Estimating correlation between unevenly spaced paleoclimate time series. Case study: abrupt climate changes during the Last Glacial

JM Polanco(1), MF Sánchez Goñi(2), MA Medina-Elizalde(3), M Mudelsee(4)

(1) Basque Centre for Climate Change (BC3), Leioa, SPAIN; (2) Ecole Pratique des Hautes Etudes (EPHE), PSL University & UMR CNRS 5805 EPOC, U. of Bordeaux, FRANCE; (3) Dept. of Geosciences, Auburn University, USA; (4) Climate Risk Analysis (CRA) & Alfred Wegener Institute (AWI), Germany. *josue.m.polanco@gmail.com

ABSTRACT

We present a computational program named BINCOR (BINned CORrelation) to estimate the correlation between two unevenly spaced climate time series, which is based on a novel estimation approach proposed by[7]. The idea is that autocorrelation (e.g., AR1) means memory, which allows to correlate values obtained on different time points. The binned correlation is performed by resampling the time series under study into time bins on a regular grid that are assigned the mean values of the variable under scrutiny within these bins. To exemplify its use, we apply BINCOR to two temporally unevenly-spaced pollen records from two marine sediments cores (MD04-2845 & MD95-2039) collected on the southwestern European margin. These series come from a global pollen & charcoal database[14] developed under the framework of the INQUA International Focus Group ACER (Abrupt Climate Changes and Environmental Responses). BINCOR works properly in detecting the well-established relationships between the compared climate records.

1 Motivation:

Several approaches are commonly used to quantify the potential association between two evenly spaced climate time series, e.g.: Pearson's and Spearman's correlation or the cross-correlation function (CCF). However, it is not straightforward when the time series are unevenly spaced

- Particularly for the case when the two time series under analysis are not sampled on identical time points, which is usual in paleoclimate research[7;16]
- The most common way to tackle this problem is to interpolate in time the original unevenly spaced climate time series in order to obtain equidistance and same times. Then, these series can be analysed using existing conventional correlation analysis techniques
- However, experience shows that interpolation has its drawbacks: it is dependent of the features of the applied method, the interpolated time series may show deviations in terms of variability or noise properties, and additional serial dependence can be introduced[4;7;8]. **Thus, interpolation should be avoided as much as possible**
- Fortunately, at least for the case of unevenly spaced climate time series sampled on identical time points, there exist some algorithms and software to carry out this task[6;8]
- However, when the two time series are not sampled on identical time points, there are few statistical techniques to estimate the correlation between them. One exception is the method of Gaussian-Kernel-based cross-correlation (gXCF) and its associated software NESTOOLBOX[9–11] and the extended version[12], or the binned correlation proposed by[7]. The software of this method, however, is not freely available on the Internet.

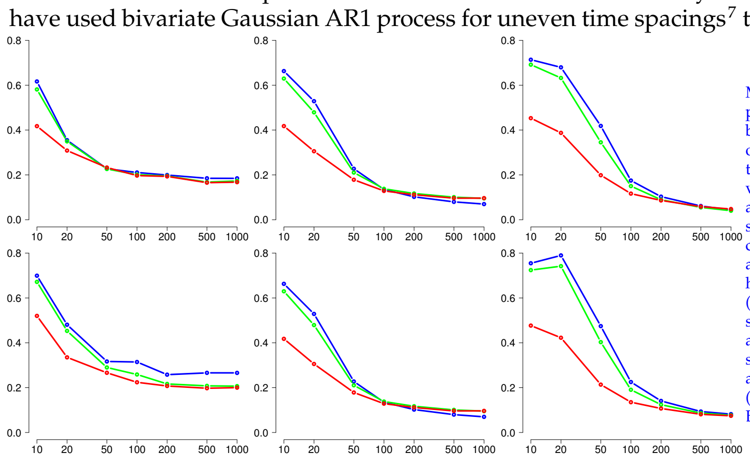
2 Method:

The procedure is described as follows[7]:

1. Input: two unevenly spaced climate time series $\{X(i), T_X\}_{i=1}^{N_X}$ and $\{Y(i), T_Y\}_{i=1}^{N_Y}$, where T_X, T_Y and N_Y, N_Y are the time domains and the sample sizes of each series, respectively.
2. Compute the average spacing among samples: $\bar{d}_X = [T_X(N_X) - T_X(1)]/(N_X - 1)$, $\bar{d}_Y = [T_Y(N_X) - T_Y(1)]/(N_Y - 1)$ & $\bar{d}_{XY} = [\bar{T}_{\max} - \bar{T}_{\min}]/(N_X + N_Y - 1)$ where $\bar{T}_{\max} = \max[T_X(N_X), T_Y(N_Y)]$ and $\bar{T}_{\min} = \min[T_X(1), T_Y(1)]$.
3. Estimate the bin-width ($\bar{\tau}$) taking into account the persistence (memory) estimated of each unevenly spaced climate time series, X and Y, which are denoted as $\hat{\tau}_X$ and $\hat{\tau}_Y$, respectively. To estimate the persistence, an AR1 model[13] is fitted to each unevenly spaced time series[5].
   - Estimate the bias-corrected equivalent autocorrelation coefficients
     - $\hat{\bar{a}}'_X = \exp(-\bar{d}_X/\hat{\tau}'_X)$, $\hat{\bar{a}}'_Y = \exp(-\bar{d}_Y/\hat{\tau}'_Y)$ & $\hat{\bar{a}}'_{XY} = \sqrt{\hat{\bar{a}}'_X \cdot \hat{\bar{a}}'_Y}$
   - Estimate the bin-width as $\bar{\tau} = -\bar{d}_{XY}/\ln(\hat{\bar{a}}'_{XY})$ (Eq. 7.48 in[5])***,
4. Determine the number of bins: $N_b = (\bar{T}_{\max} - \bar{T}_{\min})/\bar{\tau}$
5. Set: $\lim_{\inf}(n = 1) = \bar{T}_{\min}$. Then, for $n = 1, 2, \ldots, N_b$, define :
   a) $\lim_{\sup}(n) = \bar{T}_{\min} + n \cdot \bar{\tau}$
   b) $\mathrm{id}T_X$ = WHICH $[T_X \geq \lim_{\inf}(n)$ AND $T_X \leq \lim_{\sup}(n)]$;
      $\mathrm{id}T_Y$ = WHICH $[T_Y \geq \lim_{\inf}(n)$ AND $T_Y \leq \lim_{sup}(n)]$
   c) LT_X = LENGTH($\mathrm{id}T_X$); LT_Y = LENGTH($\mathrm{id}T_Y$)
      if ($LT_X > 0$ AND $LT_Y > 0$) {$F(n)$=mean($X(\mathrm{id}T_X)$); $G(n)$=mean($Y(\mathrm{id}T_Y)$); $T(n)$=[$\lim_{\inf}(n) + \lim_{\sup}(n)$]/2}
   d) $\lim_{\inf}(n) = \lim_{\sup}(n)$
6. Output: two binned climate time series $\{T_n, F(n)\}_{n=1}^{N_b}$ and $\{T_n, G(n)\}_{n=1}^{N_b}$
7. Estimate the correlation between the two binned time series through native R functions `cor and ccf or through BINCOR functions cor_ts and ccf_ts`

3 Monte Carlo Experiments:***

We conducted Monte Carlo experiments to study how the specific rules chosen for calculating the bin-width based on persistence reduces the error to arbitrarily choosing a bin-width and we have used bivariate Gaussian AR1 process for uneven time spacings[7] to generate 5000 replicates.



Monte Carlo experiments to test the impact of the rules used to calculate the bin-width and their corresponding role on the estimation of the binned correlation. The persistences for X and Y take values of 10 (column 1), 20 (column 2) and 50 (column 3), respectively. The constraints for the resampling timescales corresponds to well mixed (first row) and to wildly mixed (second row). The horizontal axis indicate the sample sizes (in log10 scale) and the vertical axis show the RMSE that is determined via averaging $(\hat{\rho}_{XY} - \rho_{XY})^2$ over 5,000 simulations. The curves in blue, green and red, indicate the rules 1 (sum), 2 (max) and 3 (the default rule option in BINCOR).

This work is under review in the *R Journal*: JM Polanco-Martínez, MA Medina-Elizalde, MF Sánchez Goñi, M Mudelsee. BINCOR: an R package to estimate the correlation between two unevenly spaced time series.

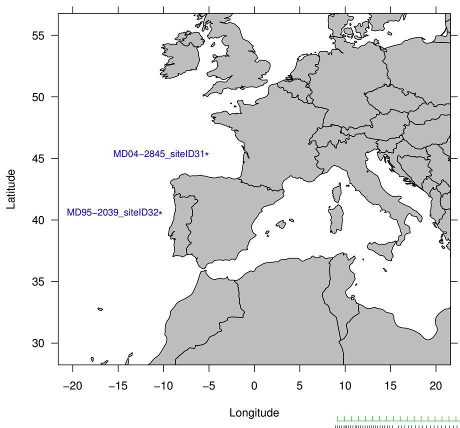
4 The BINCOR R package

developed in R version 3.1.2 runs on all major operating systems, available from CRAN repository (`http://CRAN.R-project.org/package=BINCOR`) and contains four functions:

1. **bin_cor**: the main function to build the binned time series
2. **plot_ts**: to plot and compare the "primary" and binned time series
3. **cor_ts**: to estimate the correlation between the binned time series
4. **ccf_ts**: to estimate the cross-correlation between the binned time series.

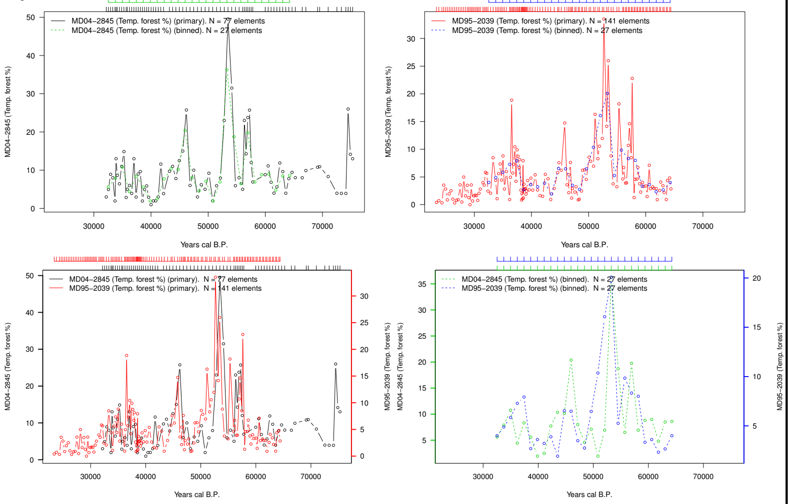
The graphical outputs can be displayed on the screen or saved as PNG/JPG/PDF. BINCOR depends on the dplR[2] and pracma[1] packages. The dplR package is used by the function bin_cor to calculate the persistence for the climate time series under study, and the pracma package is used by the functions cor_ts and ccf_ts to remove the linear trend before estimating the correlation.

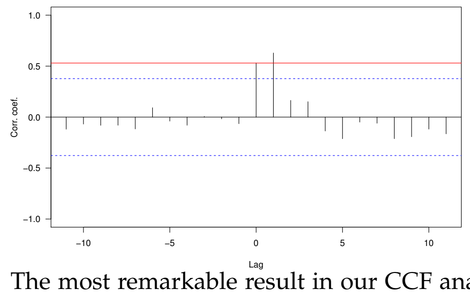
5 Case study: abrupt climate changes - Last Glacial



The pollen time series analyzed in this example span the time interval between 73,000 and 15,000 years before present (BP), thus covering the last glacial period (LGP). The climate during the LGP was characterized by millennial variability with "abrupt" transitions between cold stadials and warm interstadials known as Dansgaard-Oeschger (D-O) cycles[3;17]. The D-O cycles are characterized by rather fast atmospheric warming events over Greenland of up to 16 °C that occur within a period of approximately 40 years followed by gradual cooling leading to the cold stadials[15;17].

"Primary" (unevenly spaced) and binned pollen time series under analysis[14]. The number of elements for both time series are provided in the legend. The autocorrelation and persistence values for the time series from the site MD04-2845 are $\hat{\bar{a}}' = 0.85$ and $\hat{\tau} = 3400$ years, and for the time series from site MD95-2039 are $\hat{\bar{a}}' = 0.80$ and $\tau = 1300$ years. The horizontal top axes indicate the sampling times.





Before applying the ccf_ts function, a linear trend was removed from the binned time series activating the option rmltrd contained in ccf_ts, and the residuals were used. The CCF reveals a high correlation (r_{xy} = 0.53 between the binned time series at lag 0 (each lag is equivalent to 1220 years). The high correlation between the pollen records from sites MD04-2845 and MD95-2039 reflects similar vegetation responses to regional climate variability, particularly to precipitation and temperature changes.

The most remarkable result in our CCF analysis is that the maximum correlation (r_{xy} = 0.63) is obtained at lag 1. At face value, this result suggests that pollen variability at site MD04-2845 leads that observed at site MD95-2039 by 1220 years. Nevertheless, the most plausible explanation for this out-of-phase relationship is likely related to the chronologies of these records.

6 Final remarks:

- We presented a computational package named BINCOR (BINned CORrelation) that can be used to estimate the correlation between two unevenly spaced climate time series and is based on a novel estimation approach proposed by[7]. The package contains four functions (bin_cor, cor_ts, ccf_ts and plot_ts), is programmed in R and is available from CRAN.
- BINCOR requires the concept of nonzero persistence times, thus allowing to recover the mixing information, even when the two examined timescales differ from each other[7].
- The results from applying BINCOR to real climate data sets suggest that the BINCOR properly in detecting relationships between paleoclimate records.

Key references

[1] Borchers, H. W. *pracma: Practical Numerical Math Functions*, 2015. R package version 1.8.8.
[2] Bunn, A., Korpela, and et al. *Dendrochronology Program Library in R*, 2015. R package version 1.6.3.
[3] Dansgaard, W., and et al. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature 364*, 6434 (1993), 218–220.
[4] Horowitz, L. The effects of spline interpolation on power spectral density. *IEEE Transactions on Acoustics, Speech, and Signal Processing 22*, 1 (1974), 22–27.
[5] Mudelsee, M. TAUEST: A computer program for estimating persistence in unevenly spaced weather/climate time series. *Computers & Geosciences 28*, 1 (2002), 69–72.
[6] Mudelsee, M. Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series. *Mathematical Geology 35*, 6 (2003), 651–665.
[7] Mudelsee, M. *Climate Time Series Analysis: Classical Statistical and Bootstrap Methods*, Second ed. Springer, 2014.
[8] Ólafsdóttir, K., and Mudelsee, M. More accurate, calibrated bootstrap confidence intervals for estimating the correlation between two time series. *Mathematical Geosciences 46*, 4 (2014), 411–427.
[9] Rehfeld, K., and Bedartha, G. *NESTOOLBOX – Toolbox for the analysis of non-equidistantly sampled time series*, 2014. Matlab/Octave, version 1.01.
[10] Rehfeld, K., and Kurths, J. Similarity estimators for irregular and age-uncertain time series. *Climate of the Past 10*, 1 (2014), 107–122.
[11] Rehfeld, K., Marwan, N., Heitzig, J., and Kurths, J. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics 18*, 3 (2011), 389–404.
[12] Roberts, J., and et al. Correlation confidence limits for unevenly sampled data. *Computers & Geosciences 104* (2017), 120–124.
[13] Robinson, P. Estimation of a time series model from unequally spaced data. *Stochastic Processes and their Applications 6*, 1 (1977), 9–24.
[14] Sánchez Goñi, M. F., Desprat, S., Daniau, A.-L., Bassinot, F. C., and Polanco-Martínez, e. a. The ACER pollen and charcoal database: a global resource to document vegetation and fire response to abrupt climate changes during the last glacial period. *Earth System Science Data 9*, 2 (2017), 679–695.
[15] Sánchez Goñi, M. F., and Harrison, S. Millennial-scale climate variability and vegetation changes during the Last Glacial: Concepts and terminology. *Quaternary Science Reviews 29*, 21 (2010), 2823–2827.
[16] Weedon, G. P. *Time-series analysis and cyclostratigraphy: examining stratigraphic records of environmental cycles*. Cambridge Univ Press, Cambridge, 2003.
[17] Wolff, E. W., and et al. How has climate responded to natural perturbations? In *Understanding the earth system : global change science for application*, S. E. Cornell, I. C. Prentice, J. I. House, and C. J. Downy, Eds. Cambridge University Press, 2012, pp. 72–101.

Acknowledgments: JMPM was funded by a Basque Gov. post-doctoral fellowship. MM's work was supported by the European Commission via Marie Curie Initial Training Network LINC (project number 289447) under the Seventh Framework Programme.