it affect the analysis of extreme values? If so, how could it be incorporated into the analysis? I would welcome Professor Smith's views on these questions.

The Clustering Method

The method for forming clusters used in the paper is essentially the single link method. This has the possible disadvantage that two clusters six days apart with a single exceedance between them could be merged (Gordon, 1981). An agglomerative sum of squares method might be preferable given that the aim is to obtain compact clusters. Inspection of the dendogram could help with the choice of a cluster interval.

ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research under Contract no. N-00014-88-K-0265. The author is grateful to Richard L. Smith for providing the ozone data on which the analyses are based, to Peter Guttorp for very helpful discussions and to Michael A. Newton for useful comments.

ADDITIONAL REFERENCES

Akman, V. E. and Raftery, A. E. (1986a). Bayes factors for non-homogeneous Poisson processes with vague prior information. *J. Roy. Statist. Soc. Ser. B* **48** 322–329.

Akman, V. E. and Raftery, A. E. (1986b). Asymptotic inference for a change-point Poisson process. *Ann. Statist.* **14** 1583–1590.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.* **82** 112–122.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events*. Methuen, London.

Gordon, A. D. (1981). *Classification.* Chapman and Hall, London.

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource (with discussion). *Appl. Statist.* **38** 1–50.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford.

Raftery, A. E. (1988). Analysis of a simple debugging model. *Appl. Statist.* **37** 12–22.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73** 85–89.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 172–212.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

Comment

David Fairley

I read Dr. Smith's analysis with great interest. I was unfamiliar with the point-process approach and was impressed by how elegantly it encompasses the other extreme value theory methods. At the same time, I was impressed by how hard it is to apply this theory to the problem at hand. Beyond his expertise in extreme value theory, Dr. Smith has clearly taken great pains to take the practical issues like dependence and seasonality into account. Nevertheless, the results appear somewhat weak given the power of the theory that went into them.

My work at the San Francisco Bay Area Air Quality Management District (affectionately known as the BAAQMD) has given me an applied orientation to the analysis of ozone trends, so most of my comments are directed more toward specific practical problems than at theoretical ones. However, some of the practical considerations involved in ozone trend detection suggest ways in which the theory might be usefully extended.

In Section 2, I will try to define the underlying problem as clearly as possible. This leads to several possible ways to extend the analysis and make it more powerful (Section 3). Beforehand, though, I will suggest several reasons why the data appear to me even messier than Dr. Smith suggests.

1. MORE COMPLICATIONS

Dr. Smith dealt with some important practical difficulties, including the short-term dependence, seasonality and missing values. Several other factors complicate the picture.

Measurement error is generally overlooked in ozone trend studies, but it can affect the results substantially. The instrument measuring ozone can only measure to the nearest pphm, so the actual data are discrete. There has been more than one method for measuring ozone. The instruments used at the BAAQMD up through the mid '70's measured all

David Fairley is the statistician at the Bay Area Air Quality Management District, 939 Ellis Street, San Francisco, California 94109.

Disclaimer: The opinions expressed here are the author's own and not necessarily those of the Board of Directors or staff of the Bay Area Air Quality Management District.

oxidants instead of ozone. The earlier method tended to read higher at most sites but lower at others. Even if a single instrument has been used throughout the interval of study, there is the issue of instrument calibration. In 1976, for example, the California State Air Resources Board mandated recalibration resulting in a 4% increase in the ozone readings. In 1979, they mandated another recalibration, resulting in a 4% decrease. Oddly enough, 1978 had the highest recorded ozone levels in recent history. Then there is the issue of biases caused by manual recording of the ozone. In the Bay Area, measurements are still recorded on a strip chart and hourly averages integrated by eye by the station operator. Needless to say, errors can occur. In particular, some operators tend to read higher than others. Finally, a gradual improvement in quality control has taken place. One critic has gone so far as to suggest that all the improvements in air quality in recent years have actually been improvements in quality control that reduced the number of extreme values!

In addition to measurement error, there can be local, real changes in ozone levels which one would probably not want to treat as a trend. Ozone is consumed by NO, which is emitted by cars. If a monitoring station is located on a street with a heavy commute traffic, its ozone levels with be lower than another nearby station on a quieter street. If the amount of traffic near a site has increased dramatically over the study period, it may be responsible for what is only a very localized downtrend in ozone concentrations.

2. AN ATTEMPT TO FORMULATE THE PROBLEM

Given the multitude of difficulties in analyzing the data, it is worthwhile to be as clear and precise as possible abut what questions we are trying to answer.

High ozone concentrations cause respiratory problems. The U.S. EPA standard is that the highest hourly average ozone shall not exceed 12 pphm more than once a year on the average at any particular location (i.e., the 4th highest of the daily high hour ozone concentrations over a 3-year period must be less than or equal to 12 pphm). This standard is set to ensure reasonable safety to the most sensitive populations such as children and asthmatics. Although this standard is exceeded frequently in Houston, it is not in most other places in the U.S. Thus, typically, it is the high percentiles of the ozone distribution which are of direct regulatory interest.

The ozone standard is written in terms of a percentile of the empirical distribution of 3 year's worth of high hour ozone values. The reason for using 3 years instead of 1 is the recognition that weather affects ozone concentrations. A sequence of unusually warm, stagnant days may produce several exceedances even in a relatively clean area. The underlying intent of the regulations is to assess the marginal impact of anthropogenic emissions on ozone. That is, the intent is to find what the ozone distribution would be if one could integrate the joint distribution of ozone and weather over all weather conditions (i.e., running 100 years' worth of weather with the same emissions) yielding the marginal ozone distribution.

Ozone is measured at fixed monitoring sites, but the intent of the law is to regulate dosages. In fact, the law states that the standard shall not be exceeded *at any point*, not just at the monitoring site, to ensure that no individual is exposed to ozone concentrations above 12 pphm more than once a year.

Finally, the law is certainly intended to affect the actual ozone concentrations, not the measured.

Thus the regulatory interest focuses on a conceptual but unobservable distribution: the marginal distribution of the surface of true high hour average ozone in a region. In contrast, the data set we have to work with is the measured ozone at a finite number of not necessarily representative sites under a particular set of weather conditions.

There are two main statistical issues involved here: (1) estimating high percentiles of this conceptual distribution, particularly the upper $1/365$th percentile (called the "design value"); and (2) estimating the changes over time in these percentiles.

3. IMPLICATIONS AND EXTENSIONS

The discussion in Section 2 has several implications regarding how the present analysis might be extended. Dr. Smith alluded to one, relating ozone concentrations to weather covariates with a regression. If such a regression function were estimated, it could be used to analyze trends in ozone concentrations under one or several sets of weather conditions. The formula could be integrated over the distribution of weather covariates to yield an estimate of the upper $1/365$th percentile of the marginal ozone distribution.

Under realistic conditions, measurement error will cause an upward bias in estimates of the extreme percentiles: Suppose that Z is the true high hour ozone and $Y = Z + U$ is the measured value, where Z and Y have distribution functions F_Z and F_Y, respectively. Suppose that U is independent of Z with a density function f_U, and that $E(U) = 0$. Finally assume that $F_Z(y)$ is concave for values of y above some point y_0 with $F_Z(y_0) < 1$. (This includes the normal, lognormal, exponential, and any Generalized Extreme Value distribution with $k < 1$, i.e., all the fitted distributions.) Then

$$F_Y(y) = \int F_Z(y-u) f_U(u)\, du$$
$$= E\, F_Z(y-U) \le F_Z(y - E(U)) = F_Z(y)$$

for values of y sufficiently large that most of the distribution of U is above $y_0 - y$, where the inequality is from Jensen's inequality. Thus Y is stochastically larger than Z. This bias is not a serious consideration in trend studies, but it is when trying to estimate the design value. It would be interesting to hear how this bias can be adjusted for. But perhaps better quality control and electronic recording of measurements directly from the ozone instrument into the computer would be the most effective way to minimize this bias, as well as other measurement biases such as operator bias.

Even if the measurements were perfect, the question remains whether ozone levels at a site are representative of ozone levels beyond a very localized (i.e., one or two block) area around the site. Because ozone monitoring stations are generally miles apart, rather than blocks, a given station is generally taken to represent the ozone in a reasonably large region around the site. As mentioned previously, heavy traffic near a site can give unrepresentatively low values of ozone, so a trend toward heavier traffic may be mistaken for an improvement in ozone over a larger region. There is no way to draw legitimate inferences about the ozone in a region based on data from a single monitoring station.

One solution to this problem would be spot sampling in a neighborhood of the site. Because ozone is highly correlated spatially, only a few observations taken at randomly selected spots in the region that the monitoring station is supposed to represent could establish a baseline for drawing inferences to this larger region. This random sampling could be repeated on a yearly basis to determine whether any trend at the monitoring site represents a region-wide trend.

Since the environmental interest is in the ozone levels in a region and not just at the monitoring site, some form of averaging or combining of the data from nearby sites would be appropriate. A simple alternative is to take the average of the high hour ozone values in the neighborhood of a site. A more sophisticated idea is to use empirical Bayes estimation either directly on the percentile estimates at nearby sites, or else on the extreme value parameters. Averaging should diminish the effect of measurement error and also "random" local factors. Empirical Bayes estimation is a nice way to add power in the case where the tail distributions at neaby sites are similar (which I think should often be the case), while not forcing the unrealistic assumption of exact equality.

Another issue is how to estimate whether an air quality district is in compliance with the ozone standard. The more monitoring sites a district has, the more likely it is to be out of compliance! How can the number of monitoring sites be adjusted for? A related issue is monitor location; one district may look better than another because its monitors are cleverly located in the cleanest spots in the district. What should regulators do about this?

For trend studies, there is always a choice of which statistic to follow. The mean or median of the ozone distribution should be more precisely estimated than the upper percentiles. However Dr. Smith found that in Houston there is essentially no trend in these location parameters (nor is there in the San Francisco Bay Area). More to the point, a trend (or lack of trend) in these parameters doesn't necessarily translate into a corresponding trend in the upper percentiles. Dr. Smith found a significant downtrend in the most extreme percentiles of the Houston ozone distribution which, while not directly related to the ozone standard, is still an indication of some improvement in air quality. On the other hand, Table 5 shows an increase in the rate of exceedances of the federal ozone standard. Thus the air quality is getting worse and better at the same time! The moral is that the whole upper tail of the ozone distribution must be considered in a trend study.

A similar point can be made that trends in a whole region must be considered. In the Los Angeles area, it was found that the air quality regulations caused a delay in ozone formation, rather than a decrease, so that upwind stations got cleaner, while downwind stations got dirtier. This suggests also that there may be differing trends for different wind flow patterns, an idea which is now being studied by the California Air Resources Board.

A final comment about statistical versus practical significance. The estimated decrease in Houston's extreme ozone percentiles is statistically significant, but their air is still far from healthy. Conversely, if they made major progress toward cleaner air, a downtrend would be easy to detect, even with naive statistical techniques. In many areas of the United States, however, the air is near the borderline of the federal standard and further ozone reductions will be hard to achieve. At the same time, legislators, industry, environmentalists and the public in general will be clamoring for information on whether and how quickly the air is improving. Given the difficulties discussed so far, it may be a statistician's duty to inform them that such a determination is impossible to make.