LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

VOLKSWIRTSCHAFTLICHE FAKULTÄT

Ekkehart Schlicht:

Trend Extraction From Time Series With Structural Breaks and Missing Observations

Munich Discussion Paper No. 2008-3

Department of Economics
University of Munich

Volkswirtschaftliche Fakultät
Ludwig-Maximilians-Universität München

Online at http://epub.ub.uni-muenchen.de/2127/

Trend Extraction From Time Series With Structural Breaks and 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. Practical problems arise, however, if the time series contains structural breaks (as produced by German unification for German time series, for instance), or if some data are missing. This note proposes a method for coping with these problems.

Keywords: dummies, gaps, Hodrick-Prescott filter, interpolation, Leser filter, missing observations, smoothing, spline, time-series, trend, structural breaks, break point, break point location

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

* Department of Economics, University of Munich, Schackstr. 4, 80539 Munich, Germany, *schlicht@lmu.de*.
The final version of the paper will appear in the *Journal of the Japan Statistical Society*. I thank an anonymous referee of that journal for very helpful comments and for suggesting the issue of break point location.

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. Practical problems arise, however, if the time series contains structural breaks (as produced by German unification for German time series, for instance), or if some data are missing. This note proposes a method for coping with these problems.

2 *The Leser Filter*[1]

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

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

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 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. The smoothing constant α denotes the weight given for the trend deviations. It is typically selected in an arbitrary way but may also be estimated from the

[1] The formalization below follows Schlicht (1981).

time series (Schlicht, 2005). Minimization of (3) entails the first-order condition

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

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

Equation (4) has the unique solution

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

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

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

as the value of the criterion function (3).

3 *Formalizing Structural Breaks and Missing Observations*

Practical problem arise with time series containing structural breaks and missing observations. Equation (5) would interpret structural breaks as changes in the trend, and cannot even be applied if some data points are missing. The obvious way to generalize the filter in order to cope with this problem is to introduce dummies for the structural breaks, and to substitute missing values of the time series by numbers that minimize the criterion function (6) and generate a trend as smooth as possible. This can be done as follows.

Consider a *raw time series* x of length T with m structural breaks and n missing data points. We require $m + n \leq T - 2$ and $m \geq 0$ and $n \geq 0$. The structural breaks occur at points in time $b = (b_1, b_2, ...b_m) \in \mathbb{Z}_+^m$, where $\mathbb{Z}_+$denotes the set on non-negative integers. If the first break point is at $t = 3$, we would have $b_1 = 3$, for instance. The missing data are missing at points $c = (c_1, c_2, ...c_n) \in \mathbb{Z}_+^n$. If the first missing data point is at $t = 5$, we would have $c_1 = 5$, for instance.

Given this information, we define the *filled time series* $\tilde{x} \in \mathbb{R}^T$ by taking the raw time series x and replacing all undefined elements by zero, *viz.* $\tilde{x}_{c_i} = 0$ for all $i = 1, 2, ..., n$.

Further, we define a $T \times m$ matrix D with elements $d_{i,j} = 0$ for $i < b_j$ and $d_{i,j} = 1$ for

$i \geq b_j$. For $T = 5$ and break points at $b_1 = 2$ and $b_2 = 4$ we would have, for example

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

In a similar way, we define a $T \times n$ matrix E with elements $e_{i,1} = 1$ if the first missing variable is x_i, $e_{j,2} = 1$ if the second missing variable is x_j , etc., and all other components of E being zero. For $T = 5$ and variables 2 and 4 missing we would have, for example

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

The block matrix

$$F = (D, E)$$

combines these two matrices and will be used for further computations. It is of order $T \times (m + n)$, and we require that it is of full rank $m + n$. This assumption assures that missing observations do not mask structural breaks fully. With the above example we would have $\mathrm{rank}\,(D, E) = 4$, and the requirement would be satisfied, but if we had D as above and

$$E = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

we would have $\mathrm{rank}\,(D, E) = 3$ and the requirement would not be met; a structural break would entirely be masked by missing data.

Define further the vector of dummies $d \in \mathbb{R}^m$, where d_i is the dummy for the i-th structural break and define the vector $e \in \mathbb{R}^n$ of replacements for the missing (or empty) observations. We combine these two vectors in the vector $f' = (d', e') \in \mathbb{R}^{m+n}$. Given these definitions, the stage is set for dealing with the estimation problem.

4 *Estimation*

Define the amended vector of observations as

$$x^* = \tilde{x} + Ff. \tag{7}$$

It is a function of the filled time series $\tilde{x}$, the values assumed for the dummies d and the replacements e comprised in the vector f. The vector f can now easily be determined by replacing x by x^* in (6), and minimizing this expression. Thus we obtain the quadratic form

$$V = (\tilde{x}' + f'F')\left(I_T - (I_T + \alpha \cdot P'P)^{-1}\right)(\tilde{x} + Ff) \tag{8}$$

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

$$\frac{\partial V}{\partial s} = 2F'\left(I_T - (I_T + \alpha \cdot P'P)^{-1}\right)(\tilde{x} + Ff) = 0. \tag{9}$$

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

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

we can write

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

Hence $\left(I_T - (I_T + \alpha \cdot P'P)^{-1}\right)$ is non-negative definite of rank $T - 2$ and $F'\left(I_T - (I_T + \alpha \cdot P'P)^{-1}\right)F = \alpha F'P'\left(I_{T-2} + \alpha \cdot PP'\right)^{-1} PF$ has full rank and is positive definite. Therefore equation (9) defines the unique minimizing choice of the dummies and missing terms as

$$f^* = -\left(F'\left(I_T - (I_T + \alpha \cdot P'P)^{-1}\right)F\right)^{-1} F'\left(I_T - (I_T + \alpha \cdot P'P)^{-1}\right)\tilde{x}. \tag{10}$$

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

$$x^* = \tilde{x} + F f^*$$

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

$$y^* = \left(I_T + \alpha\, P'P\right)^{-1} \left(\tilde{x} + F f^*\right).$$

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

5 *Example: Structural Breaks*[1]

As an example for treating structural breaks, consider the time series of US unemployment (Figure 1). Beginning with period 25, a structural break has been introduced by adding 5 percentage points to the original time series. The correction obtained by the method sketched above overcorrects this break by subtracting 7.2 percentage points. The corrected trend estimation is overcorrected as well. As can be seen, the correction produces a smoother trend than the original one, as is implied by the logic of the method. A manual correction would look not very much different, or would look even worse if the adjustment is made such that the adjacent data points 24 and 25 are made to have identical values. (The correction would have been -7.8 rather than -7.2 in this case.)

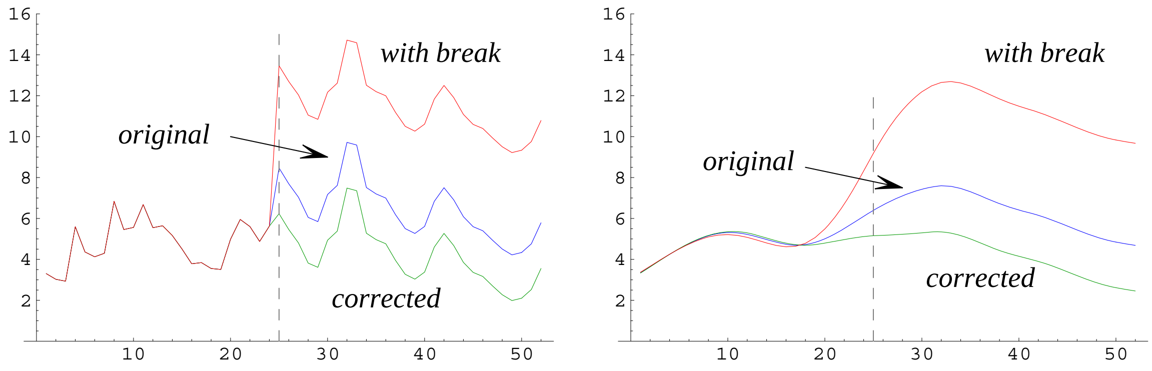
The example illustrates the functioning, as well as the problematic, of introducing dummies, as these will not only correct for structural breaks, but will also mask changes in the underlying trend.

Another illustration is provided in Figure 2(a). It depicts the time series with the structural break, as given in Figure 1(a) together with the estimated trend of the corrected series plus the estimated structural break.

6 *Digression: Locating Structural Breaks*

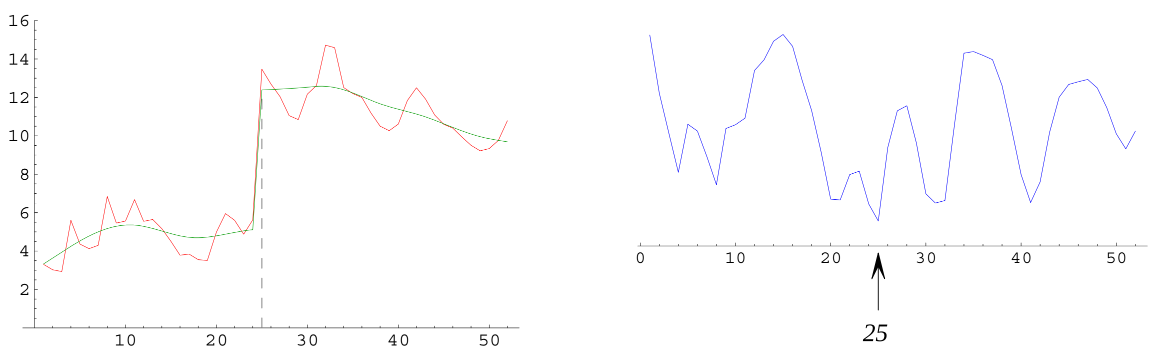
Sometimes the analyst may be in doubt about the exact positioning of the structural break. A simple way to deal with such uncertainty would be to estimate corrected time paths x^* for alternative break points and evaluate the criterion (6) for these alternative time-paths. The preferred break point would be the one giving the smallest value for the criterion. Figure

[1] All computations done with the package by Ludsteck (2004).



(a) (b)

Figure 1: (a) The original US unemployment rate 1951-2002 has been augmented by adding, beginning with period 25, five percentage points to the original series. The corrected series overcorrects the structural break in this case. (b) The smoothed series reproduce this pattern. (Smoothing constant is α =100, data from the Bureau of Labor Statistics.)



(a) Time series and estimated trend with estimated break

(b) The criterion (6) for alternative break points

Figure 2: (a) The time series with breaks (the top time series from Fig. 1(a)) together with the the estimated trend (the bottom series from Fig. 1(b)), increased by the estimated jump of 7.2 from period 25 upwards. (b) The criterion (6) for the time series depicted in Figure 2(a) corrected for alternative assumed break points. The minimum occurs at point 25, which is the correct break point.

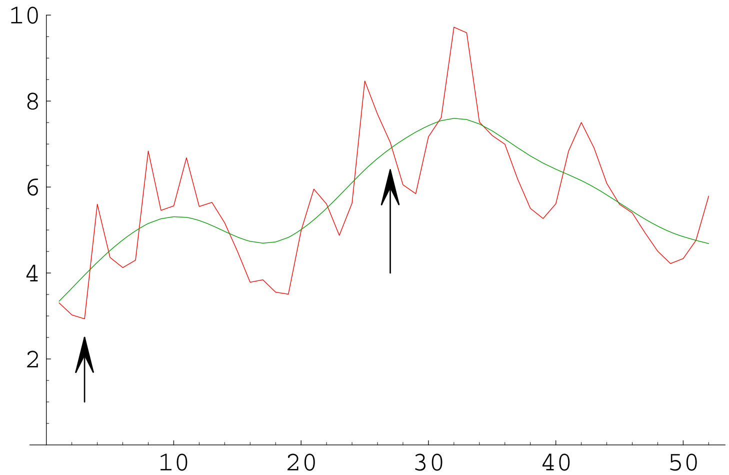


Figure 3: US Unemployment 1959-2002, original series and smoothed series, using a smoothing constant $\alpha = 100$. The arrow indicate values that have been omitted in order to produce a time series with missing data. (Data Source: Bureau of Labor Statistics.)

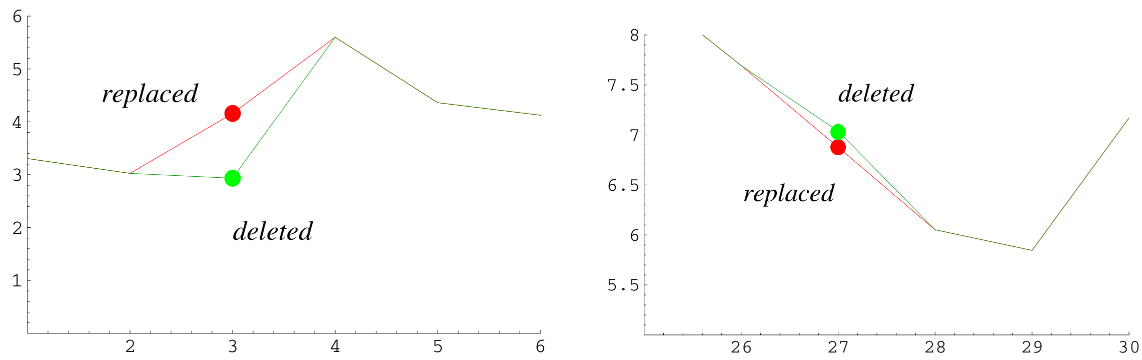
2(b) depicts the value of the criterion for alternative break points and the corresponding corrected time paths. The minimum is attained at the correct break point 25.

7 Example: Missing Observations

Consider again the time series of US unemployment (Figure 3). I have deleted two data points, numbers 3 and 27, with values of 2.9% and 7.0%, respectively, to obtain a time series with missing observations. 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.

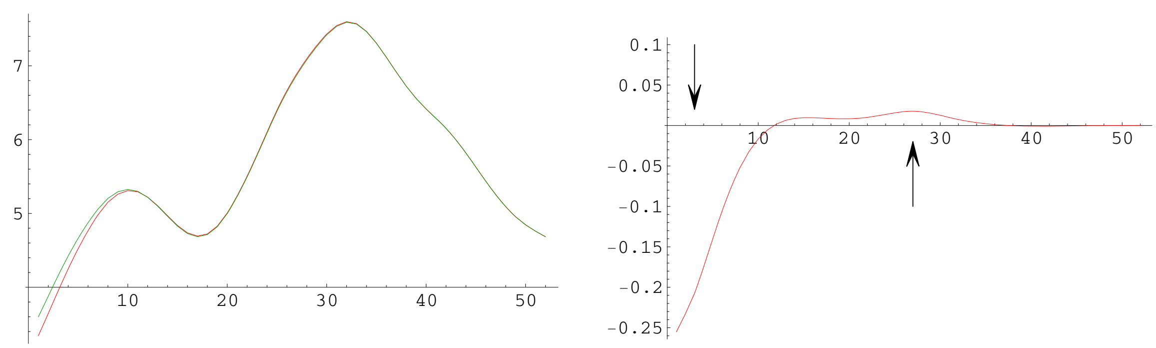
Dropping these values and estimating replacements according to the method outlined above yields estimated values 4.2% and 6.9%. This is illustrated in Figure 4. The two trend series are depicted in Figure 5. It can be seen that the omission at data point 3 has a noticeable effect, while the omission at data point 27 does not change the trend estimate in any significant way.

Looking at the substitutions illustrated in Figure 4, it may be asked whether a simple linear interpolation would not do as well. In a way, this seems a reasonable position to take.



(a) (b)

Figure 4: Deleted values and their computed replacements at data points 3 (a) and 27 (b).



(a) (b)

Figure 5: (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.

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 with the method proposed here. In this sense, the proposed method is computationally simpler.

8 *Concluding Comments*

The unified treatment of structural breaks and missing observations proposed here may not, practically speaking, be very much different from doing similar adjustments "by hand," as is common practice. The treatment proposed here can easily be automated, though, and appears less arbitrary. Further, the dummies selected here, and the substitutes for missing values can be interpreted as maximum-likelihood estimates, if the stochastic interpretation of the Leser method proposed by Schlicht (2005) is adopted, and the smoothing constant may be estimated by the method given there. In short, the method is more systematically linked to the smoothing method at hand than other methods of adjustment and interpolation 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/5161/*.

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*.