



Ekkehart Schlicht:

Trend Extraction From Time Series With Missing Observations

Munich Discussion Paper No. 2007-18

Department of Economics
University of Munich

Volkswirtschaftliche Fakultät
Ludwig-Maximilians-Universität München

Online at <http://epub.ub.uni-muenchen.de/1927/>

Trend Extraction From Time Series With Missing Observations

Ekkehart Schlicht

Abstract: Trend extraction from time series is often performed by using the filter proposed by LESER (1961), also known as the Hodrick-Prescott filter. A practical problem arises, however, when some data points are missing. This note proposes a method for coping with this problem.

Keywords: Trend extraction, missing observations, gaps, Hodrick-Prescott filter, Leser filter, spline, time-series, smoothing, interpolation.

Journal of Economic Literature Classification: C22, C32, C63, C14

1 Introduction

Trend extraction from time series is often performed by using the filter proposed by LESER (1961), also known as the Hodrick-Prescott filter, or HP-Filter. A practical problem arises, however, when some data points are missing. This note proposed a method for coping with this problem.

2 The Leser Filter¹

The idea proposed by LESER (1961) for the case where all data are available is to look for a trend $y \in \mathbb{R}^T$ such that deviation

$$u = x - y \tag{1}$$

is “small” and the trend is “smooth.” The size of the deviation is measured by the sum of squared residuals $u'u$, and the smoothness of the trend is measured by the sum of squares of changes in the direction of the trend $v'v$ where the trend disturbances $v \in \mathbb{R}^{T-2}$ are defined as

$$v_t = ((y_t - y_{t-1}) - (y_{t-1} - y_{t-2})) \quad t = 3, 4, \dots, T$$

or

$$v = P y \tag{2}$$

with

$$P := \begin{pmatrix} 1 & -2 & 1 & & 0 \\ & 1 & -2 & 1 & \\ & & \cdot & \cdot & \cdot \\ 0 & & & 1 & -2 & 1 \end{pmatrix}$$

of order $(T - 2) \times T$.

The decomposition of the original series x into trend y and residual u is obtained by minimizing the weighted sum of squares

$$V = u'u + \alpha \cdot v'v = (x - y)'(x - y) + \alpha \cdot y'P'P y \tag{3}$$

with respect to y . This gives the first-order condition

$$(I_T + \alpha \cdot P'P) y = x. \tag{4}$$

¹ The formalization below follows SCHLICHT (1981).

As $(I + \alpha P'P)$ is positive definite, the second order condition is satisfied in any case.

Equation (4) has the unique solution

$$y = (I_T + \alpha P'P)^{-1} x \quad (5)$$

which defines the Leser-Filter. It associates a trend y with the time series x , depending on the smoothing parameter α .

From (3) and (5) we obtain

$$V = x' \left(I_T - (I_T + \alpha \cdot P'P)^{-1} \right) x \quad (6)$$

as the value of the criterion function (3).

3 Missing Observations

A practical problem arises when some data are missing. Equation (5) cannot be applied such cases. The obvious way to generalize the filter in order to cope with this problem is to look for a smooth trend by substituting all missing observations by numbers that minimize the criterion function (3). This can be done as follows.

Consider a time series $x \in \mathbb{R}^T$ with $m \leq T - 2$ data points missing. The missing data are indicated by a vector $m \in \mathbb{R}^T$ with components $m_t = 0$ for all $t \in \{1, 2, \dots, T\}$ where x_t is defined, and $m_t = 1$ for all $t \in \{1, 2, \dots, T\}$ where x_t is not defined. Now define a new time series $z \in \mathbb{R}^T$ such that $z_t = x_t$ for all $t \in \{1, 2, \dots, T\}$ where x_t is defined, and $z_t = 0$ for all $t \in \{1, 2, \dots, T\}$ where x_t is not defined. Define the $(T \times m)$ -matrix with components $b_{i,1} = 1$ if the first missing variable is x_i , $b_{j,2} = 1$ if the second missing variable is x_j , etc., and all other components of B being zero. For $T = 5$ and variables 2 and 4 missing we would have, for example

$$B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Define further the vector of substitute values for the missing variables $s \in \mathbb{R}^m$, where s_i gives the substitute value for the i -th missing observation which is to be determined.

The replenished vector of observations is

$$x^* = z + Bs. \quad (7)$$

It is a function of the values assumed for the missing values s . These values can now easily be determined by replacing x by x^* in (6), and minimizing this expression. Thus we obtain the quadratic form

$$V = (z' + s'B') \left(I_T - (I_T + \alpha \cdot P'P)^{-1} \right) (z + Bs) \quad (8)$$

that is to be minimized with respect to s . The necessary condition for a minimum is

$$\frac{\partial V}{\partial s} = 2B' \left(I_T - (I_T + \alpha \cdot P'P)^{-1} \right) (z + Bs) = 0. \quad (9)$$

and the second-order condition is that $B' (I_T - (I_T + \alpha \cdot P'P)^{-1}) B$ is positive definite. As

$$(I_T + \alpha P'P)^{-1} = I_T - \alpha P'P + (\alpha P'P)^2 - (\alpha P'P)^3 + \dots$$

we can write

$$\left(I_T - (I_T + \alpha \cdot P'P)^{-1} \right) = \alpha P' (I_{T-2} + \alpha \cdot PP')^{-1} P.$$

Hence $(I_T - (I_T + \alpha \cdot P'P)^{-1})$ is non-negative definite of rank $T - 2$ and $B' (I_T - (I_T + \alpha \cdot P'P)^{-1}) B$ has full rank and is positive definite. Therefore equation (9) defines the unique maximizing choice of the substitute terms as

$$s^* = - \left(B' \left(I_T - (I_T + \alpha \cdot P'P)^{-1} \right) B \right)^{-1} B' \left(I_T - (I_T + \alpha \cdot P'P)^{-1} \right) z. \quad (10)$$

The replenished time series is obtained now by inserting (10) into (7)

$$x^* = z + Bs^*$$

and the trend is obtained by using the replenished series x^* instead of the original series x in (5):

$$y^* = (I_T + \alpha P'P)^{-1} (z + Bs^*).$$

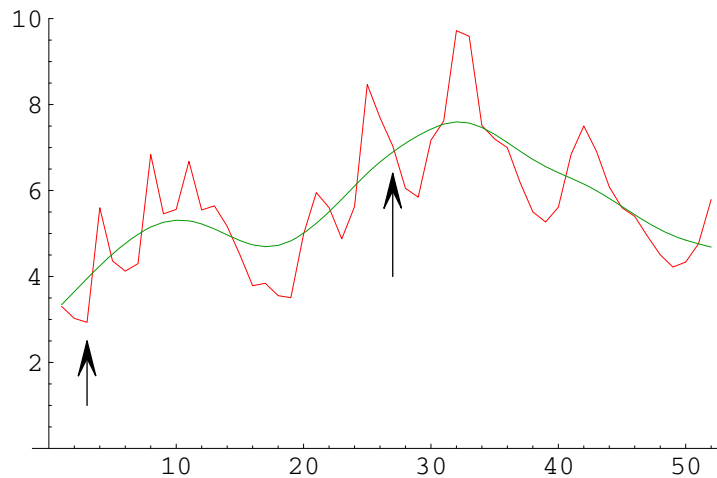


Figure 1: US Unemployment 1959-2002, original series and smoothed series, using a smoothing constant $\alpha = 100$. The arrow indicate values that have been omitted in order to produce a time series with missing data. (Data Source: Bureau of Labor Statistics.)

This gives the trend of the time series with missing observations x .

4 An Example¹

As an example, consider the time series of US unemployment (Figure 1). I have deleted two data points, numbers 3 and 27, with values of 2.9% and 7.0%, respectively. The gap at 27 is uncritical because the point sits in the middle of the data range, and assumes also a middle position between adjacent data points. The gap at 3 is critical, as it is close to the boundary of the time series, and is also extreme in its deviation from the trend.

Dropping these values and estimating replacements according to the method outlined above yields estimated values 4.2% and 6.9%. This is illustrated in Figure 2. The two trend series are depicted in Figure 3. It can be seen that the omission at data point 3 has a noticeable effect, while the omission at data point 27 does not change the trend estimate in any significant way.

¹ All computations done with the package by LUDSTECK (2004).

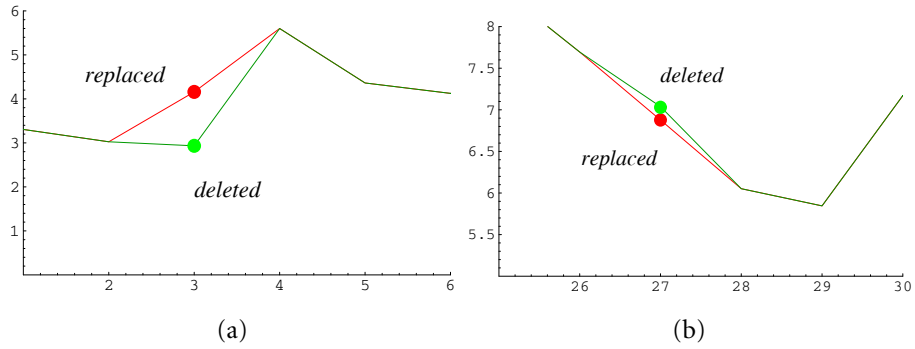


Figure 2: Deleted values and their computed replacements at data points 3 (a) and 27 (b).

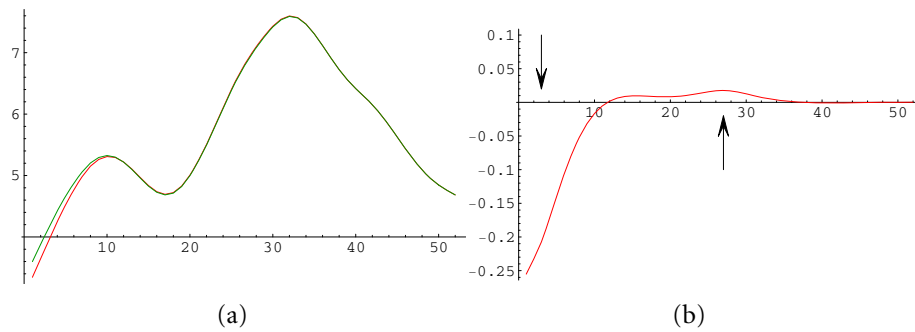


Figure 3: (a) Original trend estimation and trend estimated from incomplete data (b) the difference between these two trend estimates. The arrows indicate the position of the gaps in the data.

5 Concluding Comments

Looking at the substitutions illustrated in Figure 2, it may be asked whether a simple linear interpolation would not do as well. In a way, this seems a reasonable position to take. However, and practically speaking, with contemporary computing power, gap detection and the substitution would be done automatically in both cases. If adjacent gaps occur, linear interpolation would require case distinctions that are not necessary in the method proposed here. In this sense, the proposed method is computationally simpler. The substitution proposed here, as compared to a simple linear interpolation, offers two further advantages. First, the substitute values are points on the estimated trend. In this sense, the estimation proposed here does not distort the trend; and second, the estimated substitutes can be interpreted as Maximum-Likelihood estimates of the missing values, if the stochastic interpretation of the Leser method proposed by SCHLICHT (2005) is adopted. In short, the method is more systematically linked to the smoothing method at hand than other interpolation methods are.

References

- LESER, C. E. V. 1961, "A Simple Method of Trend Construction," *Journal of the Royal Statistical Society. Series B (Methodological)*, 23, 91–107.
- LUDSTECK, J. 2004, "HPFilter," Mathematica package, Wolfram Research, online at <http://library.wolfram.com/infocenter/MathSource/Mathematics/ProbabilityStatistics/>.
- SCHLICHT, E. 1981, "A Seasonal Adjustment Principle and a Seasonal Adjustment Method Derived from This Principle," *Journal of the American Statistical Association*, 76(374), 374–78, online at http://www.semverteilung.vwl.uni-muenchen.de/mitarbeiter/es/paper/schlicht-seasonal_adjustment.pdf.
- 2005, "Estimating the Smoothing Constant in the So-Called Hodrick-Prescott Filter," *Journal of the Japan Statistical Society*, 35, 1–19, online at <http://www.scipress.org/journals/jjss/pdf/3501/35010099.pdf>.