Notes on Estimation of Cross Correlation Function between two Time Series

Minoru Tanaka

Department of Network and Information, School of Network and Information,
Senshu University, Kawasaki 214-8580, Japan

Abstract. The purpose of this paper is discussing the estimation of cross correlation function between two time series, for instance, the monthly mean temperature deviations and monthly average CO2 levels. In order to tell if a cross correlation estimate is significantly different from zero, it should note that at least one of the series must be white noise. The method of prewhitening a time series must be very important for the cross correlation analysis. We consider the prewhitening method and discuss if there were truly a lead-lag-relationship between the real series.

Keywords: *cross correlation analysis, prewhitening, SARIMA model, monthly temperature, monthly CO2 levels*

1. Introduction

We say two time series X_t and Y_t are *jointly stationary* if they are each stationary, and the cross-covariance function

$$\gamma_{xy}(h) = \mathrm{cov}(X_{t+h}, Y_t) = E\left[(X_{t+h} - \overline{\mu}_x)(Y_t - \overline{\mu}_y)\right]$$

is a function only of lag *h, where* $\overline{\mu}_x$ and $\overline{\mu}_y$ are means of X_t and Y_t. Here $\gamma_x(h) = \gamma_{xx}(h)$ is an autocorrelation function of X_t (see, for example, Blockwell and Davis [1]).

Definition 1. The sample autocorrelation function (ACF) is defined as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},$$

where $\hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x})$, with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ *for h = 0,1,2, ..., n-1.*

Definition 2. The cross-correlation (CCF) of jointly stationary time series X_t and Y_t is defined as

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\,\gamma_y(0)}}.$$

Definition 3. The sample cross-correlation (CCF) of jointly stationary time series X_t and Y_t is defined as

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\,\hat{\gamma}_y(0)}},$$

where $\hat{\gamma}_{xy}(h) = n^{-1}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(y_t - \overline{y})$, $\hat{\gamma}_{xy}(-h) = -\hat{\gamma}_{xy}(h)$ *for h = 0,1,2, ..., n-1.*

It should note that these sample correlations make sense only when the two time series are stationary.

The sample cross-correlation is examined graphically as a function of lag *h* to search for leading or lagging relations in the two time series data. The peaks in the CCF graph are evaluated by comparing their magnitudes with their

theoretical maximum values. Then the following property concerning a large sample distribution of the sample cross-correlation function is very useful for the analyzing their leading or lagging relations (see Shumway and Stoffer [7]).

Property A. If X_t and Y_t are independent processes, then under mild conditions, the large sample distribution of the sample cross-correlation $\hat{\rho}_{xy}(h)$ is normal with mean zero and standard deviation $1/\sqrt{n}$ if at least one of the process is independent white noise.

It is well known that the sample CCF between two time series, which are not stationary, appears to show non-zero cross-correlation even though the two series are independent.

In the cross-correlation analysis one of the key words may be “prewhitening” a series. We consider the method of prewhitening a time series. Using Property A, it is seen that at least one of the series must be independent white noise. If this is not case, it must be very difficult to tell if a cross-correlation estimate is significantly different from zero.

We consider the following example (see, Example 2.33 in Shumway and Stoffer [7]). Generate a series, x_t and z_t, for t = 1, 2, ..., 120 as

$$x_t = 2\cos(2\pi t/12) + w_{t1} \quad \text{and} \quad z_t = x_{t-5} + .5\, w_{t2},$$

where $\{w_{t1}, w_{t2}; t = 1.2,..., 120\}$ are all independent standard normals, N(0,1). Then x_t leads z_t by 5 months. The generated series are shown in the top row of Figure 1.1, and the bottom row shows sample ACF of each series. The top row (left) in Figure1.2 shows the sample CCF between x_t and z_t, which shows cross-correlations but not clear the leading or lagging relation.

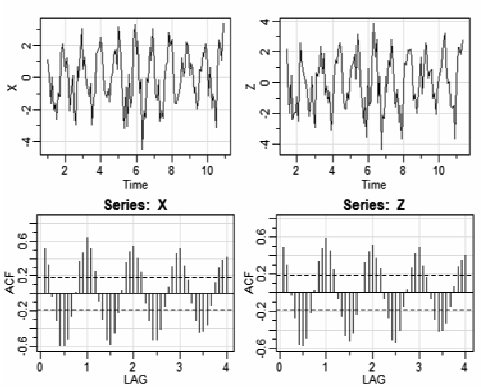

Figure 1.1 The series x_t and z_t and their sample ACFs.

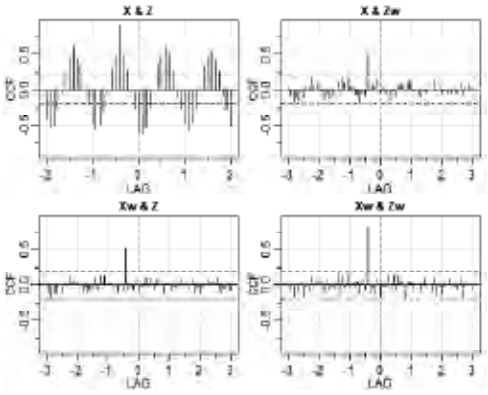

Figure 1.2 Sample CCFs between the two series.

By use of a regression of x_t on $cos(2\pi t/12)$ we define its residuals as x_{wt}, the prewhited x_t. Also we can prewhite z_t. The other graphs in Figure 1.2 show the sample CCFs between x_t and prewhited z_t, prewhitened x_t and z_t, and prewhitened x_t and prewhitened z_t. It is seen that the two series have cross-correlation estimate significantly different from zero at lag $h = -5$, which means x_t leads z_t by 5 months. The bottom row (right) in Figure1.2 shows the value of the cross-correlation (+0.75) at lag $h = -5$ is biggest among the three.

In the next section 2 we shall consider the effect of daylight hours and atmospheric CO2 concentration on mean temperature deviations at Syowa Station, Antarctica, from February 1966 to September 2019 by use of the cross-correlation analysis with a prewhitening method. We examine the *cross correlation analysis* for the group of following two series:

(1) monthly mean temperature deviations and monthly average daylight hours

(2) monthly mean temperature deviations and monthly mean CO2 levels

(3) monthly mean CO2 levels and monthly average daylight hours

(4) Ocean surface temperature deviations and yearly average sunspot number.

This paper is supported by the computer software RStudio (see Cowpertwait [2]) and Mathematica 12 (see He [3]), and by Japan Meteorological Agency [5] for provided time series data used in this paper.

2. Cross Correlation Analysis and Prewhitening

(2-1) Temperature Deviations and Daylight Hours

We consider the effect of monthly mean temperature deviations and monthly average daylight hours at Syowa Station from February 1966 to September 2019. Figure 2.1 shows the two series and the sample cross-correlation function CCF between the two series. Both of these series have the seasonal cycle. The CCF appear that the daylight hours might lead the temperature at lag 1or 2 (month). But it must be difficult to discern the lead-lag-relationship from only these graphs.

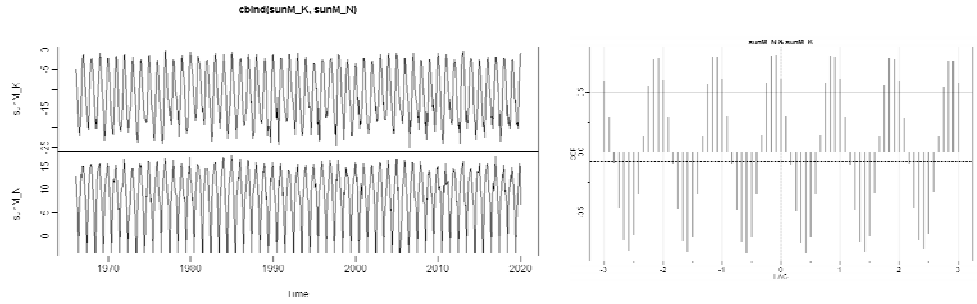

Figure 2.1 (a) monthly temperature (top) and daylight hours (bottom) (b) the sample CCF between the two series.

In Figure 2.1(b) the CCF exhibit periodicities such that observations separated by 12 months apart are positively correlated, and also observations separated by 6 months are negatively correlated.

It is difficult to tell if there were truly a lead-lag-relationship between the temperature and daylight hours, because the two series are not stationary, whose means depend on time. From the graph of the correlation function of the original series CO2 and daylight hours, spurious correlation is seen under the influence of each periodic ingredient in a cycle of 12 months.

To examine if there were truly a lead-lag-relationship between the series, which is equivalent to that if a cross-correlation estimate is significantly different from zero, Property A shows that at least one of the series must be white noise. Therefore we should consider the method of prewhitening the data (see, Section 8.5 in Shumway and Stoffer [7]). Although there are methods to whiten a series, our employing method is using a multiplicative seasonal autoregressive integrated moving average (SARIMA) model fitting (see, for example, Blockwell and Davis [1]). If X_t is an *SARIMA* series, the residuals from the fitted model must be white noise. Then we define the residuals from the fit as the prewhitened series of X_t.

The fitted SARIMA model to the temperature data is an ARIMA(2,0,1)(2,1,0)[12] with the parameters as follows

```
# Series: sunM_K
# ARIMA(2,0,1)(2,1,0)[12]
# Coefficients:
#         ar1     ar2      ma1     sar1     sar2
#      0.5816  0.0704  -0.4282  -0.7190  -0.3607
#s.e.  0.1948  0.0642   0.1941   0.0376   0.0379
```

Also the model to the CO2 data is an ARIMA(0,0,0)(1,1,0)[12] with the parameters as follows

```
# Series: sunM_N
# ARIMA(0,0,0)(1,1,0)[12] with drift
# Coefficients:
#             sar1    drift
#         -0.4147  -0.0006
# s.e.    0.0366   0.0038
# sigma^2 estimated as 2.669:  log likelihood=-1214.74
# AIC=2435.47   AICc=2435.51   BIC=2448.84
```

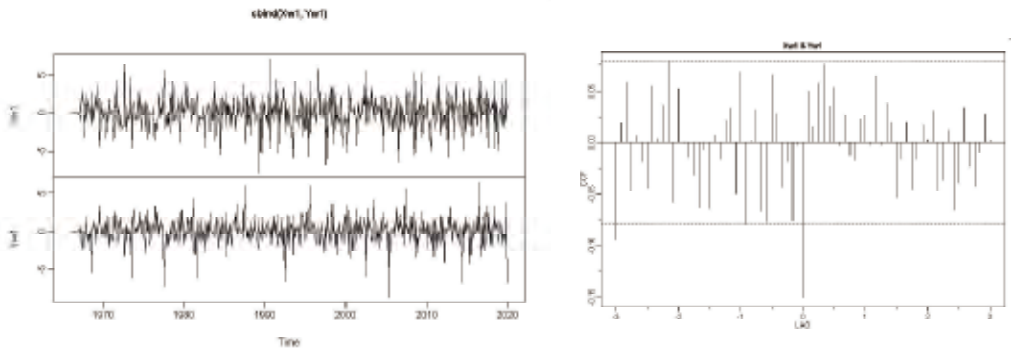

Figure 2.2 Residuals from the fitted model of temperature (top), daylight hours (bottom) and the sample CCF (left)

The cross-correlation function peaks at lag $h = 0$ ($\hat{\rho}_{xy}(0) = -0.15$), showing that the daylight hours measured at time t months is associated with the temperature at same time t. The daylight hours does not lead the temperature, which is different from the consideration above. The negative sign of the CCF at $h = 0$ implies that the two series move in different directions. It will mean that an increasing the daylight hours is decreasing the temperature in Antarctica. This is considered to be a ground heat dissipation phenomenon.

（2-2）Temperature Deviations and CO2 Levels

Figure 2.2 shows the two series x_t and y_t, monthly mean temperature deviations and CO2 levels at Syowa Station from February 1966 to September 2019.

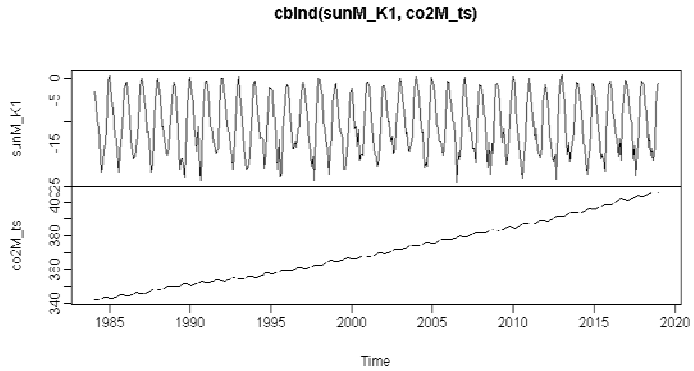

Figure 2.3 Monthly temperature deviations (top) and monthly CO2 levels (bottom) at Syowa Station.

Although the monthly average CO2 levels is increasing almost in monotone, the temperature is not increasing. It seems that both are clearly unrelated and we will confirm this quantitatively. Specifically, we calculate the sample cross correlation function between two time series by using a way of prewhitening.

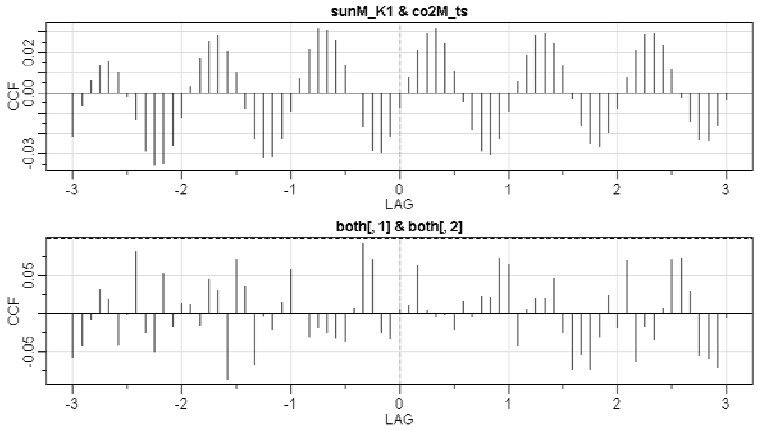

Figure 2.4 Sample CCF between the two original series.

Figure 2.3 shows that the cross correlation function between two original series has a *cycle of 12* months. This is because each has a cycle of 12 months, and it can be called spurious correlation which is not true correlation. So, in order to investigate significant (it is not *zero*) correlation, one of variables at least need to make a independent series between two stationary time series.

Method 1: Apply a SARIMA model to each series, and obtain a residuals from the model.

Here we use the prewhitening method of applying a SARIMA model to each series, and obtain the residual terms. Figure 2.5 shows the residuals from ARIMA(2,0,1)(2,1,0)[12] model of x_t (top row left), and that from ARIMA(0,1,2)(0,1,2)[12] model of y_t (top row right), and the bottoms show the sample ACF and histogram of each residuals.

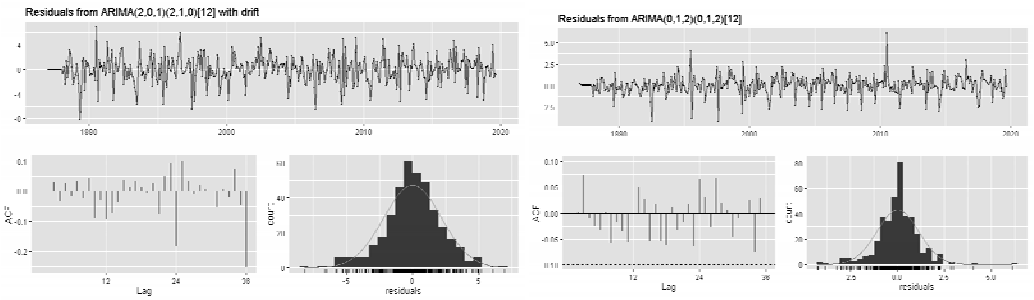

Figure 2.5 The two residuals series from the fitted models, their ACFs and histograms.

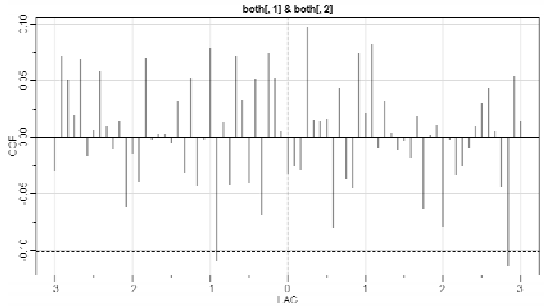

Figure 2.6 Sample CCF between the two residuals series from each fitted models

It is seen that there is no significant correlations in Figure 2.6, thus the CO2 levels may not lead the temperature and the two series will not move dependently.

Method 2: Apply the SARIMA model (fitted to y_t) to the series x_t, and obtain a residuals from the model.

We next consider the another method of prewhitening the series, that is, applying the fitted model of y_t, ARIMA(0,1,2)(0,1,2)[12], to the series x_t, and we obtain the residuals and then calculate the sample CCF between the two residuals.

Figure 2.7 shows the residuals of x_t from the model, the sample ACF, Q-Q plots and its p- values for Ljung-Box statistics. It will be seen that the residuals series is a normal white-noise. From Figure 2.8, we can also say a correlation between the temperature and the CO2 is hardly found .

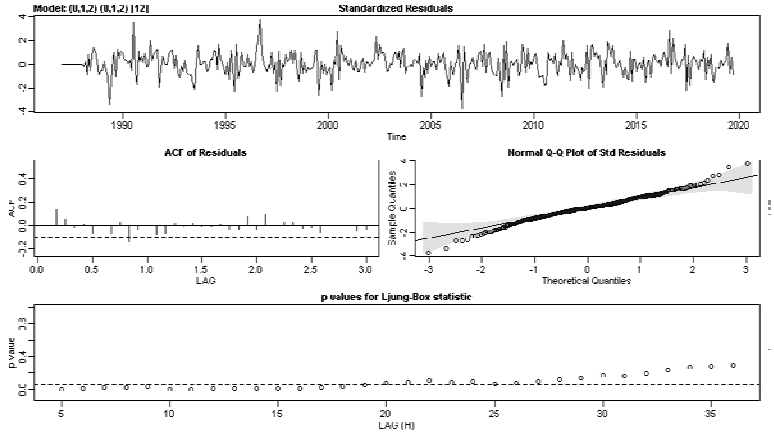

Figure 2.7 Residual analysis for the ARIMA(0,1,2)(0,1,2)[12] fit to the temperature data.

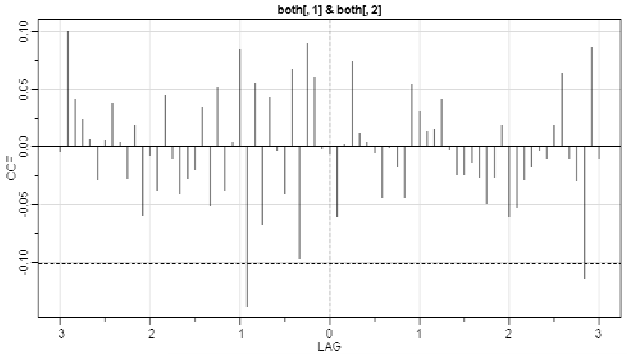

Figure 2.8 Sample CCF between the two residuals series from the same model.

(2-3) Daylight Hours and CO2 Levels

Figure 2.9 shows the two series x_t and y_t, Daylight hours and CO2 levels at Syowa Station from February 1984 to January 2019.

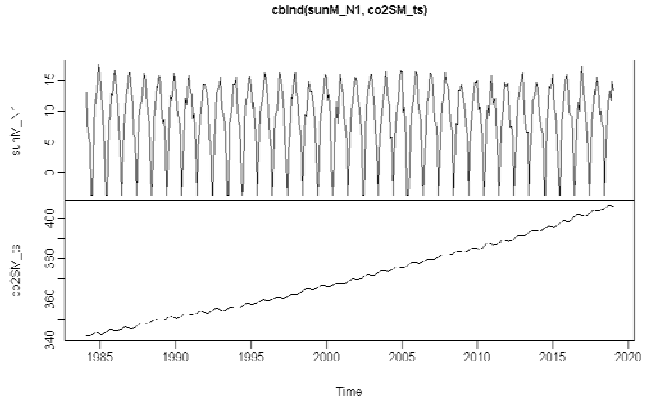

Figure 2.9 Sample CCF between the two original series.

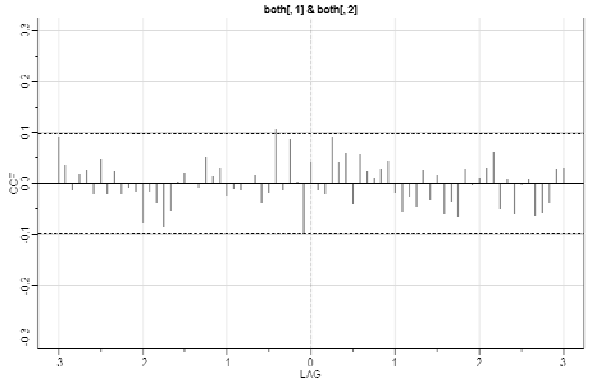

Figure 2.10 Sample CCFs between the two prewhitened series.

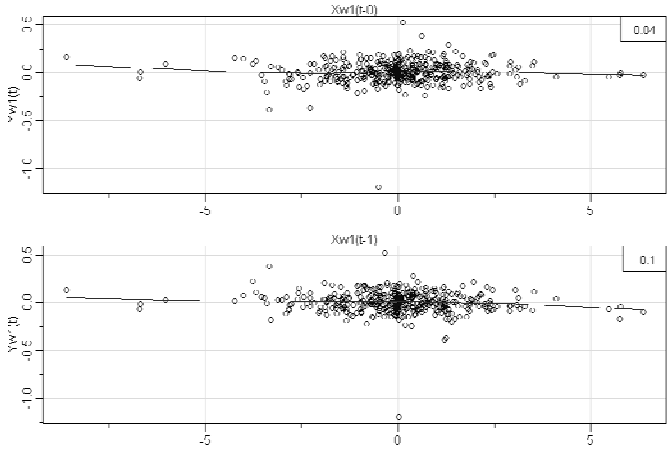

Figure 2.11 Scatter plots between the two prewhitened series $\{x_{w(t)}, y_{w(t)}\}$ (top) and $\{x_{w(t-1)}, y_{w(t)}\}$ (bottom).

Spurious correlation has come out under the influence of the periodic ingredient of each cycles 12 months. Correlation significant of the sample CCF between the two prewhitened series is not seen. However, if the number of data increases, then the cross-correlation coefficients at lag $h = -1$and lag $h = -5$ may become bigger in absolute value.

(2-4) Ocean Surface Temperature Deviations and Sunspot Number

Figure 2.12 shows the two series y_t and z_t, the Ocean surface temperature deviations and a yearly averages of the International relative sunspot number from 1880 to 2017 (see Tanaka [4], Moran [6] and Thomas [8]). The Ocean surface temperature deviations series is annual sea surface temperature anomalies averaged over the part of the ocean that is free of ice at all times (open ocean) from 1880 to 2017 (see Hansen et al.[3]).

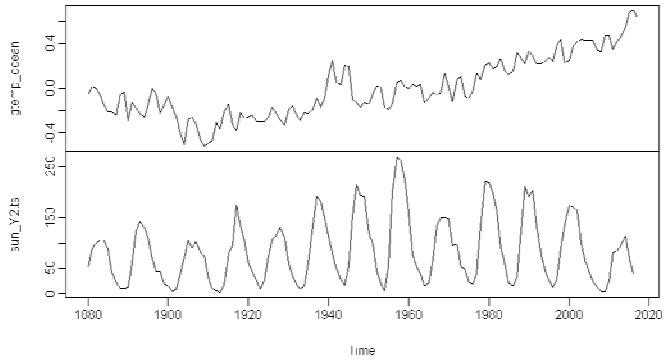

Figure 2.12 *Ocean surface temperature deviations (Y) and yearly sunspot number (Z) (1880-2017).*

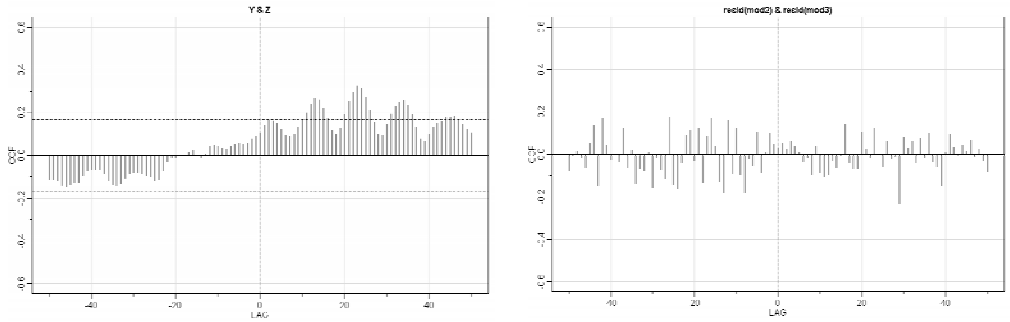

Figure 2.13 *Sample CCFs between series Y and Z (left) and between the prewhitened series (right).*

We use the prewhitening method of applying a SARIMA model to each series, and obtain two residual terms from ARIMA(1,1,3) model of Y and from ARIMA(2,1,2) model of Z. Figure 2.13 (right) shows the sample CCF between the two prewhitened series. It is seen that there is no significant correlation except for a lag $h = 30$. The cross-correlation $\hat{\rho}_{xy}(30) = -0.22$ means the S*unspot number* will lead the Ocean temperature deviations for 30 years and they move in different directions. The result of "negative sign" correlation may look similar to the case between the temperature deviations and the daylight hours at Syowa Station.

Conclusions

We have considered the effect of daylight hours and atmospheric CO2 concentration on mean temperature deviations at Syowa Station, Antarctica, from February 1966 to September 2019. The sample CCFs between the following four pairs of two series were estimated by use of the cross-correlation analysis with prewhitening methods in Section 2.

(2-1) Temperature deviations and Daylight hours:

In this case the sample CCF peaked at lag $h = 0$ ($\hat{\rho}_{xy}(0) = -0.15$), and this shows that the *daylight hours* measured at time t (months) is associated with the temperature deviations at same time *t*. The negative sign of the CCF

at lag $h = 0$ implies that the two series move in different directions.

(2-2) Temperature deviations and CO2 levels:

There was no significant correlation, thus the CO2 levels may not lead the temperature deviations and the two series may move independently.

(2-3) CO2 levels and Daylight hours:

There was no correlation significant of the sample CCF between the two prewhitened series.

(2-4) Ocean surface temperature deviations and Sunspot number:

There were two significant correlations, but $\hat{\rho}_{xy}(30) = -0.22$ means the sunspot number will lead the temperature for 30 years and they move in different directions.

By the way the obtained results were not adequate because some of the fitted SARIMA models did not passed one of goodness of fit tests (Ljung-Box). Therefore finding the best model for the prewhitening the series must be a future work for us.

References

[1] Blockwell,P.J., Davis,R.A. (2002), *Introduction to Time Series and Forecasting*, Springer Verlag, New York.

[2] Cowpertwait.P.S.P., Metcalfe.A.V. (2009), *Introductory Time Series With R*, Springer Verlag, New York.

[3] Hansen, J., Sato, M., Ruedy, R., Lo,K., Lea,D.W., Medina-Elizade,M. (2006), *Global temperature change*, Proceedings of the National Academy of Sciences, 103(39): 14288-14293.

[4] He,Y. (1995), *Time Series Pack for Mathematica*, Wolfram Research.

[5] Japan Meteorological Agency: http://www.data.jma.go.jp/cpdinfo/chishiki_ondanka/p08.html.

[6] Moran, P. (1953), *The statistical analysis of the sunspot and lynx cycles*, J. Animal Ecol., 1, 163-173.

[7] Shumway, R.H., Stoffer D.S. (2019), *Time Series: A Data Analysis Approach Using R*, CRC Press.

[8] Tanaka, M. (2010), *Time series modelling of Annual Maximum Sunspot Numbers*, Information Science and Applied Mathematics, Vol.18, 19-32.

[9] Thomas, J.H., Weiss, N.O. (2008), *Sunspots and Starspots*, Cambridge University Pres.