Journal of Official Statistics
Vol. 3, No. 4. 1987, pp. 431–438
© Statistics Sweden

Application of Spectral Analysis to Editing a Large Data Base

Khoan Tan Dinh[1]

Abstract: This paper provides a simple statistical procedure to minimize respondent reporting errors and data entry errors in a data base. The procedure has been applied successfully to a large data base consisting of monthly information on electric power plants.

Key words: Periodogram; seasonal time series; trend; white noise.

1. Introduction

Several procedures for editing data can be found in the literature. However, a spectral analysis technique has not yet been addressed. This paper discusses a method for applying spectral analysis to minimize respondent reporting errors and data entry errors in a large data base. To illustrate this method, the data collected from electric utilities by the Energy Information Administration were used. About 5 600 monthly time series were obtained from 3 000 electric plants that use different types of procedures (i.e., hydro electric, internal combustion, etc.) and fuel (i.e., light oil, natural gas, etc.). A plant may have different combinations of procedures and fuel types. These different combinations result in different time series models. Using spectral analysis, 5 600 time series were classified into three groups: white noise, seasonal, and trend. Then, for each group an appropriate statistical test to identify a new datum was employed. The classification method is discussed in Section 2. The statistical tests used for the identification of errors are provided in Section 3 and editing procedures are discussed in Section 4. Finally, some numerical examples are shown in Section 5.

2. Classification of Time Series

Given a time series, X_t, $t = 1, 2,\ldots,n$, a common method of analysis is to decompose the series into three components – a "trend component," a "seasonal component," and a "random component." That is,

[1] Mathematical Statistician, Energy Information Administration, EI 424, Washington, D.C. 20585, U.S.A.

$$X_t = \mu + T_t + \sum_{k=1}^{m} (a_k \cos \omega_k t + b_k \sin \omega_k t) + \varepsilon_t , \qquad (1)$$

where μ is the constant mean, T_t is the trend component, the summation term is the seasonal component, ε_t is the error term, and n is the number of observations,

$$m = \begin{cases} \frac{n}{2}, \text{ if } n \text{ is even,} \\ \frac{n-1}{2}, \text{ if } n \text{ is odd,} \end{cases}$$

$$a_k = \frac{2}{n} \sum_{t=1}^{n} X_t \cos \omega_k t , \; b_k = \frac{2}{n} \sum_{t=1}^{n} X_t \sin \omega_k t,$$

$$\omega_k = \frac{2\pi k}{n}, \text{ and } k = 0, 1, 2, \dots , m.$$

The periodogram is used to detect and estimate the amplitude of a cyclical component. A detailed discussion of the periodogram can be found in Fuller (1976), or Jenkins and Watts (1968). The periodogram is defined by

$$I(\omega_k) = \frac{n}{2} (a_k^2 + b_k^2) , \qquad (2)$$

where $k = 1, 2, \dots , m\text{–}1$,

$$I(\omega_m) = \begin{cases} \frac{n}{2} a_k^2, \text{ if } n \text{ is even,} \\ \frac{n}{2} (a_m^2 + b_m^2), \text{ if } n \text{ is odd.} \end{cases}$$

The quantity in (2) is the sum of two squares associated with frequency ω_k. Therefore, for an even number of observations the total sum of squares may be partitioned into $m + 1$ components. One component is associated with the mean (1 degree of freedom (df)). Each of the next m–1 components is the sum of two squares (2 df) associated with the m–1 nonzero frequencies. The last component is the single square (1 df) associated with the frequency equal to π. These results will be used to form the classification of the time series.

A time series is called white noise if it consists of the constant mean and the error term only. A trend time series consists of the constant mean, the trend component, and the error term. A seasonal time series is generally expressed in (1). In the seasonal time series, the trend component may or may not be eliminated. Based on the following tests, a time series can be classified as either white noise, trend, or seasonal. The following is a discussion of these tests.

2.1. White noise test

The white noise (wn) test is the test to determine whether a time series is white noise or not. To do that, we test the null model, $X_t = \mu + \varepsilon_t$, versus the alternative model in (1). That is, we test a white noise time series against a non-white noise time series. For this test, Fisher (1929) suggested the statistic E which is expressed as follows:

$$E = \frac{(m-1)\, M}{\sum_{k=1}^{m-1} I(\omega_k)} , \qquad (3)$$

where $M = \text{MAX}\,(I(\omega_1), \dots, I(\omega_{m-1}))$.

The percentage points for the statistic E, demonstrated by Fisher and Fuller (1976, p. 284), are listed in the column "wn" of a more general case in Table 1.

If E is less than the corresponding value of wn in Table 1, then the null model is accepted. The time series is called white noise. When the null hypothesis is rejected, the seasonal test is used to identify the category of the time series.

2.2. *Seasonal test*

The purpose of the seasonal test is to determine whether a time series is trend or seasonal. For a time series which has been rejected by the white noise test, the seasonal test is used to test a trend time series versus a seasonal time series. That is, test the null model, $X_t = \mu + T_t + \varepsilon_t$, against the alternative model in (1). The statistic for the seasonal test can be suggested as follows:

$$F = \frac{n-12}{11} \frac{\sum_{k=1}^{6} I\left(\frac{2\pi k}{12}\right)}{\sum_{k=1}^{m} I(\omega_k) - \sum_{k=1}^{6} I\left(\frac{2\pi k}{12}\right)}, \quad (4)$$

where $n = 12p$ for some integer p.

The statistic F in (4) follows the F-distribution with df (11, n–12), because the numerator and denominator are independent, and their sums of squares have 11 df and (n–1)–11 = n–12 df, respectively.

When the seasonal test is significant at the 10% level, the time series is called seasonal. Depending upon the magnitude of F in (4), we determine the type of the seasonal time series as follows:

1. If F is greater than the 1% critical value, then the time series is very cyclical. It is called "seasonal strong" (ss).
2. If F is less than the 1% critical value but greater than the 5% critical value, then the time series is somewhat cyclical. It is called "seasonal moderate" (sm).
3. If F is less than the 5% critical value but greater than the 10% critical value, then the time series is mildly cyclical. It is called "seasonal weak" (sw).

The critical values of wn, sw, sm, and ss corresponding with the n observations are derived from (3) and (4) and are shown in Table 1.

Table 1. Critical values

n	wn	sw	sm	ss
36	4.86	1.86	2.22	3.10
48	5.13	1.79	2.07	2.82
60	5.35	1.72	2.01	2.16
72	5.68	1.69	1.96	1.17

If a time series has been rejected by the white noise test, but not significant for the seasonal test, then it is called the trend time series.

3. Statistical Test for Editing

Let G_{ij} be a time series generated from a plant associated with the type of procedure and fuel collected in the ith month and jth year during the past J years. Having determined the category of the time series G_{ij} (i = 1, 2, ... , 12 and j = 1, 2, ... ,J) through the classification method, we use statistical tests as shown below to test a new datum. In the following tests, we call the value of the new datum collected in the month i of year J + 1 (current year) as "N."

3.1. *Overall t-test*

In a white noise time series, all observations vary around the overall (constant) mean. To check a new datum from this time series, the overall t-test (ott) is suggested. The statistic OT for the ott is defined as:

$$OT = \frac{|N - \hat{\mu}|}{\hat{\sigma}} , \quad (5)$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the mean and the standard deviation computed from the time series G_{ij}. That is,

$$\hat{\mu} = \frac{\sum_{i=1}^{12} \sum_{j=1}^{J} G_{ij}}{12J} \ ,$$

$$\text{and } \hat{\sigma} = \sqrt{\sum_{i=1}^{12} \sum_{j=1}^{J} \frac{g_{ij}^2}{12J - 1}} \ ,$$

where $g_{ij} = G_{ij} - \hat{\mu}$.

3.2. *Differenced t-test*

A common method to eliminate the trend component from a trend time series is to take the difference between two consecutive observations. To check a new datum from the trend time series, the differenced *t*-test (dtt) is proposed. The statistic *DT* for this test is:

$$DT = \frac{|(N - L) - \hat{\mu}_d|}{\hat{\sigma}_d} \ , \qquad (6)$$

where $\hat{\mu}_d$ and $\hat{\sigma}_d$ are the mean and standard deviation of the differenced series for two consecutive months during J years, and L is the observation preceding N.

The mathematical formulas for $\hat{\mu}_d$ and $\hat{\sigma}_d$ are shown as follows:

Let $D_{ij} = G_{ij} - G_{i-1j}$, if $i = 2, 3, \ldots, 12$ and $j = 1, 2, \ldots, J$, and let $D_{1j+1} = G_{ij} - G_{i-1j}$.

$$\text{Then } \hat{\mu}_d = \frac{D_{ij}}{12J - 1} \ ,$$

$$\text{and } \hat{\sigma}_d = \sqrt{\sum_{i=1}^{12} \sum_{j=1}^{J} \frac{d_{ij}^2}{12J - 2}} \ ,$$

where $d_{ij} = D_{ij} - \hat{\mu}_d$.

3.3. *Weighted t-test*

Depending on the seasonality of the time series, the current month's mean, the previous month's mean, and the next month's mean will have different effects on a new datum; therefore a weighted *t*-test (wtt) should be used to check a datum from a seasonal time series. To obtain the statistic *WT* for the wtt, we have to define first the weighted mean based on the current month's mean, the previous month's mean and the next month's mean. That is, let $\bar{G}_i$ be the *i*th month mean, $\bar{G}_i = \Sigma\, G_{ij}/J$. Then the weighted mean for the *i*th month is defined as:

$$\hat{\mu}_{wi} = \bar{G}_i W + \bar{G}_{i-1} \frac{1 - W}{2} + \bar{G}_{i+1} \frac{1 - W}{2} ,$$

if $i = 2, 3, \ldots, 11$,

$$\hat{\mu}_{w1} = \bar{G}_1 W + \bar{G}_{12} \frac{1 - W}{2} + \bar{G}_2 \frac{1 - W}{2}, \text{ and}$$

$$\hat{\mu}_{w12} = \bar{G}_{12} W + \bar{G}_{11} \frac{1 - W}{2} + \bar{G}_1 \frac{1 - W}{2} \ ,$$

where

$$W = \begin{cases} 1.00, & \text{if the time series is seasonal strong,} \\ 0.50, & \text{if the time series is seasonal moderate,} \\ 0.33, & \text{if the time series is seasonal weak.} \end{cases}$$

The statistic for the wtt is:

$$WT = \frac{|N - \hat{\mu}_{wi}|}{\hat{\sigma}_s} \ , \qquad (7)$$

where $\hat{\sigma}_s$ is the seasonal standard deviation of the data obtained from the monthly standard deviation, $\hat{\sigma}_{mi}$, $i = 1, 2, \ldots, 12$. That is,

$$\hat{\sigma}_s = \sqrt{\sum_{i=1}^{12} \frac{\hat{\sigma}_{mi}^2}{12}} \ ,$$

where $\hat{\sigma}_{mi} = \sqrt{\sum_{j=1}^{J} \frac{\dot{g}_{ij}^{2}}{J-1}}$ and $\dot{g}_{ij} = G_{ij} - \bar{G}_{i}$.

4. Editing Procedure

Based on the category of the time series being determined, a new observation from the time series is tested for possible respondent reporting errors and data entry errors.

4.1. From a white noise time series

Suppose a new datum is obtained from a white noise time series. To test a possible error, the ott is used. If this test is not significant, then the new datum is acceptable. Otherwise, there may be an error and it must be checked.

4.2. From a seasonal time series

Suppose a new datum is from a seasonal time series, then the wtt is used to test the new datum. If this test is not significant, then the new datum is acceptable. Otherwise, the dtt is used. Suppose that the dtt is significant, the new datum must be checked. Otherwise, this datum is acceptable.

For an observation from a seasonal time series, we might use two statistical tests (wtt and dtt) to test the new datum, because in the model of the seasonal time series, the seasonal component is significantly different from zero, but the trend component may or may not be significant. Therefore, we need the dtt to take care of the trend effect.

4.3. From a trend time series

Finally, if a new datum is associated with a trend time series, the dtt is used to test the new datum. If this test is not significant, the datum is acceptable. Otherwise, the new datum must be checked.

5. Illustrated Examples

We illustrate the classification method and the editing procedure with two examples. The first example is for a white noise time series, and the second for a seasonal moderate.

Example 1

The first example consists of data on the amount of energy, expressed in thousand kilowatthours (kWh), and produced by a plant where the procedure is the internal combustion engine and the fuel is natural gas. These data were collected in 1979–1983 and are shown in Table 2.

Table 2. Series A

Month	1979	1980	1981	1982	1983
1	5614	2746	2421	2702	1284
2	6107	2476	2583	2131	1227
3	5077	2583	2111	2279	1365
4	5266	4563	1731	1973	683
5	4706	1800	1760	1882	5
6	2058	2018	1786	2070	497
7	2426	3149	2928	3458	2069
8	3160	3356	2816	3480	2390
9	2876	2099	2730	2411	406
10	3670	1891	1904	2093	9
11	2455	1812	1871	1313	64
12	2110	2293	2307	1339	986

The data set of series A contains 60 observations. Therefore, we have 30 periodogram ordinates as shown in Table 3. The white noise test is not significant at the 10% level. Because the statistic, $E = 5.07$, in (3) is less than the 10% critical value, 5.35, from Table 1, the time series is classified into the white noise category.

Table 3. Periodogram of series A

k	Frequency	Period	Periodogram ($E+4$)
1	0.10	60.0	1 111
2	0.21	30.0	1 679
3	0.31	20.0	1 117
4	0.42	15.0	801
5	0.52	12.0	260
6	0.63	10.0	683
7	0.73	8.6	481
8	0.84	7.5	506
9	0.94	6.7	465
10	1.05	6.0	1 062
11	1.15	5.5	5
12	1.26	5.0	159
13	1.36	4.6	126
14	1.47	4.3	10
15	1.57	4.0	118
16	1.68	3.8	77
17	1.78	3.5	177
18	1.88	3.3	13
19	1.99	3.2	62
20	2.09	3.0	312
21	2.20	2.9	136
22	2.30	2.7	51
23	2.41	2.6	2
24	2.51	2.5	53
25	2.62	2.4	5
26	2.72	2.3	1
27	2.83	2.2	7
28	2.93	2.1	79
29	3.04	2.1	29
30	3.14	2.0	1

Since series A is white noise, we use the ott to test a new datum collected in 1984. The accepted region of the ott is limited by the upper and lower bounds which are computed as follows:

$$U = \hat{\mu} + 2.5\,\hat{\sigma}\ , \tag{8}$$

$$L = \hat{\mu} - 2.5\,\hat{\sigma}\ , \tag{9}$$

where U and L are the upper and lower bounds, and $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and standard deviation. The numerical values of $\hat{\mu}$, $\hat{\sigma}$, L, and U are shown in Table 4.

The monthly data collected in 1984 are tested by comparison with the upper and lower bounds in Table 4. For example, the plant associated with the internal combustion engine and natural gas reported the electric utilities (in thousand kWh) from January to August 1984 as shown in Table 5.

Table 4. Parameters for the ott

Mean	Standard deviation	Lower bound	Upper bound
2356	1275	–832	5544

Table 5. New data from series A

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Generation	1760	201	308	14	2	471	15	141

All these data are accepted, because they are between the upper and lower bounds.

Example 2

The data for the second example are also collected from the same plant with the same procedure as in the first example, but the fuel is light oil. The data set is shown in Table 6.

Table 6. Series B

Month	1980	1981	1982	1983
1	233	269	286	175
2	249	203	228	139
3	232	203	202	149
4	199	176	126	115
5	199	180	164	2
6	224	180	202	60
7	351	325	275	234
8	352	224	283	226
9	201	303	335	60
10	189	196	205	1
11	173	177	138	10
12	208	219	159	116

The data set of series B contains 48 observations. We have 24 periodogram ordinates which are shown in Table 7. The statistic, $E = 7.77$, obtained from (3) is greater than the 10% critical value, 5.13, as shown in Table 1. series B is not white noise. Therefore, the seasonal test must be used. The statistic, $F = 2.40$, derived from (4) is between the 5% critical value, 2.07, and the 1% critical value, 2.82. Hence, series B is a seasonal moderate time series.

Table 7. Periodogram of series B

k	Frequency	Period	Periodogram
1	0.13	48.0	63 439
2	0.26	24.0	45 227
3	0.39	16.0	18 539
4	0.52	12.0	11 864
5	0.65	9.6	11 707
6	0.79	8.0	3 738
7	0.92	6.9	434
8	1.05	6.0	104 672
9	1.18	5.3	10 744
10	1.31	4.8	2 855
11	1.44	4.4	4 126
12	1.57	4.0	4 616
13	1.70	3.7	4 932
14	1.83	3.4	2 246
15	1.96	3.2	2 762
16	2.09	3.0	5 602
17	2.23	2.8	3 280
18	2.35	2.7	1 203
19	2.49	2.5	378
20	2.62	2.4	638
21	2.75	2.3	2 169
22	2.88	2.2	466
23	3.01	2.1	4 089
24	3.14	2.0	6 567

Because series B is seasonal moderate, we may use the wtt and dtt to test a new datum collected in 1984. For the wtt, the weighted mean is computed from Section 3 with the weight equal to 0.50, and the boundary values corresponding to ith month of year 1984 are obtained as follows:

$$U_{wi} = \hat{\mu}_{wi} + 2.5\,\hat{\sigma}_s \quad , \tag{10}$$

$$L_{wi} = \hat{\mu}_{wi} - 2.5\,\hat{\sigma}_s \quad , \tag{11}$$

where

U_{wi}: Upper bound of the wtt for month i,
L_{wi}: Lower bound of the wtt for month i,
$\hat{m}_{wi}$: Weighted mean for the month i, and
$\hat{\sigma}_s$: Seasonal standard deviation.

The numerical values of the weighted mean, upper bound, and lower bound are shown in Table 8.

Table 8. Parameters for the wtt

Month	Mean	Standard deviation	Weighted mean	Lower bound	Upper bound
1	241	49	215	181	250
2	205	48	212	177	246
3	197	35	188	154	222
4	154	40	160	126	195
5	136	91	148	114	183
6	167	73	191	157	226
7	296	52	258	223	292
8	271	60	266	232	300
9	225	124	217	183	252
10	148	98	161	127	196
11	125	78	143	109	178
12	176	48	179	145	213

If a monthly datum of series B collected in 1984 fell outside of the accepted region determined by (10) and (11) as shown in Table 8, we use the dtt to test this datum again. The boundaries of the dtt are given as follows:

$$U_d = \hat{\mu}_d + 2.5\,\hat{\sigma}_d \quad , \tag{12}$$

$$L_d = \hat{\mu}_d - 2.5\,\hat{\sigma}_d \quad , \tag{13}$$

where

U_d: the upper bound of the dtt, and

L_d: the lower bound of the dtt.

The numerical values of $\hat{\mu}_d$, $\hat{\sigma}_d$, U_d, and L_d are shown in Table 9.

Table 9. Parameters for the dtt

Diff. mean	Standard deviation	Lower bound	Upper bound
–2	71	–179	174

Now the new data of series B are collected from January through August 1984, and their differences are calculated as reported in Table 10.

Table 10. New data from series B

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Gen.	227	38	55	9	45	69	4	27
Diff.		–189	17	–45	36	24	–64	23

In January, the observation, 227, is between the two boundaries (181, 250) of the wtt as shown in Table 8. It implies that this datum is good. The others are outside the wtt boundaries. We thus compare these differences with the dtt boundaries as shown in Table 9. Between February and January the difference data, –189, is less than the lower bound, –179, of the dtt. Therefore, one must check the February datum for possible errors. The other differences are inside the dtt boundaries and hence are accepted.

6. References

Fisher, R. A. (1929): Test of Significance in Harmonic Analysis. Proceedings Royal Society, London, Series A 125, pp. 54–59.

Fuller, W. A. (1976): Introduction to Statistical Time Series. John Wiley, New York.

Jenkins, A. M. and Watts, D. G. (1968): Spectral Analysis and Applications. Holden Day, San Francisco, California.

Received July 1986
Revised February 1988