Computers & Geosciences xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Correlation confidence limits for unevenly sampled data

Jason Roberts[a,b,*], Mark Curran[a,b], Samuel Poynter[c], Andrew Moy[a,b], Tas van Ommen[a,b], Tessa Vance[b], Carly Tozer[b,d], Felicity S. Graham[e], Duncan A. Young[f], Christopher Plummer[b,e], Joel Pedro[b,g], Donald Blankenship[f], Martin Siegert[h]

[a] Department of the Environment, Australian Antarctic Division, Kingston, Tasmania 7050, Australia
[b] Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Private Bag 80, Hobart, Tasmania 7001, Australia
[c] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Burnley, Victoria 3121, Australia
[d] University of Newcastle, Callaghan, NSW, Australia
[e] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia
[f] Jackson School of Geosciences, University of Texas at Austin, Austin, Texas, USA
[g] Centre for Ice and Climate, Niels Bohr Institute, University of Copenhagen, Denmark
[h] Grantham Institute and Department of Earth Science and Engineering, Imperial College London, London, UK

ARTICLE INFO

Keywords:
Unevenly sampled data
Autocorrelation
Bootstrapping
Gaussian kernel method
Confidence limits

ABSTRACT

Estimation of correlation with appropriate uncertainty limits for scientific data that are potentially serially correlated is a common problem made seriously challenging especially when data are sampled unevenly in space and/or time. Here we present a new, robust method for estimating correlation with uncertainty limits between autocorrelated series that does not require either resampling or interpolation. The technique employs the Gaussian kernel method with a bootstrapping resampling approach to derive the probability density function and resulting uncertainties. The method is validated using an example from radar geophysics. Autocorrelation and error bounds are estimated for an airborne radio-echo profile of ice sheet thickness. The computed limits are robust when withholding 10%, 20%, and 50% of data. As a further example, the method is applied to two time-series of methanesulphonic acid in Antarctic ice cores from different sites. We show how the method allows evaluation of the significance of correlation where the signal-to-noise ratio is low and reveals that the two ice cores exhibit a significant common signal.

1. Introduction

Sparse data correlation techniques, and the confidence limits associated with them, are a keystone of quantitative analysis in geoscience. However, uneven sampling of data is a common feature in many fields, and our inability to prescribe appropriate interpolations between data may hinder the statistical application of results. In many cases, this may come about as an inherent sampling non-uniformity. In the case of ice cores, for example, the relationship between the spatial and temporal distribution of a sample material varies with depth such that uniform spatial sampling generates non-uniform sampling on a temporal scale. Further difficulty arises from missing data or data gaps, which may be caused by physical sample size constraints, damage, or loss of samples due to contamination or analytical problems. Where numerical methods require evenly sampled data, interpolation is necessary, but must be used cautiously to avoid signal artifacts. The use of common software tools to interpolate between data points often comes at the expense of robustness, as bias may be introduced.

Rehfeld and Kurths (2014) investigated this issue in detail, benchmarking a variety of techniques to overcome the challenges introduced by irregularly-sampled time series. The use of a Gaussian kernel method gave a reliable and robust estimation in comparison to commonly-used interpolation approaches such as resampling onto a common uniform independent grid. Complications arise for irregularly-sampled data with inherent autocorrelations, however, as the estimation of a confidence interval, or some other measure of significance, requires explicit and quantitative consideration of the autocorrelations (Mudelsee, 2003; Ólafsdóttir and Mudelsee, 2014). Several methods exist for the assessment of significance, for evenly sampled data, in the presence of autocorrelation. Such methods include the effective spatial degrees of freedom method of Bretherton et al. (1999) which uses classical tests with a reduced number of degrees of freedom to account for autocorrelations in the data, and data surrogates such as bootstrapping and Fourier space methods. These latter

* Corresponding author at: Department of the Environment, Australian Antarctic Division, Kingston, Tasmania 7050, Australia.
E-mail address: Jason.Roberts@aad.gov.au (J. Roberts).

http://dx.doi.org/10.1016/j.cageo.2016.09.011
Received 3 December 2015; Received in revised form 18 July 2016; Accepted 25 September 2016
Available online xxxx
0098-3004/

methods make no assumptions on the distribution of the data (Mudelsee, 2003), so may be more appropriate for many real-world datasets. Compared to standard bootstrapping techniques, Fourier space methods have the advantage of preserving linear correlations, but lose many of their computational advantages for irregularly-sampled data.

Here, we report the development of the Gaussian kernel method, extended to provide confidence interval information, with application to airborne glacier geophysical data. An evenly-sampled, highly auto-correlated dataset of Antarctic ice thicknesses from the ICECAP (Investigating Cryospheric Evolution through Collaborative Aerogeophysical Profiling) project (see Fig. 1 for location) provides a suitable test data set to validate the approach. The correlation and confidence interval distribution is compared to a recently published method (Ólafsdóttir and Mudelsee, 2014). Data were randomly removed to simulate the effect of uneven data spacing and the resulting autocorrelation distributions compared.

As a second independent demonstration of the strength of the technique we compute the correlation between time series of methanesulphonic acid (MSA) concentration in two Antarctic ice cores (see Fig. 1 for location). MSA has been used as a proxy for Antarctic sea ice extent (Curran et al., 2003), based on the production of MSA from sea ice-associated phytoplankton which are known to be a dominant sulphur source from the sea-ice edge in Antarctica (Vance et al., 2013). Confirming that a statistically significant (at a 95% confidence interval) relationship exists between the two MSA records supports the hypothesis that the records preserve a common environmental signal.

While the Mudelsee (2003); Ólafsdóttir and Mudelsee (2014) method can be used on unevenly spaced climate time series data, in cases where the data are both unevenly spaced and on a different time base their method requires interpolation or resampling. Our Gaussian Kernel-based method removes the need for such resampling, making it well suited to computing correlations between paleoclimate records from different locations and different archives, in which different time bases are ubiquitous.

2. Method

2.1. Correlation

Correlations (C_{xy}) between unevenly and differently sampled series (x_i and y_j) are calculated using the Gaussian kernel correlation slotting (Rehfeld et al., 2011).

$$C_{xy} = \frac{1}{\sigma_x \sigma_y} \frac{\sum_{i=1}^{n_x} \sum_{j=i}^{n_y} (x_i - \bar{x})(y_j - \bar{y}) K\left(d_{x_i} - d_{y_j}\right)}{\sum_{i=1}^{n_x} \sum_{j=i}^{n_y} K\left(d_{x_i} - d_{y_j}\right)} \quad (1)$$

where the average of the two series x_i and y_j (of length n_x and n_y) is $\bar{x}$ and $\bar{y}$, respectively, and d_x and d_y are the independent variables (typically time or distance) for x and y respectively, and may differ from each other. The Gaussian kernel $K(d) = \frac{1}{\sqrt{2\pi}h} \exp(-d^2/2h^2)$ uses a width parameter (h) of one quarter of the larger of the average spacing of the two data series. Unlike Rehfeld et al. (2011) who normalise the signals to have zero mean and unit variance, we use the original signals and correct for the mean and estimate the standard deviations (σ_x and σ_y) using the same weighted summation Gaussian kernel ($K(d)$) as used in Eq. (1).

2.2. Bootstrapping

Confidence intervals (95%) are estimated using a stationary bootstrapping technique (Politis and Romano, 1994). This method accounts for persistence (serial correlation) and the associated reduction in the effective degrees of freedom in the data (Wilks, 2006) by generating

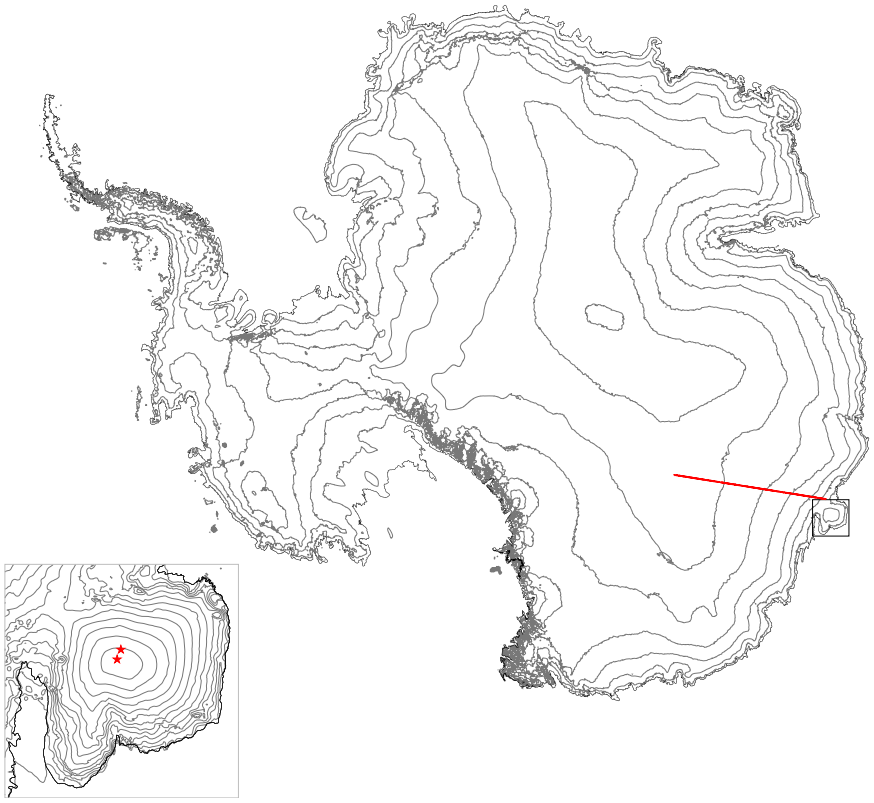

Fig. 1. Location of an airborne radar transect yielding ice thickness data (red line). Elevation contours at 500 m are from Bamber et al. (2009) (grey lines) and the ice sheet grounding line is from Bindschadler et al. (2011). Inset shows the Law Dome region of East Antarctica and the sites of the two ice cores (red stars), with DSS97 being closer to the dome summit and W10K being close to the 1300 m elevation contour.