Journal of Environmental Statistics

March 2013, Volume 4, Issue 11. http://www.jenvstat.org

# Optimal Deseasonalization for Monthly and Daily Geophysical Time Series

**A. Ian McLeod and Hyukjun Gweon**
Western University

**Abstract**

Deseasonalized geophysical time series are often used in time series models (Hipel and McLeod 1994). In this article an optimal method for selecting the deseasonalization transformation is suggested and an R package implementation (McLeod and Gweon 2012) is discussed. Our deseasonalization method may be used with the recently developed periodic autoregression model for daily river flow suggested by Tesfaye, Anderson, and Meerschaert (2011) and for the hierarchical Bayes modeling for multi-site daily temperature series discussed by Craigmile and Guttorp (2011).

Keywords: autoregressive model, harmonic regression, R, seasonality, time series modeling.

## 1. Introduction

Many geophysical time series are available on a monthly or daily basis and exhibit obvious seasonal features. These time series are often quite long. Lattice graphics capabilities available in R (R Development Core Team 2008) provide excellent multi-panel displays (McLeod, Yu, and Mahdi 2012). The built-in R function `stl()` provides a seasonal-trend decomposition that may be rendered in an attractive visual display as illustrated in McLeod *et al.* (2012) and is discussed in more detail by Cleveland (1993) for the famous (Wikipedia 2011) monthly Mauna Loa $CO_2$ time series. Long time series may also be visualized dynamically, like a movie, as illustrated in McLeod (2012b).

In this article our focus is on seasonal geophysical time series that are stationary after deseasonalizing by subtracting the seasonal mean and/or dividing by the seasonal standard deviation. [1] Suppose our time series consists of $n$ successive monthly or daily values denoted

[1] Seasonal economic/financial time series are often more complicated due to non-stationarity and weekday/holiday artifacts.

by $z_t, t = 1, \ldots, n$ where $t$ is the observation number. Then for modeling purposes it is often convenient to work with the deseasonalized version $w_t = (z_t - \mu_t)/\sigma_t$, where $\mu_t$ and $\sigma_t$ are the seasonal mean and standard deviation. In the monthly case, often $\mu_t$ and $\sigma_t$ are simply estimated by the monthly means and standard deviations. When the seasonal variances are constant, we may wish to use the detrended series, $w_t = z_t - \mu_t$.

Empirical (McLeod 1993) and theoretical (Ledolter and Abraham 1981) analyses have demonstrated that the principle of parsimony advocated for time series models by Box, Jenkins, and Reinsel (2005) is useful in selecting models that provide the best forecasts. In summary, this principle suggests that the time series model with the fewest number of parameters that adequately fits the data is preferred. Following this principle Hipel and McLeod (1994, §13.3.3) described a Fourier based approach for selecting the minimum number of Fourier components to use in the deasonalizing transformation in the case of monthly hydrological time series. In the case of monthly time the seasonal frequency and its harmonic multiples, $12k/n$ are all Fourier frequencies so the problem reduces to an orthogonal regression that allows for efficient computation (Bloomfield 2004, Ch. 4). With daily time series it is natural to take the seasonal period to be $s = 365.25$ and so, in this case, the Fourier approach cannot be used but instead we may use an harmonic regression Craigmile and Guttorp (2011).

## 2. Harmonic Regression

In general, we may use harmonic regressions to estimate $\mu_t$ and $\sigma_t$. To estimate $\mu_t$, we fit,

$$z_t = A_\mu^{(0)} + \sum_{k=1}^{F_\mu} \left( A_\mu^{(k)} \cos(2\pi k t/s) + B_\mu^{(k)} \sin(2\pi k t/s) \right) + u_t \tag{1}$$

where $A_\mu^{(0)}$ is the overall mean, $F_\mu$ denotes the number of sinusoids used, $A_\mu^{(i)}, B_\mu^{(i)}, k = 1, \ldots, F_\mu$ are the sinusoid parameters, $s$ is the seasonal period with $s = 12$ or $s = 365.25$ corresponding to the monthly and daily cases respectively, and $u_t$ is the mean-zero error that is assumed to be stationary. It is well-known that in this case the least-squares estimates for the parameters $A_\mu^{(0)}$, $A_\mu^{(k)}, B_\mu^{(k)}, k = 1, \ldots, F_\mu$ are asymptotically fully efficient (Hannan 1970, §VII, Theorem 11). The estimated seasonal mean may be written,

$$\hat{\mu}_t = \hat{A}_\mu^{(0)} + \sum_{k=1}^{F_\mu} \left( \hat{A}_\mu^{(k)} \cos(2\pi k t/s) + \hat{B}_\mu^{(k)} \sin(2\pi k t/s) \right), \tag{2}$$

where $\hat{A}_\mu^{(0)}, \hat{A}_\mu^{(k)}, \hat{B}_\mu^{(k)}$ are the least-squares estimates. Similarly the estimated seasonal variances,

$$\hat{\sigma}_t^2 = \hat{A}_\sigma^{(0)} + \sum_{k=1}^{F_\sigma} \left( \hat{A}_\sigma^{(k)} \cos(2\pi k t/s) + \hat{B}_\sigma^{(k)} \sin(2\pi k t/s) \right), \tag{3}$$

where $\hat{A}_\sigma^{(0)}, \hat{A}_\sigma^{(k)}, \hat{B}_\sigma^{(k)}, k = 1, \ldots, F_\sigma$ are the least squares estimates in the regression,

$$\hat{u}_t^2 = A_\sigma^{(0)} + \sum_{k=1}^{F_\mu} \left( A_\sigma^{(k)} \cos(2\pi k t/s) + B_\sigma^{(k)} \sin(2\pi k t/s) \right) + v_t \tag{4}$$

where $\hat{u}_t^2$ is the squared residual, $\hat{u}_t = (z_t - \hat{\mu}_t)^2$, and $v_t$ is the mean-zero stationary error term. The deseasonalized time series, $w_t = (z_t - \hat{\mu}_t)/\hat{\sigma}_t$ may then be obtained. When $F_\mu = 0$, we set $\hat{\mu}_t$ equal to the sample mean of the original series $z_t$. Similarly when $F_\sigma = 0$, $\hat{\sigma}_t$ is set to the sample standard deviation of $z_t$.

As described in Hipel and McLeod (1994, §6.3 and §13.3) we may use the AIC (Akaike 1974) or BIC (Schwarz 1978) criterion to select $F_\mu$ and $F_\sigma$, the number of harmonics used. More generally we may use generalized AIC , defined as $\text{GIC}_\alpha = -2 \log L + \alpha k$, where $L$ denotes the maximized value of the log-likelihood function and $\alpha$ is the tuning parameter with $\alpha = 2$ for the AIC and $\alpha = \log(n)$ for the BIC . Other choices for $\alpha$ have been discussed by (Taniguchi and Hirukawa 2012; Xu 2010; Xu and McLeod 2012).

For any fixed choices of $F_\mu$ and $F_\sigma$, the deseasonalized stationary time series, $w_t$, is assumed to be adequately modeled using an AR($p$). The R package `FitAR` (McLeod, Zhang, and Xu 2011) is used to automatically select $p$ and determine the value of the $\text{GIC}_\alpha$, denoted by $\text{GIC}_\alpha(F_\mu, F_\sigma)$.[2] As $F_\sigma$ changes, the scale changes for $w_t$ and so it is necessary to adjust $\text{GIC}_\alpha(F_\mu, F_\sigma)$ in order to be able to compare the effect of different choices of $F_\sigma$. This involves accounting for the transformation $w_t \longleftrightarrow z_t$ in the evaluation of the likelihood. The determinant of the logarithm of the Jacobian for the transformation $w_t \longleftrightarrow z_t$ is

$$\log J = -\sum_{i=1}^{n} \log \hat{\sigma}_t. \tag{5}$$

Hence the $\text{GIC}_\alpha(F_\mu, F_\sigma)$ corresponding to a specific choice of $F_\mu$ and $F_\sigma$ on the same scale as the original data $z_t$ is given by

$$\text{GIC}_\alpha^{(z)}(F_\mu, F_\sigma) = \text{GIC}_\alpha(F_\mu, F_\sigma) - 2 \log J, \tag{6}$$

where $\text{GIC}_\alpha(F_\mu, F_\sigma)$ is computed using the transformed series.

For monthly time series we may enumerate $\text{GIC}_\alpha^{(z)}(F_\mu, F_\sigma)$ for $F_\mu = 0, 1, \ldots, 6$ and $F_\sigma = 0, 1, \ldots, 6$ and select the optimal deseasonalization according to our chosen $\text{GIC}_\alpha$-criterion. In this case $F_\mu = 6$ would correspond to simply using the monthly means while $F_\sigma = 6$ corresponds to using the monthly seasonal standard deviations. Note that $F_\mu, F_\sigma \le 6$ to avoid aliasing (Bloomfield 2004; McLeod 2012a).

With daily time series often only a few harmonics are required for deseasonalization since the seasonal term is usually not too complicated. Hence we may choose a upper limits $U_m$ and $U_s$ and evaluate $\text{GIC}_\alpha^{(z)}(F_\mu, F_\sigma)$ for $F_\mu = 0, \ldots, U_m$ and $F_\sigma = 0, \ldots, U_s$. In practice, $U_m = U_s = 6$ is often reasonable for many daily time series.

## 3. R Package

Our R package (McLeod and Gweon 2012) implements the methods discussed in §2. By default the BIC is used to select the optimum transformation although other criteria are available in the package as well.

Often it may be of interest to compare the optimum transformation with transformations that are close to it since if there is only a small difference, a simpler transformation may be more desirable. A good way to compare possible transformations is to use the relative plausibility.

[2] Occasionally, it may happen that $\hat{\sigma}_t^2 < 0$ defined in eqn. (3) is negative. This usually happens when $F_\sigma$ or $F_\mu$ are too small or when the series needs a logarithmic or other type of transformation. When this happens we may simply set the value of the $\text{GIC}_\alpha$ to $\infty$ so this transformation will not be selected.

The relative plausibility of models, $i = 1, \ldots, I$ with generalized AIC, $\text{GIC}_\alpha(i), i = 1, \ldots, I$, is defined as

$$P_i = \exp\{-0.5(\text{GIC}_\alpha(i) - \text{GIC}_\alpha^\star)\}, \tag{7}$$

where $\text{GIC}_\alpha^\star = \min_i \text{GIC}_\alpha(i)$ (Akaike 1978; Hipel and McLeod 1994). This concept is similar to relative likelihood (Sprott 2000, §2.4). The output for the code snippets in §4 numerically illustrates the use of the relative plausibility.

All computations reported in §4 took only a few seconds. If necessary, with longer or more complicated seasonal time series or for Monte-Carlo applications, the enumeration to find the best $\text{GIC}_\alpha$ model could be vastly speeded up by using the R built-in package **parallel**.

# 4. Illustrative Examples

## 4.1. Monthly Saugeen River Flow

As an illustrative application, consider the mean monthly flow of Saugeen River at Walkerton in cumecs ($m^3$/sec) over the period from January 1915 to December 1979. There are $n = 744$ consecutive values. The ratio of maximum/minimum is about 56, so Tukey's rule-of-thumb (Tukey 1977, p. 397) suggests that a logarithmic transformation is in order and this was confirmed using a Box-Cox analysis (Hipel and McLeod 1994, §13.4.2). The lattice style boxplot in Figure 1 demonstrates that the log transformation makes the data distribution more symmetrical.

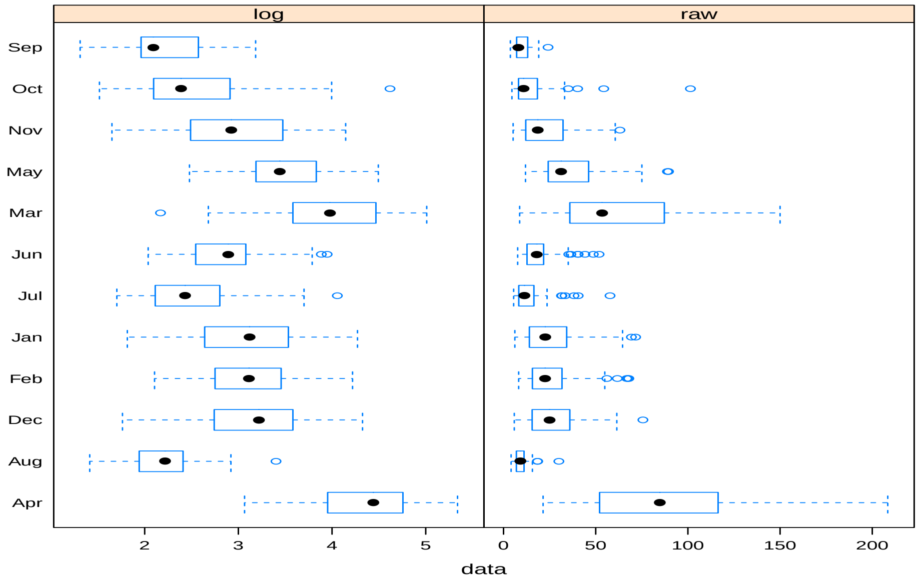


Figure 1: Comparing boxplots for original and log-transformed monthly flows for the Saugeen River

The lattice-style multipanel time series plot in Figure 2 shows the log series exhibits strong seasonality but no time trends or outliers.

Another useful plot for monthly time series obtained using the built-in R function `monthplot()` and is illustrated in Figure 3. This plot displays not only the seasonal pattern but the time

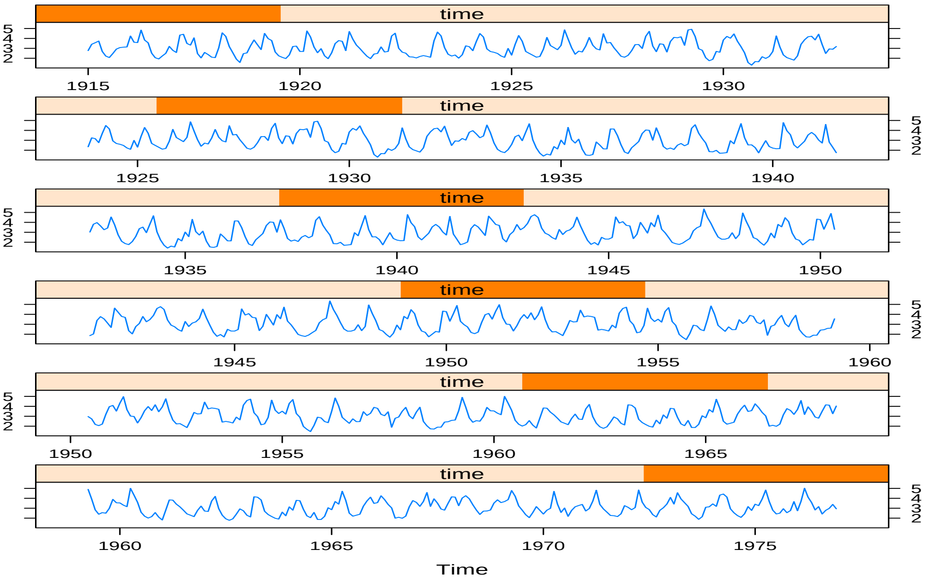


Figure 2: Lattice time series plot for monthly Saugeen flows.

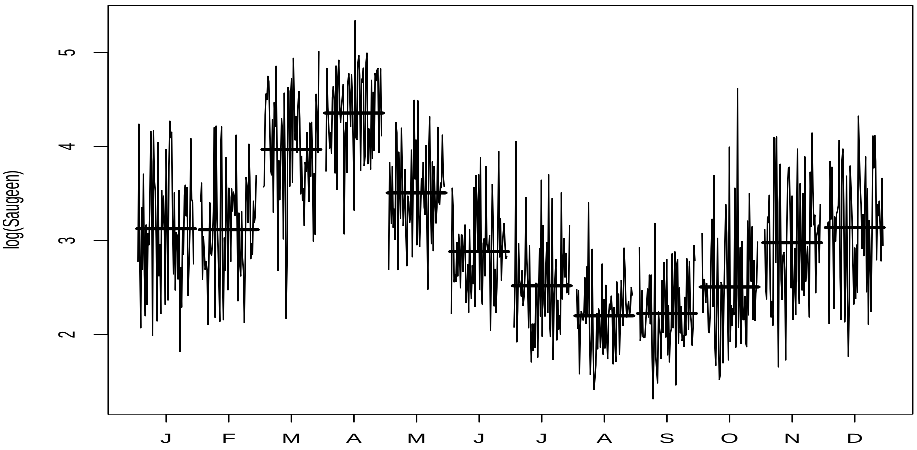


Figure 3: Monthplot for monthly Saugeen River flows.

series plot for each month separately. From this plot we don't see any evidence of trend-like changes occurring in the monthly subseries so the type of deseasonalizing transformation suggested in §2 is appropriate.

The output for the deseasonalization function, given in the code snippet below, shows that using the AIC results in $F_\mu = 5$ and $F_\sigma = 4$. This result agrees with that given by Hipel and McLeod (1994, §13.4.2) who also found that the optimal transformation was $F_\mu = 5$ and $F_\sigma = 4$ using an ARMA $(1, 1)$ model instead of selecting the best fitting AR . If the BIC criterion is used a more parsimonious deseasonalization with $F_\mu = 1$ and $F_\sigma = 1$ is obtained and only one other model has BIC-plausibility greater than 1%.

### Code Snippets

The R code snippet below generates the lattice-style boxplot in Figure 1.

```
R >require("deseasonalize")
R >require("lattice")
R >n <- length(Saugeen)
R >i<-as.vector(cycle(Saugeen))
R >m<-month.abb[i]
R >Saugeen.df <- data.frame(z=c(Saugeen,log(Saugeen)), m=c(m,m),
+ which=rep(c("raw","log"), rep(n, 2)))
R >bwplot(m~z|which, data=Saugeen.df, scales=list(x=list(relation="free")),
+ xlab="data")
```

The following R command generates Figure 2,

```
xyplot(log(Saugeen),cut=TRUE)
```

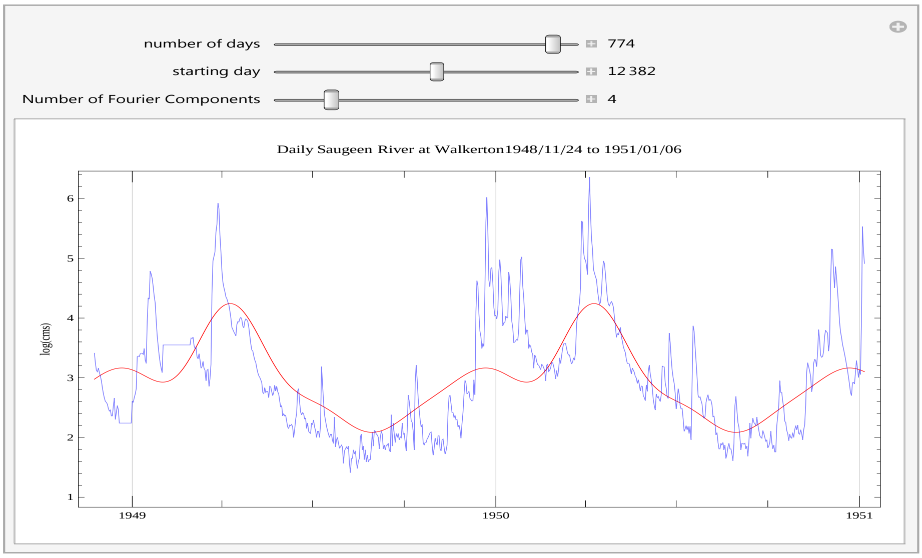
The script below shows how monthly Saugeen river flow series is deseasonalized using the AIC criterion. The optimal transformation $F_\mu = 5$ and $F_\sigma = 4$ is indicated by the $*$ in the left column. The full output has been edited to show only the best 5 models as ranked by plausibility. This script took about 32 seconds. But when the BIC was used, the time was reduced to about 12 seconds. The difference in time reflects the fact that the BIC choose more parsimonious AR models than did the AIC.

```
R >out<-ds(log(Saugeen), ic="AIC")
R >summary(out)

  Fm Fs p       AIC Plausibility %
*  5  4 3 -1171.936          100.0
   6  4 3 -1171.065           64.7
   5  5 3 -1171.029           63.5
   5  3 3 -1170.261           43.3
   6  5 3 -1170.013           38.2
```

## 4.2. Daily Saugeen River Flow

A panel from a dynamic time series plot for a subseries of the mean daily flow Saugeen River at Walkerton, Jan 1, 1915 to Dec 31, 1979 is shown in Figure 4.[3] The series is comprised on 23, 741 consecutive values.



number of days 774

starting day 12 382

Number of Fourier Components 4

Figure 4: Dynamic time series plot of fitted harmonic regression to the daily Saugeen River flows.

For this series the BIC optimal deseasonalizing transformation was found to be with $F_\mu = 4, F_\sigma = 0$. All models with relative plausibility greater than 1% agreed with the choice $F_\sigma = 0$. The choice of the parameter $F_\mu$ may be explored visually using the dynamic time series plot illustrated in Figure 4.

As an additional check, the deseasonalized series was aggregated by month and Figure 5 shows resulting the boxplot. No seasonal variation is noticeable either the means or variances, so the deseasonalization appears to be effective.

### *Code Snippet*

R script used to deseasonalize the Saugeen daily series. The output has been edited to show only the top five most plausible models. since the full output is rather length showing all combinations of $F_\mu, F_\sigma = 0, 1, \ldots, 6$. This script only took about 10 seconds which was less than for the monthly series. The reason for this is that the AR models selected were much simpler than in the monthly case.

```
R >out<-ds(log(SaugeenDay), Fm=6, Fs=6)
R >summary(out)

  Fm Fs p        BIC Plausibility (%)
*  4  0 6 -82621.80            100.0
```

[3] See the subdirectory /inst/doc located in the installation directory of our R package (McLeod and Gweon 2012) for instructions on how to view this plot dynamically on your computer.

```
5  0 6 -82617.37           10.9
3  0 6 -82615.34            4.0
6  0 6 -82613.19            1.3
```

The next script shows how the daily deseasonalized series is aggregated into months and the boxplot used to check for seasonality.

```
require("lubridate")
require("lattice")
w<-ds(log(SaugeenDay), Fm=4, Fs=0, searchQ=FALSE, standardizeQ=FALSE)$z
d<-rownames(SaugeenDay)
m<-month(d, label = TRUE, abbr = FALSE)
w.df <- data.frame(w=w, m=m)
bwplot(m ~ w, data=w.df, xlab="deseasonalized flow")
```

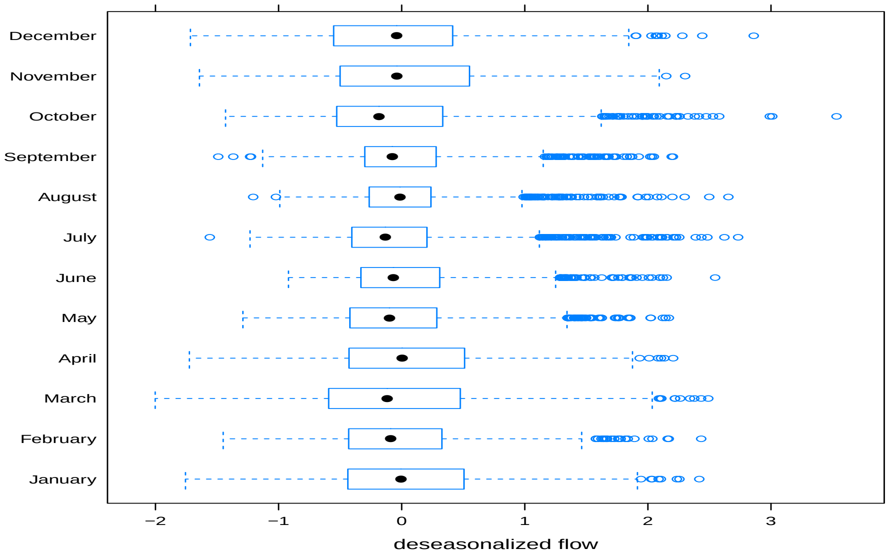


Figure 5: Boxplots of the detrended daily Saugeen River flows.

## 5. Concluding Remarks

The R function `stl()` provides another method for deseasonalizing monthly time series based on loess regression. This method is especially useful for time series that have a strong trend as well as a seasonal component but this method is less automatic and more complex than the method described in this article and it is only applicable to monthly series. The method we have described is preferable for time series models used for applications involving forecasting, simulation and intervention analysis as are described in Hipel and McLeod (1994) or in recent methods for time series modeling of daily series (Cressie and Wikle 2012; Craigmile and Guttorp 2011; Tesfaye *et al.* 2011).

Hipel and McLeod (1994); McLeod (1994) also discussed the application of periodic autoregression for modeling monthly geophysical time series. This type of correlation often occurs with river flow series when the spring runoff occurs either in March or April resulting in a reduced or even negative correlation between these two months whereas other months are positively correlated. The optimal selection using an AR($p$) could be modified to use periodic autoregression. The R package **pear** (McLeod and Balcilar 2011) is available for fitting these models. But this approach is not likely to have any noticeable impact on the final deseasonalized series.

## References

Akaike H (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**(6), 716–723. [Accessed 14-Sept-2012], URL `http://dx.doi.org/10.1109/TAC.1974.1100705`.

Akaike H (1978). "A New Look at the Bayes Procedure." *Biometrika*, **65**(1), 53–59.

Bloomfield P (2004). *Fourier Analysis of Time Series: An Introduction.* 2nd edition. Wiley.

Box GEP, Jenkins GM, Reinsel GC (2005). *Time Series Analysis: Forecasting & Control (4th Edition).* Wiley, New York.

Cleveland WS (1993). *Visualizing Data.* Hobart Press.

Craigmile PF, Guttorp P (2011). "Space-time modelling of trends in temperature series." *Journal of Time Series Analysis*, **32**, 378–395. [Accessed 14-Sept-2012], URL `http://dx.doi.org/10.1111/j.1467-9892.2011.00733.x`.

Cressie N, Wikle CK (2012). *Statistics for Spatio-Temporal Data.* Wiley.

Hannan EJ (1970). *Multiple Time Series.* Wiley.

Hipel KW, McLeod AI (1994). *Time Series Modelling of Water Resources and Environmental Systems.* Elsevier, Amsterdam. [Accessed 14-Sept-2012], URL `http://www.stats.uwo.ca/faculty/aim/1994Book/default.htm`.

Ledolter J, Abraham B (1981). "Parsimony and Its Importance in Time Series Forecasting." *Technometrics*, **23**(4), 411–414. [Accessed 14-Sept-2012], URL `http://www.jstor.org/stable/1268232`.

McLeod AI (1993). "Parsimony, Model Adequacy and Periodic Correlation in Forecasting Time Series." *International Statistical Review*, **61**(3), 387–393. [Accessed 14-Sept-2012], URL http://www.jstor.org/stable/1403750.

McLeod AI (1994). "Diagnostic Checking Periodic Autoregression Models with Application." *Journal of Time Series Analysis*, **15**(6), 221–233. Addendum, Vol. 16, No. 2, p. 647, URL http://dx.doi.org/10.1111/j.1467-9892.1994.tb00186.x.

McLeod AI (2012a). *Aliasing in Time Series Analysis*. Wolfram Demonstrations Project. Accessed 03-April-2012, URL http://demonstrations.wolfram.com/AliasingInTimeSeriesAnalysis/.

McLeod AI (2012b). *Plotting a Long Time Series*. Wolfram Demonstrations Project. Accessed 03-April-2012, URL http://demonstrations.wolfram.com/PlottingALongTimeSeries/.

McLeod AI, Balcilar M (2011). *pear: Package for Periodic Autoregression Analysis*. R package version 1.2. Accessed 22-December-2011, URL http://CRAN.R-project.org/package=pear.

McLeod AI, Gweon H (2012). *deseasonalize: Optimal deseasonalization for geophysical time series using AR fitting*. R package version 1.31. Accessed 03-April-2012, URL http://CRAN.R-project.org/package=deseasonalize.

McLeod AI, Yu H, Mahdi E (2012). *Handbook in Statistics*, volume 30, chapter Time Series Analysis with R. Elsevier.

McLeod AI, Zhang Y, Xu C (2011). *FitAR: Subset AR Model Fitting*. R package version 1.92. Accessed 22-December-2011, URL http://CRAN.R-project.org/package=FitAR.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Schwarz G (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464. [Accessed 14-Sept-2012], URL http://projecteuclid.org/euclid.aos/1176344136.

Sprott DA (2000). *Statistical Inference in Science*. Springer.

Taniguchi M, Hirukawa J (2012). "Generalized Information Criterion." *Journal of Time Series Analysis*, **33**(2), 287–297. [Accessed 14-Sept-2012], URL http://dx.doi.org/10.1111/j.1467-9892.2011.00759.x.

Tesfaye YG, Anderson PL, Meerschaert MM (2011). "Asymptotic results for Fourier-PARMA time series." *Journal of Time Series Analysis*, **32**(2), 157–174. [Accessed 14-Sept-2012], URL http://dx.doi.org/10.1111/j.1467-9892.2010.00689.x.

Tukey JW (1977). *Exploratory Data Analysis*. Addison-Wesley.

Wikipedia (2011). "Keeling Curve." [Accessed 22-December-2011], URL http://en.wikipedia.org/wiki/Keeling_Curve.

Xu C (2010). *Model Selection with Information Criteria*. Ph.D. thesis, Western University. Electronic Thesis and Dissertation Repository. Paper 46. http://ir.lib.uwo.ca/etd/46.[Accessed 03-April-2012].

Xu C, McLeod AI (2012). "Further asymptotic properties of the generalized information criterion." *Electronic Journal of Statistics*, **6**, 656–663. [Accessed 14-Sept-2012], URL http://projecteuclid.org/euclid.ejs/1334754009.

**Affiliation:**

A. Ian McLeod
Department of Statistical and Actuarial Sciences
The University of Western Ontario
London, Ontario N6A 5B7 Canada
E-mail: aimcleod@uwo.ca
URL: http://http://www.stats.uwo.ca/faculty/aim/

Journal of Environmental Statistics http://www.jenvstat.org
Volume 4, Issue 11 Submitted: 2012-09-16
March 2013 Accepted: 2012-10-5