A Note on Trend and Seasonality Estimation for Unevenly Spaced Time Series

Andreas Eckner*

First version: November 2011
Major revision: April 2017
Current version: April 3, 2017

Abstract

This note describes algorithms for estimating the trend, seasonal, and residual component of unevenly spaced times series. An implementation as an `R` package is forthcoming.

1 Introduction

When estimating a time series model such as an autoregressive moving-average (ARMA) model, it is common to first remove the trend and seasonality from the data in order to isolate the non-deterministic behavior. This paper describes such methods for unevenly spaced (also called unequally- or irregularly-spaced) time series.

There exists an extensive body of literature on trend and seasonality estimation for equally spaced time series data, see Cleveland et al. (1990), Chapter 1 in Brockwell and Davis (1991), Chapter 9 in Box et al. (2015), and a website by U.S. Census Bureau.[1] On the other hand, few methods exists specifically for unevenly spaced time series, even though such data naturally occurs in many industrial and scientific domains, such as astronomy, biology, climatology, economics, finance, geology, and network traffic analysis.

Perhaps the most common approach is to transform unevenly spaced data into equally spaced observations using some form of interpolation - most often linear - and then to apply existing methods for equally spaced time series. However, transforming time series data in such a way introduces several biases, see Scholes and Williams (1977), Lundin et al. (1999), Hayashi and Yoshida (2005), Rehfeld et al. (2011), and Eckner (2017). In particular, as shown below, linear interpolation tends to *trim down the hills and fill in the valleys* of seasonal fluctuations. In other words, seasonality estimates based on linearly-interpolated data can severely underestimate the true extent of seasonal fluctuations.

*Comments are welcome at andreas@eckner.com

[1] See www.census.gov/srd/www/sapaper/

1.1 Basic Framework

We use the notation $((t_n, X_n) : 1 \le n \le N(X))$ or $(X_{t_n} : 1 \le n \le N(X))$ to denote an unevenly spaced time series X with observation *times* $T(X) = \{t_1, \ldots, t_{N(X)}\}$ and observation *values* $V(X) = (X_1, \ldots, X_{N(X)})$, where $N(X)$ denotes the length of the time series. For a time point $t \in \mathbb{R}$, $X[t]_{\text{linear}}$ denotes the linearly-interpolated (or sampled) value of X at time t.

2 Trend Estimation

Consider an unevenly spaced time series X of the form

$$X_t = m_t + Y_t, \qquad t \in T(X),$$

where m is a deterministic trend and Y is the realization of a stationary stochastic process with mean zero. If necessary, we first apply a transformation to X, such as taking the logarithm, to achieve this form.

Method 1 (Parametric least squares) *Assume that m comes from a parametric family of functions, for example,*

$$m_t = a_0 + a_1 t + \ldots + a_q t^q. \tag{1}$$

The parameters $a_0, a_1, \ldots, a_q$ can be estimated by minimizing

$$\int_{\min T(X)}^{\max T(X)} (X[t]_{\text{linear}} - m_t)^2 \, \mathrm{d}t. \tag{2}$$

For computational convenience, (2) can be approximated by[2]

$$\sum_{i=1}^{N(X)} (X_{t_i} - m_{t_i})^2 \omega(t_i)$$

where the weights

$$\omega(t_i) = \omega(t_i, T(X)) = \begin{cases} t_{i+1} - t_{i-1} & \text{if } 1 < i < N(X), \\ t_2 - t_1 & \text{if } i = 1, \\ t_{N(X)} - t_{N(X)-1} & \text{if } i = N(X) \end{cases} \tag{3}$$

are inversely related to the local observation density of the time series.[3]

Often, a linear or quadratic trend ($q = 1$ or 2 in (1)) will be sufficient for $\hat{Y} = X - \hat{m}$ to look like a stationary time series.

[2] The approximation replaces each value of $X[t]_{\text{linear}}$ in (2) by the closest available observation value of X and uses a similar approximation for m_t.

[3] If the time series X has very large gaps, the weight ω of the last observation before and first observation after the gap could be undesirably large. In this case, one could either solve (2) exactly using numerical integration, or insert linearly-interpolated auxiliary observation into the time series X.

Method 2 (Two-sided moving average) *For a smoothing window width $\tau > 0$ and time series X, let*

$$\mathrm{TSMA}(X,\tau)_t = \frac{1}{\tau}\int_{-\tau/2}^{\tau/2} X[t+s]_{\text{linear}}\,\mathrm{d}s, \qquad t \in T(X),$$

denote the two-sided simple moving average.[4] *If m_t is linear over the interval $[t-\tau/2, t+\tau/2]$, then*

$$\mathrm{TSMA}(X,\tau)_t = m_t + \frac{1}{\tau}\int_{-\tau/2}^{\tau/2} Y[t+s]_{\text{linear}}\,\mathrm{d}s.$$

Hence, $m_t \approx \mathrm{TSMA}(X,\tau)_t$ provided that the average value of Y in the interval $[t-\tau/2, t+\tau/2]$ is close to zero. Therefore,

$$\hat{m}_t = \mathrm{TSMA}(X,\tau)_t$$

can be used as an estimate of the time series trend as long as m_t is approximately linear over time intervals of length τ.

This method calculates a local average of observation values and is therefore able to capture time-varying trends, although at the expense of an increased estimation variance compared to the previous method. However, local averages can be badly biased at the boundary (in our case, close to the beginning and end of a time series), see Hastie et al. (2009), Chapter 6.1.1. This bias can be avoided by using local regression instead of local averaging, see also Cleveland and Devlin (1988) and Cleveland et al. (1988).

Method 3 (Local linear regression) *Choose a smoothing kernel (that is, a weighting function) K_τ and kernel width $\tau > 0$, for example,*

$$K_\tau(t,s) = \frac{1}{\tau}\mathbf{1}_{|t-s|\le\tau/2}.$$

For each $t \in T(X)$ minimize

$$\int_{\min T(X)}^{\max T(X)} K_\tau(t,s)(X[s]_{\text{linear}} - \alpha_t - \beta_t s)^2\,\mathrm{d}s \tag{4}$$

with respect to α_t and β_t. For computational convenience, (4) can be approximated by the locally weighted regression problem

$$\arg\min_{\alpha_t,\beta_t} \sum_{i=1}^{N(X)} \omega(t_i)K_\tau(t,t_i)(X_{t_i} - \alpha_t - \beta_t t_i)^2, \tag{5}$$

with weights $\omega(t_i)$ given by (3). The trend estimate is given by

$$\hat{m}_t = \hat{\alpha}_t + \hat{\beta}_t t, \qquad t \in T(X).$$

See Hastie et al. (2009), Chapter 6 for an automated choice of the kernel width using cross-validation. Packages for solving local regressions problems of the form (5) are available for a variety of programming languages, such as the `loess` function in `R`, which in turn is builds on the `C` and `Fortran` implementation by Cleveland et al. (1992).

[4] We assume that sampled values before the first observation time, t_1, are equal to the first observation value, X_{t_1}. While potentially not appropriate for some applications, this assumption avoids the treatment of a several special cases.

3 Seasonality Estimation

Consider an unevenly spaced time series X of the form

$$X_t = s_t + Y_t, \qquad t \in T(X),$$

where $s : [0, d) \to \mathbb{R}$ is a deterministic seasonal component with period d and normalization $\int_{t_0}^{t_0+d} s_t \, \mathrm{d}t = 0$ for all $t_0 \in \mathbb{R}$, and Y is the realization of a stationary stochastic process with mean zero.

Method 4 (Averaging of subintervals) *Let $\phi^X : \mathbb{R} \to [0, d]$ denote the function that maps a time point to its relative position within a season. The seasonal component can be estimated via*

$$\hat{s}_t = \operatorname{avg}\{X[r]_{\text{linear}} : r \in [\min T(X), \max T(X)], \phi^X(r) = t\} \tag{6}$$

for $0 \le t \le d$. The deseasonalized time series $X - \hat{s}$ is defined by $T(X - \hat{s}) = T(X)$ and

$$(X - \hat{s})_t = X_t - \hat{s}_{\phi(t)}, \qquad t \in T(X - \hat{s}).$$

It is usually necessary to drop a sampled value $X[r]_{\text{linear}}$ from the calculation of the average in (6) if the associated time point r is not close to any observation time of X. Otherwise, linear interpolation would *trim down the hills and fill in the valleys* of seasonal fluctuations.

4 Joint Trend and Seasonality Estimation

Consider an unevenly spaced time series X of the form

$$X_t = m_t + s_t + Y_t, \qquad t \in T(X),$$

where m is a deterministic trend, $s : [0, d) \to \mathbb{R}$ is a deterministic seasonality of period d with normalization $\int_{t_0}^{t_0+d} s_t \, \mathrm{d}t = 0$ for all $t_0 \in \mathbb{R}$, and Y is the realization of a stationary stochastic process with mean zero.

Method 5 (Iterative estimation) *The trend and seasonality can be jointly estimated using a simple iterative scheme:*

Step 0: Initialize $\hat{s}_t \equiv 0$ on $[0, d)$.

Step 1: Estimate the trend of $X - \hat{s}$ using one of the methods in Section 2.

Step 2: Estimate the seasonality of $X - \hat{m}$ using the method in Section 3.

Step 3: Repeat steps 1 and 2 until convergence.

When using local linear regression in Step 2, this method essentially is a simplified version of the `STL` procedure by Cleveland et al. (1990), adapted to unevenly spaced data.

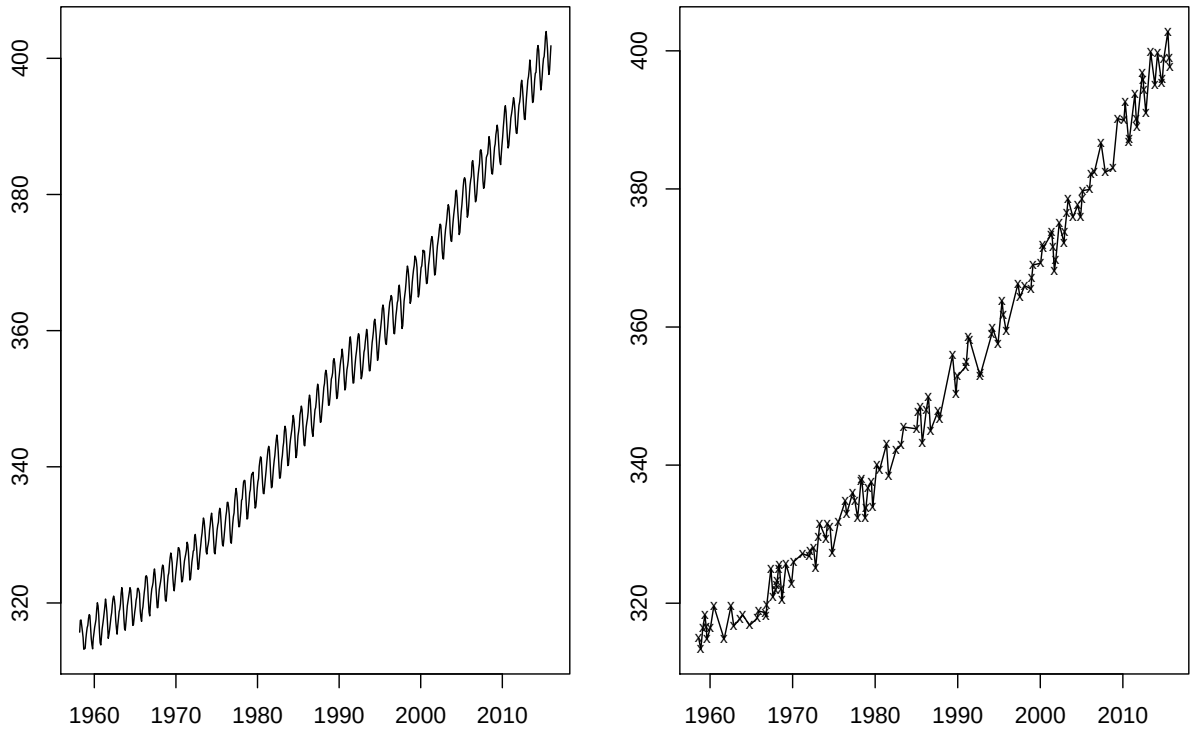

Figure 1: Atmospheric CO_2 concentration (in parts per million) at Mauna Loa Observatory, Hawaii (left-hand side). A subsampled time series of the same data with 80% of the observations randomly removed (right-hand side).

5 Example

We illustrate the joint trend and seasonality estimation using a time series of the atmospheric CO_2 concentration (measured in parts per million) at Mauna Loa Observatory, Hawaii.[5] The observations are from March 1958 through December 2015 (as of this writing) and at a monthly frequency. There are 7 missing values and 687 observations in total. Figure 1 plots the original time series (on the left-hand side) and a subsampled time series with 80% of the observations randomly removed (on the right-hand side). In both cases, individual observations are connected using straight lines, which amounts to plotting $X[t]_{\text{linear}}$ as a function of time t. In the second plot, individual observations in addition are marked by the symbol `x`.

The average spacing of observations in the subsampled time series is five months. In seven cases, the spacing is more than one year, which is more than the period length d of the seasonal component. Figure 1 illustrates that a trend-seasonality estimation procedure based on a transformation to equally spaced data via linear interpolation would severely underestimate the seasonal component. The same is true for the human eye, which has a tendency to connect dots using straight lines.

Using a log-transformation, I estimate a multiplicative decomposition of the form $X_t = m_t(1+s_t)(1+Y_t)$ using the iterative algorithm in Section 4 with local linear smoothing for the trend estimation. Figure 2 plots the estimated trend and multiplicative seasonal component for the original and subsampled time series. Given the highly irregular nature of the latter

[5] See www.esrl.noaa.gov/gmd/ccgg/trends for a detailed description and a data download section.

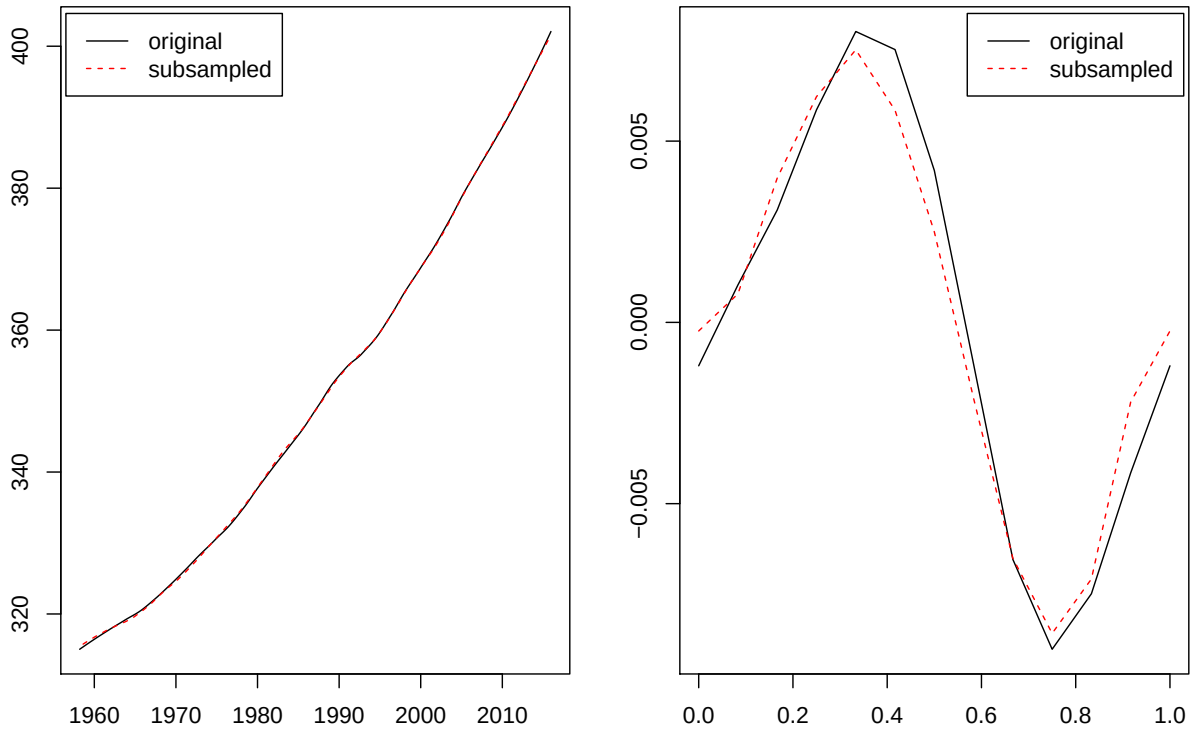

Figure 2: The estimated trend (left-hand side) and multiplicative seasonal component (right-hand side) for the original and subsampled CO_2 time series. The decomposition uses the iterative algorithm in Section 4 with local linear smoothing for the trend estimation.

time series, both decompositions are remarkably similar; the estimated trend components are virtually indistinguishable, and the peak-to-trough size of the multiplicative seasonality is roughly 1.7%.

Finally, we estimate the same multiplicative decomposition using R's `stl()` function, which is based on Cleveland et al. (1990). Figure 3 plots the estimated trend and multiplicative seasonal component for the original and *linearly-interpolated* subsampled time series. We see that `stl()` and the methodology in Section 4 produce a virtually identical decomposition for the original CO_2 data. However, for the linearly-interpolated subsampled time series, `stl()` severely underestimates the variability of the seasonal component. As a consequence, the estimated residual time series $\hat{Y}$ exhibits excess variability, because it absorbs some of the seasonal fluctuations.

Based on a simulation, Table 1 shows the expected percentage bias in the peak-to-trough size of the seasonal component as a function of the missing data fraction for (i) the iterative algorithm in Section 4 and (ii) `stl()` with linearly interpolated data.

Missing data fraction	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
% Bias of Section 4 method	0.0	0.9	1.8	2.4	3.2	3.7	4.1	4.7	4.9
% Bias of `stl()`	0.0	2.3	5.5	9.8	15.9	23.9	35.1	49.8	68.2

Table 1: Expected percentage bias in the peak-to-trough size of the seasonal component depending on the fraction of missing time series observations.

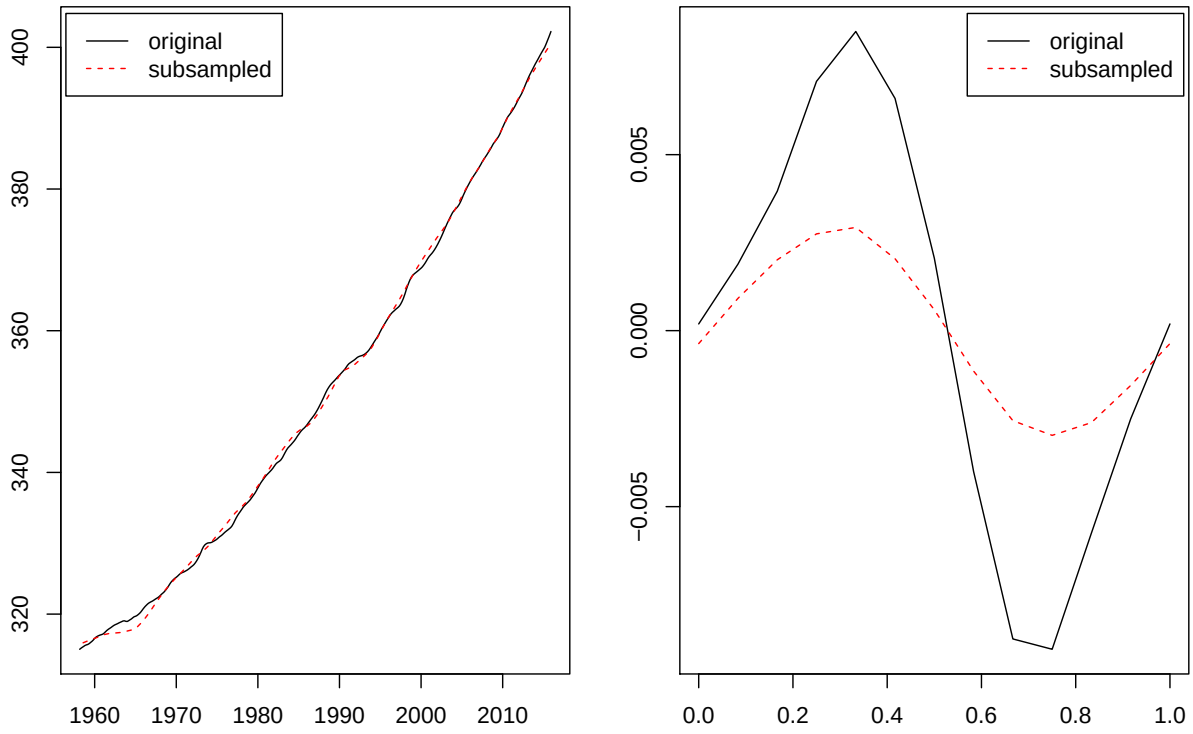

Figure 3: The estimated trend (left-hand side) and multiplicative seasonal component (right-hand side) for the original and *linearly-interpolated* subsampled CO_2 time series. The decomposition uses the `stl()` function in `R`.

To be fair, the `STL` procedure actually was designed to handle missing values, although not unevenly spaced data per se. It is widely used among statisticians due to its flexibility in capturing time-varying trends, time-varying seasonal components, and automatic choice of sensible parameters. In fact, the bias in the estimated seasonal component is entirely due to the input time series. Any trend-seasonality estimation method applied to linearly-interpolated data is bound to encounter the same problem.

6 Conclusion

We have shown how to estimate the trend, seasonal, and residual component of a time series with unevenly spaced observations. Several extension that have already been explored for equally spaced data are possible. For example, the procedures `X-13`[6] and `STL` support robust estimation, that is, they limit the influence that any single observation can have on the estimated trend and seasonal component. Adopting this feature for unevenly spaced time series would be straightforward. As already mentioned, `STL` supports time-varying seasonality, which could be implemented for unevenly spaced time series in a similar manner, specifically by using rolling estimation window in the time domain.

Finally, we focused on trend and seasonality estimation techniques that at each point in time use information about the entire time series, as opposed to only past observations. For some

[6] See www.census.gov/srd/www/x13as/

applications, such as gauging the on-the-run performance of a time series forecasting model, causality of the estimators is important. In this case, the methods discussed above could be applied in a rolling manner to a time series, where the trend and seasonal components are reestimated at periodic time intervals, or when a significant number of new time series observations become available.

References

Box, G. E. P., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley Series in Probability and Statistics.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (Second ed.). Springer.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics 6*, 3–73.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*, 596–610.

Cleveland, W. S., S. J. Devlin, and E. Grosse (1988). Regression by local fitting: Methods, properties, and computational algorithms. *Journal of Econometrics 37*(1), 87–114.

Cleveland, W. S., E. Grosse, and M.-J. Shyu (1992). A package of C and Fortran routines for fitting local regression models. Bell Labs, Technical Report.

Eckner, A. (2017). Some properties of operators for unevenly spaced time series. Working Paper.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). Springer.

Hayashi, T. and N. Yoshida (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli 11*, 359–379.

Lundin, M. C., M. M. Dacorogna, and U. A. Müller (1999). Correlation of high frequency financial time series. In P. Lequex (Ed.), *The Financial Markets Tick by Tick*, Chapter 4, pp. 91–126. New York: John Wiley & Sons.

Rehfeld, K., N. Marwan, J. Heitzig, and J. Kurths (2011). Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics 18*, 389–404.

Scholes, M. and J. Williams (1977). Estimating betas from nonsynchronous data. *Journal of Financial Economics 5*, 309–327.