Efficient inspection of data from pollution networks

L.H. Auer, J.P. Mutschlecner

Earth and Environmental Science Group, Los Alamos National Laboratory, Los Alamo, New Mexico, USA

ABSTRACT.

A package of tools for the rapid survey of data from a pollution monitoring network has been developed. It has been designed to help in the recognition of regularities as well as peculiarities in a data set. Critical features are interactive user control, real-time graphic presentation in alternative formats, and simple modification of code to permit new options. The graphics used include movies of the spatial development, time-series, and scatter diagrams modulo the diurnal or weekly cycles.

INTRODUCTION

Data from a network of detectors are crucial to the investigation of the pollution in a region. Without such data it is impossible to study either the spatial or temporal variation in the concentrations. Unfortunately, the very thing that makes such data valuable makes their use difficult: the size of the data base. Even a moderate network with 20 sites, each reporting 10 quantities hourly, generates nearly 2 million numbers annually. How does one "look at" this much data? That is the question we will address in this presentation.

There are numerous numerical methods, for example, multi-variate (Anderson [1]) and cluster analysis (Späth [5]), which attempt to identify patterns in data; however, one of the most powerful tools available is the human eye. In fact, we are so subconsciously dependent on that ability that one of the standard complaints my colleagues raise against multi-variate analysis is, "Come on. I've never seen anything like that in the data". Although one can numerically analyze the correlations in the data from a network, the results are often unsatisfying. First, because people feel that something is missing if they cannot "see the result". Second, this type of analysis is crude in that it does not permit one to easily visualize the details of the temporal and/or spatial variations.

If the information is presented correctly, a person can almost instantly recognize any underlying drifts. The key, of course, is in the words "presented correctly". Humans live in a world of motion and our senses have evolved accordingly. We are most sensitive to things that change. (Note the underlining catches our eye because it is a sudden change in the way the text is being presented.) Like our ancestors we are unable to detect the hidden "truth" in the noisy forest, but if it moves, we pounce on it. Computer technology has advanced so rapidly that we can now present the data, even on a personal computer, in a manner that reinforces the human's innate abilities.

Just showing the data as a movie, however, is not sufficient. For the efficient inspection of a large data set, flexibility in the way the data can be viewed is critical. Pollution quantities, for example, can be displayed either simply as a function of time or plotted against each other to emphasize the relative behavior. Interactive control of the presentation is, therefore, necessary for the efficient inspection of data. When a person notices something, the immediate subconscious urge is to "take a closer look." One needs to be able to switch modes rapidly and easily in order to test ideas as they come into one's head This is exactly what the tools we have developed permit the user to do.

We have not tried to develop the "definitive" analysis tool; rather, what we have created is a set of tools for "effective" and "efficient" preliminary analysis of information. The various options aid the user in the discovery of the interesting features contained in the set. The basic things we want are the recognition of 1) the regularities and 2) the peculiarities. The identification of these is the critical first step in the analysis of a data set. In particular, this is exactly the information we need for the validation of any modeling effort: Does it reproduce the known regularities? Can the model explain the origin of the peculiarities? Further, in themselves the patterns in the data can yield important insights into the nature of the pollution.

SAMPLE DATA SET

The tools for examination of network data being presented here were developed for a cooperative project between Los Alamos National Laboratory and the Intstituto Mexicano del Petroleo, Williams [6]. One of the objectives of this project is to understand the sources and mechanisms which influence the air quality in the Valley of Mexico. Part of the observational data made available to us is a long-term data base from a network of surface stations operated by the Mexican government division, SEDUE, cf. Fig. 1. These stations measure both meteorological quantities, temperature, relative humidity, and wind direction and speed, as well as pollutant concentrations including CO, NO_X, NO_2, O_3, and SO_2.

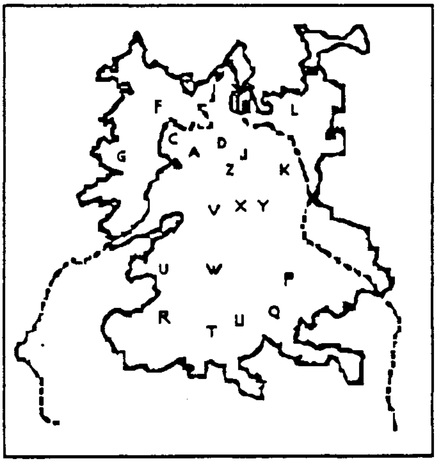

Fig. 1. SEDUE sites in the Valley of Mexico. Urban area marked by solid line.

The development of the current package was stimulated by our initial experience with two Macintosh programs, Cricket Graph [3] and MacSpin [4] which demonstrated to us the importance of rapid interactive graphics in the analysis of SEDUE data. We, like others, discovered that while extremely useful the publicly available packages are not convenient for the analysis of pollution network data chiefly because each is limited in the number of ways available to present the data. Writing a new program offered us several advantages, including the ability to add new tools and to use various computer platforms (Macs, PCs and SUNs).

The SEDUE data set is typical, in that, it is not perfect. There are significant gaps in both the spatial and temporal completeness of the data (Batterman, [2]). Indeed, it was our frustration with the difficulty of applying more traditional techniques like power series and correlation analysis that led us to develop the new tools.

All figures are for February 1991. We should note that we have found it most convenient to work with one month at time: first, because this stays well within memory limits of even modest computers, and second, because it eliminates potential complications due to seasonal effects.

METHODOLOGY

One graphic technique we use for the display of the data is to make real-time "movies" on the display screen of a personal computer or workstation of the time dependence of the variations. The speed with which the figures are presented is under the control of the user. This permits either more careful inspection of the figures or selective copying for inclusion in other documentation. The nature of the time variation is stressed by this type of presentation.

In order to reinforce the visibility of underlying order, the use of scatter diagrams and plots showing the diurnal and weekly cycles is critical. During the time covered by a data set there can be a large range in the concentrations of the pollutants, as we show in Fig. 2a,b for NO_x and NO_2 at the same site. Part of the variation is due to changes in the ambient conditions, e.g. temperature, precipitation, etc., and part due to changes in the source terms. The net result of these different modes of variation is that time series in themselves do not clearly demonstrate the relationships between quantities.

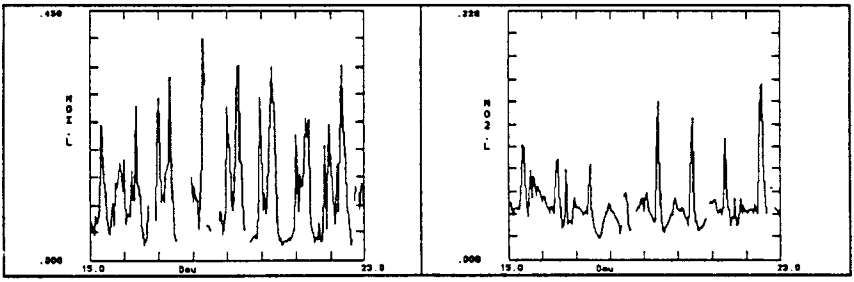

Fig. 2. Time series of a) NO_x and b) NO_2 concentrations for 15 - 23 February 1991 at SEDUE site L. From such output it is difficult to infer relative behavior.

Although the variation in the weather is irregular, there are other absolutely periodic changes which correspond to the time of day, and the day of the week.

The diurnal cycle is caused both by photo-chemistry and by the daily cycle of human activity. The weekly cycle, on the other hand, is the signature of human activity. No common natural phenomenon corresponds to this period, and certainly none that would make Sundays special.

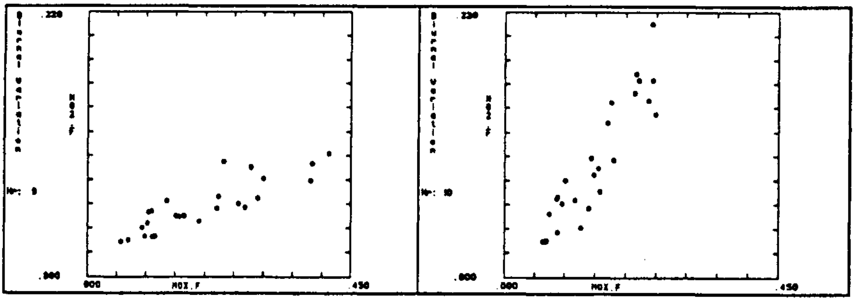

Fig. 3. Scatter diagrams of NO_x and NO_2 at F for ALL days in February but plotted only for a) 0900 and b) 1000. The bands show temporal patterns nearly independent of the pollution levels and ambient conditions.

Plotting scatter diagrams modulo these cycles filters out much of the scatter in the data. Variations due to other factors, such as weather, show up as smearing effects, increasing the randomness in the plots. In one such diagram, Fig. 3a, we show ALL the data for a given time of day, 0900 local time, and plot the concentrations of NO_2 vs. NO_x. Even though there was a variety of weather and pollution during the month, there is a well defined locus into which all the points fall. We see a pattern that at given time of day there is a roughly linear relation between these two chemicals. If we look an hour later, 1000 local, we see again a relatively well defined band but with a different slope because of the photo-chemical conversion of the NO_x mixture into NO_2, Fig. 3b. From this latter figure we see directly that the photo-chemical conversion, on most days, is complete by 10 AM, but is retarded for a few mornings.

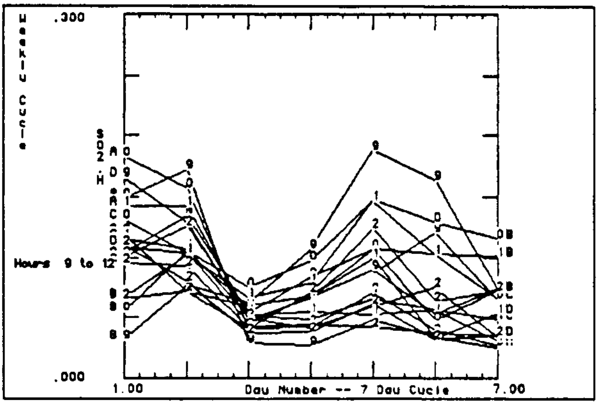

Fig. 4. The weekly cycle for station H. Days are counted starting from the beginning of the data set. Here days 3 are Sundays. Weeks are marked A, B, ..., Hours are marked by last digit in local time, i.e. 1000 is '0'.

We may similarly stress the weekly cycle by plotting the concentration of a pollutant at a given site and time of day vs. the day of the week, e.g. Fig. 4 for SO_2 at station H during February 1991. Because for non-photoactive chemicals the concentration is often not a strong function of the time of day, we may overplot several neighboring hours in the same graph to help reduce noise. The important feature is that we plot MULTIPLE weeks. If there is an intrinsic weekly cycle, presumably due to the variation of the anthropogenic sources, the shape of the curves from each week will be similar. The spread between the curves is indicative of the importance of ambient conditions in the level of pollution, but variations in the sources from one week to another could also be a contributing factor. For the data in Fig. 4 the most significant feature is the deep trough on the third day, which are Sundays in February 1991, followed by a build up of concentration during the week. The drop on Wednesday, day 6, followed by a rise to Fridays, day 1, is significant. Again note that Fig. 4 shows all four weeks and the four hours, 0900 to 1200. The similarity between the weeks during the morning hours is easily seen.

Note, in order to give a realistic impression, all figures have simply been copied from the display screen. Efficiency of the display has been given greater weight than elegance. This is a ``quick-look" system.

TOOLS FOR THE BIG PICTURE.

Given a new set of measurements our first effort must be to obtain an overlook of the behavior during the period. We want to survey all sites as rapidly as possible. Typical questions will include: What were the worst periods? What is the range in pollution among the sites? Which sites vary together? Is there spatial correlation? Can we see evidence of pollution propagating out from one region?

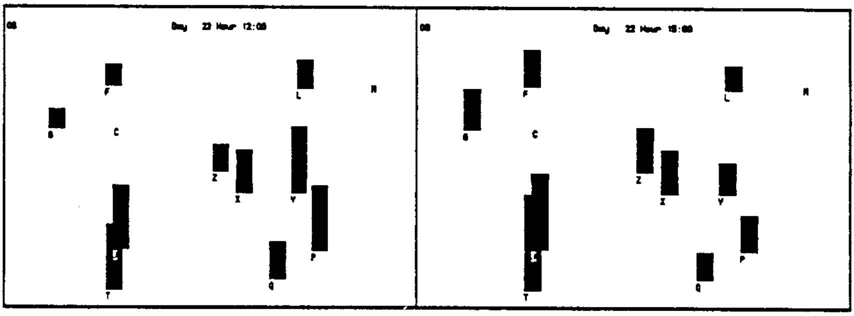

Fig. 5. Frames from a "movie" of the spatio-temporal evolution of O_3. Bar lengths are proportional to concentration at time indicated. From plots like this one can easily see changes in the gradients across the grid. Note the variation of the left (west) versus the right (east) sides in the figures

We have developed two tools for addressing these concerns. In the first, we simply show a movie of the time variation of a selected quantity, cf. Fig. 5a,b for sample snapshots of O_3. Bars, whose lengths are proportional to the instantaneous concentrations, are drawn at the correct relative geographical locations of the reporting stations. These figures compare the concentrations at 1200 and 1500 local time. (The display is actually made as a movie, which may be stopped to permit copying of the graphs for inclusion in documentation such as this paper.) This form of presentation is particularly useful for seeing structure in the temporal and spatial variations. It is not as adequate for establishing a more quantitative estimate of the relative behavior at the various sites. Bars as opposed to contour diagrams are used in this type of presentation because of the relatively low density of sites and the computational cost of interpolation. If these are not an issue, variations could be shown as a color/density movie.

Somewhat more precise comparison is permitted by the tool whose output is shown in Fig. 6. Here all values for a given day are shown in a scatter diagram plotted against the time of day. The average for all the days in the data set, in this case the whole month of February 1991, is shown as a continuous curve. A sequence of such graphs is shown for successive days at a rate which is controllable by the user. From this presentation it is easy to see: 1) which were the worst days, 2) the general form of the diurnal cycle, 3) whether there are sites which are regularly outliers in the distribution, 4) how big the spread is among the stations and 5) if there are sets of sites which show correlation in their variations.

For the efficient inspection of the data, as we noted above, flexibility in the display is critical. For example, one might look first at a movie presentation of O_3, noting which was the worst period, and then look at the scatter diagrams for that period to determine if it was simultaneously worst at all sites. Likewise, one could switch between movies of the pollutant concentrations and the winds to see if the impurity was being transported by the circulation. By providing interactive control and nearly instantaneous display, the user can switch the data being viewed rapidly enough to gain insight into the relative behavior of the various quantities.

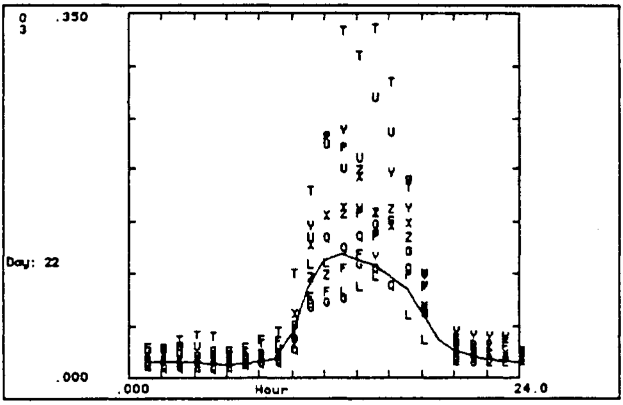

Fig. 6. Rapid display of scatter diagrams for all sites as a function of time on a succession of days is an efficient way of seeing which data cluster together as well as to determine the outliers and worst periods.

TOOLS FOR A CLOSER LOOK

When we seek effective methods for the presentation of data, in a manner that permits higher resolution, we face the fundamental problem that a single graph can only contain a certain amount of information. We argue that for inspection of the data the most effective techniques are 1) scatter diagrams, 2) plots modulo the weekly period, 3) plots and over-plots as a function of time, and 4) scatter diagrams comparing two observations filtered by the diurnal cycle. All of these permit rapid display, but, by showing information from only a few sites at a time, offer somewhat greater detail. In this section we will not only describe the tools, but also present some of the features of the SEDUE data that are revealed by their use.

The first two types of presentation stress the nature of the pollution at a site over the duration of the data base; and, therefore, the comparisons of the plots from different sites can be used to estimate the similarity or disparity across the grid. The latter two tools can be used to directly contrast the variations at sites on an hour-by-hour basis. They are particularly useful for investigating the relative nature of the diurnal cycles at the various sites. Because all the presentations are under interactive user control, the full data set may be quickly compared even though only a limited fraction is shown in any single graph.

At first it might seem that over-plots (i.e. plotting several cycles in the same graph) are not a good way to look at data in as much as they emphasize the extrema. Indeed, for very long periods this can be true, but for relatively short periods, such as the hourly observations during one month we are treating here, such plots are extremely useful. The important "trick" is to make the plots modulo a basic underlying cycle: diurnal or weekly. When this is done, the number of points being shown is reduced and the "clutter" is correspondingly smaller. More importantly, the use of the basic periods acts as an "optimal filter" for revealing patterns. That is, to the extent that there is underlying order, but with varying amplitude, the repeated over-plots quickly reveal this organization. This is true even though there may be substantial amounts of missing data.

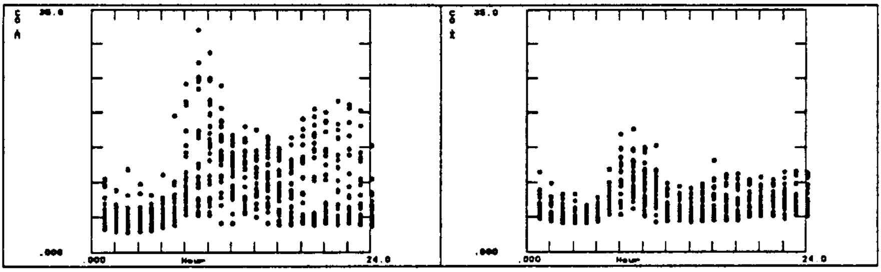

Fig. 7. Scatter diagrams of all the observations in the data set of a quantity are a "signature" of the site. Systematic differences between stations, here a) A and b) X, are easily seen from this type of presentation.

This point is demonstrated in Fig. 7a,b. In these graphs we show all the CO observations during the month of February at the two sites A and X. The existence of diurnal regularity at both sites is unquestionable, but equally important is the difference in appearance of the two plots. Presenting the data in this manner indicates directly the size of the local component in the pollution. That is, here the levels appears to be locally dominated with little exchange between these sites.

A similar use of over-plots to reinforce the underlying pattern has already been shown in Fig. 4 for the weekly period. Although there may be large variations in the pollution, plotting in this manner accentuates the relative day-to-day variation. Over-plotting partially compensates for the more random changes caused by ambient conditions. The eye averages the neighboring hours and smooths out the jumps in the curves. The practical importance of this type of presentation is that by comparing the weekly cycles at stations, one can estimate the similarity in the activities at the sites, cf. Fig. 8 a,b.

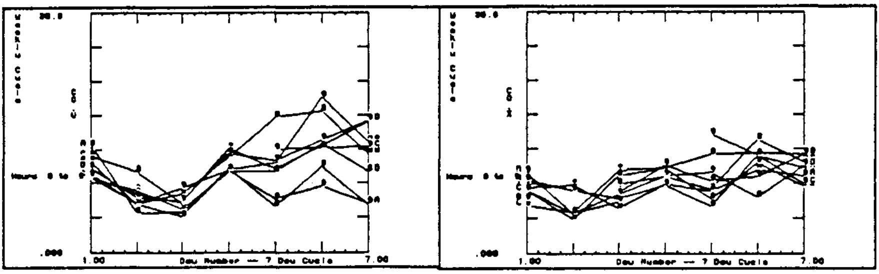

Fig. 8. Weekly cycle plots show the difference in the human activities at the stations. Here (a) site U lies toward the western edge of the grid, while (b) site X is near the center of the city.

The most familiar method for presenting data is, of course, a simple graph against time, i.e. a time series. We augment the utility of this technique by permitting the user control over 1) the time range to be displayed and 2) the stations to be included. (We do not attempt to plot different quantities on the same graph. There are inherent scaling difficulties in doing this, and, further, we have found comparison between different quantities to be dramatically simplified if they are filtered modulo the daily cycle as in the tool described next.) This method of presentation permits the closest examination of the data, both with respect to time and comparison of sites. One can choose to display any temporal sub-segment of the data, as well as which stations to include. While one can, in principle, plot many sites in the same figure, we have found that comparison is clarified by plotting data only from a restricted number of sites in one graph.

For example, in Fig. 9 we show the variation in O_3 during the period 20-23 February for stations T, U, F, and G. Both the agreement and disagreement between T and the other stations are noteworthy. (We should note that the distinction between the curves is much clearer in reality than in Fig. 9 because on the computer each curve can be displayed using a different color.) As one sees from Fig. 5, sites T and U are physically close to each other and are usually in excellent agreement, except on 22 February when ozone at T is roughly 30% more than at U–and more than a factor of two greater than at F and G. The stations F and G are likewise relatively close to each other and show similar temporal evolution. As we noted above, there is a clearly visible gradient in the O_3 with F and G usually less polluted than T and U. On 22 February, the day with the worst O_3, the divergence between the sites is likewise maximal.

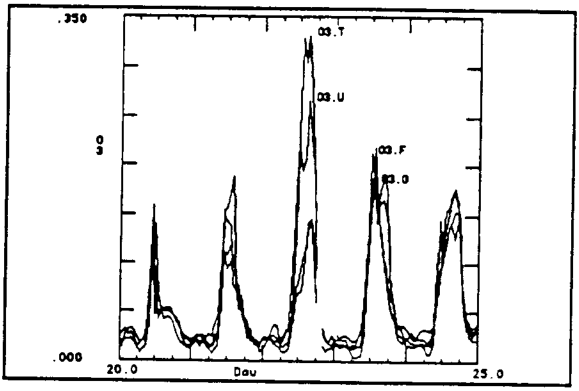

Fig. 9. Time series plots provide the most detail but interactive control over the dates and stations displayed is critical to the readability of the graphs.

Even more remarkable is the behavior in the afternoon of 20 February. Ozone normally peaks at noon and decreases smoothly in the afternoon. But on this day we can clearly see a sharp decrease at noon but with unusual persistence late into the day. What is particularly striking is that we can see this phenomenon occurred almost identically at all the sites. In terms of theoretical model validation the identification of such singular events is particularly useful. If the model can both reproduce the typical variations as well as explain unusual events one will gain confidence in its ability to predict the pollution levels.

In addition to providing detailed plots of the time variations, this tool can also be used to get an overview of the time evolution of a quantity. From Fig. 2b, for example, we can see that the NO_2 levels were somewhat worse after the 19th than before. Used in this way time series plots are an alternative to the scatter diagram as illustrated in Fig. 6, although we would argue that "big picture" technique makes it easier to compare all the sites in one graph. The choice of display depends on personal taste as well as what one is trying to find. The key point is that by providing flexibility, choice has been not dictated to the user.

The purpose of the final tool is to permit comparison of the variations in two quantities. The time-series (Fig. 9) and scatter diagrams against the time-of-day (Fig. 6) permit one to compare the "instantaneous" concentrations at various sites, but do not directly give any information on the relative behavior of two quantities. This is the problem we wish to address: Are the concentrations of quantity 'q1' at site 's1' functionally related to the concentrations of quantity 'q2' at site 's2'? Because our basic goal is to discover the regularities in the data, it is necessary to filter the comparison in an appropriate manner. For both chemical and meteorological quantities the dominant source of variation is the diurnal cycle; therefore, in order to clarify the comparisons between two quantities we remove this major factor simply by plotting at the same time-of-day. We plot scatter diagrams for two quantities for ALL the available observations in the user-selected period of days, stepping through the hours of the time to display the evolution, if any, of the relations between the two quantities.

An example of this type of plot is given in Fig. 3 a,b which shows the regularity of the conversion NO_x into NO_2. Because we are showing all days, we can see from the narrowness of the scatter in the bands that the rate of conversion depends relatively weakly on factors other than the time of day, and, further, that the conversion is essentially complete by 1000 local time no matter what the ambient weather conditions. In terms of model validation the existence of narrow loci as a function of the time-of-day is extremely useful. To be acceptable the model result must lie somewhere in the range of the scatter, no matter what the level of the pollution or weather conditions.

In fact the most perfect regularity found in this data set is demonstrated using this tool and is shown in Fig. 10 a,b for station F. This same relationship is seen, however, at all stations reporting both NO_x and CO.

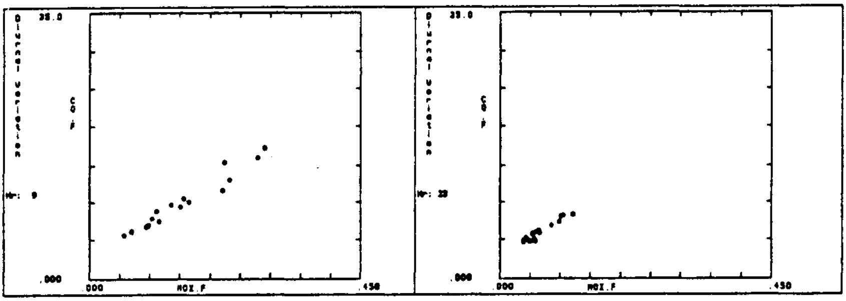

Fig. 10. NO_x and CO for all days are strongly correlated at a site F regardless of the time of day.

The linear proportionality of NO_x and CO proves that they must have both the same sources and sinks. This may provide some control over any errors in the source catalog.

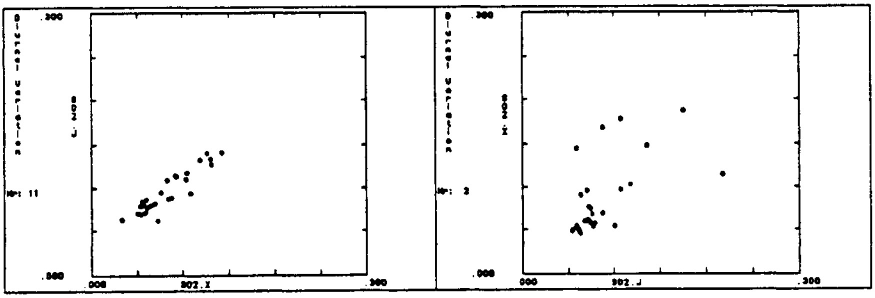

Fig. 11. The slope of the scatter diagram for SO_2 at site J and X is a function of the time of day.

Scatter diagrams for one time-of-day can be equally revealing about the nature of the pollution at two sites. In Fig. 11a,b we show the concentrations of SO_2 at stations J and X, both of which are in the central region of Mexico City. During the day at 1100, Fig. 11a, the concentrations throughout the month are proportional. Can we conclude that this equality is due to mixing? If we contend that, we must then explain why at night at 0200, Fig. 11b, the pollution level is sometimes greater at X than J. This may be due either to a greater efficiency of the mixing during the day or to different production rates in the two regions. In fact, it is a probably a combination of these and other factors.

CONCLUDING REMARKS

In describing our approach to the efficient inspection of data, we have been limited by the fact that the essential machinery we require–interactive graphics–is exactly what the printed page cannot provide. While we can show, as in Fig. 5, the existence of a regular variation, this does not give the reader the actual experience of watching the evolution take place. Further, a printed article cannot convey the importance of user control over the display. In fact, the efficiency of our approach originates in that control: first, in the choice of the type of graphical display to be used; second, in the selection of a subset of the data in order to simplify the appearance and interpretation of the display; and third, in the speed with which the information is scanned.

In itself, simple inspection of data cannot provide answers or explanations, but is an essential first step in obtaining such understanding. The next step must be to attempt to model the phenomenon. Given the regularities seen in the data set, one will be in a position to judge the adequacy of the fit. We argue that the regularities are, in some sense, more important than the details. Speaking figuratively, the model will not predict precisely the ripples on the flag, and they probably do not matter; but one had better get the flag blowing in the right direction for a given wind.

ACKNOWLEDGEMENT

This work has been supported by the Los Alamos National Laboratory which is operated by the University of California for the U.S.Department of Energy.

REFERENCES

1. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, New York, Wiley, 1984
2. Batterman, S.A. 'Optimal Estimators for Ambient Air Quality Levels', *Atmospheric Environment*, Vol. 26A, No.1, pp 113-123, 1992.
3. Cricket Graph, Malvern, PA, Cricket Software, Inc., 1987
4. MacSpin, Austin, TX, D^2 Software Inc., 1989
5. Späth, H. and Goldschmidt, J. *Cluster Dissection and Analysis*, Ellis Horwood Ltd., Chichester, 1985
6. William, M, Ruiz, M., McNair, L. et al. 'Development and Testing of an Air Quality Model for Mexico City', *Health and Ecological Effects*, Air & Waste Management Association, Pittsburgh, 1992.