Ann. Inst. Statist. Math.
32 (1980), Part B, 481-488

TREND ESTIMATION WITH MISSING OBSERVATIONS

HIROTUGU AKAIKE AND MAKIO ISHIGURO

(Received Mar. 11, 1980; revised June 9, 1980)

1. Introduction

Trend estimation is often an important problem in scientific or economic time series analysis. The main difficulty in the trend analysis is the parameterization. When a particular functional form, such as the ordinary or trigonometric polynomial, is assumed the resulting estimate is quite sensitive to the choice of this basic function. Also, to make the estimate sufficiently sensitive to an abrupt change of the trend, it is usually necessary to use a function with a large number of parameters. However, as the number of parameters increases, the stability of the estimated trend decreases.

As a solution to this problem Akaike [2] proposed the use of a general Bayesian linear model. In particular, a procedure of decomposing a time series into the trend, seasonal and irregular components was developed, where a prior distribution was assumed for the temporal differences of the trend and seasonal components to assure the smoothness of the local behavior of these components. In this model the values of the trend and seasonal components are considered as parameters and thus the model has at least twice as many parameters as the data points.

One remarkable characteristic of this parameterization is the ease of handling the case where several observations are missing. This is due to the fact that one and the same prior distribution can be assumed for the trend and seasonal components as in the case where there are no missing values. In this paper we will demonstrate by numerical examples the feasibility of trend estimation with missing observations by this model.

2. Review of the basic procedure

The procedure of Bayesian seasonal adjustment introduced by Akaike [2] is defined as follows. We consider the seasonal time series $y(i)$ for M periods with K seasons for each period, where $i=Km+j$ $(j=1, 2,$

$\cdots, K;\ m=0, 1, \cdots, M-1)$. The observation $y(i)$ is represented by

$$y(i)=T_i+S_i+I_i\ ,$$

where T_i denotes the trend, S_i the seasonal and I_i the irregular component. We assume that the irregular components I_i are independently identically distributed as Gaussian with mean 0 and variance σ^2. We further assume that the vector of parameters $a=(T_1, T_2, \cdots, T_N, S_1, S_2, \cdots, S_N)$ has a prior distribution defined by

$$\pi(a\,|\,d)=\left(\frac{1}{2\pi}\right)^N\left(\frac{1}{\sigma}\right)^{2N}|d^2D'D|^{1/2}\exp\left[-\frac{d^2}{2\sigma^2}\|Da-Da_0\|^2\right]\ ,$$

where $'$ denotes the transpose, $|\cdot|$ denotes the determinant, d is an unknown constant, $\|x\|^2$ denotes the Euclidean norm of x, a_0 is a properly chosen initial guess of a and D is a $(2N+M)\times(2N)$ matrix defined by

$$D=\begin{matrix} \updownarrow N \\ \updownarrow N \\ \updownarrow M \end{matrix}\overset{\leftarrow N\rightarrow\ \leftarrow N\rightarrow}{\left(\begin{array}{c|c} D_k & 0 \\ \hline 0 & F \\ \hline 0 & G \end{array}\right)}\ ,$$

where the matrix D_k is defined as one of

$$D_1=\begin{pmatrix} \alpha & & & & 0 \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ 0 & & & -1 & 1 \end{pmatrix},\qquad D_2=\begin{pmatrix} \alpha & & & & & 0 \\ -\beta & \beta & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ 0 & & & 1 & -2 & 1 \end{pmatrix}$$

and

$$D_3=\begin{pmatrix} \alpha & & & & & & 0 \\ -\beta & \beta & & & & & \\ \gamma & -2\gamma & \gamma & & & & \\ -1 & 3 & -3 & 1 & & & \\ & -1 & 3 & -3 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ 0 & & & -1 & 3 & -3 & 1 \end{pmatrix},$$

where α, β and γ are properly chosen small constants. D_k controls the smoothness of the trend. The flexibility of the trend can be controlled by the choice of k. The matrix F is defined by

$$F = \underset{N}{\updownarrow}\overbrace{\begin{pmatrix} eI & & & & \\ -fI & fI & & & \\ & -fI & fI & & \\ & & \ddots & \ddots & \\ & & & -fI & fI \end{pmatrix}}^{N},$$

where I denotes a $K \times K$ identity matrix. The flexibility of the seasonal component is controlled by the choice of the constants e and f. The matrix G is defined by

$$G = \underset{M}{\updownarrow}\overbrace{\begin{pmatrix} g1' & & & \\ & g1' & & \\ & & \ddots & \\ & & & g1' \end{pmatrix}}^{N},$$

where $1'$ is a K component row vector $(1, 1, \cdots, 1)$. This matrix keeps the summation of the seasonal component whithin a period close to 0.

When the constant d is known, our estimate a_* of the vector a of parameters is defined as the value of a which minimizes $\|Z(a|d)\|^2$, where $Z(a|d)$ is given by

$$Z(a|d) = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \\ dC_0(1) \\ dC_0(2) \\ \vdots \\ dC_0(L) \end{pmatrix} - \begin{pmatrix} X \\ \\ dD \end{pmatrix} \begin{pmatrix} a(1) \\ a(2) \\ \vdots \\ a(2N) \end{pmatrix},$$

where $L = 2N + M$, $C_0 = Da_0$ and

$$X = \underset{N}{\updownarrow}\left(\overbrace{\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}}^{N} \;\vdots\; \overbrace{\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}}^{N}\right).$$

Obviously, a_* is the posterior mean of a.

The choice of the constant d is realized by maximizing $\log L(d)$ which is defined by

$$(-2) \log L(d) = N \log \left[\frac{1}{N} \|Z(a_*|d)\|^2\right] + \log |d^2D'D + X'X| - \log |d^2D'D| .$$

The choice of the parameters other than d, such as e, f and/or g, can also be realized by maximizing $\log L(d)$. It is observed (Akaike, [2]) that the above procedure produces fairly satisfactory results even for a short monthly time series of six years. If α, β and γ are fixed the choice between D_k's ($k=1, 2, 3$) is also realized by maximizing $\log L(d)$.

3. Extension to the case with missing observations

The extension of the procedure defined in the preceding section to the case with missing observations is particularly simple. As was stated in Introduction, the prior distribution does not undergo any modification. The only necessary modification is to discard the components of the vector $Z(a|d)$ which correspond to the missing observations.

One interesting aspect of this procedure is that by considering the yet unobserved values as missing we can get forecasts of the trend and seasonal components without using separate computational procedures.

4. Computational scheme

For the purpose of trend estimation of a long record it would be reasonable to apply the procedure described in the preceding sections for a span of several periods and retain the values of the central portion of the estimates as our final estimates. For example, we may apply the procedure for the data up to the sixth period and keep the trend and seasonal components corresponding to the third and fourth period of the data. By shifting the data by two periods and using the retained trend and seasonal components to define the values of the first two periods of a_0 and putting the rest of the components of a_0 equal to zero, we can get a procedure which realizes the estimation of trend and seasonal components for the shifted span. By repeating this procedure we can process indefinitely long record of time series.

5. Numerical examples

The procedure described in the preceding section was applied to various real and artificial time series. To see the performance of the procedure it was first applied to a monthly time series which was published in Abe, Ito, Maruyama et al. ([1], pp. 250–251) and the result is given in Fig. 1. This data is treated in Akaike [2] but now 8 data out of the 72 original values illustrated in Fig. 1a are eliminated as shown in Fig. 1b. By comparing the estimated trend with the theoretical trend, as given in Fig. 1c, we can see that the present proce-

Fig. 1 a Original record.

Fig. 1 b Data with missing observations.

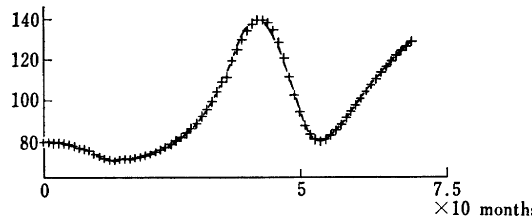

Fig. 1 c Estimated (---) and theoretical (+++) trend.

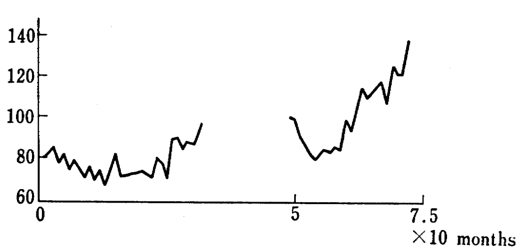

Fig. 2 a Record with heavily missing data.

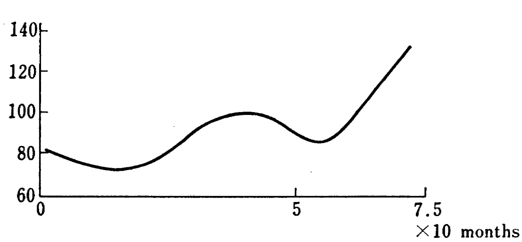

Fig. 2 b Estimated trend.

dure is producing satisfactory result. The search for the maximum of $L(d)$ was limited to the values of $d=1.0$, 2.0, 4.0 and 8.0. The values of e, f and g were fixed at 0.1, 1.0 and 10.0 respectively, and α, β and γ were set equal to 0.001. The minimum of $(-2)\log L(d)$ was attained at $k=2$ and $d=2.0$

A more drastic example is given in Fig. 2. Although rather unrealistic, this example demonstrates the potential of the present procedure in a heavily missing observations situation. The minimum of $(-2)\log L(d)$ was attained at $k=2$ and $d=16.0$.

Encouraged by the results of the preceding examples, the procedure was applied to a real record of humidity. The original record is illustrated in Fig. 3a and shows the periodic daily pattern. Since hourly readings were used for the analysis, the length of a period K was 24. The estimated series illustrated in Fig. 3b was obtained by inserting

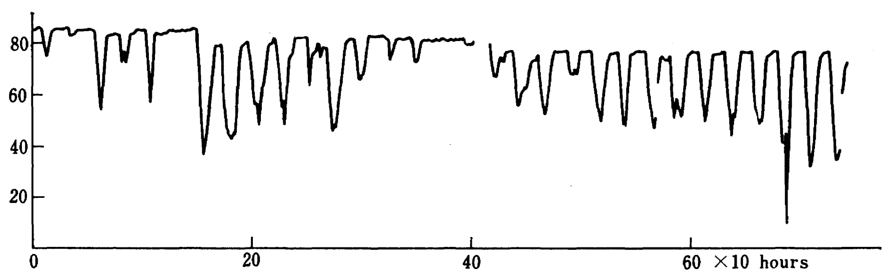

Fig. 3 a Record of humidity.

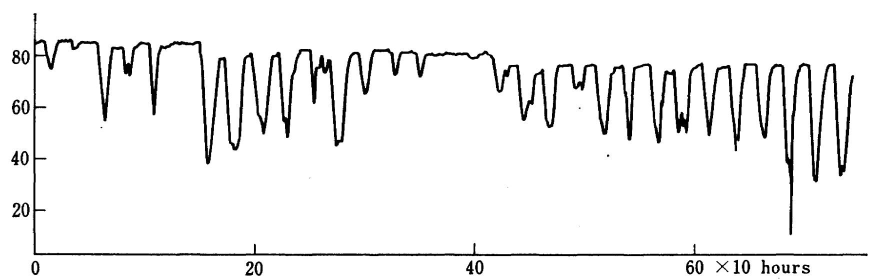

Fig. 3 b Record with the missing data estimated.

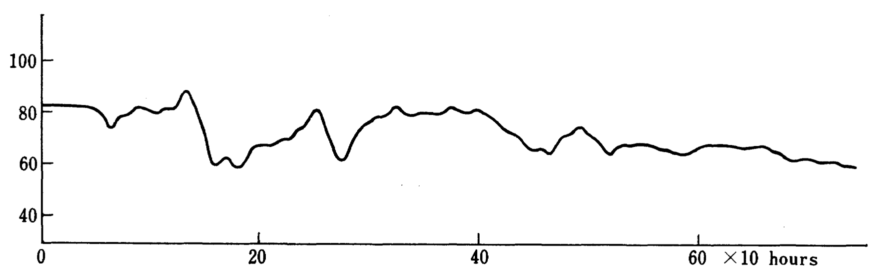

Fig. 3 c Estimated trend.

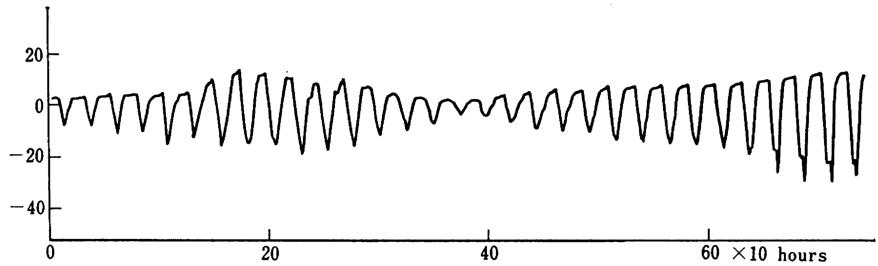

Fig. 3 d Estimated seasonal component.

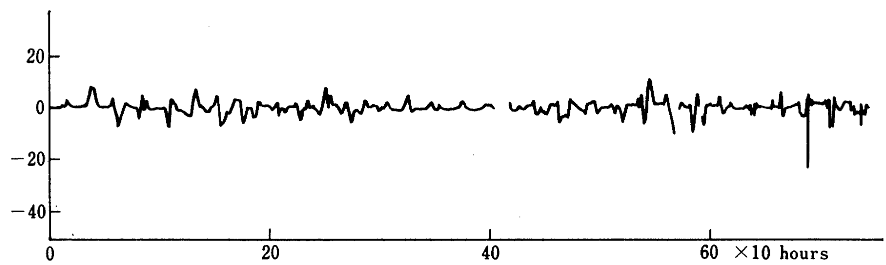

Fig. 3 e Estimated irregular component.

the estimates of T_i+S_i at the points where the original data were missing. Comparison of Fig. 3a with Fig. 3b suggests that the recovery of the missing observations is quite satisfactory. Fig.'s 3c, d, and e show the estimated trend, seasonal and irregular component, respectively. The behavior of the irregular component suggests that the abnormal observation towards the end of the record will easily be detected by applying some standard outlier detection scheme to this series.

6. Conclusion

The numerical results of the preceding section show that the proposed Bayesian type procedure of trend estimation is quite practical. The procedure will be useful for the restoration of missing values in various scientific observations.

The problem of abnormal records is always serious in the analysis of time series. Once a proper procedure of detecting the abnormality is established, the abnormal records can simply be ignored for the trend estimation by the present procedure. As is demonstrated by Fig. 3e, it seems that the present trend estimation procedure will help the detection of abnormalities and allow the development of some automatic correction procedure of abnormal observations.

A computer program which realizes the necessary computation for the Bayesian modeling discussed in this paper is given in Akaike and Ishiguro [3].

Acknowledgement

The research reported in this paper was partly supported by a grant from the Asahi Shimbun Publishing Company.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Abe, Y., Ito, S., Maruyama, A., Yoshikawa, J., Tsukada, K. and Ikegami, M. (1971). *Method of Seasonal Adjustments, Research Series* No. 22, The Research Institute of Economics, Economic Planning Agency, Tokyo (in Japanese).

[2] Akaike, H. (1979). Likelihood and the Bayes procedure. A paper presented at the International Meeting on Bayesian Statistics, May 28-Júne 2, 1979, Valencia Spain. To be published in *Trabajos de Estadistica*.

[3] Akaike, H. and Ishiguro, M. (1980). BAYSEA, A Bayesian seasonal adjustment program, *Computer Science Monographs*, No. 13, The Institute of Statistical Mathematics, Tokyo.