Setting environmental standards: A statistician's perspective

Peter Guttorp

ABSTRACT

Governmental environmental protection is commonly implemented by specifying a standard value of pollution, measured or actual, not to be exceeded. This article considers the standard for ozone pollution in the United States, interprets it using a hypothesis testing framework, and shows (in a simplified setting) how a statistician could implement this standard. The statistician's implementation is contrasted with the implementation by the U.S. Environmental Protection Agency. Some of the issues raised by these contrasting implementations are illustrated using ozone data from three areas in the United States. This article also examines potential biases in using data collected for standard compliance monitoring purposes to assess the health effects of ozone.

INTRODUCTION

To protect the population from adverse health effects caused by the pollution of air, water, and soil, many governments choose to set a standard, i.e., a value (such as daily average concentration of a particular pollutant) not to be exceeded or to be exceeded only infrequently. In addition to the standard, an implementation rule, indicating under what circumstances the standard will be considered violated, is commonly part of the regulations. Penalties and other procedures for dealing with regions out of compliance with the standard may also be part of the legislation.

This article considers the U.S. National Ambient Air Quality Standards (NAAQS) and particularly the standard for ozone. Because of a complicated legal issue, there are currently two ozone standards in effect: the 1-hr standard introduced in 1990, and the 8-hr standard introduced in 1997. The focus here is on the older one. This standard requires states to maintain an air quality such that the expected annual number of daily maximum hourly ozone averages exceeding 0.12 ppm is equal to or less than 1. The implementation rule allows the state no more than three daily maximum hourly average measurements in excess of 0.12 ppm during 3 yr at

DOI:10.1306/eg.06070504026

AUTHOR

PETER GUTTORP ~ *Department of Statistics, Box 354322, University of Washington, Seattle, Washington 98195-4322; peter@stat.washington.edu*

Peter Guttorp is professor and chair of the Department of Statistics at the University of Washington. He is also the director of the National Research Center for Statistics and the Environment, a multidisciplinary group of environmetricians. His research is focused on the methodology for scientific problems in atmospheric science, environmental science, and hematology. He has published two monographs and numerous scientific articles.

ACKNOWLEDGEMENTS

Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the agency's required peer and policy review and, therefore, does not necessarily reflect the views of the agency, and no official endorsement should be inferred. The author is grateful to the National Research Center for Statistics and the Environment standards group, in particular Mary Lou Thompson, Larry Cox, and Paul Sampson, for many illuminating discussions and Anthony Nguyen for computational help.

each approved monitoring site. The consequences of violating the standard depend on the severity of the noncompliance: if the measurements placing the state out of compliance exceed 0.18 ppm, the state must develop a comprehensive air quality model, demonstrate that the model can reproduce current data, and develop a plan for air quality improvement, which, according to the model, eventually will put the state in compliance.

Because ozone is a secondary pollutant, resulting from photochemical reactions in the atmosphere involving nitrous oxides and volatile organic compounds, any ozone abatement plan will need to address the primary pollutants. An additional consequence is that there are no point sources of ozone, and peak ozone concentrations typically occur downwind from the primary pollution sources.

Previous work looking at statistical aspects of environmental standards include Watson and Downing (1976), O'Brien et al. (1991), Symons et al. (1993), Barnett and O'Hagan (1997), Carbonez et al. (1999), and Cox et al. (1999). This article begins by outlining a statistician's first approach to the problem of determining compliance with the ozone standard. Such an approach is naturally phrased as a fairly standard hypothesis testing problem, but the choice of null hypothesis may be surprising at first. The section on the EPA compliance criterion analyzes in a similar fashion the implementation rule of the U.S. Environmental Protection Agency (EPA) when viewed as a hypothesis testing rule. The section on data analysis includes data analyses from different parts of the United States and a discussion of the validity of the simplifying assumptions made in the preceding two sections. A statistical framework for setting environmental standards has been developed by Barnett and O'Hagan (1997), and how the framework can be applied to the United States ozone standard is outlined in the section on the Barnett-O'Hagan setup. Finally, the section on network monitoring bias looks at the potential bias of using compliance monitoring networks to assess the health effects of air pollution.

A STATISTICAL SETUP

Consider a monitoring network with I sites, and let $N_{i,t}$ denote the number of daily maximum hourly ozone averages in excess of 0.12 ppm at site i ($i = 1, \ldots, I$) during year t ($t = 1, \ldots, T$). Let $\theta = EN_{i,t}$, where E stands for expected value. For simplicity, let us first consider the case $I = 1$. Then the United States 1-hr ozone standard requires that $\theta \leq 1$. A natural approach (at least for a classically schooled statistician) to the decision as to whether the standard has been met is a hypothesis test. Because the Clean Air Act (CAA) requires the EPA first and foremost to protect people from the adverse health effects of air pollution, the more serious error would be to declare a region in compliance when it is not. According to the classical Neyman-Pearson setup (e.g., Bickel and Doksum, 1977, section 5.2), the null hypothesis should be that hypothesis for which false rejection is the more serious error (than false acceptance). Hence, in the Neyman-Pearson setup, the null hypothesis must be that of noncompliance, i.e., testing H_0: $\theta > 1$ against H_A: $\theta \leq 1$. Of course, from a practical point of view, although this null hypothesis is acceptable to the EPA, it would be of some concern to individual states that run the risk of being falsely accused of violating the standard.

Assume now that different days of the year are independent. If the monitoring site lies on the boundary between the null and alternative hypotheses, i.e., has $\theta = 1$, we would have $N_{1,t} \sim \text{Bin}(365, 1/365)$, where Bin stands for the binomial distribution (Bickel and Doksum, 1977, A 13.1), i.e.,

$$P(N_{i,t} = k) = \binom{365}{k}\left(\frac{1}{365}\right)^k\left(\frac{364}{365}\right)^{365-k}$$

Because the binomial distribution is an exponential family (Bickel and Doksum, 1977, section 2.3) with a monotone likelihood ratio, the probability of type I error (rejecting H_0 when it is true) is bounded among all $\theta > 1$ by the value when $\theta = 1$. If we now, as the EPA implementation rule requires, base the decision on $T = 3$ yr of data, we have $N = \sum_{t=1}^{3} N_{1,t} \sim \text{Bin}(3 \times 365, 1/365)$, or, to a very good approximation, a Poisson distribution with parameter 3 (Bickel and Doksum, 1977, A.13.9), abbreviated Po(3), i.e.,

$$P(N = k) \approx \frac{3^k}{k!}e^{-3}$$

The optimal test (Bickel and Doksum, 1977, section 6.2) is to reject for small values of N, and a level 0.05 test rejects only if $N = 0$. In other words, from the Neyman-Pearson testing point of view, any exceedance of 0.12 ppm during a 3-yr period would render a site in violation of the standard. Using the level 0.05 is, as always in hypothesis testing, an arbitrary

choice, which must be made by the standard-setting agency. Although this test is the most powerful test at level 0.05, it is not actually very powerful. For example, if $\theta = 0.4$, the probability of declaring a state in compliance (based on a single monitor observed for 3 yr) is only 0.30.

Considering now I independent sites, sufficiency suggests basing a test on $N^I = \sum_{i=1}^{I}\sum_{t=1}^{3} N_{i,3} \sim \text{Po}(3I)$. Again, the level α test would reject for values of the test statistic below a critical value $C_{I,\alpha}$, chosen so that $P(N^I \leq C_{I,\alpha}) \leq \alpha$.

In the analysis in this section (at least) three simplifying assumptions have been made: that $N_{i,t}$ is an observable random variable, that subsequent days are independent, and that different sites in the state are independent. These assumptions are discussed in the section on data analysis.

THE EPA COMPLIANCE CRITERION

Following the same line of thought as in the previous section, first consider the EPA implementation rule for a single site in a state. The rule declares a site in compliance whenever $N \leq 3$, which has probability $\alpha = 0.647$ when $\theta = 1$ under the assumption of consecutive daily maxima being independent. Because no statistician would even consider values of α this high, one may argue that the EPA are not performing their mission under the CAA: given that the CAA requires the EPA to protect public health and that the agency has decided that 0.12 ppm daily maximum hourly average is a limit above which serious health risks to the public occur, the agency appears to make type I errors much too frequently under their implementation rule.

One naturally wonders how the implementation rule was arrived at. The explanation in the regulation (Title 40 of U.S. Code of Federal Regulations Part 50, Appendix H) says:

> The ozone standard states that the expected number of exceedances per year must be less than or equal to 1. The statistical term ''expected number'' is basically an arithmetic average. The following example explains what it would mean for an area to be in compliance with this type of standard. Suppose a monitoring station records a valid daily maximum hourly average ozone value for every day of the year during the past 3 years. At the end of each year, the number of days with maximum hourly concentrations above 120 ppb is determined and this number is averaged with the results of previous years. As long as this average remains less than or equal to 1, the area is in compliance.

In other words, the quoted section of the U.S. Code requires the law of large numbers to be applied to $n = 3$.

For a region with more than one site, the EPA implementation rule uses the test statistic $T^I = \max_{i \leq I} \sum_{t=1}^{3} N_{i,t}$, again rejecting H_0 if $T^I \leq 3$. For example, assuming again spatially independent sites, we find for $I = 7$ that $\alpha = 0.05$. The corresponding rule from the section on the statistical setup would be to reject when $N^I \leq 13$, regardless of where in the network the violations have occurred. It should be noted here that the calculation is made assuming that all the sites have $\theta = 1$, so it would be quite unlikely, for example, that one site would have 13 violations, and that all the others have none. In fact, using a simple multinomial calculation, with a frequency of about 0.36, the maximum number of violations at any of the seven sites, given that 13 violations occurred, would be three, so both implementations agree about one-third of the time.

The power in a network with seven independent sites, using the optimal test from the section on statistical setup, reaches 0.95 at $\theta = 0.4$; for instance, a state with an average of one violation every 2.5 yr at each site can be reasonably comfortable that it will not falsely be declared in violation of the standard.

DATA ANALYSIS

This section considers data from three heavily polluted regions in the United States: the Chicago area in Illinois, the South Coast region of California, and the Houston area in Texas. Previous analyses (e.g., Carroll et al., 1997; Cox et al., 1999) have indicated that a square root transformation frequently has the effect of symmetrizing the daily maximum of hourly ozone data, making a Gaussian assumption reasonable (even in the tails). The data are available from the AIRS database (Chicago and Houston; http://www.epa.gov/ttn/airs/airsaqs/index.htm) and from the California Air Resources Board (South Coast California; http://www.arb.ca.gov/aqd/aqd.htm). Table 1 contains summary statistics for the three data sets. The EPA defines

Table 1. Regional Ozone Data, 1989–1991

Region	Mean Ozone Levels*	Standard Deviation of Ozone Levels*	Number of Monitoring Stations	Number of Exceedances of 0.12 ppm	Number of Days Monitored
Chicago	0.218	0.043	10	15	642
South Coast (California)	0.250	0.068	8	661	1095
Houston	0.254	0.072	8	265	1095

*Calculated on square root scale (raw data in ppm).

the ozone season to be the entire year in California and Texas, and April 1–October 31 in Illinois.

If, as suggested above, the square root of the daily maximum of hourly ozone has a Gaussian distribution with mean μ and standard deviation σ, we have the following:

$$P\ (\text{exceedance of level } c) = \left(1 - \Phi\left(\frac{\sqrt{c} - \mu}{\sigma}\right)\right) \Big/ \Phi\left(\frac{\mu}{\sigma}\right) \tag{1}$$

where the denominator arises from the fact that ozone measurements must be positive. Using the standard deviation for the Houston network, a simple Gaussian calculation shows that one expected exceedance (for a single station) would correspond to a mean of 0.146 on the square root scale, or about 0.022 ppm on the raw scale. Hence, to bring Houston into compliance, the average daily maximum hourly readings must be reduced by a factor of 3 from the current average of 0.066 ppm. Of course, corrective action that reduces only high readings may also be possible.

The considerations so far in the article have all assumed (at least implicitly) that the quantity $N_{i,t}$, the number of exceedances at site i in year t, is an observable random variable, i.e., that we can determine without error the number of exceedances of a given level at a site from the measured daily maximum hourly ozone averages. Strictly speaking, we cannot because the measurements are made with error. To take the measurement error into account, we need to make a conditional calculation. Assume, for simplicity, a Gaussian additive measurement model on the square root scale, namely, $Y = Z + \varepsilon$, where Y is the observed square root daily maximum hourly ozone average; Z is the square root of true maximum daily hourly ozone average, assumed $N(\mu, \sigma^2)$; and ε is an independent measurement error, assumed $N(0, \tau^2)$. Here, σ^2 corresponds to the natural variability of the ozone field, and τ^2 corresponds to the uncertainty caused by imprecise measurement techniques. Then, we have, using a standard regression calculation for the case $\mu = \sqrt{0.12}$, that

$$\begin{aligned} &P(Z > \sqrt{0.12} | Z \geq 0, Y = y) \\ &\quad = P(\varepsilon < y - \sqrt{0.12} | \varepsilon \leq y, Y = y) \\ &\quad = \Phi\left(\frac{y - \sqrt{0.12}}{\sigma}\left(\frac{\Delta^2}{\sqrt{\Delta^2 + 1}}\right)\right) \Big/ \Phi\left(\frac{y}{\sigma}\left(\frac{\Delta^2}{\sqrt{\Delta^2 + 1}}\right)\right) \end{aligned} \tag{2}$$

where $\Delta^2 = \sigma^2/\tau^2$ is the signal-to-noise ratio. Equation (2) roughly corresponds to multiplying the standard deviation of the underlying pollution field by a factor of $\left(\frac{\sqrt{\Delta^2+1}}{\Delta^2}\right)$. For the values used below, the denominator is approximately 1.

The analysis in Cox et al (1999) for the California Central Valley data indicates that the standard deviation τ of the measurement errors for common instruments are about 0.020–0.027 on the square root ppm scale, corresponding to an error standard deviation of the raw measurements of about 0.002–0.003 ppm at a mean level of 0.12 ppm. Comparing these values with those in Table 1 indicates that the measurement error is a fairly large proportion of the observed variability. Using $\tau^2 = 0.000415$, a value near the lower range above, yields $\sigma^2 = 0.00381$ for the South Coast California data, and the multiplier $\left(\frac{\sqrt{\Delta^2+1}}{\Delta^2}\right)$ is equal to 0.35. Thus, taking a measurement reduces the uncertainty by a factor of 3 (but does not eliminate it). Figure 1 shows the conditional probability, given an observation of y, that the true field is actually above 0.12 ppm. For this probability to be larger than 0.95, we need an actual reading of at least 0.146 ppm. To make the probability larger than 0.99, we need to observe at least 0.157 ppm.

The assumption of independent and identically distributed data is overly simplistic. First, it fails to consider the seasonal distribution of ozone, which is very pronounced in the data considered here. For example,

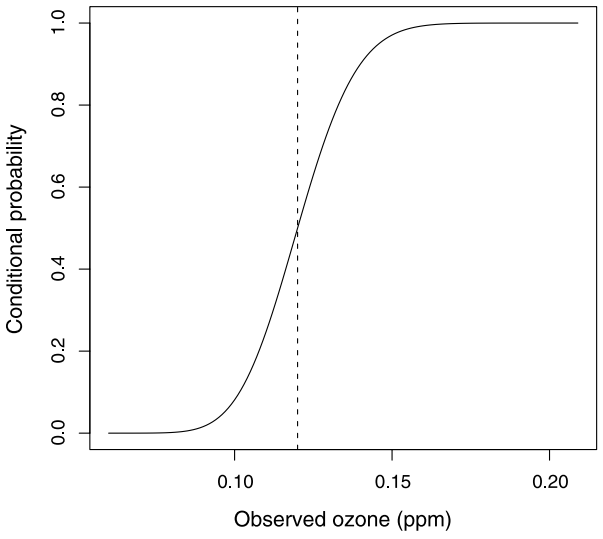

Figure 1. The conditional probability of the true concentration being above 0.12 ppm, given that the observed concentration is that shown in the *x*-axis. The parameter values chosen correspond to values suitable for the South Coast region, California. The dotted line is the value 0.12 indicated in the United States 1-hr ozone standard.

in the California Southern Coast data, the ozone levels are lower in the winter and higher in the summer. This seasonal effect can be dealt with in a more realistic fashion using time-varying mean and variance. Of more serious concern, perhaps, is the fact that the time series of daily maximum hourly average ozone shows some autocorrelation. Data analysis indicates that an autoregressive model of order 2 (Percival and Walden, 1993, p. 44) can account for most of the autocorrelation. The calculations for single-station exceedances can be redone, using simulation techniques, for a more realistic model.

Finally, the spatial correlations need to be considered. In the Chicago data set, the site-to-site correlations are 0.7 or higher. Hence, the calculations earlier in the article, assuming spatially independent stations, are not valid for the Chicago network. Simulation studies, matching the distribution of hourly maxima over the network with independent hourly maxima, indicate that the 10-station networks correspond to about two independent stations. Hence, regional spatially expressed standards would be preferable to the current formulation.

THE BARNETT-O'HAGAN SETUP

In a report written for the Royal Commission on the Environment in the United Kingdom and subsequently published as a book, Barnett and O'Hagan (1997) developed a framework for the statistical implementation of environmental standards. They distinguished between ideal standards, setting limits on the true pollution field, and realizable standards, set in terms of actual measurements. Ideal standards (the United States ozone standard is an example) are a natural approach to standard setting in that they can be related to or even based on the scientific evidence regarding health effects, crop damage, etc. However, it is impossible to implement an ideal standard. In the United States ozone case, we cannot measure the number of exceedances everywhere in the state, much less measure the expected value of this random variable. Thus, realizable standards are much easier to implement (both politically and practically) because they specify exactly what measurements constitute a violation of the standard. The downside is that it is very difficult to relate a realizable standard to the actual pollution field and consequent health effects.

It is natural to seek a compromise between these two extremes. Barnett and O'Hagan (1997) suggest a statistical implementation of an ideal standard, in their terminology, a statistically verifiable ideal standard. In the case of the United States ozone standard, such a standard amounts to specifying statistical-quality parameters for deciding whether a given region is in compliance with the standard. In the testing setup, a natural approach is to fix the type I and type II errors, the former at a value beyond which health effects are serious, and the latter at a value for which there is no evidence of health effects or at a value corresponding to peak background levels. There would be a gray area in which values are neither safe nor seriously harmful.

Another approach would be to use a Bayesian analysis, in which the statistical task is to produce a distribution of θ from which one can then judge the probability that a region is in violation of the standard. The Bayesian approach requires specifying a prior distribution, which could be based on historical data or on expert consensus.

NETWORK MONITORING BIAS

Each state is responsible for monitoring compliance with the standards in the CAA. To this effect, they operate monitoring networks, which have to be approved by the local EPA authorities. Because the network is primarily aimed at finding large values of air pollution, a site that consistently shows lower values than another is likely to be closed down. Hence, the monitoring network setup keeps changing over time, with sites selected based on high values instead of a random or systematic fashion.

The consequence of using compliance monitoring networks to study health effects can be serious. Most health effect studies (e.g., Thomas [2000]) take the ambient measurements closest to an individual's home and/or workplace as a surrogate for exposure. Clearly, if the ambient concentration measurements are from data chosen to find peaks in the mean spatial field, the exposure of an individual may be overestimated, resulting in an underestimation of the health effects of exposure to a given level of pollution. This network bias is a potentially very serious bias, particularly because the relative risk estimates in environmental epidemiology are commonly close to 1. Studies using personal monitors may be helpful to assess more precisely the health effects of a given exposure. Current technology, however, produces rather unwieldy monitors, which are likely to affect personal behaviour.

REFERENCES CITED

Barnett, V., and A. O'Hagan, 1997, Setting environmental standards: The statistical approach to handling uncertainty and variation: London, Chapman & Hall, xi + 111 p.

Bickel, P. J., and K. A. Doksum, 1977, Mathematical statistics: San Francisco, Holden-Day, 493 p.

Carbonez, A., A. H. El-Shaarawi, and J. L. Teugels, 1999, Maximum microbiological contaminant levels, Environmetrics, v. 12, p. 79–86.

Carroll, R. J., R. Chen, E. I. George, T. H. Li, H. J. Newton, H. Schmiediche, and N. Wang, 1997, Trends in ozone exposure in Harris County, Texas: Journal of the American Statistical Association, v. 92, p. 392–415.

Cox, L. H., P. Guttorp, P. D. Sampson, D. C. Caccia, and M.-L. Thompson, 1999, A preliminary statistical examination of the effects of uncertainty and variability on environmental regulatory criteria for ozone, *in* Novartis Foundation Symposium 220, Environmental statistics: analysing data for environmental policy: Chichester, Wiley, p. 122–43.

O'Brien, W., B. K. Sinha, and W. P. Smith, 1991, A statistical procedure to evaluate cleanup standards: Journal of Chemometrics, v. 5, p. 249–61.

Percival, D. E., and A. T. Walden, 1993, Spectral analysis for physical applications: Cambridge, Cambridge University Press, xxvii + 583 p.

Symons, M. J., C.-C. Chen, and M. R. Flynn, 1993, Bayesian nonparametrics for compliance to exposure standards: Journal of the American Statistical Association, v. 88, p. 1237–40.

Thomas, D. C., 2000, Some contributions of statistics to environmental epidemiology: Journal of the American Statistical Association, v. 95, p. 315–19.

Watson, W. D., and P. B. Downing, 1976, Enforcement of environmental standards and the central limit theorem: Journal of the American Statistical Association, v. 71, p. 567–73.