Synthetic running correlation coefficients

Understanding the physical world

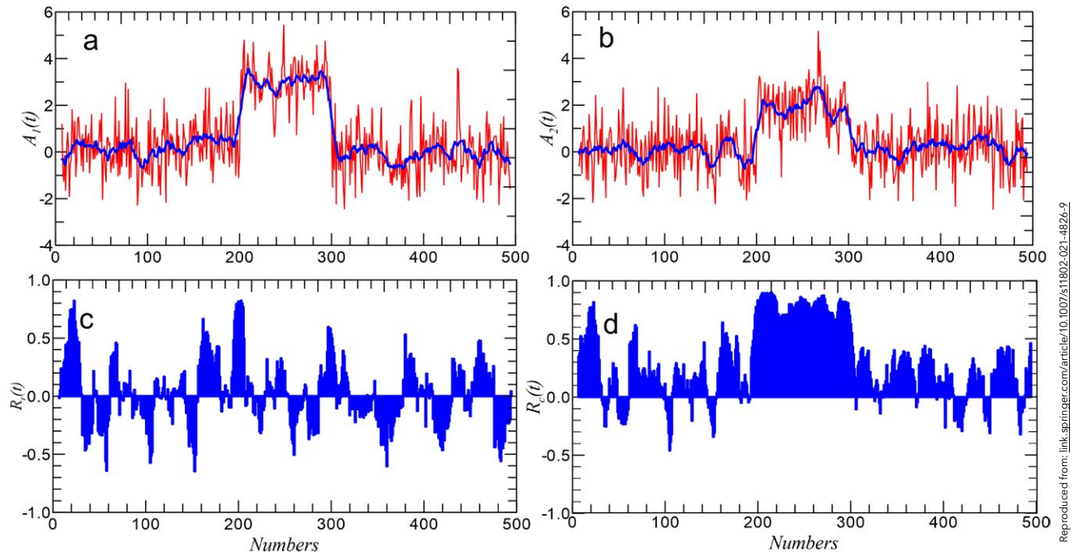
Translating data into an understanding of physical phenomena is a task that all scientists undertake, no matter what their research area is. One of the main tools that scientists have at their disposal are statistical analysis techniques. Professor Jinping Zhao at the Ocean University of China has developed the theoretical framework for the synthetic running correlation coefficient and been using this tool to explore the correlations between various physical phenomena, including the amount of cloud coverage and Arctic Sea ice levels.

Whether you want to understand how to prevent the melting of the Arctic ice sheets, predict the weather, or make a timetable for the changing tides, you need the same thing – information on how the physical phenomena that control these processes work. However, for the types of process that occur in the atmosphere or ocean, getting a detailed understanding of the underlying physics can be challenging.

Phenomena in the natural world are generally highly complex. This complexity comes from the number of contributing processes. For example, predicting weather conditions means understanding numerous variables, the most important of which are the temperature, cloudiness, humidity, precipitation, and wind conditions. Each of these contributing factors then contains its own complex set of variables that also need to be understood.

When exploring new phenomena, it is also not always obvious which variables need to be considered or how important different contributing factors are. This is why scientists often start by making a series of observations of an event, collecting lots of different data, and then try to see which information in the dataset correlates to each other and how those correlations change with time.

Two strongly correlated events have a strong statistical relationship with each other. A simple example would be height and weight – taller people are generally heavier. However, there is a famous saying in statistics that 'correlation does not equal causation'. It is not necessarily true that because one person is heavier than another, they will be taller.



(a) and (b) are two different white noise datasets (red) and their local means (blue), with a constant added in the 200-300 interval of each dataset. (c) and (d) are the local and the synthetic running correlation coefficients, respectively.

Understanding the relationships between variables and whether there is true causality between variables is a challenging statistical task. There are a number of different statistical tests that can be applied to data to try and determine robustly which events are linked and which are correlated by pure coincidence.

Professor Jinping Zhao and his colleagues at the Ocean University of China have been investigating the quality of previous statistical tests and developing new tools for the analysis of various kinds of datasets that show the evolution of events over time. Zhao's new approach, using the synthetic running correlation coefficient, seems to be a better way of identifying real correlations in a variety of data types, including atmospheric and oceanic information.

TIME-SERIES DATA

A correlation coefficient measures the strength of a correlation between two variables. A value of one is equal to perfect correlation: when a change in one variable maps exactly to a change in another. Zhao and many scientists are interested in looking for correlations in a type of data known as a time-series dataset. Time-series data is any measurement that is recorded as a function of time.

A common example of time-series data is weather records. Time-series measurements can be over short or long timescales. Most records will look at how a particular variable, such as the highest given temperature in a day, has changed over days, months, or years.

The analysis they had been performing was not quite the standard local running correlation coefficient, but a new algorithm.

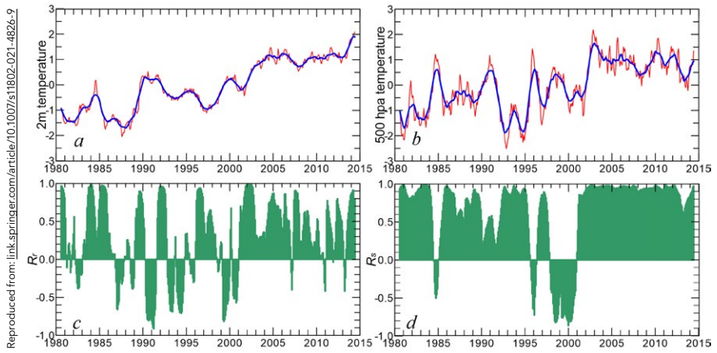
Depending on the amount of historical data available, time-series datasets can become very long and cumbersome to analyse. Sometimes it also makes sense when looking for correlations in a dataset to only look at a particular region, or a spell of time, of the dataset. For example, if the data is recorded over the course of a calendar year, it might make sense to break down the analysis into seasonal windows. There may be events that are correlated in the spring and summer that are not correlated in the winter; therefore, analysing the year of data as a whole may obscure those results.

Using a moving window to analyse chunks of a dataset for correlations is known as obtaining the local running correlation coefficient, where 'local' was added by Zhao to distinguish it from the synthetic running correlation coefficient described below. Using the local running correlation coefficient to analyse time-series datasets has been a common approach for nearly ninety years, but Zhao has started to have some doubts about the validity of certain results seen with this method.

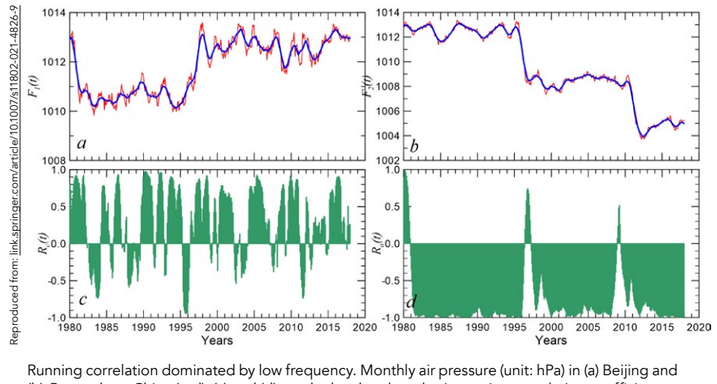
SYNTHETIC RUNNING CORRELATION COEFFICIENT

The researcher notes that he obtained 'very interesting results with the running correlation coefficients' but that 'once you changed the algorithm to the local running correlation coefficient, those important phenomena disappeared'. The team's surprise at these results prompted them to investigate further.

On closer investigation of their results, Zhao and his team realised that the



Air temperature anomalies (red) at (a) 2 m and (b) 500 hPa averaged for the North Atlantic over the period 1980–2015. (c) and (d) are the local and the synthetic running correlation coefficients, respectively.



Running correlation dominated by low frequency. Monthly air pressure (unit: hPa) in (a) Beijing and (b) Guangzhou, China (red); (c) and (d) are the local and synthetic running correlation coefficients, respectively.

If datasets include obvious low-frequency signals, the synthetic running correlation coefficient has a great advantage.

analysis they had been performing was not quite the standard local running correlation coefficient, but actually a new algorithm they have termed the 'synthetic running correlation coefficient'. By looking at the correlation of monthly averaged geological data, the team realised that the small mathematical differences between the two models were causing some important correlations to be missed with the standard local running correlation coefficient approach.

Normally, when looking for correlations in time-series data, a local average is calculated with each time window, known as the local mean. With the synthetic running correlation coefficient, a global average of the whole time series is used – so the average value is not fluctuating as the window moves. In many scenarios, this does not make an appreciable difference to the correlation coefficients calculated.

How often an event happens over time is known as its frequency, with high-frequency events occurring more often in a given time period. In the resulting data, this produces a signal that oscillates with a repeating pattern in time. For running correlations, high-frequency events occur on periods shorter than the time window sampled, whereas low-frequency events happen on a timescale longer than the sampling window.

What the researchers found is that, for phenomena that had both low- and high-frequency signals, the synthetic running correlation coefficient would reflect the real correlations more accurately, whereas the local running correlation coefficient was only sensitive to the high frequency events.

DATA SAMPLING

Zhao's findings show the importance of considering sampling windows in statistics and how even a small, localised region of a dataset should include information about the larger dataset as a whole. While the difference between the two approaches is minimal for many seasons and sub-seasonal events that reflect high-frequency processes in oceanographic and geological data, Zhao has found several examples where the appropriate statistical method is key.

He has found that the relationship between cloudiness and ice concentrations in the Arctic is much more complex than previously thought. It is generally assumed that clouds reflect solar radiation, reducing the amount of sea ice melt, but with his new analysis, Zhao realised that the correlation between these two variables actually changes depending on the year, with a negative correlation becoming positive sometimes.

Finding anomalous events in complex datasets is key for understanding the world around us, and with Zhao's new tools, there is now an additional approach to do just that, whether the changes occur on a short or longer timescale.

Behind the Research




Dr Jinping Zhao
Dr Yanyue Shi
Dr Yong Cao
Xin Wang

E: jpzhao@ouc.edu.cn **T:** +86 1385 3238 188

Research Objectives

Professor Zhao developed the theoretical framework for the synthetic running correlation coefficient.

Detail

Address
Ocean University of China
College of Oceanic and Atmospheric Sciences
239 Songling Road, Qingdao, 266100
China

Bio
Jinping Zhao, PhD, is professor at the Ocean University of China and physical oceanographer. His research is focused on the Arctic change in ocean, climate and sea ice.

Funding
This study is supported by the National Natural Science Foundation of China (Grant No. 41941012 and 41976022).

Collaborators

- Yong Cao, Senior Experimentalist, physical oceanography
- Yanyue Shi, Professor, mathematics
- Xin Wang, PhD student, physical oceanography

References

Zhao, J, Cao, Y, Shi, Y, Wang, X, (2021) Mathematical proof of the synthetic running correlation coefficient and its ability to reflect temporal variations in correlation, *Journal of Ocean University of China*, 20(3), 562–572. doi.org/10.1007/s11802-021-4826-9

Zhao, J, Cao, Y, Wang, X, (2018) The physical significance of the synthetic running correlation coefficient and its applications in oceanic and atmospheric studies, *Journal of Ocean University of China*, 17(3), 451–460. doi.org/10.1007/s11802-018-3798-x

Personal Response

What do you think will be the most important findings with your new approach?

"The standard local running correlation coefficient (LRCC) presents the correlation of datasets dominated by high-frequency signals. However, if the datasets include obvious low-frequency signals, the synthetic running correlation coefficient (SRCC) has a great advantage. Analysis resulting from the SRCC algorithm always presents rich information on low-frequency processes, which is almost invisible when applying the LRCC algorithm. Therefore, SRCC is capable of providing something new and exciting. It is especially important for climate change studies where low-frequency information needs to be extracted. Comparing results from your data using both the SRCC and LRCC algorithms will lead to a wonderful picture."