# ANALYSIS OF SO-CALLED BLANK-OFF DATA

Jamie Riggs

## INTRODUCTION

Time series is an ordered sequence of observations. In our blank-off data, the sequence is through time, and we make the assumption that each consecutive observation is separated by the same time interval. There is no evidence in the original data set to validate this assumption.

The objectives of the analysis that follows is to describe the behavior of the series through plots of the data and descriptive measures of the main properties of the series, and explain the behavior of the series in terms of other time series.

## RESULTS

1. All four channels (0 – 3) are nonstationary time series.
2. Channel 0 is described by an ARIMA(0,1,2) model, channels 2 and 3 are described by an ARIMA(0,1,1) model, and channel 1 is described by an ARIMA(2,1,1)x$(1,0,0)_{25}$ seasonal model.
3. The paired channels have various combinations of cross correlations.

Differencing is necessary to obtain a stationary time series. Stationary series are required to understand autocorrelation, which is whether an observation is dependent upon previous observations. If there is no dependence on previous observations, it is highly likely to be white noise. If there is dependence, as these blank off data are, then it is not white noise, and, in the case of these data, some type of system behavior is suspected of causing the autocorrelation.

The cross-correlation analysis indicates that there is interaction from one channel to another. The autocorrelation and cross-correlation shown with the analysis below is that any observations of, say, the Sun, must be corrected for this systematic error before useful science can be derived.

## DESCRIPTIVE ANALYSIS

*Table 1: Channel output summary statistics.*

| Statistic | Channel 0 | Channel 1 | Channel 2 | Channel 3 |
|---|---|---|---|---|
| **Mean** | -20,748 | -31,886 | -17,180 | -19,747 |
| **Std Dev** | 8,022 | 13,965 | 8,003 | 8,479 |
| **Min** | -34,572 | -796 | -33,618 | -35,043 |
| **Max** | -6,227 | -61,941 | -2,673 | -2,791 |

This portion of the analysis begins with plots of the original data as presented in the file 071027blnkoffCOTDM.csv. These plots (Illustrations 1 and 2) show that the series contain at the very least a downward trend, excepting channel 1, which shows no clear trend. The most obvious feature of each series is their apparent lack of stationarity (for any two arbitrarily chosen intervals of equal length, the mean response within the intervals is statistically the same, has finite variance, and has the same autocorrelation structure), which is apparent by the slow decay across the lags in the channel autocorrelation plots in Illustration 3. Notice in Illustration 3 that, when the series are superimposed one upon the other, we see that channel 1 is clearly different channels 0, 2, and 3 (Snap0Jy, Snap1Jy, Snap2Jy, and Snap3Jy). Table 1 holds the series summary statistics.

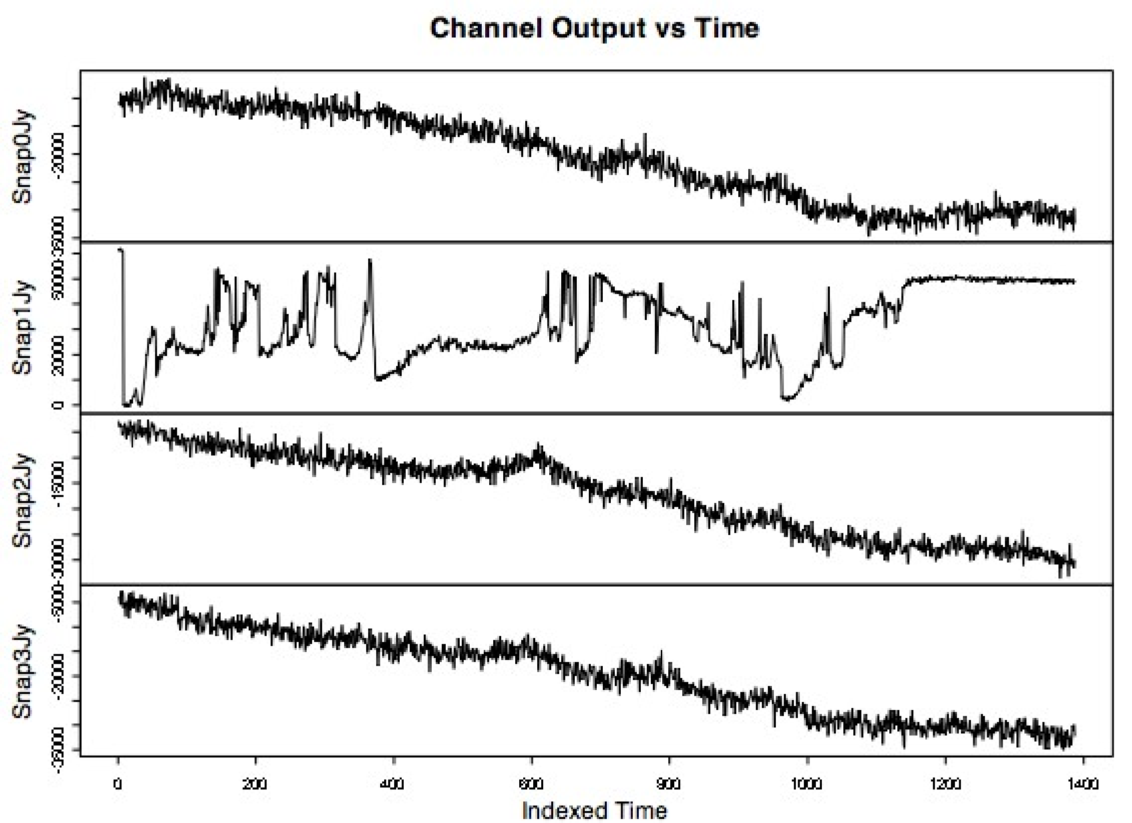


*Illustration 1: Channels 0 - 3 output. See Table 1 for these series statistics.*

We now take the first differences of each series to produce stationary series. In addition, we test whether the transformed series are normally distributed as, if they are, then we have Gaussian time series, which facilitates prediction bounds determination. Illustrations 4 – 7 are graphical displays of the fit of normal distributions to the series histograms. The Shapiro-Wilk test statistics for normally distributed data are shown in the captions of each of these Illustrations, which indicate that, except for channel 1, the first-differenced series are distributed as a normal distributions (we do not reject the null hypothesis as W is not sufficiently small). The channel 1 distribution is skewed to the left and non-normal (we reject the null hypothesis).

The evidence that our time series are not stationary comes from the AutoCorrelation Function (ACF). Simplistically, the ACF describes how an observation at time $t$ is related to the observations at times $t$-1 though $t$-$h$. A slowly decreasing ACF, as we see in Illustration 3, is indicative of a non-stationary time series.

And just what do these normally-distributed, first-differenced series look like? Illustrations 8 and 9 show the separate and the superimposed series, respectively. The salient feature is the lack of an obvious trend, and, though we need to test explicitly for cyclical behavior, there is no obvious periodic behavior. These plots lead us to suspect that the first-differenced transformations have given us stationary time series in addition to the normal distributions. The superimposed plot (Illustration 8) shows practically no difference among the channels.

The next and, perhaps, the most important question about these series is whether each channel is autocorrelated; and whether the series are cross-correlated. The next section is this parametric analysis.

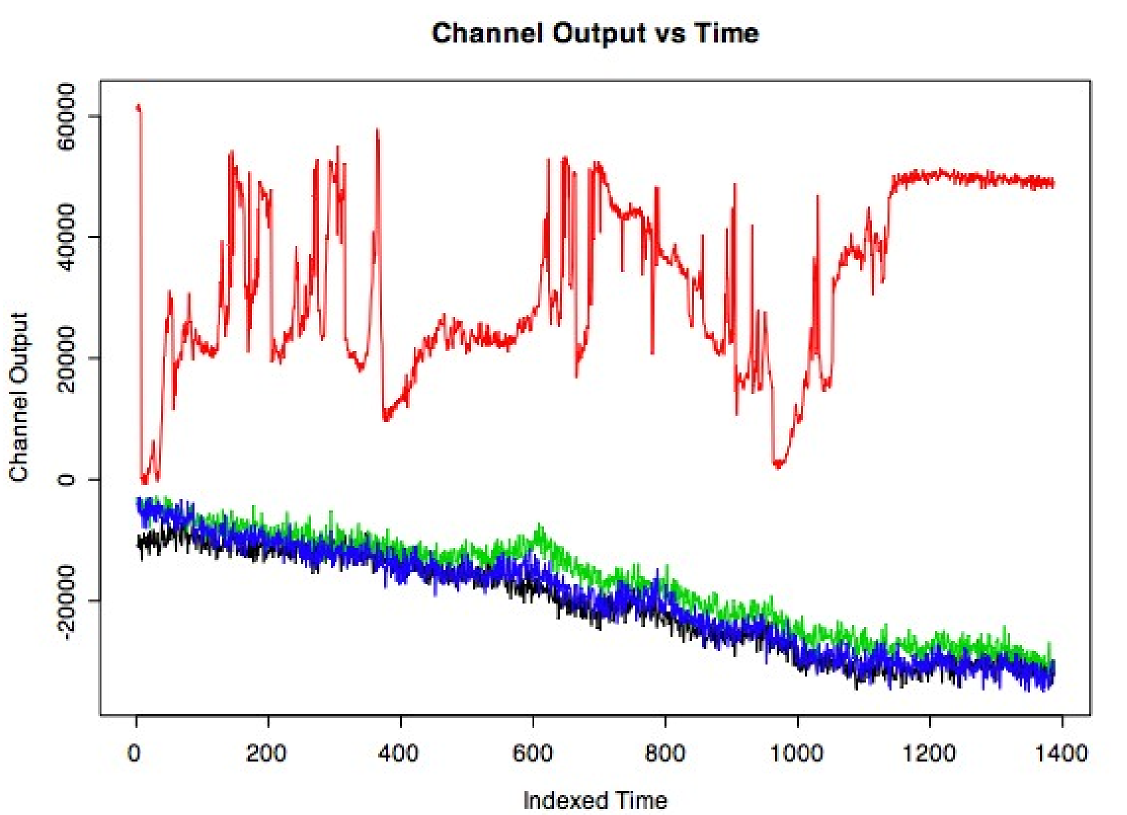


*Illustration 2: Channels 0 - 3 superimposed upon each other. Note channel 1 (red) is clearly differently from the others. The question of whether a significant difference exists among all the channels is determined below.*

## TIME SERIES MODELS – ARIMA MODELS

Our first-differenced time series exhibit no apparent deviations from stationarity as we shall see shortly from the corresponding rapidly decreasing ACF plots. This leads us to consider the class of time series models known as AutoRegressive Integrated Moving Average (ARIMA) models. As we saw above, the slowly decaying ACF of the non-differenced data indicated these data were not stationary. We also saw that by taking the first differences of each series, we obtained stationary time series. Thus, we have satisfied a mandatory condition for building ARMA models (a sub class of ARIMA models) for our data. An ARMA model is a causal model with 2 values $p$ and $q$ indicating the order of the autoregressive portion ($p$) of the the ARMA model, and the moving average portion ($q$). ARIMA models are specified by $p$, $q$, and $d$, where $p$ and $q$ are as before, and $d$ is the order of differencing

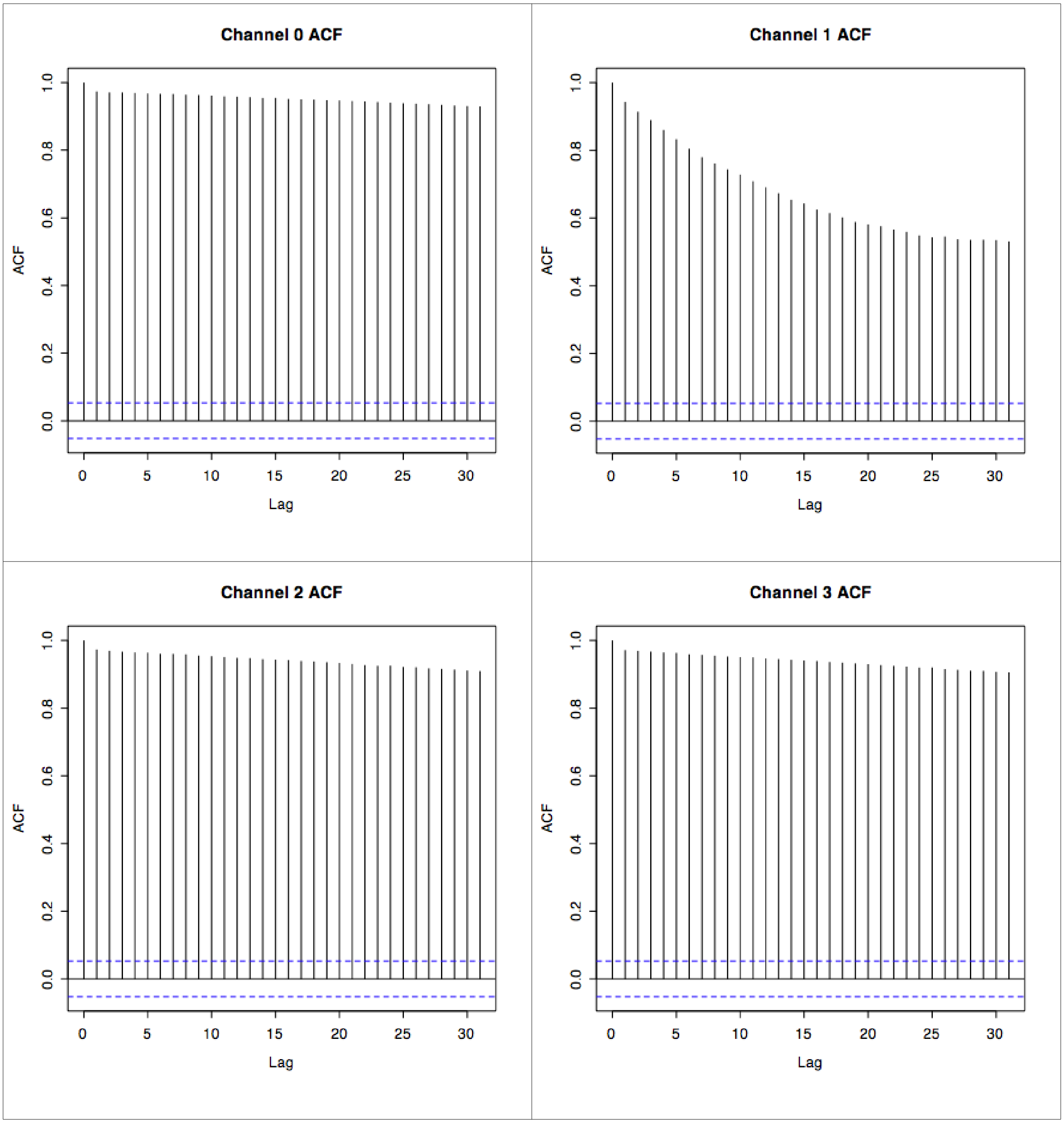


*Illustration 3: ACF for channels 0 - 3. The slowly decreasing amplitudes over the lags signifies nonstationary series.*

where $d = 1$ in our case, allowing us to use the ARMA class models.

The autoregressive specifier indicates the number of time units over which an observation $t$ is dependent on preceding observations. The moving average specifier indicates the number of white noise observations preceding $t$ that combine to equal the combined autoregressive terms. We can rearrange this equality such that future observations may be forecasted.

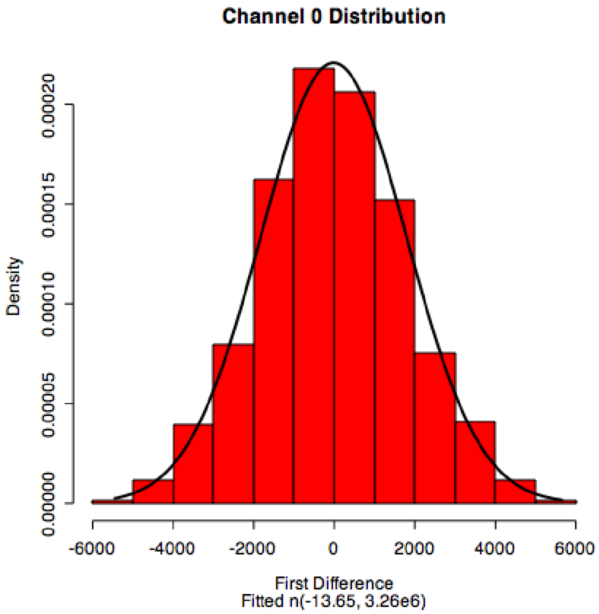


*Illustration 4: W = 0.999, p = 0.31.*

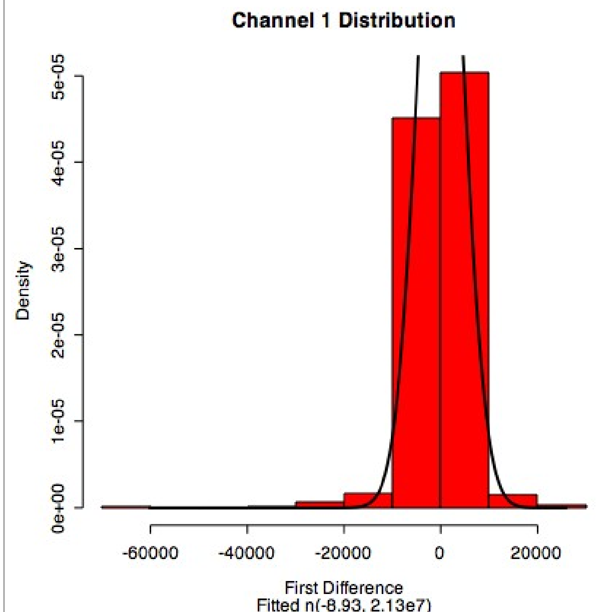


*Illustration 5: W = 0.610, p = 0.00.*

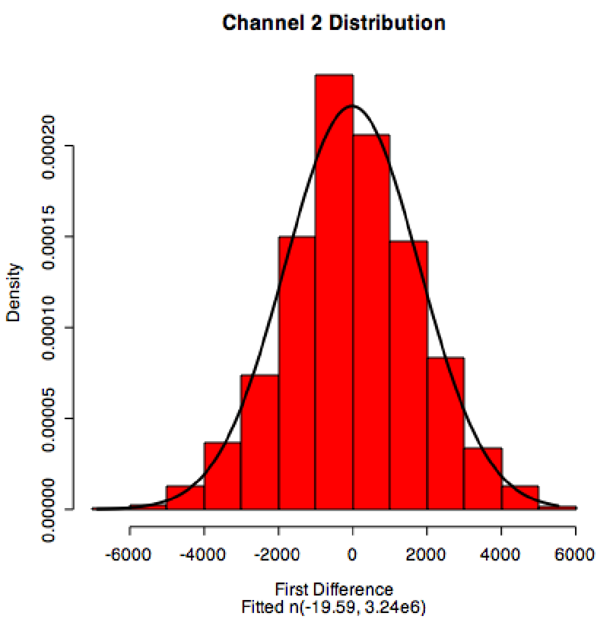


*Illustration 6: W = 0.999, p = 0.34.*

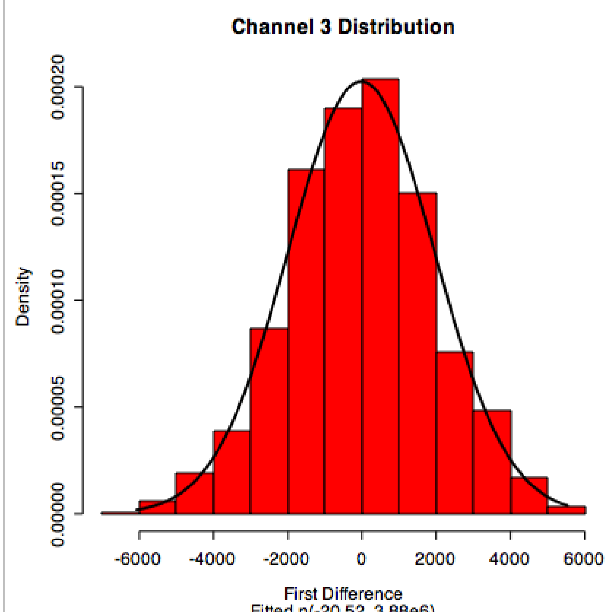


*Illustration 7: W = 0.999, p = 0.42.*

We now determine the ARMA($p$,$q$) specifiers for each of the four channels. The combination of the ACF and the Partial Autocorrelation Function (PACF) provide the infomation needed to find appropriate values. The ACF clearly shows that the times series are stationary as discussed above (Illustration 10). The specifier for the MA part ($q$) may be determined from the ACF plots. Examination

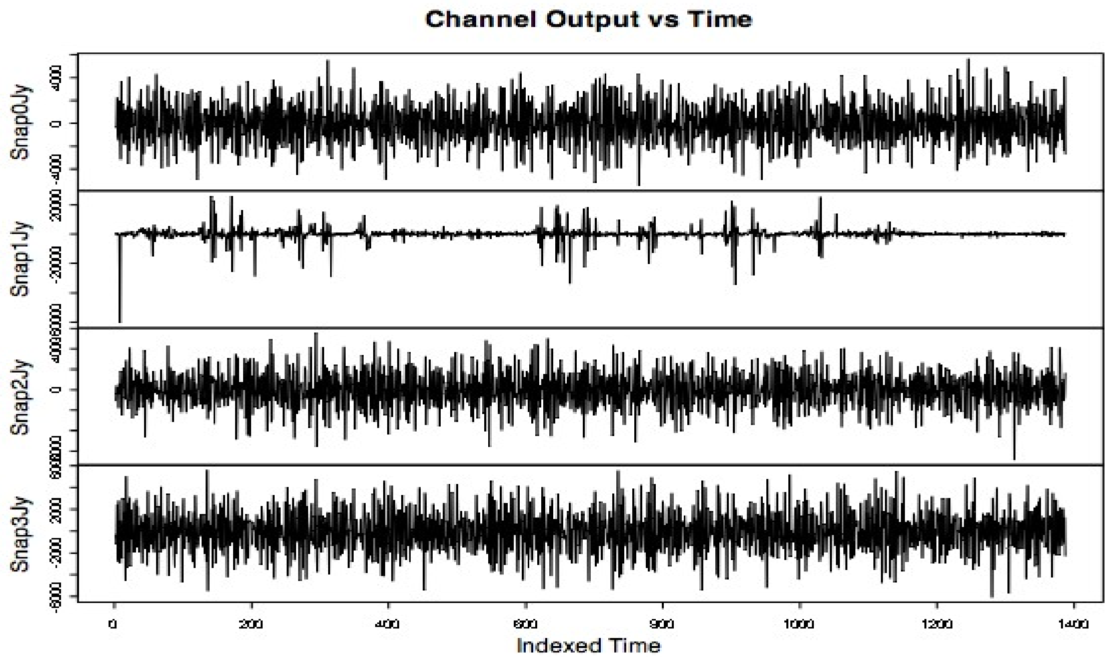


*Illustration 8: Channels 0 - 3 first differences (lag 1). See Table 2 for these series statistics.*

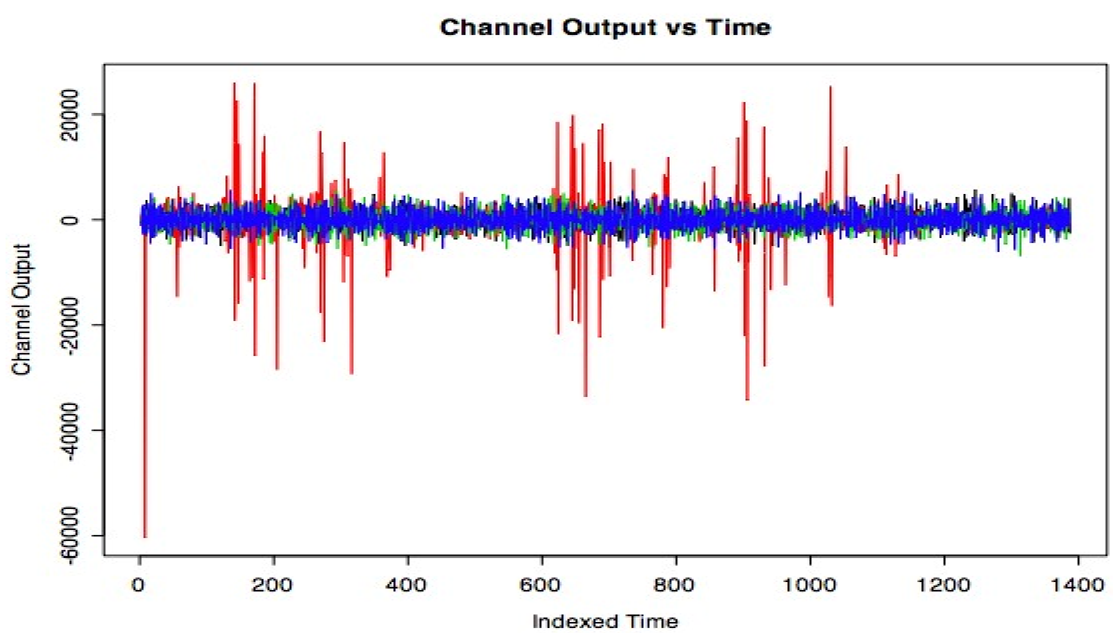


*Illustration 9: Channels 0 - 3 superimposed upon each other. Note channel 1 (red) no longer clearly falls below the others, though the variability periodicity appear quite different.*

of the PACF plots (Illustration 11) show that weighted observation-to-observation contribution to the current observation is ambiguous except for channel 1, which eliminates possible AR components, Thus, the ACF plots lead us to test an ARMA(0,2) model [an ARIMA(0,1,2) model] which is abbreviated to a MA(2) (Moving Average with $q = 2$) process for channel 0, and an ARMA(0,1) model [an ARIMA(0,1,1) model] which is abbreviated to a MA(1) (Moving Average with $q = 1$) process for channels 2 and 3. Channel 1 is more involved as there is strong evidence of a cyclical (called seasonal in statistical science) behavior, which we shall discuss in more detail below.

A moving average of order 1 averages the observations from the current and previous time intervals, while the moving average of order 2 averages the from the current and the previous two time intervals. Thus, after removing the mean, we have, respectively the MA(1) and MA(2) equations:

$$X_t = Z_t + \theta Z_{t-1}$$
$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$$

where $X_t$ is the $t^{th}$ observation, and $\{Z_t\}$ is white noise with mean 0 and finite variance $\sigma^2$ if and only if $\{Z_t\}$ are not correlated, and is denoted $\{Z_t\} \sim WN(0, \sigma^2)$. The parameter $\theta$ in the MA(1) equation and the parameters $\theta_1$ and $\theta_2$ are to be determined from each series data. The model parameters are, by channel, in Table 2.

*Table 2: ARIMA(0,1,1) model parameters for channels 0, 2, and 3.*

```
arima(x = tmp[, 1], order = c(0, 0, 2))

Coefficients:
          ma1     ma2  intercept
      -0.9038  0.0095   -15.3604
s.e.   0.0290  0.0286     3.8325

sigma^2 estimated as 1797403:  log likelihood = -11947.94,  aic = 23903.88
```

```
arima(x = tmp[, 3], order = c(0, 0, 1))

Coefficients:
          ma1  intercept
      -0.8945    -18.996
s.e.   0.0109      3.864

sigma^2 estimated as 1836986:  log likelihood = -11963.03,  aic = 23932.07
```

```
arima(x = tmp[, 4], order = c(0, 0, 1))

Coefficients:
          ma1  intercept
      -0.8872   -20.0735
s.e.   0.0127     4.5163

sigma^2 estimated as 2196322:  log likelihood = -12086.81,  aic = 24179.63
```

The ARMA(0,1) model works well for channels 2 and 3, as is seen in the model diagnostics in Illustration 13, and the diagnostics in Illustration 12 shows a good fit for channel 0 of an ARMA(0,2). However, as is seen in the ACF and PACF plots for channel 1 (Illustrations 10 and 11), a definite enhancement is required. Closer inspection of these plots show a significant lag at 25, which implies that the channel 1 data are better fitted to a seasonal model. We can then consider ARIMA models for the sub-series sample 25 intervals apart. So we may include both seasonal and non-seasonal terms by extending the differencing technique used above to achieve stationarity. We introduce the lag-$s$ difference as $X_t - X_{t-s} = (1-B^s)X_t$, where $B^s$ is the back-shift operator that selects $X_{t-s}$. Thus,

$$\nabla = (1-B)^d (1-B^s)^d X_t = (1-B)^1 (1-B^{25})^1 X_t = X_t - X_{t-1} - X_{t-25} + X_{t-1} X_{t-25}$$

where $(1-B)^1$ is the first difference operator we use to de-trend the series. This leads us to consider an ARIMA model of the general form of

$$\Phi_{AR}(B)\Phi_{SAR}(B^s)X_t = \Theta_{MA}(B)\Theta_{SMA}(B^s)Z_t,$$

where *SAR* is the seasonal autoregressive component, and *SMA* is the seasonal moving average component. As before, $\Phi_{AR}$, $\Phi_{SAR}$, $\Theta_{MA}$, and $\Theta_{SMA}$ are parameters to be estimated from the channel 1 data.

The diagnostics (see Illustration 12) confirm that channel 1 is modeled well by an $ARIMA(2,1,1)\times(1,0,0)_{25}$. After removing the mean of the differenced series, the model is

$$X_t = \phi_1 X_{t-1} - \phi_2 X_{t-2} - \phi_3 X_{t-25} + Z_t + \theta_1 Z_{t-1}$$

where, as before, $X_t$ is the $t^{th}$ observation, and $\{Z_t\}$ is white noise with mean 0 and finite variance $\sigma^2$ if and only if $\{Z_t\}$ are not correlated, and is denoted $\{Z_t\} \sim WN(0,\sigma^2)$. The parameters $\phi_1$ through $\phi_5$ and $\theta_1$ through $\theta_3$ are to be determined from the series data. These parameter estimates are in Table 3.

*Table 4:* $ARIMA(2,1,1)\times(1,0,0)_{25}$ *model parameters for channel 1.*

```
arima(x = ch1s, order = c(2, 1, 1), seasonal = list(order = c(1, 0, 0), period = 25))

Coefficients:
          ar1      ar2      ma1     sar1
      -0.2430  -0.1008  -0.0356  -0.0828
s.e.   0.2582   0.0703   0.2596   0.0288

sigma^2 estimated as 19536135:  log likelihood = -13600.71,  aic = 27211.42
```

## CROSS-CORRELATION

The final portion of our analysis examines the question as to whether the channel series are cross-correlated, as these series are measured over the same time period and intervals. Cross-correlation among the channels suggests one or more channels may be leading indicators. Leading indicators may be real, or they may be an artifact of the electronics.

Illustration 14 shows the lag 1 pairwise channel scatter plots with loess smoothing. We see that channels 0 and 1 don't appear to be correlated between themselves or with channels 2 and 3. Channels 2 and 3 appear to be correlated between themselves. To see if these relations indeed hold, we examine

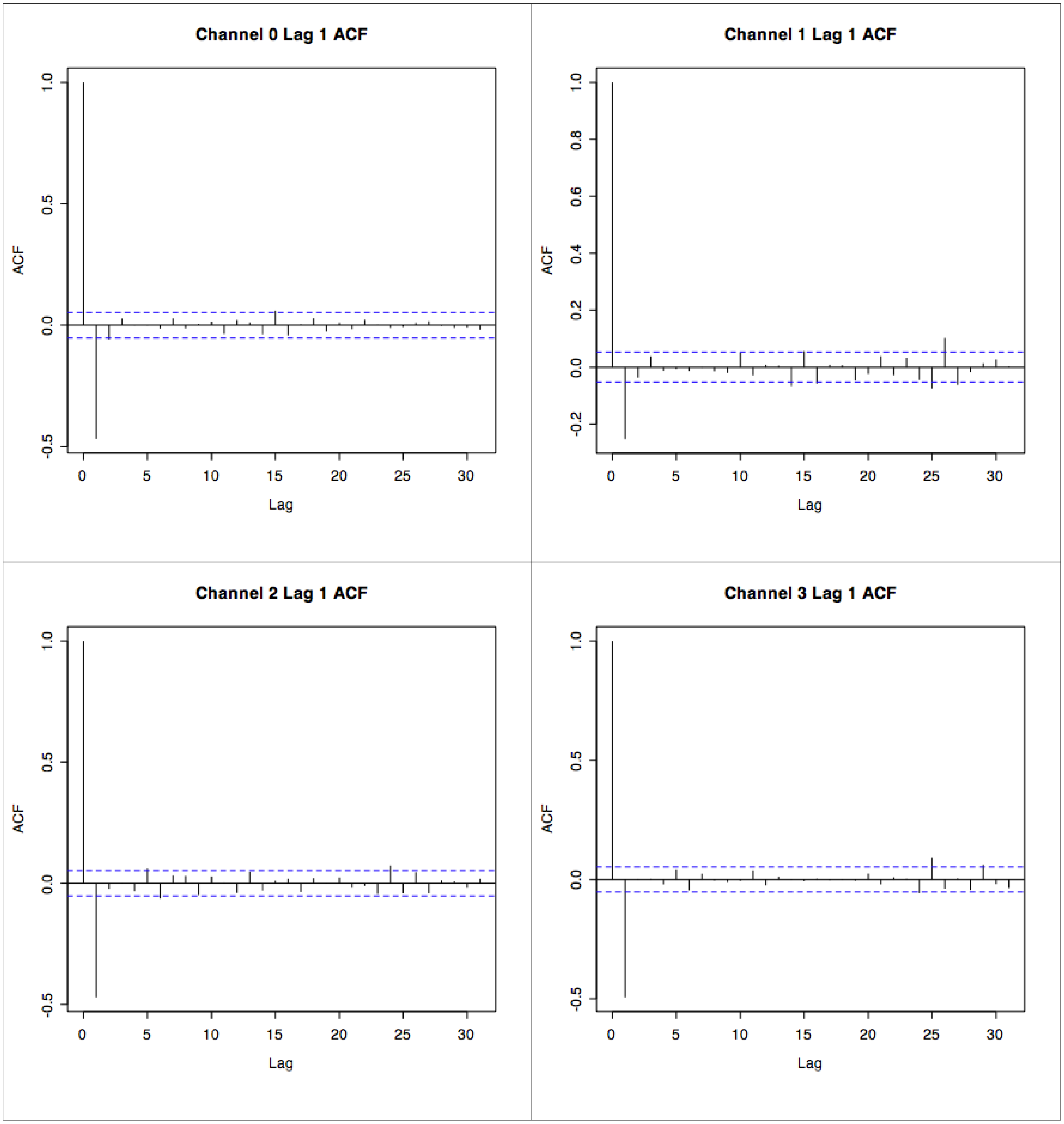


*Illustration 10: ACF for channels 0 - 3. The rapidly decreasing amplitudes over the lags signifies stationary series. Also, the order of the ARMA model (MA) can be determined.*

the multivariable ACF and PACF plots. The ACF shows at best weak correlation between channels 2 and 3, but the PACF plots show significant leading or lagging between channel pairs. Channel 0 lags channels 1, 2, and 3. Channel 1 leads channels 0, 2, and 3. Channel 2 leads channels 0 and 3. The reason for these apparent cross-correlations needs to be determined.

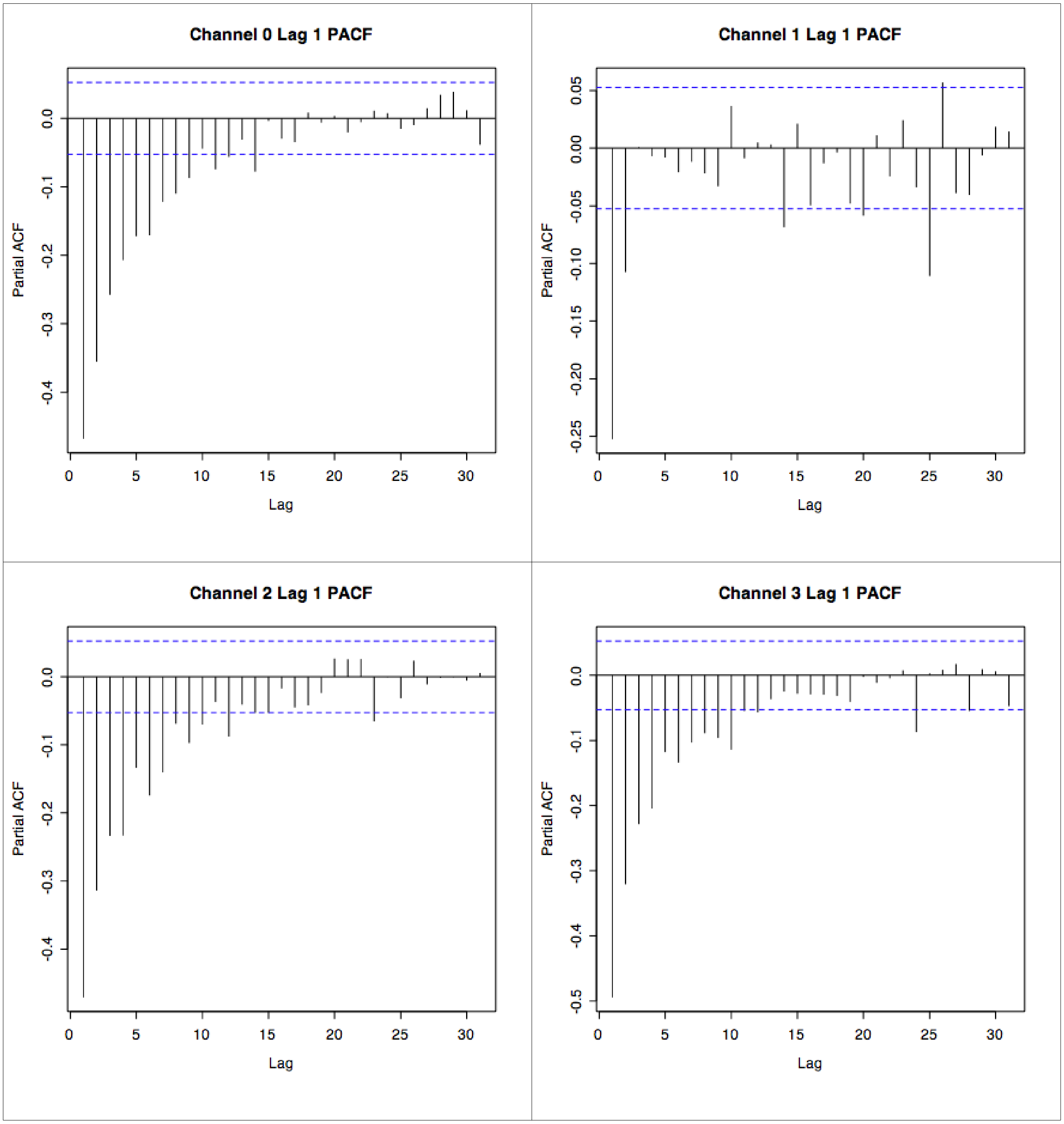


*Illustration 11: PACF for channels 0 - 3. The gradually changing amplitudes over the lags indicates these series have no autocorrelated behavior.*

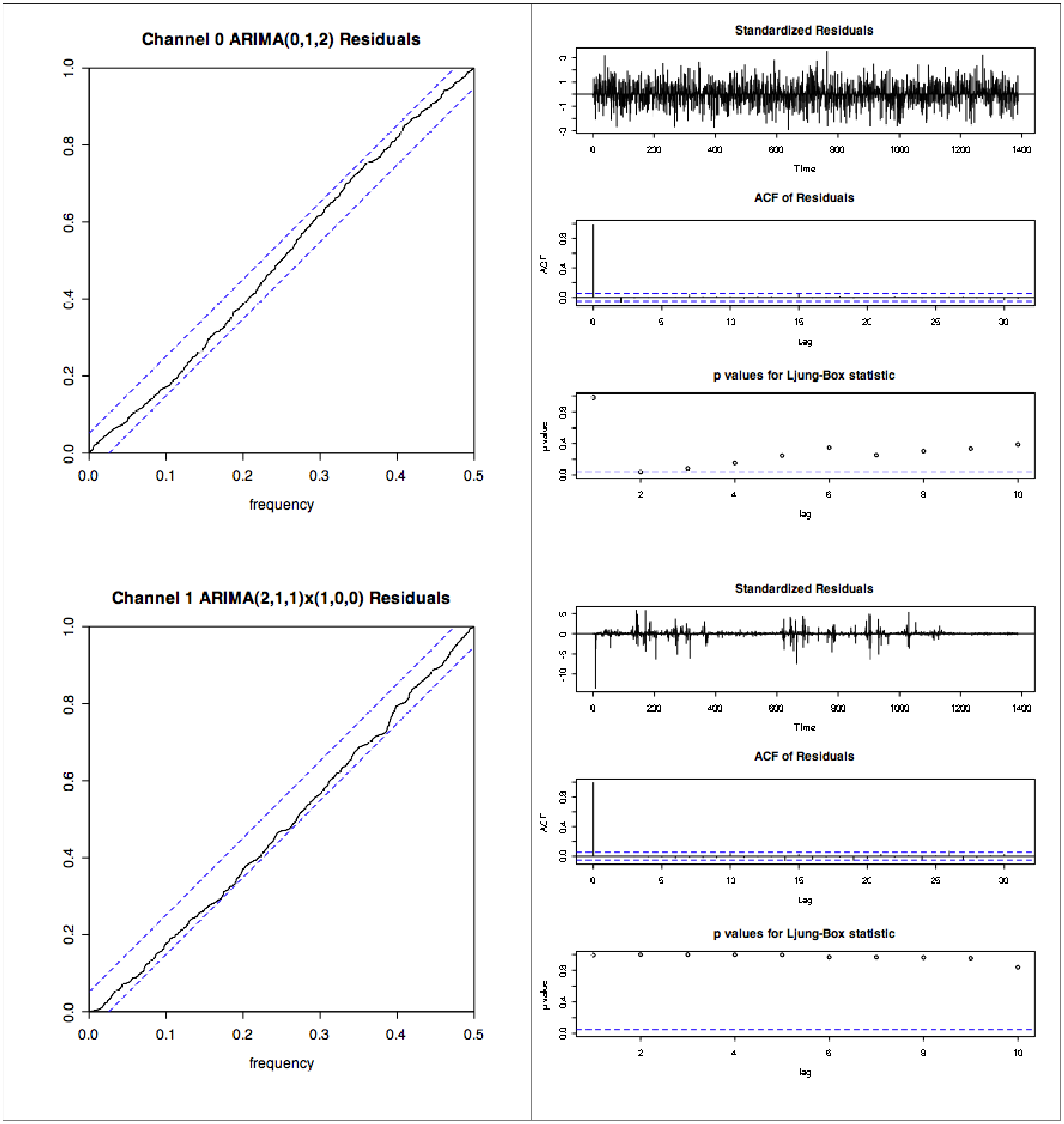


*Illustration 12: ARMA (ARIMA) model diagnostics support an ARMA(0,1) model for channel 0, and an ARMA(0,2) model for channel 1.*

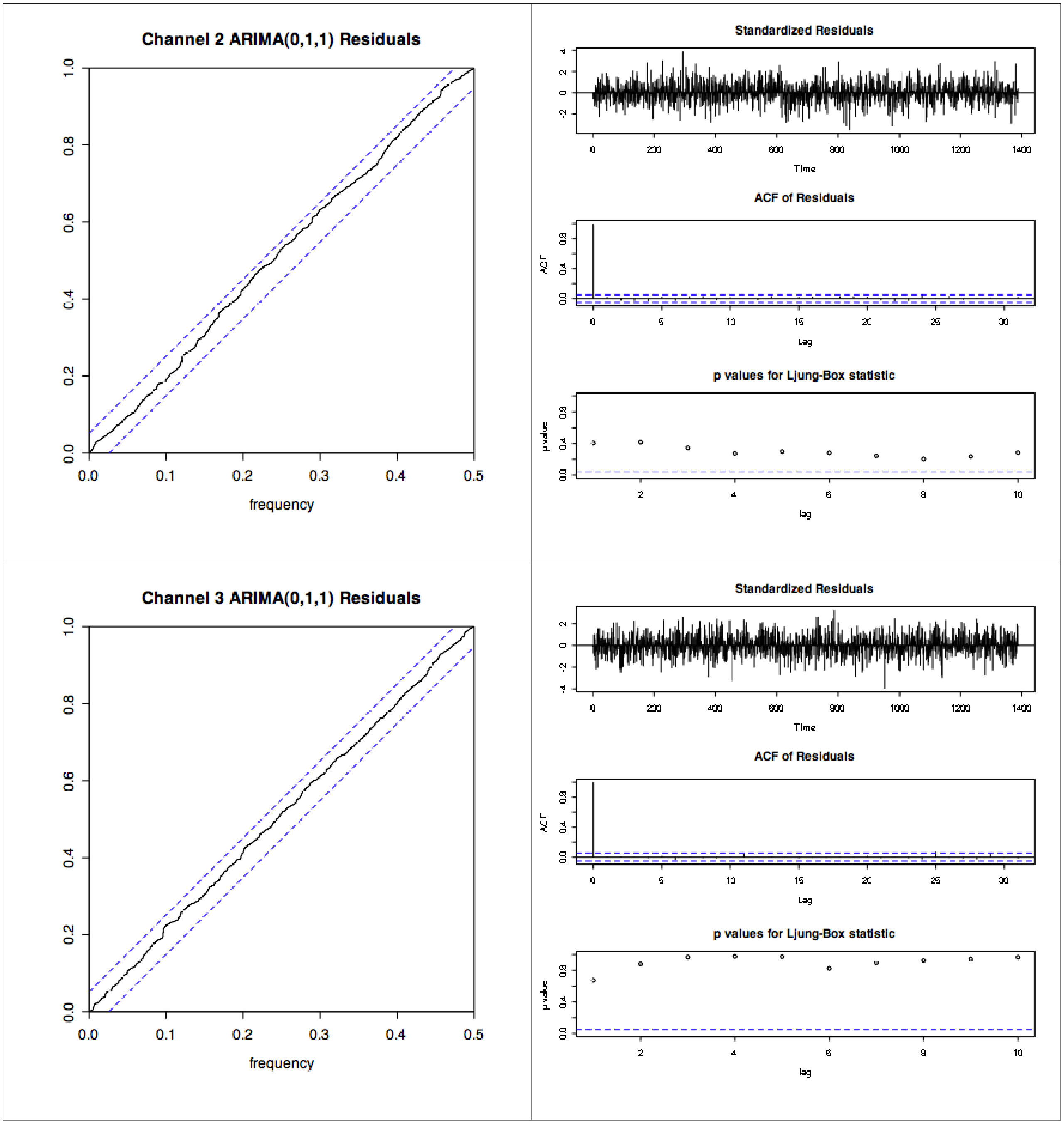


*Illustration 13: ARMA (ARIMA) model diagnostics support an ARMA(0,1) model for channels 2 and 3.*

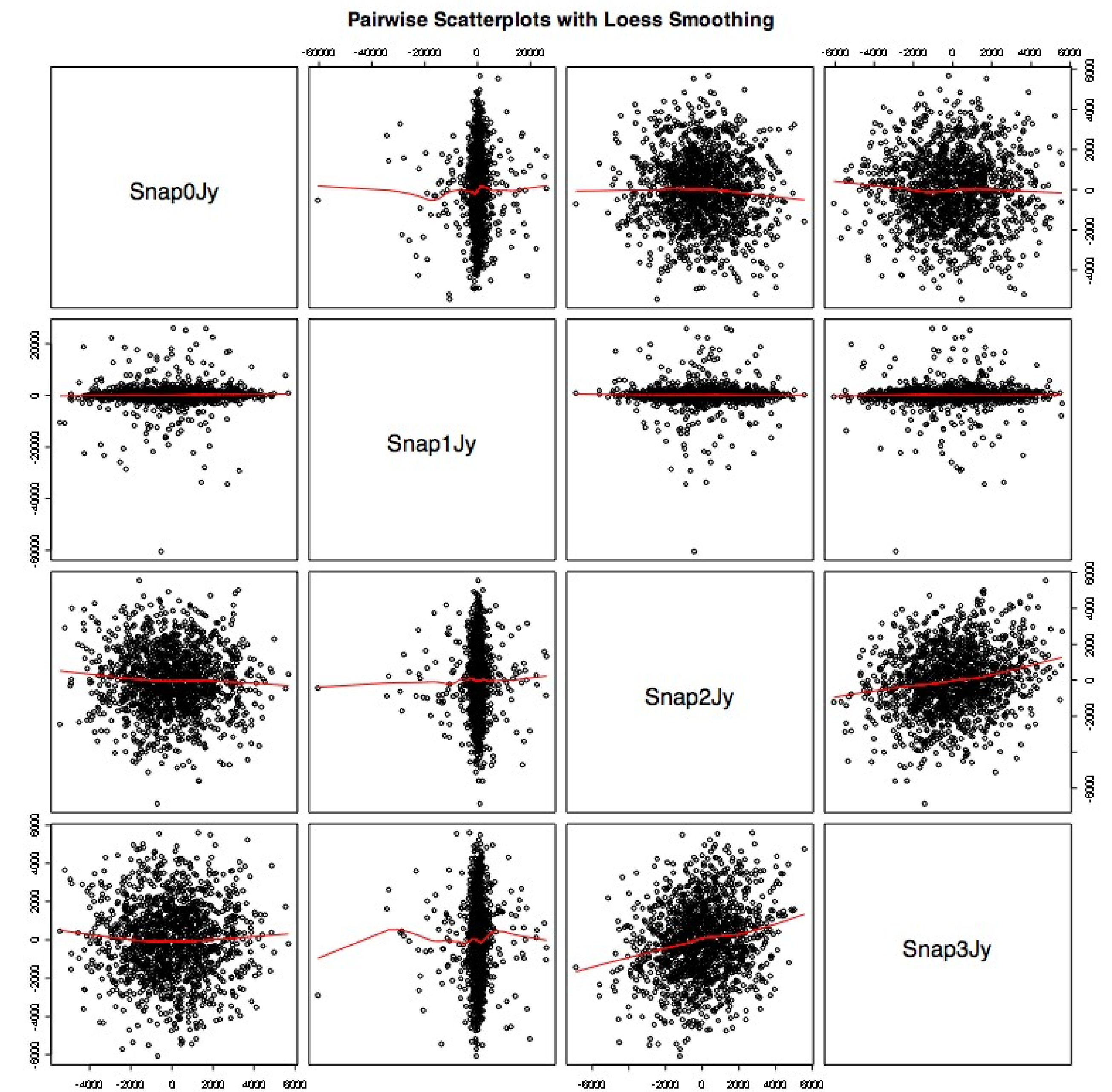


*Illustration 14: Pairwise scatter plots of the channel lag 1 data with Loess smoothing.*

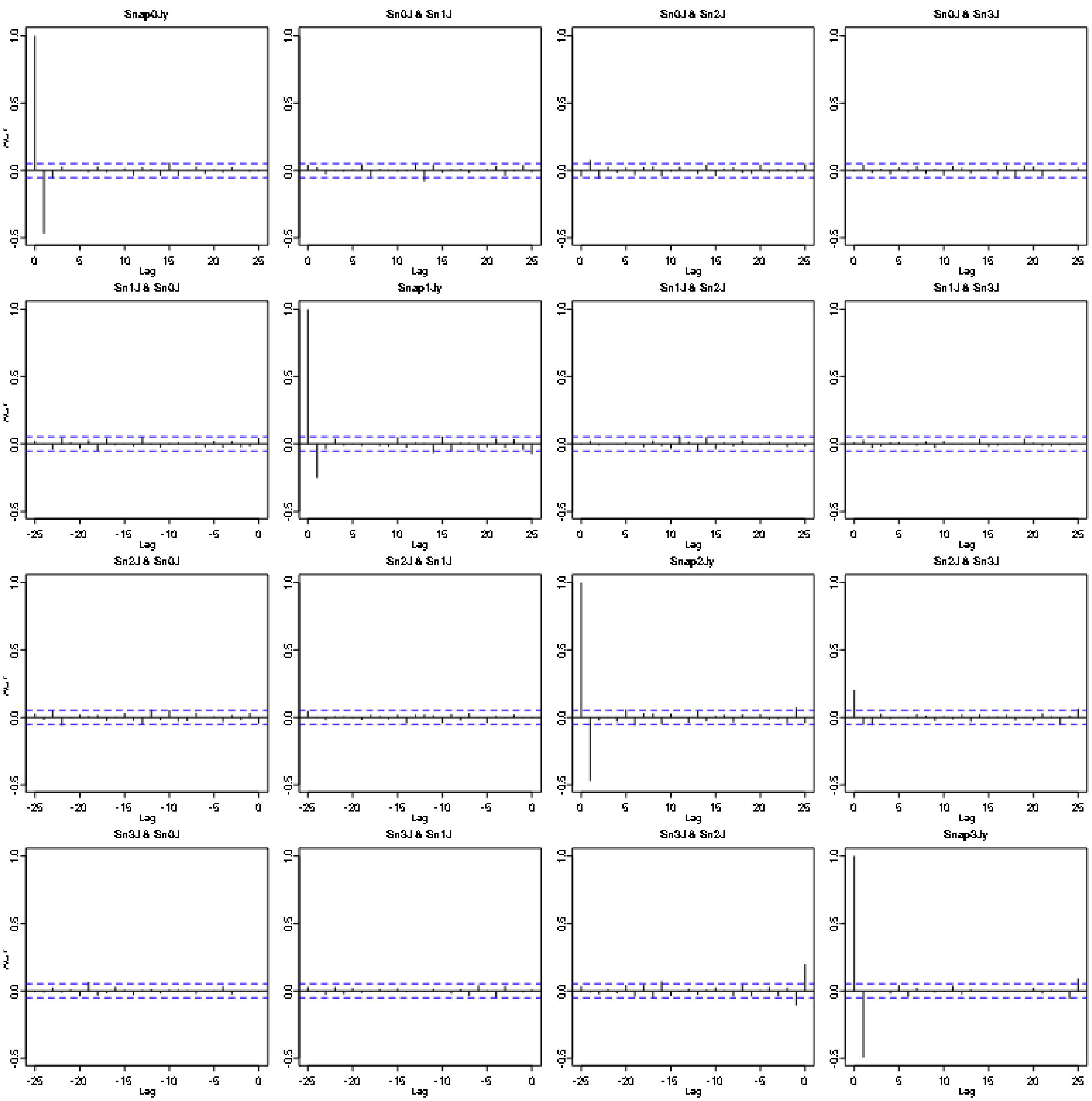


*Illustration 15: Pairwise cross-correlated channel ACF plots.*

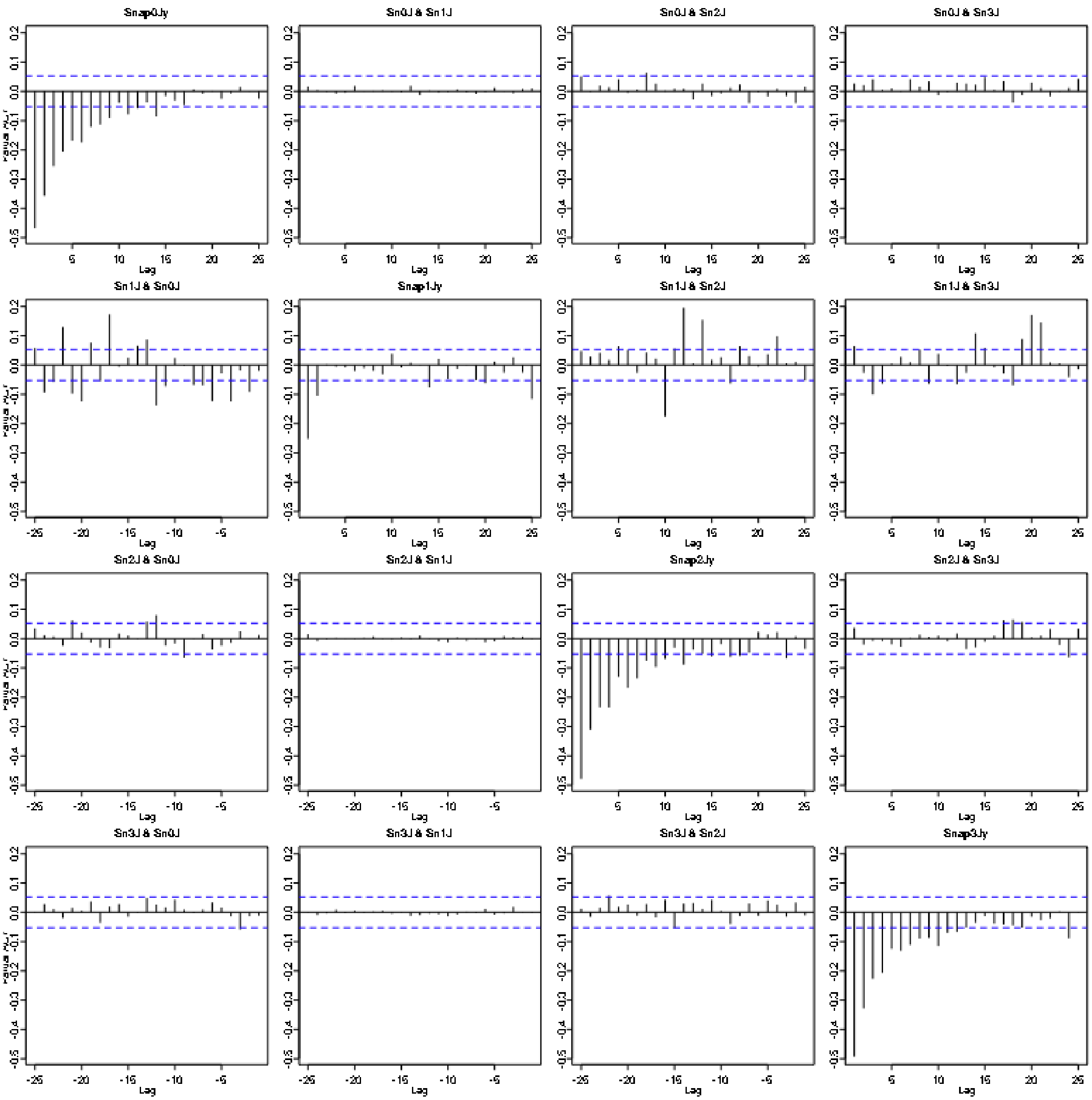


*Illustration 16: Pairwise cross-correlated channel PACF plots.*