CovMax indicates the procedure based on the maximization of the cross-covariance function between N_2O and vertical wind speed (W) or sonic temperature (T); PWB indicates the procedure based on the assessment of the cross-correlation function after pre-whitening with bootstrap using either W and T variables.

A comparison of time lags detected by several approaches over longer period is depicted in Figure 2. Results indicate that time lags detected by PWB are more stable (around 1.6 sec) than those obtained by CovMax. As reported in Table 1, 60% of 456 time lags were detected between +1.30 and +2.20 sec, a plausible range for the characteristics of this EC system. The use of T in place of W facilitates the detection of the optimal time lag via CovMax, however, there are not a few cases where the detected time lag diverges at the boundaries of the temporal window (Figs. 2a and 2b), mainly due to the presence of spurious correlations. Such cases are difficult to distinguish during the processing of massive raw data files and may introduce significant biases in flux estimates, if not preliminary removed. Setting a smaller temporal window for time lag detection may alleviate the problem, but not ensure that the detected time lag converges to the expected ones. Regarding the time lags detected by PWB (Fig. 2c), large deviations from 1.6 sec are often characterized by a larger uncertainty (50% HDI greater than 10 time steps, i.e., 0.5 seconds). We found that in most of these situations, such uncertainty is indicative of zero N_2O fluxes or symptomatic about the presence of instrumental errors affecting raw EC data.

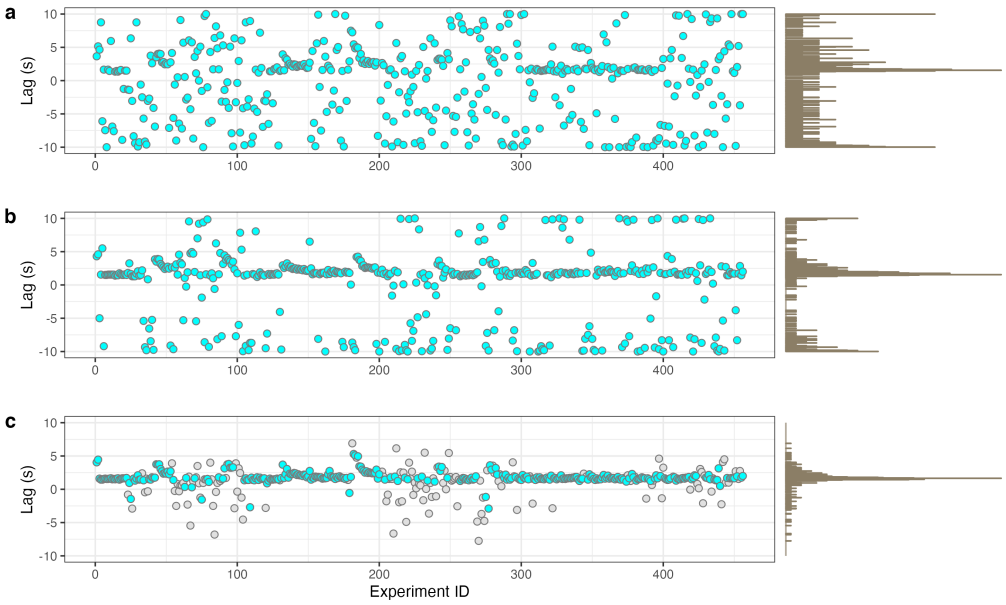

Figure 2: Time lags detected for nitrous dioxide (N_2O) atmospheric concentrations by means of cross-covariance maximization approach using vertical wind speed (panel a), sonic temperature (panel b) and the assessment of the cross-correlation function after pre-whitening with bootstrap (PWB, panel c). Grey points in panel c indicate time lags highly uncertainty (range of the 50% HDI >0.5 sec).

4 Final remarks

GHGs monitoring is crucial to fight against climate change. Beyond new instrumentations with increased accuracy and precision, the development and the application of advanced statistical tools can facilitate the analysis of such complex phenomena. In this work a completely data driven procedure for the detection of time lag for raw EC data was presented. The proposed approach, based on the assessment of the cross-correlation function after pre-whitening with bootstrap (PWB), is designed to overcome the limitations of existing procedures when the correlation between variables is of low order of magnitude (i.e. for EC fluxes with low SNR). We expect that the proposed procedure will become useful for the centralized data processing pipelines of research infrastructures (e.g. ICOS-RI) where the use of completely data-driven procedures constitutes an essential prerequisite.

Acknowledgments. The Integrated Carbon Observation System - Research Infrastructure (ICOS ERIC, `https://www.icos-cp.eu/`) and the ICOS ETC funding from the Italian Ministry of Research. We thank Eiko Nemitz and Carole Helfter for providing the data.

References

[1] Cryer, J. D., Chan, K. S. (2008). *Time series analysis: with applications in R*. Springer New York.

[2] Foken, T., Aubinet, M., Leuning, R. (2012). The eddy covariance method. In: Aubinet, M., Vesala, T., Papale, D. (eds): *Eddy covariance: a practical guide to measurement and data analysis*, pp 1–19. Springer Science & Business Media.

[3] Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.

[4] Langford, B., Acton, W., Ammann, C., Valach, A., Nemitz E. (2015). Eddy-covariance data with low signal-to-noise ratio: time-lag determination, uncertainties and limit of detection. *Atmospheric Measurement Techniques* **8**, 4197–4213. doi: 10.5194/amt-8-4197-2015

[5] Rebmann, C., Kolle, O., Heinesch, B., Queck, R., Ibrom, A., Aubinet, M. (2012) Data acquisition and flux calculations. In: Aubinet, M., Vesala, T., Papale, D. (eds) *Eddy covariance: a practical guide to measurement and data analysis*, pp 59–83. Springer Science & Business Media.

[6] Yule, G. (1926). Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* **89**, 1–64.