United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

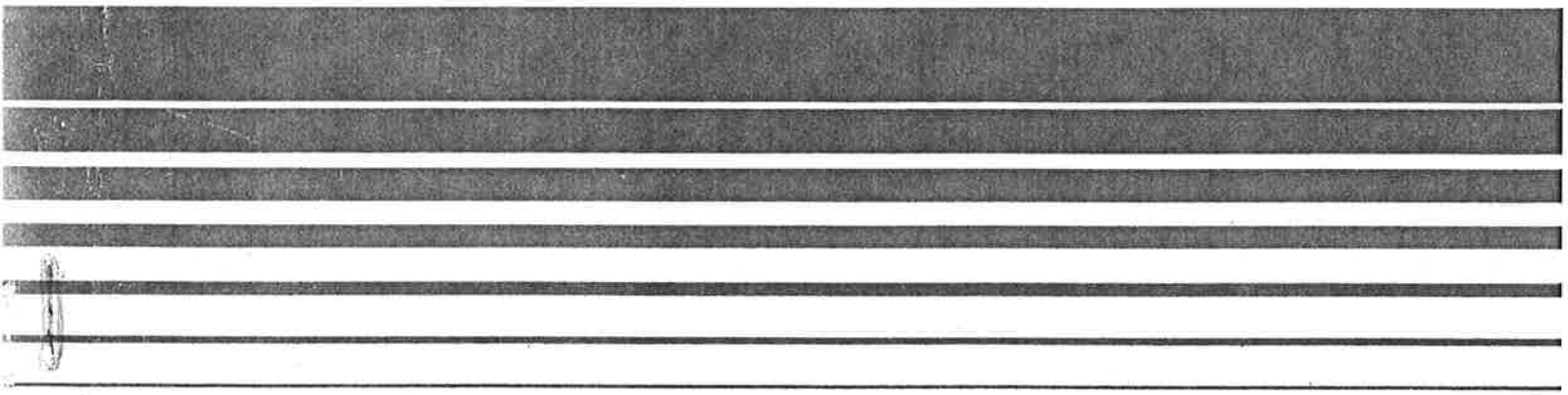
EPA-450/2-78-037
OAQPS No. 1.2-092
July 1978

Air

EPA

Guideline Series

Screening Procedures for Ambient Air Quality Data



EPA-450/2-78-037
OAQPS No. 1.2-092

Screening Procedures for Ambient Air Quality Data

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air, Noise, and Radiation
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

July 1978

OAQPS GUIDELINE SERIES

The guideline series of reports is being issued by the Office of Air Quality Planning and Standards (OAQPS) to provide information to state and local air pollution control agencies; for example, to provide guidance on the acquisition and processing of air quality data and on the planning and analysis requisite for the maintenance of air quality. Reports published in this series will be available - as supplies permit - from the Library Services Office (MD-35), U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711; or, for a nominal fee, from the National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.

Publication No. EPA-450/2-78-037
(OAQPS No. 1.2-092)

TABLE OF CONTENTS

FIGURES AND TABLES

1. INTRODUCTION

This guideline discusses screening procedures to identify possible outliers in ambient air quality data sets. The Standing Air Monitoring Work Group (SAMWG) has emphasized the need for ensuring data quality as an integral part of an air monitoring program.[1] The purpose of this document is to present data screening techniques to be applied to ambient air quality data by the Regions (or States) before the data are entered into SAROAD. Although the primary emphasis is on computerized techniques, the summary briefly discusses which procedures are feasible to implement manually. These screening techniques have proven to be effective in identifying "atypical" concentrations which often are found to have been miscoded or otherwise invalid. The meaning of the word "atypical" will become more apparent in the actual discussions of these procedures, but on an intuitive level it describes an event with very low probability and therefore, one that is unlikely to occur.

The purpose of these screening procedures is to identify specific data values that warrant further investigation. The fact that a particular data value is flagged by these tests does not necessarily mean that the value is incorrect. Therefore, such values should not be deleted from the data set until they have been checked and found to actually be erroneous.

The screening procedures discussed in this guideline are primarily intended to examine the internal consistency of a particular data set. For this reason, they are not designed to detect subtle errors that may result from incorrect calibration or a variety of other factors that can result in incorrect values that superficially appear consistent. That is perhaps, the easiest place to contrast these screening procedures with an overall quality assurance program. A quality assurance program usually examines all phases of the monitoring effort from data collection to the data set that is finally produced. Such an effort is much more comprehensive than the techniques presented here and is discussed in more detail elsewhere.[2] Thus, the

techniques presented here may be considered as one part of the overall quality assurance program. However, they have been shown to be a cost-effective means of eliminating the more obvious errors and thereby improving data quality.

In selecting screening procedures for this guideline, emphasis has been given to those techniques that have actually been used to examine air quality data sets.[3-7] Although some other approaches are briefly discussed, the intended purpose of this document is to present techniques that have been used successfully rather than to merely propose possible approaches that may some day prove useful.

This document is organized so that this introduction is followed by a brief discussion of the background of the problem and then a section presenting the screening procedures followed by a conclusion and a series of appendices. In addition to a summary of the recommendations, the conclusion contrasts the initial step of identifying a possible outlier with the final step of actually deleting the value and also discusses the proper place for these tests in the overall data handling scheme. The appendices consist of articles discussing the application of these tests to air quality data and computer programs to perform the tests. This structure was chosen so that the screening procedures could be presented in various levels of detail. The discussion in the main body of the document is intended to give a general overview and an intuitive understanding of what each test is designed to do. The appendices provide more detail and would be of interest to those concerned with the actual implementation of these screening procedures. Those readers interested in more details on the underlying statistical theory will find the appropriate articles included in the references.

2. BACKGROUND

It is a truism to say that data quality is important. Virtually no one will argue that data quality is not important, but the key question is "how important?" Obviously, the degree of data quality required depends upon the intended use of the data. This is why air pollution data sets present some interesting practical problems.

One use of air quality data is to assess compliance with legal standards such as the National Ambient Air Quality Standards (NAAQS).[8] The form of these standards frequently re-enforces the need for data quality. For example, the NAAQS for total suspended particulate, sulfur dioxide, carbon monoxide, and oxidant all specify upper limit concentrations that are not to be exceeded more than once per year. In such cases, it is the second highest value for the year that becomes the decision-making value. With this application in mind, the need for data quality is obvious.

Another factor that must be considered in air monitoring programs is the volume of data involved. Continuous instruments can produce as many as 8760 hourly observations for the year. Intermittent monitoring schedules for 24-hour data routinely produce 60 or so values per year. When these numbers are accumulated for several pollutants for an entire network, State, or for the Nation, the total number of data values quickly becomes cumbersome. For example, it is estimated that EPA's National Aerometric Data Bank is currently expanding at the rate of 20 million values per year. Therefore, maintaining a data bank for air pollution measurements involves the basic conflict of having to routinely process large volumes of data and yet at the same time ensure an almost zero defect level of data quality. Because of the nature of the standards, many users may only be interested in the two highest values at each site for each pollutant. It should be noted that two values from a data set of 8760 observations constitutes 0.023 percent of the data. This means that the user's perception of data quality may be entirely different from the

true data quality. For example, if only 0.05 percent of the data points were too high due to errors, this would still be sufficient to have the user complain that, "the data are useless." On the other hand, if elaborate editing checks are introduced, the sheer volume of data may result in high costs or processing delays, and the user may now complain that the data are not sufficiently current for him to make timely decisions.

With this background in mind, it is apparent that an ideal air quality data screening system must be able to process large volumes of data in an inexpensive fashion while flagging virtually every error. Also, because it is frequently difficult and time consuming to verify suspect data points, every flagged value should be a genuine error. Unfortunately, while these characteristics are obviously desirable, they are also almost impossible to attain. However, the techniques presented here represent a useful first step to identify and eliminate the more glaring errors in air quality data sets.

3. SCREENING PROCEDURES

As discussed in the previous section, the choice of an appropriate screening procedure depends not only upon the desired data quality, but also the volume of data to be processed and the amount of resources available for screening the data. For example, a very effective means of identifying outliers is to have an experienced air pollution analyst visually inspect graphs of the data. While this may initially appear reasonable, it is seldom practical because the volume of data quickly becomes too large and other demands are frequently made on the analyst's time. The primary purpose of these screening procedures is to make efficient use of the analyst's time by identifying suspect values that should warrant closer examination.

For the purposes of this discussion, tests for 24-hour data and hourly data are presented separately. This distinction is made because of the difference in the volume of data to be processed. The 24-hour data are commonly obtained by every-sixth-day monitoring, while hourly data are obtained from continuous instruments that may produce over 700 values per month. Therefore, there can be as much as a hundred-fold difference in the volume of data. This difference affects the types of tests that can be efficient to screen the data.

The tests are discussed in terms of screening one month's worth of data. The choice of one month is somewhat arbitrary and could be varied. However, there were several factors that made this choice seem reasonable. To process less than a month's worth of data would result in very small data sets for every-sixth-day sampling schedules and yet more than one month would result in very large data sets for hourly data. Another consideration was to maintain a fairly short time interval between data collection and flagging suspect values. This was done with the idea that the sooner a suspect value is identified, the easier it would be to verify the measurement. The use of months rather than quarters also lessens the effect of seasonality for tests that involve comparisons with data from previous time periods.

One point worth noting in the discussion of these statistical tests concerns the validity of the underlying assumptions. As a general rule, these types of tests assume that the observations are independent. To some extent, this may be approximately correct in the case of every-sixth-day sampling, but obviously there are seasonal and diurnal patterns associated with air quality levels that make this assumption questionable in general. This problem could be approached by the use of time series models to minimize the auto-correlation (interdependence of successive values), but from a practical viewpoint, the tests discussed here have been shown to work reasonably well. In a sense, the viewpoint taken here is to use the simplest test that has been successfully demonstrated and have that fact substantiate the claim that the underlying assumptions are "approximately satisfied."

3.1 Twenty-Four Hour Data Tests

There are several statistical tests that may be used to screen 24-hour air quality data sets. Tests attributed to Dixon,[9] Grubbs,[10] and Shewhart[11] have been considered for identifying suspect air quality values.[2-4, 7] Conceptually, all these tests yield a probability statement that provides a measure of the internal consistency of the data set. The Dixon and Shewhart test procedures have been applied to air quality data sets.[7]

The Dixon test may be conveniently used to examine one month's worth of 24-hour data. Basically, this test is used to examine the relative spread within the data set and is quite easy to compute. For example, if there were five values in the month, it is only necessary to rank the data from smallest to largest. Then the difference between the highest and second highest values is divided by the difference between the highest and lowest values. This ratio gives a fraction ranging from zero to one. A graphical presentation of this test is given in Figure 1 for two data sets that have four points in common, but the second data set contains a value of 420 $\mu g/m^3$ instead of the 42 $\mu g/m^3$ in the first data set, i.e., a possible tran-

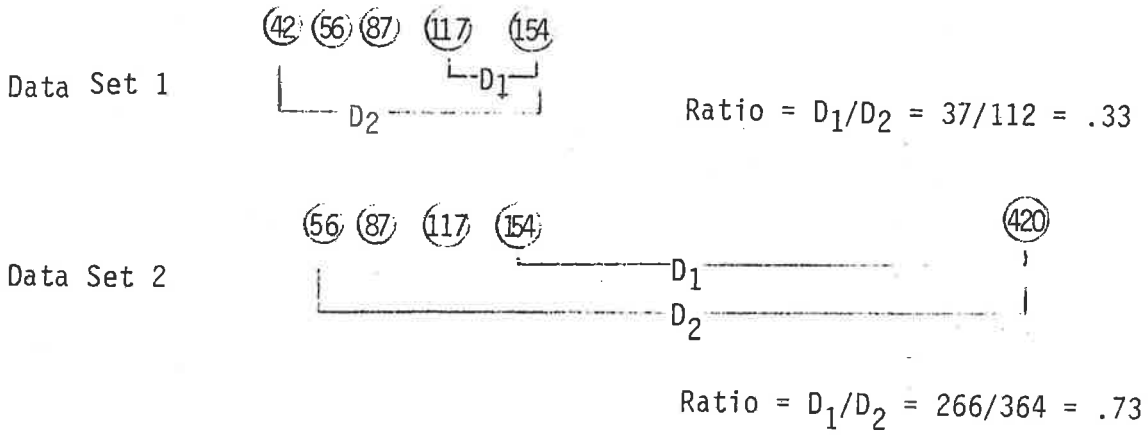


Figure 1. Illustration of Dixon Ratio Test for two TSP monthly data sets.

scription error. The computed ratio in the first cases is .33 which is acceptable while the second ratio is .73 which would be flagged at the 5 percent level as a possible outlier. The closer this ratio is to one, the more likely it is that the high value is an outlier rather than a correct value. Tables are available to determine the probability associated with this computed ratio.[3,9] The Grubbs test is conceptually similar although the ratio used is the difference between the highest value and the mean divided by the standard deviation. This requires slightly more computation, but again tabulated values for the associated probabilities are available.[2,10]

One characteristic of these types of tests is of particular interest in terms of their possible use with air quality data. These tests implicitly assume that at least one value in the data set is correct. If all of the values in Figure 1 were multiplied by 10, the computed ratios would remain unchanged. The key point is that these tests merely check for internal consistency and consequently, it is possible to have a data set that is entirely wrong and yet internally consistent. Initially, it may appear perfectly reasonable to expect that at least one value in the data set will be correct. However, in evaluating these tests it became apparent that the data handling schemes involved can occasionally produce an entire month of data that is incorrectly coded and therefore, improperly scaled. With this in mind, it becomes apparent that it is not sufficient to check for internal consistency; some type of comparison must also be made to ensure that the values fall within a reasonable range.

This can be accomplished by the use of the Shewhart test.[7,12] This test compares the monthly mean and range with those from the past few months. Again, tabulated values are available to determine the associated probabilities.[12] However, the main point is that the test is basically a two-fold screening procedure. If a monthly range differs appreciably from past monthly ranges, then it suggests an outlier within the month. On the other hand, if the monthly mean differs appreciably from past monthly means, then a scaling problem is likely.

This test has been applied to air quality data using comparisons with the past three months of data. Although some differences would be expected because of the seasonality of certain air pollutants, the use of data from the previous three months seems satisfactory.[7]

The application of the Shewhart test involves the computation of upper and lower control limits for both the monthly mean and range based upon the average of the three previous monthly means and ranges. The formulas for these upper control limits (UCL) and lower control limits (LCL) are:

$$UCL_R = D_4\bar{R} \qquad LCL_R = D_3\bar{R} \quad \text{(for the range)}$$

and

$$UCL_{\bar{x}} = \bar{\bar{x}} + A_2\bar{R} \qquad LCL_{\bar{x}} = \bar{\bar{x}} - A_2\bar{R} \quad \text{(for the mean)}$$

where $\bar{R}$ is the average of the previous three monthly ranges and
$\bar{\bar{x}}$ is the average of the previous three monthly means.

The values A_2, D_3, and D_4 depend upon sample size and may be obtained from tables.[12] An abbreviated table is contained in the "Quality Assurance Handbook for Air Pollution Measurement Systems" - Volume 1 (Table H-3).

The Shewhart test can be illustrated graphically as shown in Figure 2. This actual example involved data that were submitted with all values for October miscoded and too large by a factor of ten.[7] The upper and lower control limits are indicated on the graph and the October values for the mean and the range are obviously questionable.

Although the required use of information from previous months may complicate the data processing, results to date[7] suggest that the Shewhart test is sufficiently effective to make it the method of choice for twenty-four hour data.

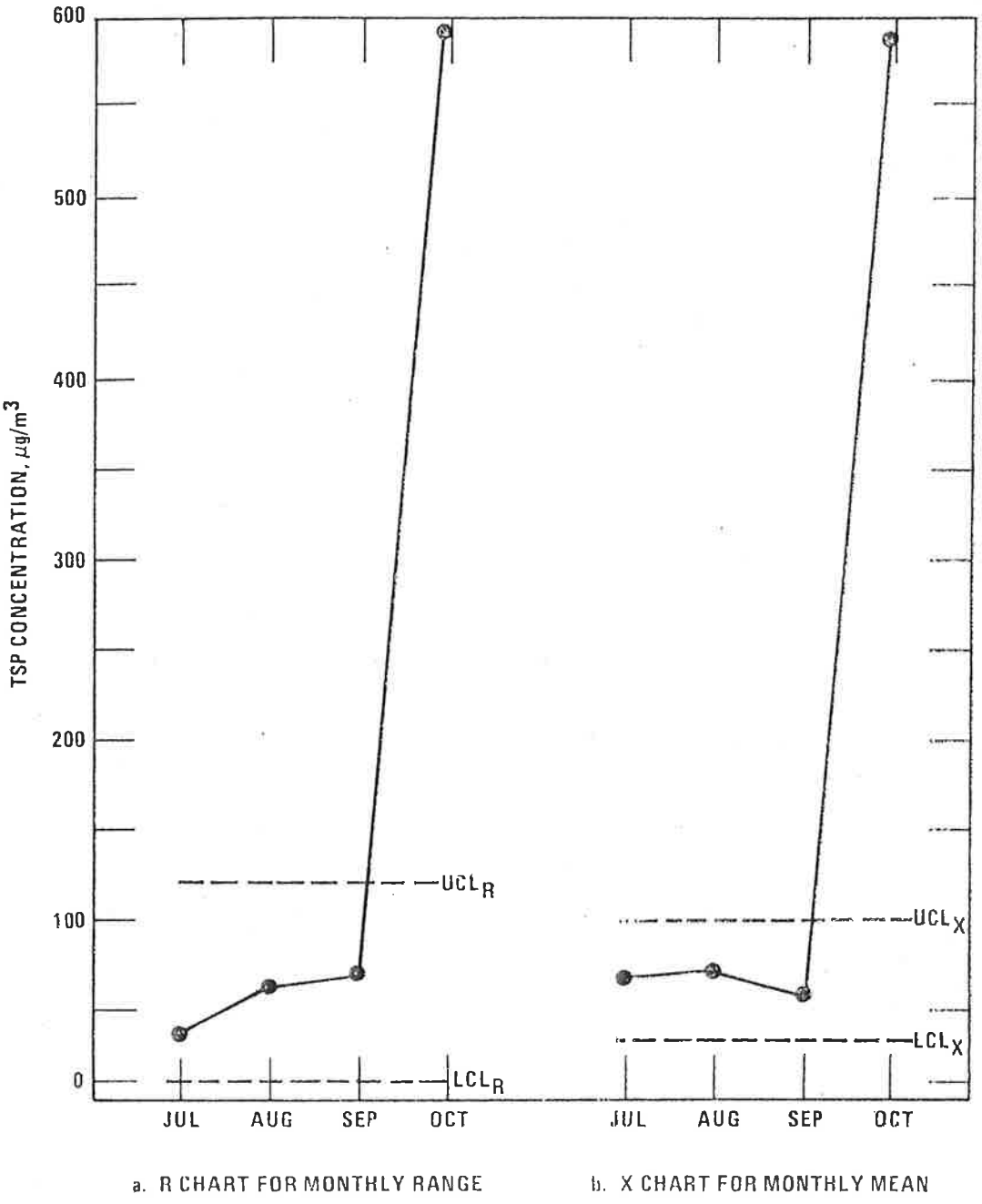


Figure 2. Example of Shewhart Control Chart Test applied to data with multiple transcription errors in month of October.

3.2 Hourly Data Tests

Hourly air pollution data sets present the most obvious conflict of desired data quality versus data volume. Just one month's worth of data may contain over 700 observations, and yet there is still a need to ensure virtually perfect data quality. The volume of data is a prime factor affecting the choice of a feasible screening procedure.

An obvious starting point is to consider the seasonal and diurnal patterns in the data. This was done and a computer program was developed to check for departures from typical patterns.[5] These typical patterns were determined from historical air quality data.

The computerized version of these tests checks four basic characteristics of the data set. The first is a maximum upper limit for an individual value. The second is an upper limit on the difference between adjacent hourly values. The third check examines spikes i.e. where the middle hour of three consecutive hourly values is much higher (or lower) than the two adjacent hours. These spikes are checked both in terms of absolute change in concentration and percent change. The fourth check specifies an upper limit for the average of four consecutive hourly values. The purpose of the first three tests is fairly obvious. The last check was introduced to identify small clusters of high values. This type of pattern was seen occasionally in sample data sets where no single value was above the maximum value limit but yet a cluster of data values appeared too high. The choice of cut-off values for these patterns tests is inherently subjective. Typical values that were selected on the basis of empirical tests on actual data sets are presented in Table 1. As indicated, the cut-off may vary for different stratifications of the data. For example, higher cut-offs are used for carbon monoxide during rush-hours compared to the rest of the day. Similarly, the values for ozone vary from season to season and from day to night. These variations reflect the seasonal and diurnal patterns of each pollutant. As discussed, the choice of these cut-offs

is subjective, but this should be viewed in terms of the purpose of these tests. The results of these tests are not sufficient grounds to eliminate data values, but only serve to identify values that require further examination. Viewed in this perspective, these cut-offs are satisfactory.

Table 1.

SELECTED QUALITY CONTROL TESTS

Typical Cut-Off Values for Patterns Test on Hourly Values
(Concentration in $\mu g/m^3$)

Pollutant	Data Stratification	Maximum Hour Test	Adjacent Hour Test	Spike Test	Consecutive 4-hr Test
Ozone					800
Total Oxidant ($\mu g/m^3$)	Summer-day	1000	300	200(300%)	500
	Summer-night	750	200	100(300%)	500
	Winter-day	500	250	200(300%)	500
	Winter-night	300	200	100(300%)	500
Carbon Monoxide (mg/m^3)	Rush traffic hours	75	25	20(500%)	40
	Non-rush traffic hours	50	25	20(500%)	40
Sulfur Dioxide ($\mu g/m^3$)	None	800*	200*	200(500%)*	1000*
Nitrogen Dioxide ($\mu g/m^3$)	None	1200	500	200(300%)	1000

NO_x

* Higher values may be used for sites near strong point sources.

Although these patterns tests are fairly simple and easy to understand, there are certain inherent deficiencies. Basically, the values are flagged on the basis of a yes-no decision with no probability value associated with the flagged value. Another problem, which is probably more serious from a practical standpoint, is the need to vary the amount of allowable departure from site to site or area to area.

With this in mind, a statistical screening test for hourly data was developed[6] that would mimic the decisions made by an experienced analyst. The reason for this was an attempt to avoid a black-box approach where the screening procedure was viewed as a mysterious oracle delivering arbitary decisions. Values that appear to be quite unlikely from a statistical viewpoint may actually be quite likely in the real world. For example, massive traffic jams do happen and may result in high carbon monoxide levels. Windstorms can mean high total suspended particulate levels. Sudden shifts in wind direction can mean that a monitor near a point source goes from a zero reading to almost full scale and back in a few hours. The high variability associated with peak hourly air pollution values makes it almost impossible to develop a screening procedure that does not occasionally flag values that are correct. But it seemed desirable to avoid the situation where an air pollution analyst would tire of repeatedly checking flagged values that turned out to be correct. Therefore, emphasis was given to a test that would flag values that an air pollution analyst would want to investigate. An effective way to accomplish this was to develop a test that would mimic experienced human judgment so that the analyst would understand why the value was flagged.

Experienced analysts frequently use the approach of looking for unusual jump discontinuities between successive hourly values or departures from expected diurnal or seasonal patterns. It became apparent that many of the outliers could be detected by a simple approach. In most cases, unusually high values could be detected by examining the frequency distribution of the hourly

data for a given period of time, such as a month, quarter, or year. Suspect values would be associated with large gaps in the frequency distribution. The length of the gap and the number of values above the gap afforded a convenient means of detecting possible errors. With this simplification of the problem, it becomes possible to develop a probabilistic framework for this problem.[6]

Figure 3 displays a histogram of actual carbon monoxide data for one month. As indicated, there is one hourly value equal to 30 mg/m^3, but no other values above 12 mg/m^3. It is relatively easy to compute the probability associated with such a gap by assuming that the data may be approximated by an exponential distribution. This type of approximation has been examined and appears to be adequate for the upper tail of the distribution, i.e., the higher concentration ranges.[13] The actual formula for the probability of this gap is quite simple[6], and as would be expected, the probability of this particular gap occurring is quite small (.0006). In fact, the value of 30 mg/m^3 was merely a keypunch error, and the correct value was 3.0 mg/m^3.

It should be noted that the gap test is designed to identify unusually high values. Errors that produced unusually low values will not necessarily be detected. A possible option is to also employ the previously discussed pattern test[5] which will flag unusually low values if they result in a departure from the typical pattern. Both tests are fairly efficient, and on EPA's UNIVAC-1110 computer the computerized versions of these tests can process 25,000 hourly values for approximately $1.00.

4. CONCLUSION

For twenty-four hour data, the Shewhart test is a convenient means of identifying possible errors. As discussed in the previous section, this test checks not only internal consistency within a month, but also consistency with adjacent months. This second check necessitates an added file of historical information, but experience suggests that this extra step is warranted. For hourly

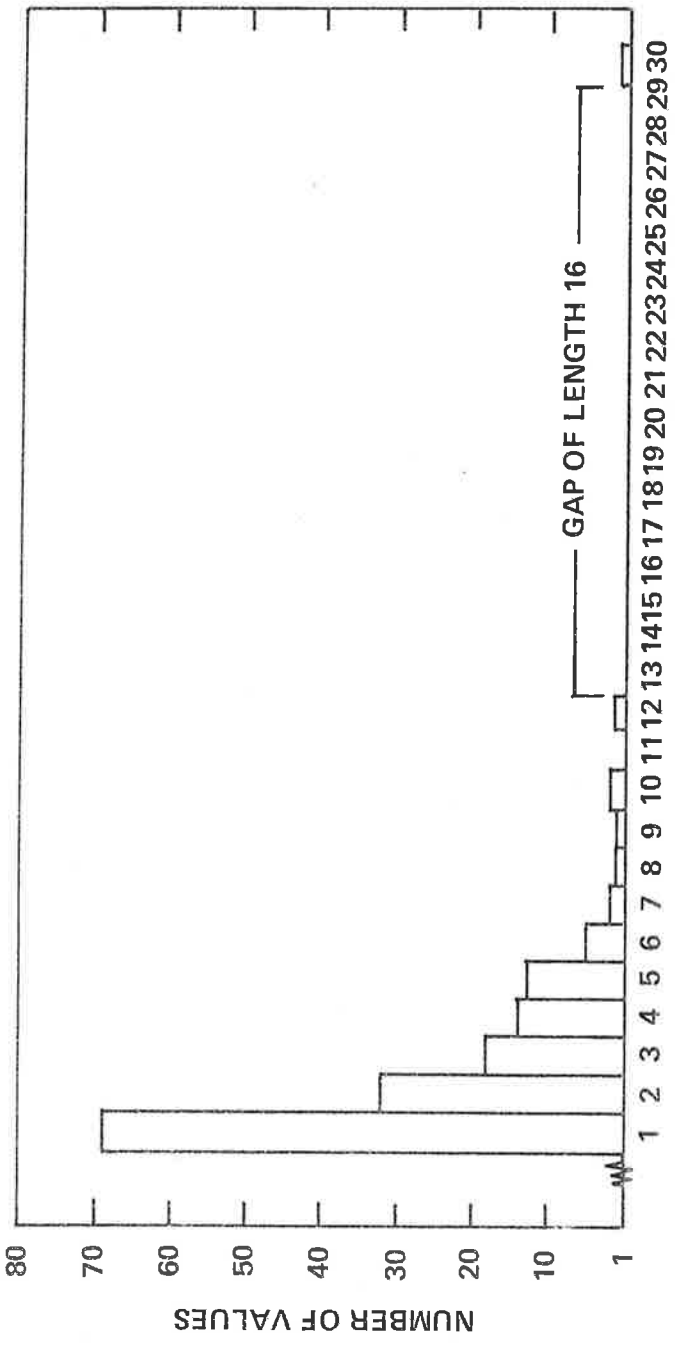


Figure 3. Illustration of Gap Test for hourly carbon monoxide data.

data, the gap test is convenient and effective for identifying potentially anomalous high values. Based upon empirical results, gaps with probabilities less than .01 should be identified. The gap test can also be supplemented by incorporating the pattern type tests into the screening process. Although the Shewhart test could also be used on monthly data sets of hourly data, the volume of data involved and the need for additional historical data files may complicate this type of implementation. Therefore, the gap test would have the advantage of being easier to implement and yet appears to work reasonably well in actual practice.

It should be noted that the use of the gap test requires the computation of a frequency distribution for the hourly data for the month in question. Although this is relatively easy to determine with a computerized system, it could be very time consuming to do manually. Therefore, if computer processing is not possible it may be more convenient to use the Shewhart test for hourly data. While manually computing the monthly mean of the hourly data would also be time consuming, the determination of the monthly range and the comparison with previously monthly ranges may be done quite easily. Thus, without any computer processing capabilities, the Shewhart test on ranges affords a convenient method for screening hourly data.

In recommending these particular tests for use in screening air quality data, it should be noted that this guideline is not intended to state that these are the only tests that may be used. These procedures are intended to represent a baseline screening program for air quality data. They focus primarily on identifying unusually high values and are not ideal for detecting unusually low values. In a sense, these may be regarded as the minimum requirement. Obviously, there are no restrictions against using procedures that are better. In some cases, State and local agencies may have well established screening programs that are efficient and have been proven to be effective. There are certain refinements that can be made in screening these type of data sets. For example, a test such as the Shewhart procedure could be

used to detect changes in the standard deviation at a site. Time series models and the use of associated data, such as meteorological variables, would be expected to increase sensitivity and possibly result in even better data quality. However, it remains to be seen if these more elaborate approaches are cost effective when processing vast quantities of data from locations throughout the Nation.

An important consideration is the proper placement of these procedures in the overall data handling scheme. As a general rule, the tests should be applied as close to the data collection step as possible. This will minimize the time lag before the potential outlier is identified and thereby make it easier to check the value in question and still ensure that the data is submitted to EPA in a timely fashion. Procedures for handling data anomalies and suspect data identified in EPA's National Aerometric Data Bank are discussed in the AEROS User's Manual.[14] However, the main thrust of a data screening program is to detect and correct any such errors before the data are submitted to EPA.

As a final comment, it should be noted that once a value is flagged as a possible anomaly, it cannot be arbitrarily dropped from the data set. It must first be verified that the data point actually is incorrect. The fact that the data point is statistically unusual does not necessarily mean that it did not occur. There are a variety of factors that should be examined to determine whether the data point should be deleted. In general, the data screening tests presented here would detect only very gross errors. For example calibration errors can produce data sets that are internally consistent and consequently would pass these tests. The data sets flagged by these tests will usually contain a few values that are much higher than the rest of the data. In many cases these will obviously be the result of a transcription or coding error. Simple, but effective, steps in examining these flagged values include comparisons of adjacent hourly values at the same site, comparisons with other pollutant or meteorological data for the site in question, and comparisons with data for the same pollutant recorded at other nearby monitoring sites for the same time period.

REFERENCES

1. Air Monitoring Strategy for State Implementation Plans. Prepared by the Standing Air Monitoring Work Group, May 1977.

2. Quality Assurance Handbook for Air Pollution Measurement Systems Volume I - Principles U.S. Environmental Protection Agency. Environmental Monitoring and Support Laboratory, Research Triangle Park, North Carolina 27711. EPA-600/ 9-76-005, January 1976.

3. Hunt, W.F., Jr., T. C. Curran, N. H. Frank, and R. B. Faoro. Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air. Transactions of the Joint European Organization for Quality Control/International Academy for Quality Conference, Venice Lido, Italy. September 1975.

4. Hunt, W. F., Jr., R. B. Faoro, and S. K. Goranson. A Comparison of the Dixon Ratio Test Applied to the National Aerometric Data Bank. Transactions of the 30th Annual ASQC Conference, Toronto, Ontario, Canada. June 1976.

5. Hunt, W. F., Jr., R. B. Faoro, T. C. Curran, and W. M. Cox. The Application of Quality Control Procedures to the Ambient Air Pollution Problem in the U.S.A. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Monitoring and Data Analysis Division, Research Triangle Park, North Carolina 27711. Presented at the 20th Annual European Organization for Quality Control Conference in Copenhagen, Denmark, June 15-18,1976.

6. Curran, T. C., W. F. Hunt, Jr., and R. B. Faoro. Quality Control for Hourly Air Pollution Data. Transactions of the 31st Annual Technical Conference of the American Society for Quality Control, Philadelphia, Pennsylvania, May 16-18, 1977.

7. Hunt, W. F., Jr. J. B. Clark, and S. K. Goranson. The Shewhart Control Chart Test - A Recommended Procedure for Screening 24-Hour Air Pollution Measurements. APCA Paper No. 77-61.2, presented at the 70th Annual Meeting of the Air Pollution Control Association, Toronto, Ontario, Canada. June 1977.

8. Title 40 - Protection of Environment. Requirements for Preparation, Adoption, and Submittal of Implementation Plans, Federal Register. 36 (158): 15490, August 14, 1971.

9. Dixon, W. J., Processing Data for Outliers, Biometrics, Vol. 9 No. 1, March 1953, pp. 74-89.

10. Grubbs, F. E. and G. Beck. Extension of Samples, Sizes and Percentage Points for Significance Tests fo Outlying Observations, Technometrics, Vol 14, No. 4, November 1972, pp. 847-854.

11. Shewhart, W. A. Economic Control of Quality of Manufactured Product. D. Van Nostrand Company, Inc., Princeton, N.J., 1931, p. 229.

12. Grant, E. L. Statistical Quality Control. McGraw-Hill Book Co., New York, 1964, p. 122-128.

13. Curran, T. C. and N. H. Frank. Assessing the Validity of the Lognormal Model When Predicting Maximum Air Pollutant Concentrations. Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston, Massachusetts, 1975.

14. AEROS Manual Series Volume II: AEROS User's Manual. U.S. Environmental Protection Agency, Office of Air and Waste Management, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina EPA-450/2-76-029 (OAQPS No. 1.2-039) December 1976.

APPENDIX A - Gap Test for Hourly Data

This appendix contains additional information on the gap test for hourly data. The following material is included:

(1) A copy of the paper, "Quality Control for Hourly Air Pollution Data," which explains the details of the test,

(2) A brief description of the computer program for this test

(3) A listing of the FORTRAN computer program

QUALITY CONTROL FOR HOURLY AIR POLLUTION DATA

Thomas C. Curran, Mathematical Statistician
William F. Hunt, Jr., Chief, Data Analysis Section
Robert B. Faoro, Mathematical Statistician

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Monitoring and Data Analysis Division
Research Triangle Park, North Carolina 27711

Presented at the 31st Annual Technical Conference of the
American Society for Quality Control
Philadelphia, Pennsylvania
May 16-18, 1977

1977 ASQC TECHNICAL CONFERENCE TRANSACTIONS—PHILADELPHIA

QUALITY CONTROL FOR HOURLY AIR POLLUTION DATA

Thomas C. Curran, Mathematical Statistician
William F. Hunt, Jr., Chief, Data Analysis Section
Robert B. Faoro, Mathematical Statistician

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Monitoring and Data Analysis Division
Research Triangle Park, North Carolina 27711

Previous papers[1-3] have discussed techniques for screening air pollution data sets with particular attention given to 24-hour measurements. The present paper focuses upon the use of screening procedures for hourly ambient air quality measurements. As with any quality control procedure, it is useful to consider the nature and intended use of the data before discussing the screening technique.

Hourly air pollution data sets present some interesting practical problems when one considers the use of a screening procedure. The most obvious feature is the volume of data. For example, 24-hour air pollution measurements are usually obtained by every-sixth-day sampling resulting in approximately 60 samples per year. In contrast, hourly measurements are obtained from continuous monitors that operate every day and, therefore, may produce as many as 8,760 values per year. Thus, hourly data sets are commonly 100 times larger than those for daily measurements. The reason that the volume of data is important becomes apparent when the use of the data is examined. For the most part, air pollution data is collected to determine status with respect to certain legal standards, such as the National Ambient Air Quality Standards.[4] These standards specify upper limits for air pollution concentrations. Of particular interest for this paper are the standards for oxidants or carbon monoxide which indicate hourly values "not to be exceeded more than once per year."[4] In these situations it is the second highest value from a data set of 8,760 observations that becomes the decision-making value. Obviously, this places a premium on ensuring data quality.

From a practical viewpoint, maintaining a data bank for air pollution measurements involves the basic conflict of having to routinely process large volumes of data and yet at the same time ensure an almost zero defect level of data quality. Many sites monitor for several pollutants so that on the national level, thousands of sites are routinely submitting tens of thousands of data points each year. However, because of the nature of the standards, many users may only be interested in the two highest values at each site for each pollutant. It should be noted that two values from a data set of 8,760 observations constitutes 0.023 percent of the data. This means that the user's perception of data quality may be entirely different from the true data quality. For example, if only 0.05 percent of the data points were too high due to errors, this would still be sufficient to have the user complain that "the data are useless." On the other hand, if elaborate editing checks are introduced, the sheer volume of data may result in high costs or processing delays, and the user may now complain that the data are not sufficiently current for him to make timely decisions.

With this background in mind, it is apparent that an air quality data screening program must be able to process large volumes of data in an inexpensive fashion while flagging virtually every error. Also, because it is frequently difficult and time consuming to verify suspect data points, every flagged value should be a genuine error. Unfortunately, while these characteristics are obviously desirable, they are also almost impossible to attain. The approach presented here is primarily intended to eliminate the more glaring errors from these hourly data sets. The major emphasis is on screening the higher concentration values to check for general internal consistency within the data set.

RATIONALE FOR SCREENING PROCEDURE

In our initial development of a screening procedure for hourly data, a computer program was developed that checked for departures from typical patterns.[3] These typical patterns were selected on the basis of experience with various types of air pollution data. Basically, the values were flagged on a yes-no decision, and there was no probability statement associated with the rejected values. One stage in this development was

to give sample data sets to experienced air pollution data analysts to see what values they would reject. There were two reasons for this step. The most obvious was to ensure that the computerized screening procedure was consistent with so-called expert judgment. However, another reason was the need for a test that would mimic the decisions made by an experienced analyst. The reason for this was an attempt to avoid a black box approach where the screening procedure was viewed as a mysterious oracle delivering arbitrary decisions. The point here is that it can be quite time consuming for the data analyst to check flagged data points. Values that appear to be quite unlikely from a statistical viewpoint may actually be quite likely in the real world. For example, massive traffic jams do happen and may result in high carbon monoxide levels. Windstorms can mean high total suspended particulate levels. Sudden shifts in wind direction can mean that a monitor near a point source goes from a zero reading to almost full scale and back in a few hours. The high variability associated with peak air pollution values makes it almost impossible to develop a screening procedure that does not occasionally flag real values. But it seemed desirable to avoid the situation where an air pollution analyst would tire of repeatedly checking flagged values that turned out to be correct. Therefore, emphasis was given to developing a test that would flag values that an air pollution analyst would want to investigate. An effective way to accomplish this was to develop a test that would mimic experienced human judgment so that the analyst would understand why the value was flagged.

To a large degree the preliminary test on patterns was successful. Experienced analysts used the same basic approach of looking for unusual jump discontinuities between successive hourly values or departures from expected diurnal or seasonal patterns. However, there were two main deficiencies in this computerized procedure based upon departures from suspected patterns. One was the lack of a probabilistic framework. The second, and probably the more serious from a practical standpoint, was the need to vary the amount of allowable departure from site to site. The probabilistic framework could be provided by a time series model, and the parameters varied from site to site. However, it became apparent during the preliminary investigation that many of the outliers could be detected by a much simpler approach. In most cases, unusually high values could be detected by examining the frequency distribution of the hourly data for a given period of time, such as a month, quarter, or year. Suspect values would be associated with large gaps in the frequency distribution. The length of the gap and the number of values above the gap afforded a convenient means of detecting possible errors. With this simplification of the problem, it becomes possible to develop a probabilistic framework for the problem as discussed below.

PROBABILITY OF A GAP

In order to compute the probability of a gap in the empirical frequency distribution, it is necessary to assume some type of underlying distribution. Although this involves an oversimplification because it ignores dependency between successive hourly values, such approaches have traditionally been used with success in air pollution data analysis.[5] The lognormal distribution has customarily been used for this purpose. However, the exponential distribution has also been found to provide a reasonable approximation for the upper tail, or higher concentrations, of hourly air pollution data.[6] Because the higher concentration values were of primary interest and the exponential distribution is mathematically convenient, it was used as the underlying distribution. As with any measurements, although the approximating distribution is continuous, the air pollution values are discrete valued. For simplicity, they may be assumed to be intergers because this involves merely a change of scale. A gap in the frequency distribution may then be described in terms of its length, the number of values above the gap, and at what concentration the gap begins. Therefore, if a monthly empirical frequency distribution of hourly values has n values greater than concentration c but no values between c, and c+k, this would be a gap of length k starting at c with n observations above the gap. To compute the probability of this event, consider the following:

Let X be an exponential random variable.

Then $Pr(X \cdot c) = 1-e^{-\lambda(c-0)}$ where $\lambda \cdot 0$, $c \cdot 0$.

Thus, $Pr(X \cdot c) = e^{-\lambda(c-0)}$.

The probability that X is greater than c+k given that X is greater than c is

$$Pr(X \cdot c+k \mid X \cdot c) = \frac{Pr(X \cdot c+k)}{Pr(X \cdot c)} = \frac{e^{-\lambda(c+k-0)}}{e^{-\lambda(c-0)}} = e^{-\lambda k}$$

Because X is distributed exponentially, this expression is independent of the concentration c.

Assuming independence, the probability that n values are greater than ck given that these n values are greater than c is

$$(e^{-\lambda k})^n = e^{-n\lambda k}.$$

Thus, the probability of a gap of length k with n values above the gap is $e^{-n\lambda k}$. This probability then becomes the criteria for rejecting suspect data.

APPLICATION

A relatively simple FORTRAN program was written to process hourly data, compute the empirical frequency distribution, and examine any gaps. Because of the manner in which the data is routinely submitted to the U.S. Environmental Protection Agency's National Aerometric Data Bank, the program was written to check the data on a monthly basis (744 hourly values). The parameter λ obviously varies from one data set to another. For simplicity, λ was determined from the 50th and 95th percentiles of the data. This was computationally convenient and also emphasized the fit for the upper tail. Results to date in evaluating this test indicate that this approach is adequate.

Past experience has indicated that an occasional source of error is the miscoding of units so that an entire month of data would be internally consistent yet too high by some scale factor. To account for this, a second estimate of λ was computed using an assumed value for the 99.9th percentile, i.e., a value that historically should not be exceeded more than one time in a thousand.

RESULTS

In order to provide a realistic test of this screening procedure, actual data sets were used. One of particular interest involved carbon monoxide data that had been quickly key-punched and then manually edited for a specific study. This provided a preliminary and corrected version of the file. The preliminary file had known errors and the corrected file was presumably valid. The first test run on the preliminary file processed 21,362 hourly values from 40 monthly data sets. Eight of these monthly data sets were flagged. Hourly carbon monoxide values would be expected to mostly fall in the range of 0 to 50 ppm. In this first test, values of 900, 800, 700, and 500 were found resulting in gap lengths greater than 100 and associated probabilities of less than 1 in 10,000. These results are shown in Table 1. Of the eight flagged data sets,

TABLE 1. Rejected Site Months From Sample Data Set

Site	Month/year	Number of values	Maximum	2nd high	Gap length	Starting at	Number of values above	Probability
33	Oct. 1974	530	30	13	16	14	1	.0006
33	Nov. 1974	604	500	300	100	15	3	<.0001
33	Dec. 1974	671	800	500	100	41	4	<.0001
33	Jan 1975	653	500	500	100	20	2	<.0001
33	Feb. 1975	510	33	18	14	19	1	<.0001
39	June 1975	707	900	700	100	27	3	<.0001
901	July 1974	620	15	14	3	11	3	.0056
901	Aug. 1974	334	800	800	100	11	5	<.0001

seven had keypunch errors. The one remaining month was flagged on the basis of a gap of length 3 and the data appeared to be reasonable. This presented no difficulty for the analyst because the computer printout was sufficient to indicate that these data were in an intuitively acceptable range and probably did not warrant further investigation.

It took less than 30 seconds on EPA's UNIVAC 1110 to process these 21,362 hourly values, and the total cost was approximately $1. It should be noted that the program does several other editing checks so that this cost includes more than the screening procedure for gaps.

CONCLUSIONS

Using gaps in monthly frequency distributions appears to be a convenient means of screening hourly air pollution data sets for outliers. Results to date indicate that it satisfies the criteria of being easy and economical to implement while producing output that is intuitively understandable to an air pollution data analyst. The test successfully spots the more obvious errors. As expected, the initial results also suggest that these types of data sets do have a much lower error rate than the user perceives because of the emphasis on only the few highest values.

There are certain refinements that can be made in screening these type of data sets. Time series models and the use of associated data, such as meteorological variables, would be expected to increase sensitivity and possibly result in even better data quality. However, it remains to be seen if these more elaborate approaches are cost effective when processing vast quantities of data from locations throughout the nation.

As a final comment, it should be noted that once a value is flagged as a possible anomaly, it cannot be arbitrarily dropped from the data set. It must first be verified that the data point actually is incorrect. The fact that the data point is statistically unusual does not necessarily mean that it did not occur.

REFERENCES

1. Hunt, W. F., Jr., and T. C. Curran. An Application of Statistical Quality Control Procedures to Determine Progress in Achieving the 1975 National Ambient Air Quality Standards. Transactions of the 28th Annual ASQC Conference, Boston, Massachusetts, May 1974.

2. Hunt, W. F., Jr., T. C. Curran, N. H. Frank, and R. B. Faoro. Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air. Transactions of the Joint European Organization for Quality Control/International Academy for Quality Conference, Venice Lido, Italy, September 1975.

3. Hunt, W. F., Jr., R. B. Faoro, and S. K. Goranson. A Comparison of the Dixon Ratio Test and Shewhart Control Chart Test Applied to the National Aerometric Data Bank. Presented at the 30th Annual Conference of the American Society for Quality Control. Toronto, Ontario, Canada, June 1976.

4. Title 40 - Protection of Environment. National Primary and Secondary Ambient Air Quality Standards. Federal Register. 36:(84):8186-8201, April 30, 1971.

5. Larsen, R. I. A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards. U.S. Environmental Protection Agency, Research Triangle Park, N.C. Publication No. AP-89. 1971.

6. Curran, T. C. and N. H. Frank. Assessing the Validity of the Lognormal Model when Predicting Maximum Air Pollutant Concentrations. Presented at the 68th Annual Meeting of the Air Pollution Control Association, Boston, Massachusetts, 1975.

LCS:130:70:991

Description of Gap Test Computer Program

Overview

This FORTRAN program may be used to read SAROAD format raw data cards and screen hourly data for the criteria pollutants according to the gap test. Each monthly data set is screened for gaps and also for the number of hourly values exceeding a user supplied upper limit (SMAX( )). This latter feature is incorporated into the program to protect against an entire month, or portion of a month, being too high due to incorrect scaling. The user supplied upper limit is a concentration that should not be exceeded more than one time in a thousand. The program counts the number of values above this limit and uses the Poisson approximation to compute an associated probability.

The gap test is calculated by fitting two different exponential distributions to the data. One estimate is obtained from the 50th and 95th percentiles of the data while the other uses the 50th percentile of the data and the specified upper limit as the 99.9th percentile. These two different estimates are employed to protect against different types of errors. Output may be obtained for each monthly data set or PCUT( ) may be varied to suppress printing of acceptable data.

The program contains certain editing features to prevent arrays from being over-subscripted. Summary results of the processing are printed at the end of each run.

Program Input

SAROAD raw data cards (cards for non-hourly data are ignored)

Program Execution

On EPA's UNIVAC 1110, the following runstream will execute the program.

```
@ASG,A TRRP*ADSS.
@XQT TRRP*ADSS.GAP
@ADD (your data file - cards)
@FIN
```

Program Statistics

On EPA's UNIVAC 1110, this program will process 25,000 hourly values or 2000 cards in approximately 30 seconds and a cost of $1.

```
      DIMENSION VIN(12),KDAYS(31,2),AVIN(12),SMAX(5)
      DIMENSION INCELL(102),NREM(102),SCALE(5,4)
      DIMENSION PTEST(5)
      DIMENSION NOTMI(102),XI(4)
      DIMENSION ICUT(10)
      DIMENSION PCUT(5)
      DIMENSION NOTOUN(100,2)
      DATA PCUT/.01,.01,.01,.01,.01/
      DATA ABLANK/'    '/
      DATA AMISS/'9999'/
      DATA SMAX/75.,100.,60.,50.,50./
      DATA SCALE/1000.,26.2,13.8,19.6,19.5,1.00,.0262,.019,.0196,.0136,1
     &.00,.010,.010,.010,.010,10.0,10.0,10.0,10.0,10.0/
      33X CTIO
      DATA MDAY1,MHR1,MDAY2,MHR2/0,0,0,0/
C
C ------------------------------------------------
C INCELL(I) IS FOR CONCENTRATION I-1
C INCELL(102) IS FOR MISSING VALUES
C VAL IS FOR THE SET OF HOURLY VALUES
C SCALE(I) IS SCALE FACTOR FOR POLLUTANT I
C ------------------------------------------------
C
      DATA MAX1,MAX2/0,0/
      DATA ICUT/5,5,5,5,5,5,5,5,5,5/
      DATA KNTREC,KNTERR,KNTBAD,KNTDUP/0,0,0,0/
      DATA KNTVAL,KNTSMP,KNTFLG,KABORT/0,0,0,0/
      DATA KNTMHR/0/
      DATA KREAD,KWRITE,NOMORE/0,0,0/
      DATA S1,S2,S3,S4/0.,0.,0.,0./
      DATA IYR,IMO,INPL,IMETH/0,0,0,0/
      DATA INOTON,INPLTL/0,0/
      DO 5 I=1,100
      DO 5 J=1,2
5     NOTOUN(I,J)=0
      WRITE(6,450)
      KNTIO=1
      GO TO 900
C INPUT SECTION
10    CONTINUE
      S1L=S1
      S2L=S2
      S3L=S3
      S4L=S4
      IYRL=IYR
      IMOL=IMO
      INPLL=INPL
      IMETHL=IMETH
15    CONTINUE
      KNTREC=KNTREC+1
      READ(5,401,ERR=1200,END=998)ITYPE,S1,S2,S3,S4,ITPER,IYR,IMO,IDAY,I
     &CARD,INPL,IMETH,IUNITS,IDP,(VIN(I),I=1,12),(AVIN(I),I=1,12)
401   FORMAT(I1,A2,A4,A3,A3,I1,I2,I2,I2,I2,I5,I?,I2,I1,12F4.0,T33,12A4)
C
C ----------------------
C   EDIT CHECKS
C ----------------------
C
      IF(ITYPE.NE.1)GO TO 1300
      INPLT=0
      IUN=0
      IF(INPL.EQ.42101)INPLT=1
      IF(INPL.EQ.42401)INPLT=2
      IF(INPL.EQ.42602)INPLT=3
      IF(INPL.EQ.44101)INPLT=4
      IF(INPL.EQ.44201)INPLT=5
      IF(INPLT.EQ.0) GO TO 1315
      IF(ITPER.NE.1) GO TO 1310
      IF(IUNITS.EQ.1)IUN=1
      IF(IUNITS.EQ.5)IUN=2
      IF(IUNITS.EQ.7)IUN=3
      IF(IUNITS.EQ.8)IUN=4
      IPUCH=IUN*INPLT
      IF(IPUCH.EQ.0) GO TO 1320
      IF(ICARD.NE.0) GO TO 35
```

```
      ICARD=1
      GO TO 35
   35 IF(ICARD.NE.12)GO TO 1325
      ICARD=2
   36 CONTINUE
      IF(IDAY.GT.31) GO TO 1350
      IF((IMO.LT.1).OR.(IMO.GT.12))GO TO 1350
      IF(SAL.NE.S1)KWRITE=1
      IF(SAL.NE.S2)KWRITE=1
      IF(SAL.NE.S3)KWRITE=1
      IF(SAL.NE.S4)KWRITE=1
      IF(IYRL.NE.IYR)KWRITE=1
      IF(IMOL.NE.IMO)KWRITE=1
      IF(INPLL.NE.INPL)KWRITE=1
      IF(IMETHL.NE.IMETH)KWRITE=1
      IF(KREAD.EQ.0)KWRITE=0
      IF(KWRITE.EQ.1)GO TO 200
   50 CONTINUE
      KDAYS(IDAY,ICARD)=KDAYS(IDAY,ICARD)+1
      KCLEAR=0
      INPLTL=INPLT
C DETERMINE SCALE FACTOR
      SF=SCALE(INPLT,IUN)
C FIND MAX1,MAX2,NVAL
      DO 120 I=1,12
      IF(AVIN(I).EQ.AMISS)GO TO 120
      IF(AVIN(I).EQ.ABLANK)GO TO 120
      NVAL=NVAL+1
      VAL=VIN(I)/(SF*10.**IDP)
      IVAL=VAL
      IF(IVAL.GE.0)GO TO 110
      WRITE(6,105)KNTREC
  105 FORMAT(1H ,'NEGATIVE CONCENTRATION ON CARD ',I10,' SET TO 0.')
      IVAL=0
  110 CONTINUE
      IF(IVAL.LT.MAX2) GO TO 115
      MAX2=IVAL
      IF(IVAL.LT.MAX1) GO TO 115
      MAX2=MAX1
      MDAY2=IDAY
      MHR2=(ICARD-1)*12+I
      MAX1=IVAL
      MDAY2=MDAY1
      MHR2=MHR1
      MDAY1=IDAY
      MHR1=(ICARD-1)*12+I
      IF(VAL.GT.SMAX(INPLT))NHIGH=NHIGH+1
  115 IF(IVAL.GT.100) IVAL=100
      INCELL(IVAL+1)=INCELL(IVAL+1)+1
  120 CONTINUE
      KREAD=1
      GO TO 10
C
C ---------------------------------
C THIS SECTION COMPUTES TO PRODUCE OUTPUT
C ---------------------------------
C
  200 CONTINUE
C COMPUTE NUMBER OF VALUES IN CELLS >=I
C NLAST=LAST OCCUPIED @@@X CTIO
      NLAST=101
      NREM(101)=INCELL(101)
      DO 210 IMINUS=1,100
      I=101-IMINUS
      NREM(I)=NREM(I+1)+INCELL(I)
      IF(NREM(I+1).EQ.0) NLAST=I
  210 CONTINUE
C
C ---------------------------------
C THIS LOOP FINDS FIRST EMPTY CELL ABOVE CUTOFF
C NOEMP=0 IF NO EMPTY CELLS
C ---------------------------------
C
      NOEMP=0
      NSTART=ICUT(INPLTL)
```

```
      DO 230 I=NSTART,101
      IF(INCELL(I).EQ.0) GO TO 232
230   CONTINUE
      NOEMP=1
232   NFIRST=I
C COMPUTE NCEMP(I) = NUMBER OF CONSECUTIVE EMPTY CELLS
C STARTING WITH AND INCLUDING I
      DO 250 I=NFIRST,NLAST
      DO 240 J=I,NLAST
      IF(INCELL(J).NE.0) GO TO 245
240   CONTINUE
245   NCEMP(I)=J-I
250   CONTINUE
C COMPUTES HIGHEST FACTOR FOR EXPONENTIAL OUTLIER TEST
      NLAST1=NLAST-1
      TEX=0
      DO 270 I=NFIRST,NLAST1
      TEMP=NCEMP(I)*NREM(I)
      IF(TEMP.LT.TEX) GO TO 270
      IF(NCEMP(I).LT.3)GO TO 270
      TEX=TEMP
      INTEX=I
270   CONTINUE
C COMPUTE LAMBDAS FOR EXPONENTIAL OUTLIER TESTS
      PNVAL=NVAL
      P50=.5*PNVAL
      P95=.95*PNVAL
      C95=101.
      C50=101.
      DO 330 IMINUS=2,101
      I=103-IMINUS
      IF(NREM(I).LT.P95)C95=I-2
      IF(NREM(I).LT.P50)C50=I-2
330   CONTINUE
      C999=SMAX(INFIL)
      XL(1)=0.
      IF(C95-C50)332,332,331
331   XL(1)=2.3026/(C95-C50)
332   XL(2)=0.
333   XL(2)=6.2146/(C999-C50)
      IF(XL(1).LE.0.)ITP1=1
      IF(XL(2).LE.0.)ITP2=1
C PERFORM OUTLIER TESTS
      DO 350 I=1,2
      PTEST(I)=0.
      XARG=TEX*XL(I)
      IF(XARG.GT.50.0) GO TO 350
      PTEST(I)=EXP(-XARG)
      IF(PTEST(I).GT.1.0000) PTEST(I)=1.0
350   CONTINUE
      IF(ITP1.EQ.1)PTEST(1)=-.9999
      IF(ITP2.EQ.1)PTEST(2)=-.9999
C POISSON APPROXIMATION
      XNHIGH=NHIGH
      XLPOIS=PNVAL/100.
      PROD=1
      SUM=0
      NHIGH1=NHIGH+1
      DO 370 I=1,NHIGH1
      SUM=SUM+PROD
      XI=I
      PROD=PROD*XLPOIS/XI
370   CONTINUE
375   CONTINUE
      PTEST(3)=1.0
      IF(XLPOIS.GT.50) GO TO 378
      PTEST(3)=1-EXP(-XLPOIS)*(SUM)
378   CONTINUE
      IF(NHIGH.EQ.0)PTEST(3)=1.0
      GO TO 400
C
C ---------------------------
C ERROR MESSAGES
C ---------------------------
C
```

```
      1200  CONTINUE
            IF (ITYPE.NE.1) KNTNHR=KNTNHR+1
            IF (ITYPE.NE.1) GO TO 15
            WRITE(6,1201)KNTREC,S1,S2,S3,S4
 1201 FORMAT(1H ,'INPUT FORMAT ERROR ON CARD ',I8,'. SITE ',A2,A4,A3,A
     &3,' BEING PROCESSED.')
            KNTIG=KNTIG+1
            KNTBAD=KNTBAD+1
            IF(KNTBAD.GT.100)GO TO 1000
            GO TO 15
 1300  CONTINUE
            KNTNHR=KNTNHR+1
            GO TO 15
 1310  CONTINUE
            WRITE(6,1311)
 1311 FORMAT(1H ,'NOT HOURLY DATA')
            GO TO 1398
 1315  CONTINUE
            IPX=0
            IF(INOTDN.EQ.0)GO TO 1317
            DO 1316 IP=1,INOTDN
 1316 IF(INFL.EQ.NOTDUN(IP,1)) IPX=IP
 1317 IF(IPX.EQ.0) INOTDN=INOTDN+1
            IF(IPX.EQ.0) IPX=INOTDN
            NOTDUN(IPX,1)=INFL
            NOTDUN(IPX,2)=NOTDUN(IPX,2)+1
            GO TO 15
 1320  CONTINUE
            WRITE(6,1321)
 1321 FORMAT(1H ,'PROGRAM DOES NOT ALLOW FOR THIS POLLUTANT-UNIT COMBINA
     &TION')
            GO TO 1398
 1325 CONTINUE
            WRITE(6,1326)ICARD,KNTREC
 1326 FORMAT(1H ,I2,' IS AN INCORRECT START HOUR ON CARD',I10,1H.)
            GO TO 1398
 1330  CONTINUE
            WRITE(6,1331)IDAY,KNTREC
 1331 FORMAT(1H ,I2,' IS AN INCORRECT DAY ON CARD',I10,1H.)
            GO TO 1398
 1340  CONTINUE
 1341 FORMAT(1H ,'EXTRA CARDS FOR SITE ',A2,A4,A3,' IN ',I2,3H/19,I2,
     &' FOR POLLUTANT ',I5,1H.)
 1350  CONTINUE
            WRITE(6,1351)IMO,KNTREC
 1351 FORMAT(1H ,I2,' IS AN INCORRECT MONTH ON CARD',I10,1H.)
 1398  CONTINUE
            WRITE(6,1399)ITYPE,S1,S2,S3,S4,ITPER,IYR,IMO,IDAY,ICARD,INFL,IME
     &TH,IUNITS,IDP,(AVIN(I),I=1,12)
 1399 FORMAT(1H ,I1,A2,A4,A3,A3,I1,I2,I2,I2,I2,I5,I2,I2,I1,12A4)
            KNTIG=KNTIG+2
            KNTERR=KNTERR+1
            IF(KNTERR.GT.100)GO TO 1000
            GO TO 15
C
C ---------------------
C OUTPUT SECTION
C ---------------------
C
 400  CONTINUE
            IOBYPS=0
            DO 410 I=1,3
            IF(PTEST(I).LT.PCUT(ID))IOBYPS=1
 410  CONTINUE
            KNTVAL=KNTVAL+NVAL
            KNTSMP=KNTSMP+1
            IF(IOBYPS.EQ.0)GO TO 900
            KNTFLG=KNTFLG+1
            IF(KNTIC.NE.0) GO TO 440
            WRITE(6,450)
 450  FORMAT(1H1,'SAROAD SITE',2X,'POLLUTANT',2X,'MO/YEAR',3X,'OBS',3X,1
     &X,'MAX(DAY/HOUR)',1X,1X,'2ND HI(DAY/HOUR)',2X,'GAP',1X,'START',1X,
     &'ABOVE',3X,'50/250',4X,'500/999',4X,'CUTOFF')
 440  CONTINUE
            IF(INTEX.NE.0) GO TO 441
```

```
      INTEX=1
      NCEMP(INTEX)=0
      NREM(INTEX)=0
 441  INTEX1=INTEX-1
      WRITE(6,451) S1L,S2L,S3L,S4L,INCL,IYSL,  NVAL,MAX1,
     &MDAY1,MAX2,MDAY2,
     &NCEMP(INTEX),INTEX1,NREM(INTEX),(INTST(I),I=1,3)
 451  FORMAT(1H ,A2,A4,A3  ,3X,I5,3X,I2,1H/,I2,1H-,I2,2X,I4,2X,  I5,
     &1H(,I2,1H/,I2,4H:00),2X,I5,1H(,I2,1H/,I2,4H:00),2X,
     &I3,2X,I3,2X,I4,4X,F6.4,4X,F6.4,4X,F6.4,4X,F6.4)
      KNTTC=KNTTC+1
      IF(KNTIC.EQ.50)KNTIC=0
C TEST OUTPUT
      IF(NOMORE.EQ.1) GO TO 9993
C
C EXECUTE TO CLEAR ARRAYS
 900  CONTINUE
      KCLEAR=1
      IF(NOMORE.EQ.1) GO TO 9993
      KWRITE=0
      KERROR=0
      INTEX=0
      DO 950 I=1,31
      DO 950 J=1,2
      IF(KNTREC.EQ.0)GO TO 945
      IF(KDAYS(I,J).GT.1)KERROR=1
 945  CONTINUE
      KDAYS(I,J)=0
 950  CONTINUE
      IF(KERROR.EQ.0)GO TO 960
      WRITE(6,1341)S1L,S2L,S3L,S4L,IMOL,IYRL,INPL
      KNTTC=KNTTC+1
      KNTDUP=KNTDUP+1
      KERROR=0
 960  CONTINUE
      ITP1=0
      ITP2=0
      MAX1=0
      MAX2=0
      NPTIC=0
      NPTRCT=0
      NLAST=0
      NVAL=0
      DO 430 I=1,102
      INCELL(I)=0
      NREM(I)=0
      NCEMP(I)=0
 400  CONTINUE
      IF(KNTREC.EQ.0)GO TO 10
      GO TO 50
 998  CONTINUE
      KNTREC=KNTREC-1
      NOMORE=1
      IF(KCLEAR.EQ.1)GO TO 1100
      GO TO 2000
 999  CONTINUE
      GO TO 1100
1000  KABORT=1
1100  CONTINUE
      WRITE(6,1010)
      IF(KABORT.EQ.1)WRITE(6,1012)
      WRITE(6,1022)KNTREC
      WRITE(6,1024)KNTBAD
      WRITE(6,1026)KNTERR
      WRITE(6,1027)KNTNHR
      WRITE(6,1028)KNTVAL
      WRITE(6,1030)KNTSHP
      WRITE(6,1032)KNTDUP
      WRITE(6,1034)KNTFLG
      KNTNFG=KNTCMP-KNTFLG
      WRITE(6,1036)KNTNFG
      WRITE(6,1038)
      IF(INCTON.EQ.0)GO TO 9999
      WRITE(6,1041)
      DO 1110 I=1,INCTON
```

```
      WRITE( 6,1053)NOTDUN( I, 2) ,NOTDUN( I, 1)
 1110 CONTINUE
 1010 FORMAT(1H1,28X,'***** PROCESSING SUMMARY *****')
 1012 FORMAT(1H ,'---- PROGRAM ABORT --TOO MANY INPUT ERRORS -- CHECK
     &  INPUT FILE ----')
 1022  FORMAT(1H ,I10,' INPUT CARDS WERE READ.')
 1024  FORMAT(1H ,I10,' CARDS WERE NOT PROCESSED BECAUSE OF FORMAT ERRORS
     &.')
 1026  FORMAT(1H ,I10,' CARDS WERE NOT PROCESSED BECAUSE OF UNACCEPTABLE
     & DATA.')
 1027 FORMAT(1H ,I10,' CARDS DID NOT HAVE A 1 IN THE FIRST COLUMN.')
 1023  FORMAT(1H ,I10,' HOURLY VALUES WERE PROCESSED.')
 1030  FORMAT(1H ,I10,' SITE-MONTH-POLLUTANT DATA SETS WERE PROCESSED.')
 1032  FORMAT(1H ,I10,' SITE-MONTH-POLLUTANT DATA SETS HAD REDUNDANT CARD
     &S.')
 1034  FORMAT(1H ,I10,' SITE-MONTH-POLLUTANT DATA SETS WERE FLAGGED.')
 1036  FORMAT(1H ,I10,' SITE-MONTH-POLLUTANT DATA SETS WERE NOT FLAGGED.
     & ')
 1038  FORMAT(1H ,17X,'***** ***** ***** ** ** * ***** ** ** *')
 1051 FORMAT(1H ,'NOTE**CARDS FOR THE FOLLOWING POLLUTANTS WERE IGNORED:
     & ')
 1053 FORMAT(1H ,5X,I10,' CARDS FOR POLLUTANT ',I5,1H.)
 9999 CONTINUE
```

CARD	CARD	SITE	POLLUTANT	MO/YEAR	OBS	MAX(DAY/HOUR)	2ND HI(DAY/HOUR)
7	79	39	81	03	005	04 25 1	1/1 97 0	592	39(1/12:00)	7(1/1:00)
7	79	39	91	03	006	04 29 1	1/1 97 0	592	39(1/12:00)	17(1/1:00)

CAP	START	ABOVE	50/95	50/999	CUTOFF
91	0	1	.0000	.0000	.1193
81	13	1	.0000	.0000	.1193

```
                ***** PROCESSING SUMMARY *****
   152 INPUT CARDS WERE READ.
     0 CARDS WERE NOT PROCESSED BECAUSE OF FORMAT ERRORS.
     0 CARDS WERE NOT PROCESSED BECAUSE OF UNACCEPTABLE  DATA.
     0 CARDS DID NOT HAVE A 1 IN THE FIRST COLUMN.
  1410 HOURLY VALUES WERE PROCESSED.
     4 SITE-MONTH-POLLUTANT DATA SETS WERE PROCESSED.
     0 SITE-MONTH-POLLUTANT DATA SETS HAD REDUNDANT CARDS.
     2 SITE-MONTH-POLLUTANT DATA SETS WERE FLAGGED.
     2 SITE-MONTH-POLLUTANT DATA SETS WERE NOT FLAGGED.
          ***** ***** ***** ***** ***** *****
```

APPENDIX B - Pattern Test for Hourly Data

This appendix contains additional information on the pattern tests for hourly data. The following material is included:

(1) A copy of the paper "Automated Screening of Hourly Data,"

(2) A brief description of the computer program for these tests

(3) A listing of the FORTRAN Computer Program

1978 ASQC TECHNICAL CONFERENCE TRANSACTIONS—CHICAGO

First Line, (for all papers except Title Page)

AUTOMATED SCREENING OF HOURLY AIR QUALITY DATA

Robert B. Faoro, Mathematical Statistician
Thomas C. Curran, Mathematical Statistician
William F. Hunt, Jr., Chief, Data Analysis Section

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Monitoring and Data Analysis Division
Research Triangle Park, North Carolina 27711

INTRODUCTION

Over the past several years a number of different automated methods to screen air quality data for errors have been proposed.[1-4] Basically these techniques were developed to detect the more obvious data errors resulting primarily from keypunch, transcription, or periodic malfunctioning instruments. More subtle errors from inadequate calibrations procedures or similar problems resulting in measurement bias will not be detected by these procedures. The goal of these techniques was to ensure a high quality data product for the higher concentration levels because in many cases these higher values determine an areas status with respect to the various ambient air quality standards and the amount of emission controls needed. For example, the second highest hourly observation out of a possible 8760 hours in a year is used to determine compliance for carbon monoxide and ozone. Other pollutants have the second highest day as the decision making statistic. Pollutants having annual mean standards such as total suspended particulate, sulfur dioxide, and nitrogen dioxide, would require that more attention be given to the complete annual data set.

Basically the techniques which have been developed can be classified by their application into two main categories: 24 hour (intermittent systematic sampling) and hourly data (continuous sampling). Procedures for screening 24-hour data, will not be discussed in this paper. They have been described in previous papers.[2-4] A guideline document[6] has been prepared describing the complete air quality data screening package together with summary documentation of both tests described in this paper. The purpose of this paper is to evaluate two different schemes for screening hourly air quality data. These two procedures will be referred to as the typical pattern test and the monthly gap test.

These screening procedures were developed to be both simple and yet effective discriminators between "good and "bad" data. Another requirement was that these tests could be done efficiently by a computer. Being simple and computer-efficient was most important because of the sheer magnitude of data requiring screening. At the present time, for example, there are over 2000 continuous monitoring sites located throughout the country who submit data to the National Aerometric Data Bank (NADB) located in Research Triangle Park, North Carolina. If each of these sites collected a complete year of data (8760 hours), the total annual data submission to the data bank from these sites would be over 17 million measurements. Being effective discriminators of "good" and "bad" data is of course important since it would be time consuming and costly to flag "good" data and of course, disastrous to miss flagging "bad" data.

DESCRIPTION OF SCREENING TESTS

Although air quality is difficult to predict, generally it behaves within certain natural bounds and exhibits fairly regular geographical, seasonal, weekly, and diurnal concentration patterns depending upon emission and meteorological factors. The screening tests discussed here attempt to discover inconsistencies in the data that warrant further scrutiny. For example, the pollutant ozone, which is formed when hydrocarbon and oxides of nitrogen emissions predominantly from motor vehicles are irradiated by sunlight generally exhibits lower concentrations during the nighttime hours and during the winter months. Nitrogen dioxide does not show as distinct as seasonal pattern as ozone, but still has a well defined diurnal pattern. Generally, nitrogen dioxide exhibits a distinct morning peak (8-10 a.m.) resulting from the oxidation of nitric oxide emissions from motor vehicles during the morning commuter rush. Pollutant concentration patterns usually behave fairly regularly and do not exhibit, except when under the influence of a strong local source, extreme hour to hour variation patterns. Likewise, high (low) pollutant concentrations usually result

from a gradual increase (decrease) in concentrations rather than a sudden rise (fall). Table 1 shows six typical days of nitrogen dioxide (NO_2) hourly concentrations from a site in Los Angeles, California. Note that the hours immediately following the morning rush hour are typically the highest for this pollutant and that the concentrations show gradual changes in concentrations from one hour to another.

These data screening procedures look for different types of inconsistencies in the data. The pattern test look for extremely high concentrations never or very rarely exceeded in the past and other types of unusual pollutant behavior. O'Reagan[5] discusses some very interesting screening concepts along these same lines. The gap test looks for breaks in the monthly frequency distribution of the hourly pollutant observations. An example for ozone of a significant break in the three highest observations in a month would be:

HIGHEST HOUR	2nd HIGHEST HOUR	3rd HIGHEST HOUR
929 $\mu g/m^3$	929 $\mu g/m^3$	374 $\mu g/m^3$

A brief description of the two screening procedures will be presented before they are applied to actual air quality data. The typical pattern tests are not statistical tests in that probalistic statements cannot be made about a rejected data point. They instead represent simple and practical ways to check for obvious errors in the data. Basically these tests can be classified into two main categories:

- tests which look for unusual pollutant behavior, such as exceeding some extremely high concentration, either never before exceeded or exceeded only very rarely based on past "good" data and
- a test which looks for unusually high values in the day with respect to the other values in the day.

More specifically, the tests look for the following types of errors:

- hourly values exceeding an upper limit empirically derived from prescreened historical data (Max Hour)
- differences in adjacent hourly values exceeding an empirically derived upper limit difference (Adjacent Hour)
- a single value being much different than the other values in the day using a modification of the Dixon Ratio Test
- differences and percent differences between the middle value and its' adjacent values in a 3-hour interval exceeding certain pre-derived limits, (Spike) and
- averages of four or more consecutive hours exceeding some pre-derived concentration limit (Consecutive Hour).

Table 2 gives typical upper limit check values used in the various pattern tests for EPA's Region V, consisting of the States of Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin. One of the main drawbacks of these kinds of tests is that ideally these limits values would reflect a particular site, or a group of sites, having common air pollution characteristics. It is impossible to have individual limits for each and every site. Therefore, some discrimination is sacrificed by merely having a given set of parameters for all sites. Of course, if you are only interested in screening data from a small number of sites, it may indeed be feasible to have site specific parameters. The pattern test outputs each day that contains at least one hour that violates a particular test and gives the tests which were violated.

The frequency distribution gap test was developed to provide an even simpler means of screening hourly data. The two main advantages of this approach were that the results could be put in a probalistic framework and that it could be applied universally to all data without modification. In order for the pattern test to be optimally effective, the limit checks would need to be varied on a site by site basis.

The theory behind the gap test is that unusually high values could be detected by examining the frequency distribution of the hourly data for a given period of time, such as a month, quarter or year. The test will be employed on a monthly basis in

TABLE I - SELECTED DAYS OF HOURLY NO_2 CONCENTRATIONS IN PARTS PER MILLION FROM A LOS ANGELES CENTER CITY SITE IN AUGUST, 1976

DAY	12	1	2	3	4	5	6	7	8	9	10	11[a]	12	1	2	3	4	5	6	7	8	9	10	11	DAILY MEAN NO.	
22	.01	.01	.01	.01	.01	.01	.02	.05	.07	.09	.06		.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.03	23
23	.01	.01	.01	.01	.03	.03	.04	.10	.21	.33	.21		.08	.06	.06	.07	.06	.05	.05	.05	.04	.04	.04	.04	.08	23
24	.04	.04	.04	.04	.05	.05	.11	.17	.26	.29	.22		.07	.05	.05	.05	.09	.07	.07	.07	.05	.04	.05	.05	.09	23
25	.05	.03	.03	.03	.03	.04	.05	.11	.16	.19	.14		.07	.04		.07	.07	.07	.07	.05	.05	.05	.05	.04	.09	23
26	.05	.05	.05	.06	.07	.07	.08	.14	.25	.21	.17		.10	.07	.06	.07	.07	.07	.06	.05	.05	.05	.06	.07	.09	23
27	.07	.08	.05	.05	.06	.05	.06	.10	.19	.24	.14		.09	.07	.06	.07	.06	.05	.09	.13	.11	.10	.11	.11	.10	23

a. The values for this hour are missing because of calibration checks.

TABLE II - SUMMARY OF LIMIT VALUES USED IN EPA REGION V FOR PATTERNS TESTS

POLLUTANT	DATA STRATIFICATION	MAX. HOUR	ADJ. HOUR (last hour)	SPIKE (center hr)	CONS. 4-HOUR
Ozone-Total Oxidant ($\mu g/m^3$)					
	summer-day	1000	300	200(300%)	500
	summer-night	750	200	100(300%)	500
	winter-day	500	250	200(300%)	500
	winter-night	300	200	100(300%)	500
Carbon Monoxide (mg/m^3)					
	rush traffic hours	75	25	20(500%)	40
	non-rush traffic hours	50	25	20(500%)	40
Sulfur Dioxide ($\mu g/m^3$)					
	EPA Region	2600	500	200(500%)	1000
Nitrogen Dioxide ($\mu g/m^3$)					
	None	1200	500	200(300%)	1000

TABLE III - EXAMPLES OF DATA[a] FLAGGED BY BOTH TESTS

SITE	DATE (MO/ DAY)	POL	MIDN	1	2	3	4	5	6	7	8	9	10	11	N	1	2	3	4	5	6	7	8	9	10	11
1	12/12	NO_2	55	55	45	45	55	45	45	36	45	36	55	64	750	64	64	73	73	73	64	64	64	64	73	64
2	2/10	NO_2	83	92	83	73	73	73	73	83	83	92	83	83	64	64	64	73	83	83	760[c]	83	83	64	73	73
3	1/12	NO_2	56	MSG[b]	MSG	38	38	38	38	38	38	376	113	56	38	75	56	75	75	75	75	75	75	75	75	75
3	1/24	NO_2	56	75	MSG	MSG	56	56	75	75	75	75	56	56	75	75	56	56	56	75	75	56	56	301	94	75
3	1/25	NO_2	75	75	MSG	MSG	75	75	75	56	56	56	56	56	56	75	75	56	56	56	56	338	113	75	75	56
4	2/3	O_3	51	47	41	41	18	22	8	12	12	8	8	8	22	18	31	267	51	61	47	51	41	51	61	61
4	2/8	O_3	67	71	61	61	257	41	47	47	51	51	47	47	41	41	57	61	61	61	57	57	61	51	51	51
4	2/17	O_3	90	67	419	47	57	41	57	41	41	18	22	22	31	31	41	41	47	47	37	37	4	MSG	MSG	MSG
5	9/7	O_3	1	1	1	1	1	1	1	1	20	80	90	150	940	930	940	920	840	860	810	730	690	630	630	650
6	4/13	O_3	0	0	0	0	0	0	0	0	0	0	0	100	0	100	563	664	357	10	306	100	0	153	357	768
7	11/18	O_3	MSG	0	6	2	2	2	0	4	4	10	6	6	510	523	MSG	MSG	6	6	8	6	6	8	8	8
8	9/17	O_3	290	290	250	360	450	540	370	410	420	520	660	800	1000	1000	1000	1000	1000	1000	1000	1000	630	550	400	320

a. All data reported in micrograms of pollutant per cubic meter of air sampled.

b. MSG means the hourly value is missing.

c. Suspicious value.

these applications. The length of the gap and the number of values above the gap afford a convenient means of detecting possible errors. The exponential distribution was used to describe the upper tail of the hourly pollutant concentrations and thereby, provided the underlying theory for detecting significant gaps in the frequency distribution of the hourly data. An example of the days flagged from the gap test can be found in Table 4 of this paper. A detailed description of the test and its' application to some actual air pollution data has been discussed previously.[1]

DATA BASE

Two sets of actual hourly air quality data taken from the NADB were screened using both techniques. The two sets represent the pollutants nitrogen dioxide (NO_2) and ozone (O_3) from about 40 randomly selected sites located throughout the country for the year 1976. There were approximately 100,000 hourly values for each pollutant. It took less than 1 minute of computer time costing about $2.00 on the UNIVAC-1110 system for each data set to do both screening procedures

RESULTS

Overall the two tests rejected basically the same data from both pollutant data sets. Of the 23 specific instances of rejected data, 18 of these were rejected by both tests while the remainder (5) had one but not both tests rejecting. The instances where the tests substantiated each other are almost without question true data anomalies while the rest of the cases are more doubtful anomalies. Table 3 gives examples of some of the data which were flagged by both tests. These data represent days with either a single hourly anomaly or in some cases multiple data errors. All told, 87 days (0.8%) out of over 10,000 days of data were rejected by the pattern test while 21 months (5.0%) of data out of 409 site months screened were rejected by the gap test.

Table 4 gives several examples of days flagged by the pattern test and months flagged by the gap test where the two procedures did not flag the same data. All of the days flagged by the pattern test with the exception of the Los Angeles day (June 23rd) probably contain errors. The specific hour identified as in error are underlined. The reason that the gap test did not flag these data is because in each of these cases the errors represent hourly concentrations which were not unusual for the month and therefore no significant gap in the monthly frequency distribution of observations occurred. These types of data errors then represent typical values for the month as a whole but they were unusual when they were compared with the data values recorded around the data value in question. Both examples of data flagged by the gap test will require further examination. The San Diego NO_2 data for August is unusual, however, because of the missing data immediately following the specific data in question.

CONCLUSION

Based on a limited, but yet representative set of continuous hourly NO_2 and O_3 data, it has been shown that the pattern and gap screening tests mimic each other very well in terms of the data rejected. There were only minor discrepancies between the two tests. What is even more important is that both tests rejected data which in most cases contained real errors. This was particularly true when both tests rejected the same data. The overall rejection rate was quite low for both tests. Although all of the hourly data passing the tests were not reviewed, what data was reviewed did not reveal any obvious data errors that were missed by the tests. It is recommended that the gap test be used as the initial means of screening large hourly data sets because its' printed output is much less than the pattern test generates, particularly, of course, in the case where a lot of data is in error. There is also a slight savings in the amount of computer time for the gap test. The pattern test then can be used as a backup to substantiate the results of the gap test or to provide more specific output about the days which contain errors.

It is recommended that these procedures be used by the agency collecting the data instead of being used at the Regional or National (NADB) level. The problems of verification and correction of data flagged can be done more efficiently and effectively nearest the source of the data. Presently, the States of Minnesota, Ohio, and Wisconsin are using these procedures on a regular basis.

TABLE IV - EXAMPLES OF DATA[a] FLAGGED BY ONE TEST BUT NOT THE OTHER

SITE	DATE (MO/ DAY)	POL	FLAGGED BY	MIDN	1	2	3	4	5	6	7	8	9	10	11	N	1	2	3	4	5	6	7	8	9	10	11
9	6/23	NO_2	PATTERN	75	56	75	94	113	169	244	546	752	376	75	MSG[b]	MSG	94	94	94	94	94	132	132	132	94	75	94
10	12/13	NO_2	PATTERN	149	149	130	130	111	92	149	111	205	130	111	337	130	130	130	130	160	186	186	167	167	167	186	149
11	1/8	NO_2	PATTERN	50	50	50	50	50	50	60	60	70	70	480	150	60	60	60	60	60	80	80	90	90	100	100	110
11	1/9	NO_2	PATTERN	120	110	110	100	80	80	80	90	90	100	410	200	100	80	70	70	90	100	100	100	100	90	90	90
11	12/5	NO_2	PATTERN	50	50	60	60	30	30	20	10	0	10	100	500	50	10	10	10	10	10	10	10	10	10	10	10
12	6/30	O_3	PATTERN	25	41	18	39	37	47	61	55	31	76	MSG	125	145	MSG	MSG	272	243	47	292	218	155	194	57	0
13	8/8	O_3	PATTERN	4	0	0	0	0	0	0	6	39	73	120	143	174	151	433	161	147	196	153	84	49	8	16	31
14	8/30	NO_2	GAP	148	130	148	92	148	167	205	MSG	412	393	186	167	MSG	MSG	MSG	MSG	MSG	MSG	MSG	MSG	MSG	MSG	MSG	MSG
15	2/27	NO_2	GAP	MSG	MSG	MSG	MSG	MSG	MSG	MSG	MSG	320	320	508	432	MSG	MSG	94	94	113	132	132	132	132	132	150	113

a. All date reported in micrograms of pollutant per cubic meter of air sampled.

b. MSG means the hourly value is missing.

REFERENCES

1. Curran, Thomas C., W. F. Hunt, and R. B. Faoro, Quality Control for Hourly Air Pollution Data, Presented at the 31st Annual Technical Conference of the American Society for Quality Control, Philadelphia, Pennsylvania, May 16-18, 1977.

2. Hunt, W. F., and T. C, Curran. An Application of Statistical Quality Control Procedures to Determine Progress in Achieving the 1975 National Ambient Air Quality Standards. Transactions of the 28th Annual ASQC Confrence, Boston, Massachusetts, May 1974.

3. Hunt, W. F., T. C. Curran, N. H. Frank, and R. B. Faoro. Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air. Transactions of the Joint Eurpoean Organization for Quality Control/International Academy for Quality Conference, Venice Lido, Italy, September 1975.

4. Hunt, W. F., R. B. Faoro, and S. K. Goranson. A comparison of the Dixon Ratio Test and Shewhart Control Chart Test Applied to the National Aerometric Data Bank. Presented at the 30th Annual Conference of the American Society for Quality Control. Toronto, Ontario, Canada. June 1976.

5. O'Reagan, Robert T. Practical Techniques for Computer Editing of Magnitude Data. Unpublished paper, Department of Commerce, Bureau of the Census, Washington, D.C. 20223, 1972.

6. Curran, T. C., Guidelines for Screening Ambient Air Qaulity Data. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina 27711. (In preparation)

DESCRIPTION OF PATTERN TEST COMPUTER PROGRAM

This FORTRAN program consists of a main program and five subprograms to screen hourly air quality data for unexpected departures from typical patterns. The typical pattern tests are not statistical tests in that probabilistic statements cannot be made about a rejected data point. They instead represent simple and practical ways for checking for various possible, and in most cases, obvious errors in the data. The tests specifically look for the following types of errors:

- hourly values exceeding an empirically derived upper limit
- difference in adjacent hourly values exceeding an empirically derived upper limit difference
- a value in a day being much different than the other values in the day using a modification of the Dixon Ratio Test
- differences and percent differences between the middle value and it's adjacent values in a 3-hour interval exceeding certain pre-derived limits, and
- consecutive values of four or more hours exceeding some pre-derived concentration limit.

The main program reads the standard hourly SAROAD card format, calls the subprograms, and outputs to the printer the results of the screening procedure. Listings of the main program and the subprograms are included following this discussion. The input cards must be ordered by the date (year, month and day) within each site, pollutant-method combination. Any number of site pollutant-method combinations can be run back to back without any means of separation. An end file (@ E o F) indicator or other end of file indicators on tape is used to signal the end of the input data set. The screening checks are

performed in the subprograms. There is a separate subprogram for each of the pollutants considered: carbon monoxide, sulfur dioxide, nitrogen dioxide, and photochemical oxidants. The fifth subprogram is used for checking the data sequence of the inputted data cards. An example of the printed output is shown in the table enclosed. The output consists of the site code, pollutant-method code, year, month, and day, the hourly values for the day in question, and the test or tests which the data violated. Also, following the completion of a site, pollutant-method combination a line is printed out showing the number of days screened.

Program Input

SAROAD raw data cards

Program Execution

On EPA's UNIVAC 1110, the following runstream will execute the program.

```
@ ASG, A TRRP*ADSS.
@ XQT TRRP*ADSS.PATTERN
@ ADD (your data file-cards)
@ Fin
```

Program Statistics

On EPA's UNIVAC, this program, like the gap test will process about 25,000 hourly values or 2,000 cards in approximately 30 seconds at a cost of about $1.00.

```
      C
      C
      C ***** AIR DATA SCREENING SYSTEM (ADSS) CONTINUOUS TESTS *****
      C
      C
            DIMENSION ITEST(5),IS(48),IDATA(12),XDATA(12),ADATA(12),XSITE(4),
           1XSITEO(4),AT(5)
            COMMON ITEST,IS,XSITE,IMO,NUM,SUM,SUM2,XSITEO,IYR,IHDAY,IX
            DATA ABLANK/'    '/
            DATA AB/' '/
            DATA ANEX/'X'/
            ISW=0
            IDYST=0
            IDST=0
            ISW3=0
            IP6=0
            ISW2=0
            ISW4=0
          2 N=1
            NUM=0
            SUM=0
            SUM2=0
      C
      C ***** READ SARCAD HOURLY DATA CARD *****
      C ***** BUILD TWO DAY ARRAY *****
      C
            GO TO 5
          4 XSITEO(1)=XSITE(1)
            XSITEO(2)=XSITE(2)
            XSITEO(3)=XSITE(3)
            XSITEO(4)=XSITE(4)
            IPOLO=IPOL
            IMTDO=IMTD
            IYRO=IYR
            IMOO=IMO
            IDAYO=IDAY
          5 READ(5,200,END=175) (XSITE(I),I=1,4),ITME,IYR,IMO,
           1IDAY,ISTHR,IPOL,IMTD,IUNIT,IDP,(IDATA(J),J=1,12),(ADATA(J),J=1,12)
           2,IX
        200 FORMAT(1X,A2,A4,2A3,I1,4I2,I5,2I2,I1,12I4,T33,12A4,T2,I2)
            DO 7 J=1,12
            IF(ADATA(J).NE.ABLANK) GO TO 7
            IDATA(J)=9999
          7 CONTINUE
            IF((IPOL.NE.42101).AND.(IPOL.NE.42401).AND.(IPOL.NE.42602).AND.
           1(IPOL.NE.44101).AND.(IPOL.NE.44201)) GO TO 5
            IF((IUNIT.EQ.77).OR.(IUNIT.EQ.78)) GO TO 351
            IF(IDP.EQ.0) GO TO 6
            GO TO 350
          6 IF(ISW3.EQ.0) GO TO 178
            IF((XSITEO(1).EQ.XSITE(1)).AND.(XSITEO(2).EQ.XSITE(2))
           1.AND.(XSITEO(3).EQ.XSITE(3)).AND.(XSITEO(4).EQ.XSITE(4))
           2.AND.(IPOLO.EQ.IPOL).AND.(IMTDO.EQ.IMTD)) GO TO 8
            ISW2=1
            GO TO 112
          8 IF(N.EQ.1) GO TO 12
            GO TO 14
         12 IF(ISTHR.NE.12) GO TO 17
            GO TO 19
      C
      C ***** CARDS OUT OF ORDER *****
      C
         17 ISW3=0
            GO TO 4
         18 IC=0
            DO 16 J=1,12
         16 IS(J+IC)=IDATA(J)
            IHDAY=IDAY
            N=2
            GO TO 4
         14 IF(N.EQ.2) GO TO 21
            GO TO 22
         21 IF(ISTHR.NE.12) GO TO 17
            IF(IHDAY.NE.IDAY) GO TO 25
            GO TO 24
```

APPENDIX C - Shewhart Test

This appendix contains additional information on the Shewhart Test for 24-hour data. The following material is included:

(1) A copy of the paper, "The Shewhart Control Test - A Recommended Procedure for Screening 24-Hour Air Pollution Measurements,"

(2) A brief description of the computer program for the Shewhart Test

(3) A listing of the Cobol computer program

THE SHEWHART CONTROL CHART TEST - A RECOMMENDED PROCEDURE FOR SCREENING 24-HOUR AIR POLLUTION MEASUREMENTS

Introduction

At the present time there are over 8,000 air monitoring sites operated throughout the United States by the Federal, state, and local governments.[1] These sites collect approximately 20,000,000 ambient air pollution values annually, which are sent to the U.S. Environmental Protection Agency's (EPA) National Aerometric Data Bank (NADB) in Durham, North Carolina. The data are primarily collected to measure the success of emission control plans in achieving the National Ambient Air Quality Standards (NAAQS). As one might expect with data sets this large, anomalous measurements slip through the existing editing and validation procedures. Because of the importance that is attached to violations of the NAAQS, a quality control test to ensure the validity of the measurement of both short- and long-term concentrations is extremely important.

A series of quality control tests have been examined[2-4] to check ambient air quality data for anomalies, such as keypunch, transcription, and measurement errors. The Shewhart Control Chart Test[5] has been selected to screen 24-hour air pollution measurements. This paper discusses its application to three major pollutants--total suspended particulate (TSP), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). The Shewhart Test is applied to data from monitoring instruments which generate one measurement per 24-hour period and are operated on a systematic sampling schedule of approximately once every 6 days. In the cases of SO_2 and NO_2, there are also continuous monitoring instruments, which monitor the pollutants constantly; but our discussion here is concerned only with 24-hour data. The application of the test results in flagged data which need to be verified as either valid or invalid.

A computer software program, the Air Data Screening System, has been written in the computer languages COBOL and FORTRAN. This program incorporates the Shewhart Control Chart Test. It has been successfully applied to data collected in EPA's Region V, which encompasses the states of Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin. In terms of population, it is the largest of EPA's regions, and there is extensive monitoring of the above pollutants. The purpose of the Region V evaluation is to determine whether the data flagged by the Shewhart test are valid or invalid and to identify, if possible, the source of the error.

This paper will discuss the flow of data from the state and local governments; the data-editing process; the basic characteristics of the data; the application and evaluation of the Shewhart Test; and the computer software program, the Air Data Screening System (ADSS); it will conclude with our recommendations.

Data Flow

Most ambient air quality data are collected by state and local air pollution control agencies and are forwarded via EPA's Regional Offices to the NADB. A considerable amount of data is forwarded--approximately 20 million air quality measurements a year. The data are sent quarterly in a standard format[6] that specifies the site location; the year, month, and day

of sampling; and the measurement itself (24-hour or 1-hour value) in micrograms or milligrams per cubic meter ($\mu g/m^3$ or mg/m^3) or parts per million (ppm). A corresponding site file contains descriptive information on the sampling-site environment. EPA edits the submitted data, checking for consistency with acceptable monitoring methods, and other identifying parameters. In the data-editing program, air quality data with extremely high values are flagged. Data that do not pass these checks or that have values exceeding certain predetermined limits are returned to the originating agency via the Regional Office for correction and resubmittal.

Unfortunately, with data sets this large, there are still anomalous measurements that slip through the existing editing and validation procedures. Therefore, there is a need for a simple cost-effective statistical test that can be applied to the air quality data by which to detect, primarily, obvious transcription, keypunch, and measurement errors. Statistical tests do not eliminate, however, the need for more intensive quality assurance at the local level. For example, inadequate calibration procedures or similar problems that result in measurement bias will not be detected by our statistical procedures, which are intended primarily for macroanalysis.

Basic Characteristics of TSP, SO_2, and NO_2 Data

Basic characteristics of the TSP, SO_2, and NO_2 data were considered in selecting the quality control test being used. To begin with, the test was applied to data which were obtained from monitoring instruments that generate one measurement per 24-hour period.[7] For such monitoring methods, EPA recommends that a systematic sampling procedure of once every 6 days, or 61 samples per year, be used at a minimum to collect the data.[8] Such a sampling procedure generates data, which for our purposes, may be considered as approximately independent.

In examining the distributional properties of the data, past research has shown that ambient TSP concentrations are approximately lognormally distributed.[9,10] This is sometimes true for SO_2 and NO_2, also, but is not always the case.

In selecting the quality control tests, the averaging times which correspond to the NAAQS are important. The values of interest are the peak concentrations (24-hour average measurements) for TSP and SO_2, and the annual means for TSP, SO_2, and NO_2.

The final data characteristic of importance is the seasonality of the pollutants. As an example, in some areas of the country, TSP and SO_2 measurements are highest in the winter months and lowest in the summer months. Therefore, the factor of seasonality had to be considered in the selection of the quality control test to minimize this as a possible source of error.

Shewhart Control Chart Test

The Shewhart Control Chart Test[5] can be used to examine both shifts in monthly averages, as well as shifts in the monthly range. From the former it can detect possible multiple errors and from the latter, single anomalous values. In this test the data can be divided up into what Shewhart called rational subgroups.[11] In a manufacturing process the subgroups would most likely relate to the order of production. Ambient air quality measurements