DETECTING OUTLIERS IN MICROMETEOROLOGICAL TIME SERIES

*Dr. Rudy S. Tan**

1. Introduction

The collection of large masses of micrometeorological time series by completely automatic systems now used in air pollution monitoring has raised new problems in statistical methodology, particularly on methods to screen and improve the quality of the data. It is a well-known fact that no amount of sophisticated statistical computing can make the result worth the time and effort if the data are of dubious quality. Unfortunately, the automation of data acquisition has made the problem of obtaining good quality micrometeorological time series much more complex. The reasons appear to be not only the enormously large number of variates and observations that can be recorded, but also the various technical factors affecting the computerized systems. It is rare that all instruments in a large monitoring network function continuously for over a period of three days. Furthermore, although instrument performance in recent years has improved considerably, most instruments have drift characteristics requiring very frequent calibrations. In view of the high cost involved in setting up and maintaining an automatic continuous monitoring system, data quality control methods have to be developed in order to make maximum use of the available observations and also to assess the performance of the instruments, e.g., to determine if the instrument should be checked prior to its normal scheduled maintenance. The solution to the problem of data quality control for micrometeorological time series is not simple since the measurements are dependent. Otherwise, quality control and editing methods used in sample surveys (see Naus, 1975) are readily applicable.

*The author is an associate professor, Statistical Center, University of the Philippines. This paper is from a Chapter of the author's doctoral dissertion completed at North Carolina State University, U.S.A.

In general, the quality of micrometeorological time series collected by completely automatic systems is affected by gross errors, outliers, and systematic errors. Gross errors, as distinguished from random measurement errors, are obviously incorrect observations. These should be discarded since they can adversely affect the estimates of mean and variance of the time series and may result in errors when making critical decisions like shutting down or cancelling certain plant operations under certain atmospheric conditions. Severe weather is the most frequent cause of gross errors. Lightning, hail, and strong winds accompanying a heavy thunderstorm can damage the instruments mounted on towers. Furthermore, electrical "surges" during thunderstorm activities are the common cause of noise burst in the communication link between the instruments and electronic processing equipment resulting in garbled data. During cold weather, thick ice accumulation on the instruments can make them either inoperable or record bad data. Fortunately, gross errors are easy to detect and few deterministic tests usually would be able to screen them out.

"Outliers" are defined as values which appear to depart markedly from the rest of the data. They are often called "stragglers", "sports", or "mavericks" in the literature (Anscombe, 1960). Spectral analysis which is now widely used in the physical sciences is very sensitive to outliers. An outlying observation can introduce a spectral peak in the frequency domain and several can make the spectrum appear multipeaked. When many outliers are present, it may be impossible to obtain any meaningful analysis of the spectrum of interest (see Koopmans, 1974, Section 9.5). Outliers may also be caused by severe weather, by occasional gusts especially during periods of light winds, and by certain atmospheric phenomena like breaking gravity waves. Even flying objects like birds when passing very close to the measuring instrument are likely to cause some observations to go wild momentarily and introduce outliers. Unfortunately, they are difficult to detect because of the highly subjective nature of outlier rejection procedures (Collett and lewis, 1977). If questionable values are really bad observations, then they should be discarded and imputed by some means before any analysis of the data is undertaken. However, if they are just extreme values, a decision must be made as to whether or not to include them in the analysis. In the later case, extreme caution must be exercised when employing any criterion for rejecting outliers. It may happen that the outlying observations occurred naturally, and only by including them in the analysis can the process under study be fully described.

Finally, systematic errors are those which cause the measurements to differ from the "true" values in a constant or regular manner. They are generally much more difficult to detect than outliers even if repeated observations are taken. Averages based on data subject to systematic errors will normally be biased. There are various causes of systematic errors in micrometeorological time series. Although proper calibration is a necessary prerequisite before undertaking any measurement, the instrument eventually looses its reliability after a period of time. If the zero point has shifted significantly, then systematic errors are introduced into the measurements which may not be evident. For this reason, frequent calibration of the instrument is recommended in any automatic monitoring system. Asymmetrical ice coating can change the response characteristic of the instrument. Thus, data obtained when icing occurs may have systematic errors (Alexeiev *et al.*, 1974). Sheltering effects due to the location of the instrument can also introduce systematic errors. It has been observed that there is a significant reduction in the measured wind speed when the anemometer is located close to the tower structure (Angell *et al.*, 1976, and Wieringa, 1976). Other causes of systematic errors are improper leveling of the wind direction transmitter, faulty potentiometers, improper conversion of the output signal, and misalignment of instruments. The latter can introduce systematic errors into the measurements after a certain period of time. It has been observed that aerodynamic lifting due to the orientation of the bivane tail as it is mounted on the wind sensor can cause an appreciable error in the elevation angle measurements after the tail had deteriorated (Pendergast, 1975).

This paper is concerned only with the detection of outliers in micrometeorological time series after first screened for gross errors by simple deterministic tests. Methods for detecting systematic errors or instrumental drift involves modelling the time series with stochastic time-varying parameters and are beyond the scope of this study.

2. Review of Literatures on Detection of Outliers in Time Series

There seems to be a lack of publications on the detection of outliers in time series. In recent years, time series is increasingly becoming an important area of applied statistics especially in the physical sciences, where for a long time curve fitting by ordinary least squares was the most frequently used (or "abused") method of statistical analysis. Unfortunately, a time series is more outlier prone than a random sample and the problem of outlier detection is more difficult to handle.

Probably, the first to consider the problem of outliers in time series was Fox (1972). He considered two types of outliers that may occur in a time series and named them type I and type II. A type I outlier corresponds to a gross error in which the error affects only a single observation. Let the time series be represented by the p*th* order autoregressive model

$$\alpha_0 Z + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \ldots + \alpha_p Z_{t-p} = e_t, \quad t = p + 1 \ldots, N,$$

where the $e_t's$ are independently and normally distributed random variables with mean 0 and variance σ^2. Due to the presence of an outliers in the q*th* observation, the observed time series Y_t is such that

$$Y_t = \begin{cases} Z_t & \text{if } t \neq q \\ Z_q + \delta & \text{if } t = q \end{cases}$$

On the other hand, a type II outlier corresponds to an extreme value. The error δ affects not only the observation Y_q, but also the subsequent observations $Y_{q+1}, \ldots, Y_N$. The p*th* order autoregressive model representation of the observed time series Y_t is

$$\alpha_0 Y_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \delta_t = e_t, \qquad t = p + 1, \ldots, N,$$

where

$$\delta_t = \begin{cases} 0 & \text{if } t \neq q \\ \delta & \text{if } t = q \end{cases}$$

and the $e_t's$ are independently and normally distributed with mean 0 and variance σ^2. Four situations are possible, namely: the outliers are all of type I, the outliers are all of type II, all of the outliers are of the same but unknown type, and the outliers are a mixture of both types. The author then derived likelihood ratio and approximate likelihood criteria for testing whether a particular observation is an outlier.

Abraham and Box (1975, 1976) approached the problem of outliers in time series by the non-classical Bayesian method. First, the authors considered two situations: (1) there is no specific information to distinguish the degree of goodness of one observation from another, but there is a general possibility that r $\geqslant$ 1 of the observations may be an outlier; (2) there is information that r specific observations obtained under known conditions are outliers, but the extent of their distributions is unknown. They listed two objectives: (1) to make inferences about the parameters of the model in the possible presence of outliers, and (2) to determine whether a particular observation is an outlier and estimate its deviation from expectation for a good observation. Finally, they characterized the outlier problem by two models in the context of the p*th* order autoregressive process. The first, called the aberrant innovation model, is given by

$$\alpha_0 Y_t + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \delta_t = e_t,\ t = p+1, \ldots, N,$$

where

$$\delta_t = \begin{cases} \delta & \text{if } t = q_1, \ldots, q_r \\ 0 & \text{otherwise} \end{cases}$$

and the e_t's are independently and normally distributed random variables with mean 0 and variance σ^2. In this model, the error δ stays in the process through time $t = q_1, \ldots, q_r$, where $q_1, \ldots, q_r$ or r may not be known. Now, the second, called the aberrant observation model, is given by

$$\alpha_0 Z_t + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \ldots + \alpha_p Z_{t-p} = e_t,\ t = p+1, \ldots, N$$

where Z_t is the time series and the e_t's are independently and normally distributed random variables with mean 0 and variance

σ^2. The observed time series Y_t in the presence of outliers can be written as

$$Y_t = \begin{cases} Z_t + \delta & \text{if } t = q_1, \ldots, q_r \\ Z_t & \text{otherwise} \end{cases}$$

Here, the error may be thought of as an observational error having an immediate effect only on Y_t at $t = q_1, \ldots, q_r$, where q_1 , . . ., q_r or r may not be known. These two models are analogous to the two types of outliers proposed by Fox. The different situations, objectives, and model characterizations of Abraham and Box may be combined to give eight different outlier problems.

It is assumed before applying the method of Fox or Abraham and Box that any trend or seasonal variation is either negligible or has been removed. This assumption requires that the deterministic trend component of the time series is removed by a robust method, i.e., one that is not easily influenced by the presence of outlying observations. The conventional method of ordinary least squares is not robust. Minimizing the sum of the absolute values of the errors is probably the only robust method of regression in the technical sense of the word. However, this method requires the use of linear programming which is not feasible when the number of observations is large. The various methods for multiple linear regression (see, for example, Andrews, 1974, or Hinich and Talwar, 1975) are not applicable to time series since the data is decimated by removing all observations which are suspected as outliers. In a time series, the residuals have autocorrelation structure and it would not be advisable to discard any observation without first imputing its value by some means. Roughan and Evans (1970) proposed to fit a trend curve to a time series in the presence of outliers by an iterative scheme. Their criterion for fitting is simply having an equal number of observations on either side of the curve for all the segments into which the time series was arbitrarily divided. Although the method appears to be robust, it has no statistical basis and the algorithm is not entirely satisfactory. In addition to the problem of trend in the time series, the method of Fox or Abraham and Box assumes that the outliers all belong to the same population and that the error σ is a constant. This assumption may be true with some physical data, but in general it will not be.

3. A Regression Method for Detecting Outliers in Time Series

The two method for detecting outliers discussed in the previous section have been found impractical for screening large masses of micrometeorological data due to the problem of trend in the data and the complex computations involved. For example, the posterior distribution of the autoregressive parameters given the observations, in the aberrant innovation model of Abraham and Box, is a weighted average of multivariate t-distributions whose exact evaluation is computationally difficult. Therefore, a method is developed in this study which is based entirely on multiple linear regression analysis. This commonly employed statistical analysis is readily adaptable to time series modelling and efficient computer programs are available even in some textbooks in statistics (e.g., Burford, 1970).

3.1. *The Model*

In order to avoid the problem of first removing any trend in the data before modelling the time series, the following second-order autoregressive model with a trend line is considered:

$$Y_t = \mu + \alpha t + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + e_t, \quad t = 3, \ldots, N \qquad (3.1)$$

where the e_t are normally and independently distributed random variables with mean 0 and variance σ^2_e. The proposed model is simply a reparameterization of the following simple linear model with autocorrelated errors:

$$\begin{aligned} Y_t &= \mu^* + \alpha^* t + Z_t, \quad t = 3, \ldots, N, \\ Z_t &= \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + e_t, t = 3, \ldots, N, \\ \mu^* &= \frac{\mu(1-\beta_1-\beta_2) - \alpha(\beta_1 + 2\beta_2)}{(1-\beta_1-\beta_2)^2}, \end{aligned} \qquad (3.2)$$

$$\alpha^* = \frac{\alpha}{1-\beta_1-\beta_2} \qquad (3.3)$$

It is assumed that the parameters of the model μ, α, β_2 and β are fixed constants, and the roots of the characteristic equation m^2

$\beta_1 m - \beta_2 = 0$ are less than one in absolute value. The second assumption is satisfied if Z_t is a stationary time series.

The $(N - 2)$ equations in (3.1) can be written in matrix form as follows:

$$\underset{\sim}{Y} = \underset{\sim}{\chi}\,\underset{\sim}{\theta} + \underset{\sim}{e},$$

where

$$\underset{\sim}{Y}' = (Y_3, \ldots, Y_N),$$

$$\underset{\sim}{\chi} = \begin{bmatrix} 1 & 3 & Y_2 & Y_1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & N & Y_{N-1} & Y_{N-2} \end{bmatrix}$$

$$\underset{\sim}{\theta}' = (\mu, \alpha, \beta_1, \beta_2),$$

and

$$\underset{\sim}{e}' = (e_3, \ldots, e_N).$$

The ordinary least squares estimator of $\underset{\sim}{\theta}$ is

$$\begin{aligned} \underset{\sim}{\hat{\theta}} &= (\underset{\sim}{\chi}'\underset{\sim}{\chi})^{-1}\underset{\sim}{\chi}'\underset{\sim}{Y} \\ &= \underset{\sim}{\theta} + (\underset{\sim}{\chi}'\underset{\sim}{\chi})^{-1}\underset{\sim}{\chi}'\underset{\sim}{e} \end{aligned}$$

Clearly, $\underset{\sim}{\hat{\theta}}$ is a biased estimator of $\underset{\sim}{\theta}$ since $\Sigma[(\underset{\sim}{\chi}'\underset{\sim}{\chi})^{-1}\,\underset{\sim}{\chi}'\underset{\sim}{e}] \neq 0$ because e_t, $t = 3, \ldots, N$, are not independent of Y_t for $t - 3, \ldots, N$. Indeed the small sample distribution of $\underset{\sim}{\hat{\theta}}$ has not yet been obtained. However, $\underset{\sim}{\hat{\theta}}$ is consistent and asymptotically normal, i.e.,

$$\operatorname{plim} \underset{\sim}{\hat{\theta}}_n = \underset{\sim}{\theta}$$

and

$$\sqrt{n}\,(\underset{\sim}{\hat{\theta}} - \underset{\sim}{\theta}) \;\dot{\sim}\; N(\underset{\sim}{\hat{\theta}}, \underset{\sim}{Q}^{-1}\sigma_e^2),$$

where $\underset{\sim}{Q} = \operatorname{plim} \frac{1}{N}\,(\underset{\sim}{\chi}'\underset{\sim}{\chi})$

and plim denotes limit in probabilıty (see Fuller, 1976; Section 8.2). In other words, if the sampling time is sufficiently long, the distribution of θ would be approximated by a normal distribution with mean $\underset{\sim}{\theta}$ and variance $Q^{-1}\sigma_e^2/N$. This implies that it is reasonable to employ multiple regression to estimate the parameters of the model (3.1) and also that the usual tests of significance are not too misleading.

The choice of the model was based on the fact that micrometeorological variables like temperature, wind speed, and wind direction angles recorded at short intervals of time (5 seconds or less) display strong persistence, and when the sampling time is long (1 hour or more) a "trend" in the mean may exist. After examining several correlograms from the pilot data used in this study, it was evident that the second-order autoregressive model would adequately represent the observed time series of temperature, wind speed, azimuth, and elevation angle under the atmospheric conditions encountered. This model has been widely used to describe various time series in practice (Stralkowski, *et. al.*, 1970).

The deterministic linear trend in the model is to take into account any change in the level of the time series due to diurnal variations. The slope of the trend line is significant only during certain periods. During most of the day and night, the mean of the process generating the time series is generally constant. A straight line trend was found satisfactory in describing the mean of the process up to three hours sampling time. However, a more flexible trend curve would have to be considered if the time duration was much longer.

3.2. *The Screening Procedure*

The steps in screening the data for outliers are as follows:

1. The parameters μ, α, β_1, and β_2 in the model (3.1) are estimated by the method of ordinary least squares.

2. Let $\hat{Y}_t$ be the predicted value of Y_t, $\hat{e}_t = Y_t - \hat{Y}_t$ be the residual, and $\hat{\sigma}_e^2$ be the estimated mean square error.

3. If $|e_t| > \delta\hat{\sigma}_e$, then Y_t is replaced by Y_t and this new value used to predict Y_{t+1}. The rejection criterion δ is predetermined.

4. After all the suspected outliers have been replaced by their predicted values, the parameters in the model are reestimated using the recently created data set.

5. Now, using the revised estimates of the parameters in the model and mean square error, steps 2, 3, and 4 are repeated using the original observations.

6. Finally, the cycle is terminated when a certain number of iterations is reached, or if

$$\left|\hat{\sigma}_{e(\mathrm{i})} - \hat{\sigma}_{e(i-1)}\right| \leqslant k,$$

where $\hat{\sigma}_{e(i)}$ is the value for $\hat{\sigma}_e$ in the *ith* iteration and k is a specified small number.

Note that the predicted value of Y_t is a weighted linear combination of "good" observations and the weights are the coefficients of the autoregressive model. Although $\hat{Y}_t$ may not be very close to the true value of Y_t, it will at least minimize the influence of the suspected observation. In order to start the prediction, it is required that the first two observations are not outliers.

3.3. *Estimates of the Model Parameters*

The estimates of the parameters μ, α, β_1 and β_2 in (3.1) are obtained directly from the computer output of a multiple linear regression program. However, the degrees of freedom for the estimated mean square error $\hat{\sigma}_e^2$ has to be corrected for the number of outliers. Thus

$$\hat{\sigma}_e^2 = \frac{SSE}{N\text{-}M\text{-}6}, \tag{3.4}$$

where SSE is the error sum of squares, N is the number of observations, M is the number of outliers, and 6 is the number of parameters in the model [4] plus the number of observations deleted. The standard errors of the estimated coefficients are obtained by multiplying $\hat{\sigma}_e^2$ by the positive square-root of the diagonal elements in the $(\underset{\sim}{\chi}'\underset{\sim}{\chi})^{-1}$ matrix. Specifically, the standard errors of the estimated autoregressive coefficients β_1 and β_2, are respectively,

$$SE(\hat{\beta}_1) = \hat{\sigma}_e\sqrt{a_{33}}$$

and $$SE(\hat{\beta}_2) = \hat{\sigma}_e\sqrt{a_{44}},$$

where a_{33} and a_{44} are the *3rd* and *4th* diagonal elements in the $(\underset{\sim}{\chi}'\underset{\sim}{\chi})^{-1}$ matrix. To test the hypothesis that $\beta_2 = 0$ i.e., the time series is a first-order autoregressive process, the following t-statistics

$$t = \hat{\beta}_2 / SE(\hat{\beta}_2)$$

is computed.

From (3.2) and (3.3), the estimates of the trend line coefficients μ^* and a^* are, respectively,

$$\hat{\mu}^* = \frac{\hat{\mu}(1-\hat{\beta}_1\hat{\beta}_2) - \hat{\alpha} \quad (\hat{\beta}_1 + 2\hat{\beta}_2)}{(1 - \hat{\beta}_1 - \hat{\beta}_2)^2}$$

and

$$\hat{\alpha}^* = \frac{\hat{\alpha}}{1-\hat{\beta}_1 - \hat{\beta}_2}$$

Unfortunately, the variances of $\hat{\mu}^*$ and $\hat{\alpha}^*$ involve nonlinear combinations of the estimates $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ which are mathematically intractable. Therefore, an indirect approach has to be considered in order to obtain the standard errors of $\hat{\mu}^*$ and $\hat{\alpha}^*$. Durbin (1960) has suggested a two-state procedure to estimate μ^* and α^* which will give estimates with asymptotically the same mean vector and dispersion matrix as the OLS estimates obtained by the direct minimization of $\Sigma\, e_t^2$ in (3.1). The procedure is as follows: After estimating the autoregressive parameters β_1 and β_2 in (3.1) by OLS, the transformed variables

$$W_t = Y_t - \hat{\beta}_1 Y_{t-1} - \hat{\beta}_2 Y_{t-2}$$

and

$$T_t = \hat{\beta}_1 + 2\hat{\beta}_2 + (1 - \hat{\beta}_1 - \hat{\beta}_2)\, t$$

are computed. Then, W_t is regressed on a column of 1's and T_t giving the following estimates of μ^* and α^*:

$$\hat{\mu}^* = \frac{\sum_{t=3}^{N} W_t - \hat{\alpha}^* \sum_{t=3}^{N} T_t}{(N-2)\ (1-\hat{\beta}_1\ \hat{\beta}_2)}$$

$$\hat{\alpha}^* = \frac{(N-2)\sum_{t=3}^{N} W_t T_t - (\sum_{t=3}^{N} W_t)(\sum_{t=3}^{N} T_t)}{(N-2)\sum_{t=3}^{N} T_t^2 - (\sum_{t=3}^{N} T_t)^2}$$

These estimates are equivalent to (3.5) and (3.6), respectively. Thus, the approximate standard errors of the trend line coefficients are simply the following:

$$SE(\hat{\mu}^*) = \frac{\hat{\sigma}_e}{1-\hat{\beta}_1-\hat{\beta}_2}\sqrt{\frac{\sum_{t=3}^{N} T_t^2}{(N-2)\sum_{t=3}^{N} T_t^2 - (\sum_{t=3}^{N} T_t)^2}}$$

and

$$SE(\hat{\alpha}^*) = \hat{\sigma}_e\sqrt{\frac{N-2}{(N-2)\sum_{t=3}^{N} T_t^2 - (\sum_{t=3}^{N} T_t)^2}}$$

3.4 *Measure of Model Adequacy*

It is possible that the model (2.1) does not provide an adequate fit to the observed time series. This will result either in the rejection of unusual number of "outliers" with some good observations among them, or not detecting the really outlying observations. One way to measure the adequacy of the model is to construct an index based on the autocorrelations of the residuals. Theoretically, if the fit of the proposed model is the appropriate one, the residuals should just be "white noise" with autocorrelations equal to zero for all lags greater than or equal to 1. Such an index is proposed in this study and is given by

$$G = 1 - \sum_{h=1}^{L} r_h^2 ,$$

where $$r_h = \frac{\sum_{t=3}^{N-h} e_t\, e_{t+h}}{\sum_{t=3}^{N} e_t^2}$$

and L is the number of sample autocorrelations. Clearly, if the fit is perfect, $G = 1$. However, $\sum_{h=1}^{L} r_h^2$ is seldom equal to 0 since the residuals from a fitted model tend to be slightly autocorrelated. Box and Pierce (1970) have suggested a statistic to test the smallness of $\sum_{h=1}^{L} r_h^2$. This is the Q statistics and is related to G as follows:

$$Q = (N-2)(1-G)$$

The distribution of Q is approximately a chi-square with (L-4) degrees of freedom. Box and Jenkins (1976) refer to Q as the "portmanteau" lack-of-fit test statistic. A significant value of Q would indicate that the fitted model is not adequate and may not be satisfactory for screening the data for outliers. This test is valid only if L is at least 20. Chatfield and Prothero (1973) showed that Q is not a very powerful statistic for detecting specific departures from white noise behavior in the residuals. Nevertheless, it is a useful diagnostic check on the adequacy of the model to represent the data.

4. Application

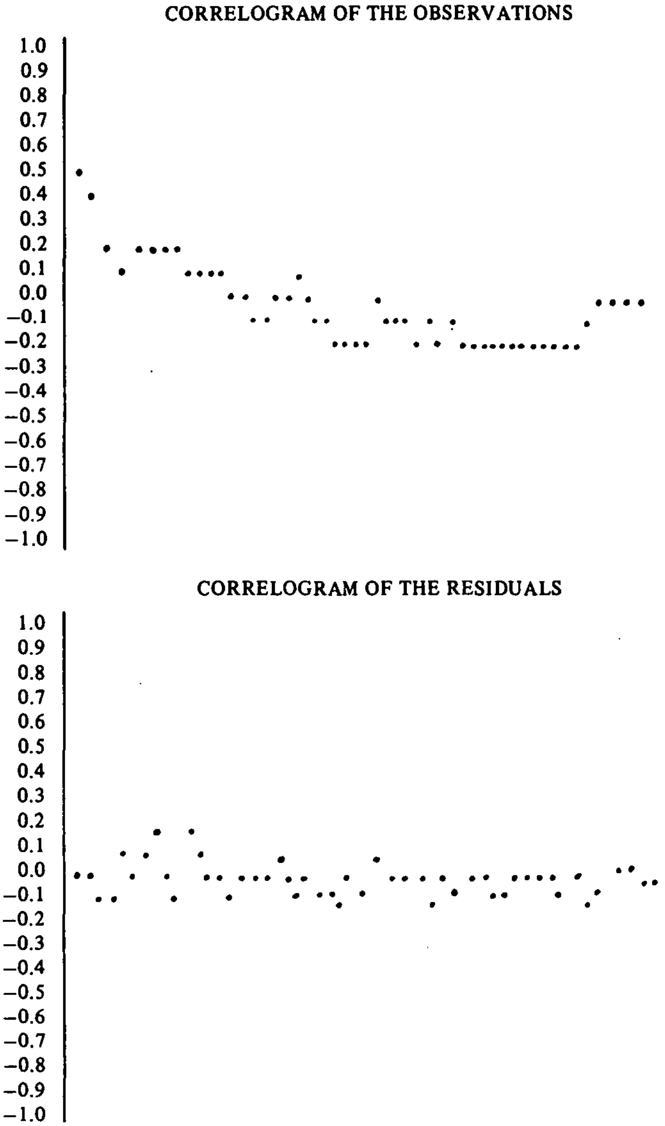
4.1 *Some Aspects of the Computer Program*

The basic computer program for the screening procedure was written in Fortran IV level G. For the purpose of this study, linkage routines were added to the program so that the procedure will interface with the Statistical Analysis System (SAS). The procedure was given the name SCREEN and its SAS implementation is discussed in Appendix 1.

To test the procedure, several artificial time series were generated and bad observations introduced. One example is discussed here. The observations were generated from the model $Y_t = 270.00 + 0.05t + Z_t$ where $Z_t = 0.5Z_{t-1} + 0.1Z_{t-2} + e_t$, where e_t is $N(0,1)$. Seven observations were replaced by values whose magnitude was $\pm 8\sigma_e$ from the mean. All of the 7 outliers were detected and also the true values of the parameters were within less than 2 standard errors of the estimates (see Fig. 1). The plots of the 200 observations together with their upper and lower control limits, adapting the terminology in industrial quality control, are presented in Fig. 2.

```
                         NO. OF OBSERVATIONS = 200
                         NO. OF ITERATIONS     =   4
                         NO. OF OUTLIERS       =   7

         T              YBAD              YHAT              EHAT
        48         278.00000         272.69653           5.30347
        49         278.00000         272.65082           5.34918
        50         278.00000         272.69283           5.30717
        51         278.00000         272.72194           5.27806
       160         270.00000         277.45851          -7.45851
       161         270.00000         277.68024          -7.68024
       162         270.00000         277.78203          -7.78203

   COEFFICIENTS OF                  STANDARD
    THE TREND LINE                   ERRORS          T-VALUES
        270.29603                   0.34017          794.59509
          0.04707                   0.00288           16.32333

    COEFFICIENTS OF                 STANDARD
   THE AUTOREG. MODEL                ERRORS          T-VALUES
         0.39172                    0.07154            5.45258
       _ 0.17220                    0.07156            2.40641
   CHARACTERISTIC ROOTS
         0.65473
         0.26301
     MEAN SQUARE ERROR             DEGREES OF FREEDOM
         1.02294                    187
            MEAN                 APPROXIMATE STD ERROR
       M1 = 275.05854                   0.16483
       STANDARD DEVIATION        APPROXIMATE STD ERROR
          SD1 = 1.16940                 0.08027
          SD2 = 1.16549

M1 IS SIMPLE MEAN COMPUTED FROM THE DATA AFTER REPLACING
   THE OUTLIERS BY THEIR PREDICTED VALUES.
SD1 IS COMPUTED FROM THE DATA AFTER REMOVING THE TREND
   AND REPLACING THE OUTLIERS BY THEIR PREDICTED VALUES.
SD2 IS COMPUTED FROM THE COEFFICIENTS OF THE FITTED SECOND-
   ORDER AUTOREGRESSIVE MODEL.
-                                                              -
DEGREES OF FREEDOM = NO. OF OBSERVATIONS - NO. OF PARA-
METERS - NO. OF OUTLIERS - 2
```

Fig. 1 Results of procedure SCREEN using 200 observations with 7 outliers generated from the model $Y_t = 108.035 + 0.02t + 0.5\ Y_{t-1} + 0.1\ Y_{t-2} + e_t$, where $e_t \sim N(0, 1)$.



```
NOTE: EACH POINT CORRESPONDS TO 1 LAG

MEASURE OF MODEL ADEQUACY
          G = 0.68411
PORTMANTEAU LACK OF FIT TEST
CHISQUARE = 83.1777  DF = 46
```

FIG. 1 (Continued)

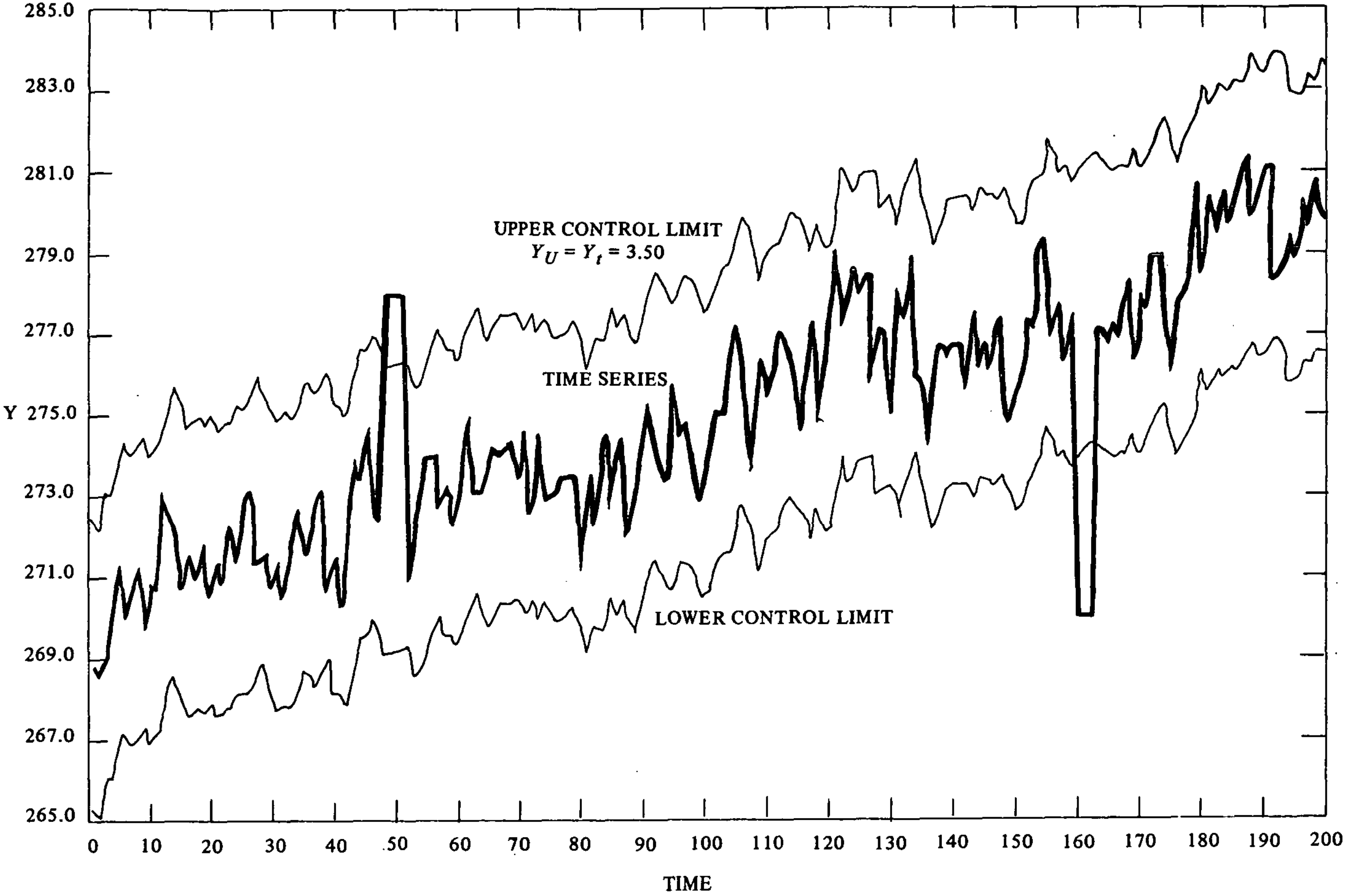


Fig. 2 Plots of 200 observations with 7 outliers generate from the model $Y_t = 108.035+0.02t + 0.5\ Y_{t-1} + 0.1\,Y_{t-2} + e_t$, where $e_t \sim N(0.1)$

The rejection criterion used for this example was 3.5. Assuming that the e_t's are independently and normally distributed random variables with mean 0 and variance 1, only values really far outside the 99.9% confidence limits of the predicted values of Y_t were replaced as outliers.

Presented in Fig. 3, 4, 5, and 6 are the results from procedure SCREEN compared with that of procedure GLM and AUTOREG using actual data*. The last two procedures are described in SAS76 (Barr *et al.*, 1976). The observations were temperature measurements from the *5th* level of the WJBF TV tower taken on April 5, 1976 at 1400H, EST (Eastern Standard Time).

Wind speed measurements from the SRL TV tower taken at the same time and level as the temperature measurements were used in the example in Appendix 1. Here, the rejection criterion used was 4.5 and 5 outliers were detected. The measure of model adequacy was $G = 0.91$, which is very high. Also, the computed chi-square was 64.56. This can be compared with the tabulated chi-square values for 46 degrees of freedom which are 62.8 and 71.2 at 5% and 1% levels of significance, respectively. Thus, it can be concluded at the 1% level of significance that there is no evidence to reject the hypothesis that the model provides an adequate representation of the behavior of the observed time series.

4.2 *Outlier Rejection Rates of the SRL Micrometerological Data*

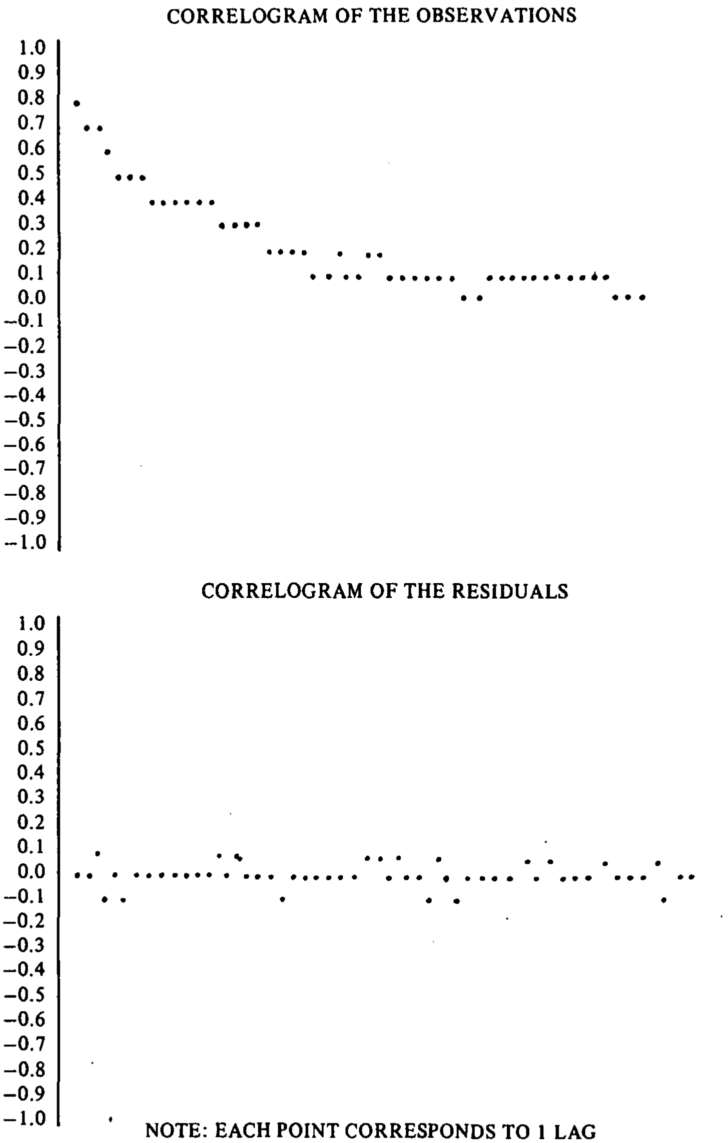
Two days of temperature and wind speed measurements from the TV tower and about three days of wind speed, azimuth and elevation angle measurements from the 7 SRL towers were screened for outliers. The total number of observations screened was 1,496,800. This is equivalent to 2,079 data-hours. Each data-hour consists of 720 observations since the sampling interval was 5 seconds.

Presented in Table 1 are the percentage distributions of outliers in the hourly measurements of temperature and wind speed from the TV tower based on 336 data-hour each. The percentage of hourly

*The data for this study were provided by the Savannah River Laboratory (abbreviated SRL) of E.I. duPont de Nemours and Company at Aiken, South Carolina, U.S.A. Measurements on temperature, wind speed, azimuth and elevation angles were obtained from 7 meteorological towers within the Savannah River Plant Site and from 7 levels of the instrumented WJBF television tower located about 21 kilometers away. Hereafter, the 7 towers will be simply referred to as the SRL towers and WJBF television tower as the T.V. tower.

```
                                   Y6
                    NO. OF OBSERVATIONS              =  720
                    NO. OF ITERATIONS                =    1
                    NO. OF OUTLIERS                  =    0

        COEFFICIENTS OF             STANDARD         T-VALUES
         THE TREND LINE              ERRORS
          16.24188                   0.04560         356.20807
           0.00504                   0.00011          46.77785

        COEFFICIENTS OF             STANDARD         T-VALUES
       THE AUTOREG MODEL             ERRORS
           0.54432                   0.03554          15.31710
           0.30756                   0.03552           8.65809

       CHARACTERISTIC ROOTS
           0.88992
          -0.34560

         MEAN SQUARE ERROR     DEGREES OF FREEDOM
           0.00784                   714

                 MEAN          APPROXIMATE STD ERROR
         M1 = 18.06344               0.02232

         STANDARD DEVIATION   APPROXIMATE STD ERROR
         SD1  =  0.15049             0.01075
         SD2  =  0.15059

M1  IS SIMPLE MEAN COMPUTED FROM THE DATA AFTER REPLACING
    THE OUTLIERS BY THEIR PREDICTED VALUES.
SD1 IS COMPUTED FROM THE DATA AFTER REMOVING THE TREND
    AND REPLACING THE OUTLIERS BY THEIR PREDICTED VALUES.
SD2 IS COMPUTED FROM THE COEFFICIENTS OF THE FITTED SECOND-
    ORDER AUTOREGRESSIVE MODEL.

      DEGREES OF FREEDOM = NO. OF OBSERVATIONS - NO. OF
              PARAMETERS - NO. OF OUTLIERS - 2
```

Fig. 3 Results of procedure SCREEN on one hour of wind speed measurements from the TV tower



CORRELOGRAM OF THE OBSERVATIONS

CORRELOGRAM OF THE RESIDUALS

NOTE: EACH POINT CORRESPONDS TO 1 LAG

```
MEASURE OF MODEL ADEQUACY

G =  0.90318

PORTMANTEAU LACK OF FIT TEST

CHISQUARE = 69.7120  DF =  46
```

FIG. 3 (Continued)

```
                         STATISTICAL   ANALYSIS   SYSTEM
                       GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y6
SOURCE                DF     SUM OF SQUARES    MEAN SQUARE     F VALUE      PR > F    R-SQUARE          C.V.
MODEL                  3       794.95441181   264.98480394    33782.00      0.0001    0.993004        0.4903
ERROR                714         5.60059112     0.00784397                 STD DEV                   Y6 MEAN
CORRECTED TOTAL      717       800.55500293                             0.08856616               16.66344011

SOURCE                DF          TYPE I SS      F VALUE     PR > F     DF    TYPE IV SS   F VALUE    PR > F
T                      1       784.38790756     99998.90     0.0001      1    0.30982993     39.50    0.0001
YLAG 1                 1         9.97850182      1272.12     0.0001      1    1.84030093    234.61    0.0001
YLAG 2                 1         0.58800244        74.96     0.0001      1    0.58800244     74.96    0.0001

                                    T FOR HO:        PR > |T|     STD ERROR OF
PARAMETER            ESTIMATE     PARAMETER = 0                   ESTIMATE
INTERCEPT          2.41155179              4.37        0.0001     0.37855109
T                  0.00074596              6.28        0.0001     0.0001869
YLAG 1             0.54432019             15.32        0.0001     0.03553676
YLAG 2             0.30756200              8.66        0.0001     0.03552311
```

$$\hat{\mu}^* = \frac{\hat{\mu}(1-\hat{\beta}_1-\hat{\beta}_2)-\hat{\alpha}(\hat{\beta}_1 + 2\hat{\beta}_2)}{(1-\hat{\beta}_1-\hat{\beta}_2)^2}$$

$$= \frac{2.41155179(1-0.54432019-0.30756200) -0.00074596(0.54432019 + 2 \times 0.30756200)}{(1-0.54432019-0.30756200)^2}$$

$$= 16.241885$$

$$\hat{\alpha}^* = \frac{\hat{\alpha}^*}{1-\hat{\beta}_1-\hat{\beta}_2}$$

$$= \frac{0.00074596}{1-0.54432019-0.30756200}$$

$$= 0.005036$$

Fig. 4 Checks on the values of μ^* and α^* using procedure GLM.

```
                          S T A T I S T I C A L   A N A L Y S I S   S Y S T E M
                               GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE:
SOURCE               DF     SUM OF SQUARES      MEAN SQUARE    F VALUE      PR>F    R-SQUARE         C. V.
MODEL                 1       17. 16395245      17.16395245    2194.30    0.0001    0.763977        3.2986
ERROR               716         6.60059112       0.00782205                STD DEV                   W MEAN
CORRECTED TOTAL     717        22.76454357                              0.08844236               2.68121487
SOURCE               DF       TYPE I SS    F VALUE     PR>F       DF      TYPE IV SS     F VALUE      PR>F
X                     1     17.16395245    2194.30   0.0001        1     17.16395245     2194.30    0.0001
PARAMETER          ESTIMATE      T FOR HO:    PR>|T|         STD ERROR OF
                               PARAMETER = 0                    ESTIMATE

INTERCEPT         2.40571266       356.71     0.0001           0.00674422
X                 0.00503623        46.84     0.0001           0.00010751
```

$$W = Y_t\text{-}0.54432019Y_{t\text{-}1}\text{-}0.30756200Y_{t\text{-}2}$$

$$X = t - 0.54432019(t\text{-}1) - 0.30756200\,(t\text{-}2)$$

$$\hat{\mu}^* = \frac{a}{1 - \hat{\beta}_1 - \hat{\beta}_2} = \frac{2.40571266}{1\text{-}0.54432019\text{-}0.30756200} = 16.241885$$

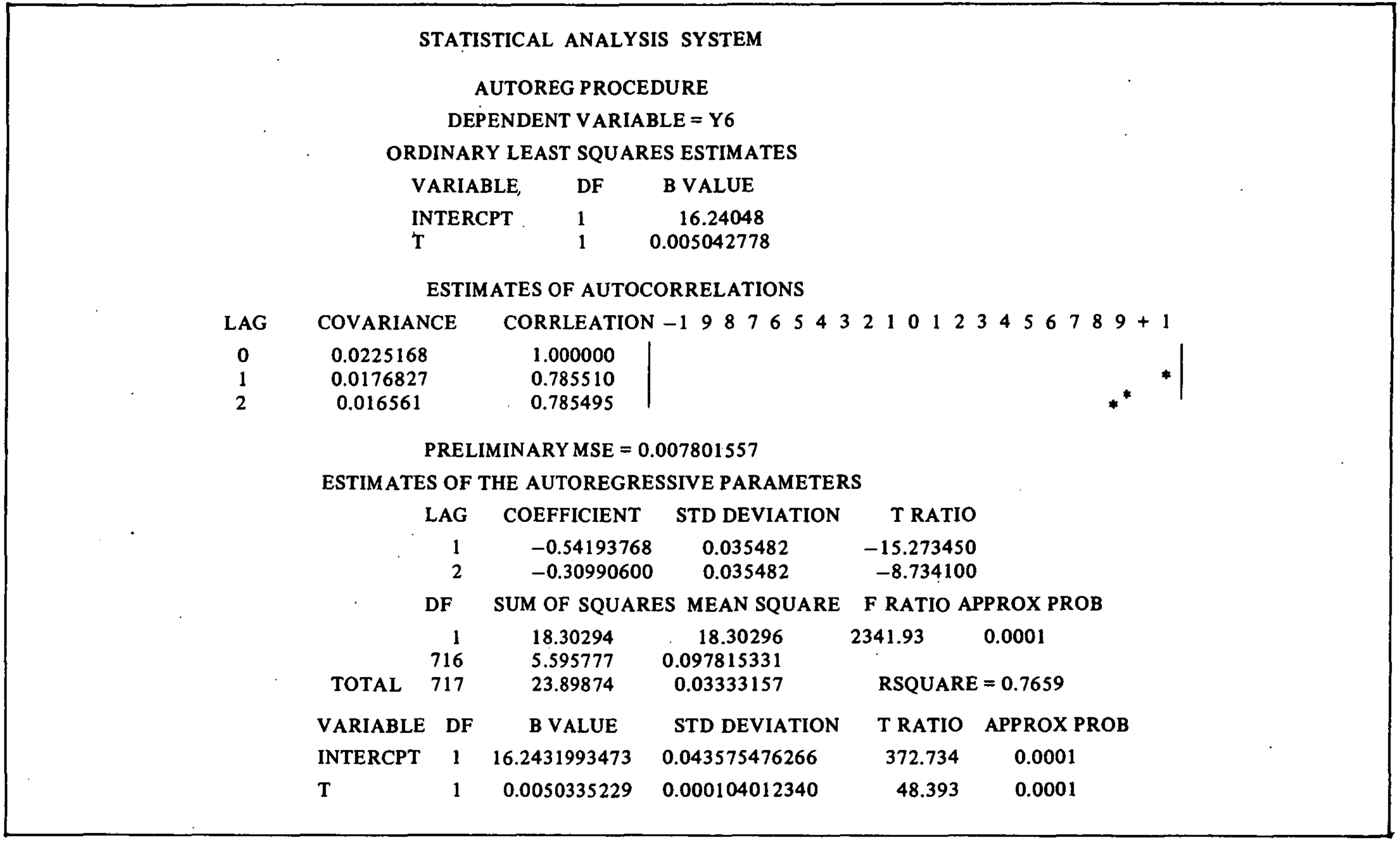
$$\hat{a}^* = b = 0.005036$$

$$SE\,(\hat{\mu}^*) = \frac{SE\,(a)}{1\text{-}\hat{\beta}_1\text{-}\hat{\beta}_2}\,.\,\text{cf} = \frac{0.00674422}{1\text{-}0.54432019\text{-}0.30756200}\sqrt{\frac{716}{714}} = 0.045597$$

$$SE\,(\hat{\alpha}^*) = SE\,(b)\,.\,cf = 0.00010751\sqrt{\frac{716}{714}} = 0.000107$$

Note: The correction factor denoted by *cf* is to adjust the estimated mean square error for 2 degrees of freedom in estimating β_1 and β_2.

Fig. 5 Checks on the values of *SE* (μ^*) and *SE* (α^*) using procedure GLM.



```
                    STATISTICAL  ANALYSIS  SYSTEM

                         AUTOREG PROCEDURE
                      DEPENDENT VARIABLE = Y6
                  ORDINARY LEAST SQUARES ESTIMATES
                   VARIABLE,      DF      B VALUE
                   INTERCPT       1        16.24048
                   T              1      0.005042778

                    ESTIMATES OF AUTOCORRELATIONS
LAG     COVARIANCE     CORRLEATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 + 1
 0       0.0225168       1.000000  |                                          |
 1       0.0176827       0.785510  |                                        * |
 2       0.016561        0.785495  |                                      **

                    PRELIMINARY MSE = 0.007801557
           ESTIMATES OF THE AUTOREGRESSIVE PARAMETERS
                 LAG    COEFFICIENT    STD DEVIATION      T RATIO
                  1     -0.54193768     0.035482        -15.273450
                  2     -0.30990600     0.035482         -8.734100
                 DF    SUM OF SQUARES  MEAN SQUARE   F RATIO APPROX PROB
                  1     18.30294        18.30296     2341.93     0.0001
                716     5.595777      0.097815331
         TOTAL  717     23.89874       0.03333157        RSQUARE = 0.7659
         VARIABLE   DF    B VALUE       STD DEVIATION     T RATIO   APPROX PROB
         INTERCPT    1  16.2431993473  0.043575476266     372.734     0.0001
         T           1   0.0050335229  0.000104012340      48.393     0.0001
```

Fig. 6 Checks on the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ using procedure AUTOREG.

TABLE 1

PERCENTAGE DISTRIBUTIONS OF OUTLIERS IN HOURLY MEASUREMENTS OF TEMPERATURE AND WIND SPEED FROM THE TV TOWER.

Number of Outliers	*Temperature*	*Wind Speed*
0	76.8	63.1
1	7.7	18.4
2	3.6	6.3
3	1.2	3.3
4	1.5	2.1
5	0	0.9
6–10	4.2	4.5
11–30	2.7	0.9
31–50	0.9	0
>50	1.5	0.6

TABLE 2

PERCENTAGE DISTRIBUTIONS OF OUTLIERS IN HOURLY MEASUREMENTS OF WIND SPEED, AZIMUTH AND ELEVATION ANGLES FROM THE 7 SRL TOWERS.

Number of Outliers	*Wind Speed*	*Azimuth Angle*	*Elevation Angle*
0	81.7	53.3	51.6
1	9.0	12.2	23.2
2	1.7	6.8	9.4
3	2.6	4.3	6.8
4	0.2	2.1	2.1
5	0.2	0.6	1.3
6–10	0.4	4.9	1.9
11–30	1.3	5.8	1.1
31--50	0.4	3.2	0.2
> 50	2.6	6.8	2.4

measurements with at least one outlier was 37% for wind speed compared to only 23% for temperature. The percentage of wind speed data with 1 to 5 outliers rejected was about twice that for temperature. Data with rejection rates of more than 7% (containing at least 50 outliers) were less than 1% for wind speed and less than 2% for temperature.

The percentage distribution of outliers in the hourly measurements of wind speed, azimuth and elevation angles from the 7 SRL towers are presented in Table 2. The number of data-hours for each distribution is 469. Almost 50% of the wind azimuth and elevation angle hourly data have at least one outlier compared to only about 18% for wind speed. The percentage of elevation angle with low rejection rates (between 1 to 5 outliers rejected) was higher than for the wind azimuth angle. However, the wind azimuth angle have higher percentage with more than 50 outliers rejected.

5. Conclusions

The regression method for data quality control proposed in this study had been found satisfactory for screening the SRL tower measurements. When really "bad" observations were few and scattered, they were detected most of the time and imputed by some reasonable values based on the fitted second-order autoregressive model. However, in situations where there was a long run of bad observations, or the wind azimuth angle shifted level by more than 90°, the method was only useful for indicating that there was something wrong with the data. The indications were the unusually high outlier rejection rates (more than 50%) and obviously wrong estimates of the model parameters and other statistics. The hourly SRL measurements sampled every 5 second used in this study have mostly very low outlier rejection rates (1 to 5 outliers rejected per 720 observations).

LIST OF REFERENCES

Abraham, B. and G. E. P. Box. 1975. Outliers in time series, Tech. Report No. 440, Dept. of Statistics, Univ. of Wisconsin, Madison.

————. 1976. Bayesian analysis of some outlier problems in time series. Dept. of Statistics. Univ. of Wisconsin, Madison.

Alexeiev, J. K., P. C. Dalrymple, and H. Gerger. 1974. Instrument and observing problems in cold climates. Tech Note No. 135. World Meteor. Org.

Andrews, D. F., 1974. A robust method for multiple linear regression. Technometrics, Vol. 16, No. 4, pp. 523-531.

Angell, J. and A. Bernstein. 1976. Evidence for a reduction in wind speed on the upwind side of a tower, J. Appl. Meteor., Vol. 15, No. 2, pp. 186-188.

Anscombe, F. J. 1960. Rejection of outliers, Technometrics, Vol. 2, No. 2, pp. 123-147.

Barr, A. J., J. H. Goodnight, J. P. Sall, and J. T. Helwig. 1976 *A User's Guide to SAS 76.* Sparks Press, North Carolina.

Barten, A., 1962. Note on unbiased estimation of the squared mutliple correlation coefficient. Statistics Neerlandia, Vol. 16, No. 2, pp. 151-163.

Bloomfield, P. 1976. *Fourier Analysis of Time Series: An Introduction.* John Wiley & Sons, New York.

Box, G. E. P. and G. M. Jenkins. 1976. *Time Series Analysis, Forecasting and Control,* 2nd Ed., Holden-Day, San Francisco.

Box, G. E. P. and D. A. Pierce. 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J. Amer. Statist. Assoc., Vol. 65, No. 332, pp. 1509-1526.

Burford, R. L. 1970. *Basic Statistics for Business and Economics.* Charles S. Merrill, Ohio.

Briggs, G. A. 1969. *Plume Rise.* U. S. Atomic Energy Commission, Div. Tech. Info., TID 25075.

–––– . 1975. Plume rise predictions, lectures on air pollution and environmental impact analysis. Amer. Meteor. Soc., pp. 59-104.

Chatfield, C. 1975. *The Analysis of Time Series: Theory and Practice.* Chapman and Hall, London.

Chatfield, C. and D. L. Prothero. 1973. Box-Jenkins seasonal forecasting: problems in a case-study. J. Roy. Statis. Soc, A., Vol. 136, pp. 295-336.

Collette, D. and T. Lewis. 1976. The subjective nature of outlier rejection procedures. Appl. Statist., Vol. 25, No. 3, pp. 228-237.

Cooley, J. W. and J. W. Tuckey: 1965. An algorithm for the machine calculation of complex Fourier Series. Math. Comput., Vol. 19, pp. 297-301.

Cooper, R. and B. Rusch. 1968. The SRL meteorological program and off-site dose calculations. E. I. du Pont de Nemours & Co., South Carolina.

Draper, N. R. and H. Smith. 1966. *Applied Regression Analysis.* John Wiley *& Sons, New York.

Durbin, J. 1960. Estimation of parameters in time series regression models. J. Roy Statist. Soc., B, Vol. 22, pp. 139-153.

–––– . 1967. Tests of serieal independence based on the cumulative periodogram. Bull. Int. Statist. Inst., Vol. 42, pp. 1039-1049.

Erickson, E. 1975. On the representation of frequency spectra in meteorology. Boundary-layer Meteor., Vol. 8, No. 2, pp. 221-226.

Fox, A. J. 1972. Outliers in time series. J. Roy. Statist. Soc., B, Vol. 34, No. 3, pp. 350-363.

Fuller, W. A. 1976. *Introduction to Statistical Time Series.* John Wiley & Sons, New York.

Granger, C. W. J. 1966. The typical spectral shape of an economic variable. Econometrica, Vol. 34, No. 1 pp. 150-161.

Hay, J. S. and F. Pasquill. 1959. Diffusion from a continuous source in relation to the spectrum and scale of turbulence. *Atmospheric Diffusion and Air Pollution,* ed. F. Frenkiel and P. Sheppard, *Advances in Geophysics,* Vol. 6, pp. 345-365, Academic Press, New York.

Hinich, M. and P. Talwar. 1975. A simple method for robust regression. J. Amer. Statist. Assoc., Vol. 70, No. 349, pp. 113-119.

Jenkins, G. M. and D. G. Watts. 1968. *Spectral Analysis and Its Application.* Holden-Day, San Francisco.

Jones, R. H. 1965. A Re-appraisal of the Periodogram Analysis. Technometrics, Vol. 7, No. 4, pp. 531-542.

Kendall, M. G. and A. Stuart. 1963. The Advanced Theory of Statistics, Vol. I, London, Griffin.

Kenny, J. F. and E. S. Keeping. 1951. *Mathematics of Statistics, Part II.* D. Van Nostrand, New York.

Koopmans, L. H. 1974. *The Spectral Analysis of Time Series,* Academic Press, New York.

Lumley, J. L. and H. A. Panofsky. 1964. The Structure of Atmospheric Turbulence. Interscience Publishers, New York.

Mood, A. M. F. A. Graybill, and D. C. Boes. 1974. *Introduction to the Theory of Statistics.* McGraw-Hill, New York.

Naus, J. 1975. *Data Quality Control and Editing.* Marcel Dekker, New York.

Ogura, Y. 1957. The influence of finite observation intervals on the measurement of turbulent diffusion parameters. J. Meteor., Vol. 14, No. 2, pp. 176-181.

-----. 1959. Diffusion from a continuous source in relation to a finite observation interval. *Atmospheric Diffusion and Air Pollution,* ed. F. Frenkiel and P. Sheppard, *Advances in Geophysics,* Vol. 6, pp. 149-159. Academic Press, New York.

Panofsky, H. A. 1962. Scale analysis of atmospheric turbulence at 2 meters. Quart. J. Roy. Meteor. Soc., Vol. 88, No. 375, pp. 57-69

Parzen, E. 1972. Some recent advances in time series analysis. Statistical Models and Turbulence, ed. M. Rosenblatt and C. Van Atta. Springer-Verlag, New York.

Pasquill, F. 1974. Atmospheric Diffusion. 2nd Ed., John Wiley & Sons, New York.

---- . 1975. Some topics relating to modelling of dispersion in boundary layer. U. S. Environmental Protection Agency, North Carolina.

Pendergast, M. M. 1975. A cautionary note concerning aerodynamic flying of bivane wind direction indicators. J. Appl. Meteor., Vol. 14, No. 4, pp. 626-627.

Pendergast, M. M. and T. Crawford. 1974. Actual standard deviations of vertical and horizontal wind direction compared to estimates from other measurements. Symp. Atmospheric Diffusion and Air Pollution. Santa Barbara, Amer. Meteor. Soc., 1-6.

Porch, W. and M. Dickerson. 1976. Statistical comparisons of Savannah River anemometer data applied to quality control of instrument networks. National Tech. Inf. Services.

Renig, W. C. 1963. The 1951 preoperational environmental survey for the Savannah River Plant – in retrospect. Health Phys., Vol. 9, pp. 83-85. Pergamon Press, New York.

Roughan, J. L. and H. H. Evans, 1970. Editing time series. Austral. J. Statist., Vol. 12, No. 3, pp. 141-149.

Sethuraman, S. and R. Borwn. 1976. Validity of the log-linear relationship over a rough terrain during stable conditions, Boundary-layer Meteor., Vol. 10, No. 4, pp. 489-501.

Sethuraman, S. and J. Tichler. 1977. Statistical hypothesis tests of some micrometeorological observations. J. Appl. Meteor., Vol. 16, No. 5, pp. 445-461.

Slade, D. H. 1968. *Meteorology and Atomic Energy.* U. S. Atomic Energy Commission. Div. Tech. Infl, TID 24190.

Smith, F. B. 1962. The effect of sampling and averaging on the spectrum of turbulence. Quart. J. Roy. Meteor. Soc., Vol. 88, No. 376, pp. 177-180.

Stralkowksi, C. M., S. M. Wu, and R. E. DeVor. 1970. Charts for the interpretation and estimation of the second-order autoregressive model. Technometrics, Vol. 12, No. 3, pp. 669-685.

Sutton, O. G. 1932. A theory of eddy diffusion in the atmosphere. Proc. Roy. Soc., A, Vol. 135, pp. 143-165.

Taylor, G. I. 1921. Diffusion by continuous movement. Proc. London Math. Soc., A, Vol. 20, pp. 196-211.

Tennekes, H. and Lumley, J. L. 1972. *A First Course in Turbulence.* The MIT Press, Cambridge.

Thiel, H. 1961. *Economic Forecasts and Policy.* North-Holland, Amsterdam.

Tillman, J. E. 1972. The indirect determination of stability, heat and momentum fluxes in the atmospheric boundary layer from simple scalar variables during dry unstable conditions. J. Appl. Meteor., Vol. II, No. 5, pp. 783-792.

Türner, D. B. 1970. *Workbook of Atmospheric Dispersion Estimates.* U. S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Weber, A. H., J. S. Irwin, J. P. Kahlen, and W. B. Petersen. 1975. Atmospheric turbulence properties in the lowest 300 meters. U. S. Environmental Protection Agency. Research Triangle Park, North Carolina.

Wieringa, J. 1976. An objective exposure correction method for average wind speed measured at a sheltered location. Quart. J. Roy. Meteor. Soc., Vol. 102, No. 431, pp. 241-253.