NISS

Model Selection in Environmental Statistics

Richard L. Smith

Technical Report Number 32
June, 1995

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

MODEL SELECTION IN ENVIRONMENTAL STATISTICS

RICHARD L. SMITH [1]

June 19 1995

Paper presented at the International Workshop on Model Uncertainty and Model Robustness, Bath (England), June 30 - July 2, 1995.

Abstract

Environmental problems have become a major focus for modern applied statistics, and there are a number of reasons why model selection issues tend to be of particular importance in analysing environmental data. One reason is that there are often vast amounts of data to be analysed, but it may well be that only a small part of the data are relevant to the questions being asked, which often have to do with issues such as long-term trends or extreme values. Thus, the results obtained may well depend critically on the model selected. A second reason for thinking critically about model selection issues in environmental data is the desire to obtain results that are at least compatible with physical models, such as those constructed by atmospheric scientists. These general issues are illustrated by two areas of research from the environmental statistics group at the (U.S.) National Institute of Statistical Sciences, one concerned with trends in tropospheric ozone, the other with the health effects of atmospheric particles.

1. Introduction

This paper has two objectives. The first is to give some general discussion on the role of model selection in environmental statistics. My comments here will be neither very detailed nor very profound, but I hope they will serve as something of a stimulus to discussion. The second objective is to present two examples that I have worked on recently myself, both of which required some ingenuity in identifying the right model. The first of these concerns modelling high-level exceedances of ground-level ozone, a topic of considerable importance in the USA in the light of measures currently being enforced to reduce ozone levels. The models fitted in this case were suggested by extreme value theory, but the verification of those models was based on an adaptation of work on predictive model diagnostics by Dawid and Seillier-Moiseiwitsch. My second example is concerned

[1] Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, UK (rsmith@statslab.cam .ac.uk). Much of this work took place at the National Institute of Statistical Sciences, P. O. Box 14162, Research Triangle Park, NC 27709-4162, USA, where it was supported by the U.S. Environmental Protection Agency under Cooperative Agreement #CR819638-01-0.

with the influence of small particles in the atmosphere (so-called pm10) on human health and mortality. A series of papers in the epidemiological literature has claimed a very strong relationship between pm10 and mortality among the elderly population, and this has led to considerable public discussion in both the USA and the UK. However, the statistical basis of this claim is by no means clear-cut, and hinges quite a bit on what model is chosen to fit to the data. It therefore seems a good example to introduce to a conference on model selection. Both examples were developed at the (US) National Institute of Statistical Sciences and I would like to acknowledge the contributions of others who worked on the problems: Peter Bloomfield, Li-Shan Huang and Françoise Seillier-Moiseiwitsch (ozone); Jerry Sacks, Trish Styer, Nancy McMillan, Feng Gao and Jerry Davis (pm10).

2. General Remarks

Typical problems in environmental statistics involve the interpretation of data on environmental variables, such as meteorological data, air and water pollution monitors, river flows and sea levels, etc., and their "consequences" as measured by human mortality and health statistics, animal and fish population sizes, etc. Of course, not all of these elements will be present in any particular application. One characteristic of environmental problems is that, particularly on the "environmental variables" side of the problem, vast amounts of data are typically available. Thus there is considerable scope for comparing different models. However, in many cases the questions being asked are ones that depend critically on the model chosen. Examples include trends in climatological or environmental pollution data, extreme values, low-dose extrapolation of toxicological effects, etc. These are all quantities that cannot be measured in any direct way and so require the identification of a suitable model.

A particular issue in many environmental applications is the interplay between statistical and physical modelling techniques. In the context of climate change, for example, most of the quoted predictions on the consequences of an enhanced greenhouse effect are based on general circulation models or GCMs, which are large-scale physical models of the earth's atmosphere. Little of the evidence for global warming comes directly from statistical analysis, though there is evidence in the climate record of a gradual but steady rise in global temperatures over the last 150 years. Thus one challenge in this area is to see to what extent the available data support or contradict the predictions of GCMs. The relevance of this to a discussion of model selection is that the GCM – or whatever other physical model is appropriate to a particular application – may help to define the model to be studied. An example of this is Bloomfield's (1992) study of the "climate sensitivity" parameter (a measure of the rate of change of temperature with respect to atmospheric carbon dioxide) in which a family of GCM outputs, parametrised by climate sensitivity, was fitted to the observed climate record. On the other hand, even in this example, the GCM did not dictate the whole of the model; for instance, it tells us nothing about the kinds of time series correlations that are appropriate. Bloomfield and Nychka (1992) and Smith (1993) had further discussions of that aspect. In the context of ground-level ozone, much work has been done on regional oxidant models (ROMs) which monitor the transportation and chemistry of ozone and its precursors over a time scale of anything between a few hours and about 14 days. Expertise gained from these models has been an enormous help in identifying relevant meteorological variables and the general form of the statistical model to be fitted (Bloomfield *et al.*, 1993) but this is still only one part of the overall problem; ROMs are unable to model long-term trends (over a period of several years) so in this context some statistical modelling is essential. The moral of the story is that, where physical models are available, they can and should be used to provide valuable guidance to the choice of a statistical model; but it is rare for the physical model to dictate the statistical model, and there is still plenty of scope for the use of more conventional statistical model selection and model verification techniques.

Other examples of the need for an interrelation between physical and statistical models include the "downscaling" (also called "disaggregation") problem in climatology, which arise from the inability of GCMs to give reliable predictions over small areas of the earth's surface. Typically, the GCMs provide predictions averaged over grid cells whose sides are in the range of 3^o–5^o latitude and longitude. This is of little use in predicting, say, the pattern of rainfall over a radius of 50 miles, so one needs statistical techniques to relate the short-range spatial behaviour of rainfall to long-term characteristics as might be measured by temperature, rainfall or air circulation indices compiled at the level of a GCM grid cell. One can interpret the paper of Handcock and Wallis (1994), who developed spatial-temporal models for temperature, from this point of view: for example, their model allows predictions of the trend at individual locations, with an associated assessment of the quality of prediction. On the rainfall side, models such as those of Rodriguez-Iturbe *et al.* (1987a, 1987b) and Cox and Isham (1988) are relevant to this kind of question. Smith and Robinson (1995) have made a preliminary attempt to reinterpret the latter theme from the point of view of Bayesian model fitting using MCMC methods.

Aside from all this discussion is the possibility that information derived from physical models might be used *directly* to motivate the formulation of a statistical model. I mention this only tangentially here because it will not be developed in the ensuing discussion, but it is a major theme in some areas of environmental research, see e.g. Young and Lees (1993).

3. High-threshold Exceedances of Ground-level Ozone

Ozone is produced at ground level as a result of chemical processes involving nitrous oxides and hydrocarbons, which are emitted by vehicles or in industrial processes. However, the complex nature of the chemical reactions involved means that there is no easily identifiable cause-effect relationship; there is a particular tendency for high ozone concentrations to arise on hot, still days and so any attempt to monitor trends in ozone must take account of the weather. As far as the effects are concerned, high ozone levels are known to be particularly harmful to people with asthma or other respiratory conditions, and until recently (see section 4) ozone was widely regarded as the leading human health hazard arising from atmospheric pollution. In the USA, an ozone standard, based on a maximum daily level of 120 ppb, has existed since 1979; the standard permits this level to be exceeded only three times in any three-year period at any single monitoring station. In practice some large cities exceed the standard many times per year, and the 1991 Clean Air Act Amendment of the US Congress set tough timetables for these exceedances to be reduced. In the UK, concern about ozone has never reached the same level as in the USA, partly because with our colder climate, ozone levels are not nearly so high. Fowler *et al.* (1993) recommended that there should be a UK ozone standard, something which the British government has yet to act upon, but widespread public concern about air pollution levels during a recent (May 1995) hot spell in Britain, associated with calls for lowered speed limits and other measures to reduce air pollution, suggests that such a standard may well be considered in the future. The study reported here (Smith and Huang 1994, Smith, Huang and Seillier-Moiseiwitsch 1995) was motivated by efforts to measure ozone trends in the USA, where the responsible agency is the Environmental Protection Agency (EPA). Previous or parallel studies utilising an extreme values approach to high ozone exceedances are contained in papers of Smith (1989), Shively (1991) and Smith and Shively (1995).

Table 1: Measured meteorological variables

Variable	Meaning
OPCOV	Opaque cloud cover (%)
PR	Barometric pressure (mb.)
T	Temperature (^{o}F)
TD	Dewpoint Temperature (^{o}F)
RH	Relative humidity (%)
Q	Specific humidity (g./kg.)
VIS	Visibility (km.)
WSPD	Wind speed (m./sec.)
WDIR	Wind direction (^{o} from North)

The raw data consisted of hourly ozone readings for 1981–1991 at a network of 45 monitoring stations around the city of Chicago. Exceedances of the 120 ppb threshold varied greatly between monitoring stations, but the worst station had 41 exceedances within the 11-year period, indicating a definite lack of compliance with the standard, though not nearly as bad as that in some southern cities such as Los Angeles and Houston. For most of our study, we considered "network maxima" formed by maximising the daily values over a subset of stations in the network, making suitable allowances for missing data. Bloomfield *et al.* (1993) described the method precisely. We also considered a suite of meteorological variables, which are listed in Table 1. To these were added some constructed variables, listed in Table 2. In addition to the meteorological covariates, YEAR was used as a covariate (=1 for 1981, through to 11 for 1991), and two variables CDAY and SDAY, defined by

$$\text{CDAY} = \cos(2 \times \pi \times \text{DAY}/365.25), \quad \text{SDAY} = \sin(2 \times \pi \times \text{DAY}/365.25),$$

were included in the analyses. Here DAY represents the day within the year (=1 for January 1, etc.). The inclusion of CDAY and SDAY is intended to reflect the fact that there remains a residual seasonal effect even when the direct influence of season on meteorology has been taken into account, while the coefficient of YEAR in the final model will be interpreted as a residual trend when all seasonal and meteorological effects have been taken into account. For some analysis we also used a variable YEAR2, defined to be the square of YEAR, so allowing for the possibility of a quadratic trend. Further discussion of the choice of covariates is given by Bloomfield *et al.* (1993) and was, as discussed in Section 2, influenced to no small extent by the expertise of physical modellers within the EPA.

Table 2: Additional variables created from the data

WIND.U	$-\text{WSPD} \times \sin(2 \times \pi \times \text{WDIR}/360)$
WIND.V	$-\text{WSPD} \times \cos(2 \times \pi \times \text{WDIR}/360)$
T2	$(\text{T-60})^2/10$
T3	$(\text{T-60})^3/1000$
T.WSPD	WSPD × (T-60)

3.1 Exceedances of a single threshold

The natural starting point is to consider exceedances of just a single threshold, i.e. the ozone standard of 120ppb, and the initial analysis was performed on the assumption that different days are independent. This assumption was thought justifiable in our initial analysis because of the belief (reinforced by the physical modellers) that the main cause of correlation in the data is the persistence of meteorology, so that once we adjust for meteorology the residuals should be independent. However, as we shall see, matters are not in fact quite so simple as that, and one simple say to include dependence in the model is to introduce a variable PDAY (=1 if the previous day was an exceedance, 0 otherwise) as an additional covariate.

Once these assumptions are made, the natural approach is to fit the logistic model for binary data,

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_j x_{ij}\beta_j, \tag{1}$$

where p_i denotes the probability that the threshold is exceeded on day i, x_{ij} is the value of the j'th covariate on day i and β_j is the corresponding coefficient. We always assume

$x_{i1} = 1$, i.e. there is always a constant term in the model, but the other covariates are selected from those described above.

One additional point to note is that all the ozone values are recorded to the nearest ppb, and a level of exactly 120 is counted as an exceedance.

With the model and covariates set out as described, standard variable selection techniques were used to identify a specific model for the network maxima, with results shown in Table 3. It should be noted that some variables were included in pairs, specifically CDAY and SDAY were included together, as also were WIND.U and WIND.V, which explains why some of the variables are retained in the model even though their t ratio was less than 2.

Table 3: Network maxima, threshold 120

Variable	Estimate	Stand. error	t Ratio
CONST	−22.9	2.357	−9.72
YEAR	.3416	.1632	2.09
YEAR2	−.04268	.01395	−3.06
CDAY	−.8855	.7036	−1.26
SDAY	.5628	.2911	1.93
PDAY	.6829	.2736	2.50
T	.2928	.02747	10.66
RH	−.01799	.01058	−1.70
WIND.U	−.03688	.04011	−.92
WIND.V	−.1887	.0449	−4.20
VIS	−.06959	.02065	−3.37
T.WSPD	−.01945	.002992	−6.50

It will be noted that the trend variables YEAR and YEAR2, as well as the dependence variable PDAY, are all significant in this model, and this is reinforced by Table 4, in which NLLH (negative log likelihood) values are given for various combinations of these parameters.

Table 4: Comparison of models using NLLH

Trend/PDAY	No	Yes
None	319.299	313.211
YEAR only	309.023	305.025
YEAR+YEAR2	302.984	299.934

Table 4 reinforces the conclusion that YEAR, YEAR2 and PDAY are all significant. Adding additional variables to account for lags of greater than one day did not improve the fit.

The coefficients of YEAR and YEAR2, interpreted as a quadratic trend throughout the 11-year period of the study, lead to the conclusion that, after taking meteorology into account, there was a slight increase in ozone up to around 1984, but thereafter a steady decrease. However, this conclusion is based only upon searching within rather a narrow class of models, and leaves open the question of whether the chosen model actually fits the data. We turn to this question next.

3.2 Predictive assessment of model fit

For normal linear models, there is an extensive literature on residuals and diagnostics, well covered by all the standard textbooks. For binary data, one approach to model verification is simply to try to adapt the techniques for linear models, and this approach is taken for example, by Collett (1991). However, there is a much older literature on the assessment of forecasting schemes for binary data, which originates in the context of weather forecasting. This approach provided the background for Phil Dawid's theory of "prequential analysis"; see Dawid (1982, 1984) for expositions of this theory and Dawid (1986) for a review of the background on probability forecasting. Recent works by Seillier-Moiseiwitsch and Dawid (1993) and Seillier-Moiseiwitsch (1995) have focussed attention on goodness of fit criteria, which we develop here.

The usual paradigm is forecasting rainfall. Each day, a forecaster quotes the percentage probability of rainfall, in practice expressed to the nearest 10%. After many days, we are able to compare the sequence of forecasts with the sequence of binary variables representing the observation of whether or not it rained on each day. How should we assess how well the forecaster is performing?

Most criteria are based on a combination of two different measures. On the one hand, the forecaster should be *well-calibrated.* This means, for example, that if we grouped together all days on which the forecaster quoted a 40% probability of rain, then the observed proportion of rain days within that grouping should be close to 40%. Clearly this is a desirable attribute of a good forecaster, but it is not sufficient in itself. It is possible for forecasts to be well calibrated but practically useless, e.g. if the forecaster always quotes the same probability. A second quality called *resolution* or *refinement* is also needed. This has to do with the extent to which the forecasts succeed in distinguishing between wet and dry days.

There are various ways on constructing goodness of fit tests to measure calibration. The main technical difficulty is the justification of an asymptotic χ^2 distribution when the forecasts are sequential and possibly dependent on past data. However, much is now known about this problem, cf. Seillier-Moiseiwitsch & Dawid (1993). It is not so easy to decide whether a forecaster has good refinement. Dawid (1982, 1986) suggested that a suitable criterion for a forecaster to be well calibrated and to have good resolution should be that his or her forecasts remain well-calibrated when restricted to any subsequence of the data, subject to a selection rule that requires the decision whether or not to include a particular day in the subsequence should depend only on the information available to the forecaster, which in most cases means the data available from previous days. In practice, of course, one cannot consider all such subsequences simultaneously, but if one suspects a forecaster to have poor resolution, one can try to construct subsequences on which the calibration test might fail.

An older idea is that of *scoring rules*. A scoring rule is simply a measure or score of how well a forecaster is performing. There are a number of such rules known, but the best known are the *Brier scoring rule* $\sum(a_i - p_i)^2$ and the *logarithmic scoring rule* $\sum\{-a_i \log p_i - (1 - a_i)\log(1 - p_i)\}$. Here a_i is the observed value (0 or 1) on day i. However, while scoring rules might serve to decide whether one forecaster is better than another, they are not so easily turned into formal tests of whether to accept or reject a certain model.

In the context of the present study, prediction of ozone exceedances is not one of our primary aims. We are much more concerned with long-term issues such as whether there is a trend in the data or whether a particular subset of the data represents unusual conditions when judged against long-term climatology. Nevertheless, logistic models for binary data are very much of a predictive nature, since they allow us to quote a "probability of exceedance" based on current weather conditions and (if PDAY is included) past ozone exceedances. Therefore, we can use ideas from the probability forecasting literature to assess their goodness of fit. In the models with PDAY, this is a genuinely sequential problem, so we do need the theory of sequential tests developed by Dawid (1984) and Seillier-Moiseiwitsch & Dawid (1993).

There is, however, one further complication in applying these ideas here. The literature we have described is concerned with an "honest" forecasting procedure in which the forecast for day i depends only on data observed preceding day i. This is not the case if we are considering a parametric model whose parameters have been estimated from the whole data set. The issue is similar to the familiar one in goodness of fit testing, that tests constructed on the assumption that the model is known are not valid without some adjustment if the model depends on parameters which are estimated from the data.

One version of this problem has been considered by Seillier-Moiseiwitsch (1995) for the case of a linear logistic model. She considers a procedure in which, before making a forecast for day i, the parameters of the model are re-estimated based on all the data up to time $i - 1$. Under this set up, the asymptotic properties of the procedure are identical to those of a sequential forecasting scheme as in Seillier-Moiseiwitsch & Dawid (1993). The difference between the two papers is that the analysis of Seillier-Moiseiwitsch & Dawid (1993) is based on the null hypothesis that the sequential forecasting scheme is indeed the exact model that generated the data, whereas Seillier-Moiseiwitsch (1995) assumes that the parametric model is correct and makes explicit allowance for the fact that the parameters at each stage are estimated.

In the present context, it is not practicable to update the model after every observation but we have employed a variant which seems to achieve the same effect: for each year, the model was refitted to the data omitting that year and then used to produce probability forecasts for the year. All the resulting forecasts were then combined to assess how well they performed on the observed data on threshold exceedances. Thus the analysis is honest in the sense that no forecast probability of exceedance depends on the exceedance itself, or on nearby correlated values. It can be shown that such a modification preserves the asymptotic properties of the procedure (F. Seillier-Moiseiwitsch, personal communication).

As an example of these ideas, Table 5 is concerned with the calibration of probability forecasts produced by the model of Table 3. This table is very similar to Table 1 of Seillier-Moiseiwitsch & Dawid (1993). Each row of the data represents a specific interval of forecast probabilities, denoted $(p_{min}, p_{max}]$. The frequency n denotes the number of days for which the forecast lay in that interval, and r is the observed number of exceedances based on forecasts within the interval. The next value e is the expected number of exceedances $\sum p_i \, I(p_{min} < p_i \leq p_{max})$ and w is its variance $\sum p_i(1-p_i) \, I(p_{min} < p_i \leq p_{max})$, under the assumption that the p_i do indeed represent true forecast probabilities. The final column gives the test statistic $z = (r-e)/\sqrt{w}$. In large samples, these will have approximately standard normal distributions, independent for non-overlapping probability intervals. The main difference from Seillier-Moiseiwitsch & Dawid (1993) is that we have used unequal probability intervals to reflect the fact that the forecast probabilities are heavily weighted towards 0. Finally, the last row of Table 5 gives an overall assessment of calibration by combining the intervals together.

Table 5: Calibration of probability forecasts

p_{min}	p_{max}	n	r	e	w	z
0.000	0.025	1660	13	7.724	7.621	1.911
0.025	0.050	182	10	6.574	6.327	1.362
0.050	0.075	102	3	6.282	5.890	−1.352
0.075	0.100	65	5	5.640	5.147	−0.282
0.100	0.200	132	18	18.949	16.131	−0.236
0.200	0.300	70	21	17.692	13.161	0.912
0.300	0.400	45	15	15.832	10.227	−0.260
0.400	0.500	28	11	12.629	6.913	−0.620
0.500	0.750	46	26	28.813	10.551	−0.866
0.750	1.000	24	21	20.889	2.614	0.069
0.000	1.000	2354	143	141.024	84.581	0.215

A second way of classifying the data is by year. Table 6 shows a table constructed similarly to Table 5 but where the rows represent individual years of data. Recall that each year's entry is based on a model fit excluding that year's data.

In Table 5, the z values are generally small, indicating a good fit, the only seemingly significant value being in the first row, but the overall value of $\sum z^2$ (9.51) is clearly not significant against its nominal χ^2_{10} distribution. Table 6 is a little more disturbing since the model has clearly overpredicted the exceedances for 1989 and underpredicted for 1991, though for the most important years, 1983 and 1988, the agreement of observed and predicted numbers of exceedances is remarkably good. The underprediction for 1991 should make us cautious about extrapolations based on the quadratic trend.

Table 6: Calibration of probability forecasts by year

Year	n	r	e	w	z
1981	214	8	11.119	7.642	−1.128
1982	214	11	11.129	7.847	−0.046
1983	214	26	28.326	12.358	−0.662
1984	214	18	12.870	8.695	1.740
1985	214	13	8.398	6.597	1.792
1986	214	9	10.058	7.907	−0.376
1987	214	18	19.274	10.312	−0.397
1988	214	24	23.433	11.509	0.167
1989	214	3	11.909	7.750	−3.200
1990	214	3	2.207	2.026	0.557
1991	214	10	2.301	1.938	5.530

This analysis was repeated for all six models in Table 4, and in all six cases the basic calibration table was similar to Table 5, with no significant discrepancies. For the analysis by year, the model with no trend showed a distinct pattern of z values, with a significant underprediction of exceedances in 1984 and 1985 and a significant overprediction in 1989 and 1991. The model with linear trend showed a pattern of negative z values at the beginning, then positive, and then negative again, with the most significant z values being −2.823 (1981), 2.546 (1984), 2.886 (1985) and −3.192 (1989). This can be taken as indicating that a quadratic trend really is needed. These results are all for models including PDAY; the results without PDAY showed identical patterns in each of the three cases, but generally more extreme z values. Finally, the Brier scores and associated z values (nominally standard normal test statistics) were:

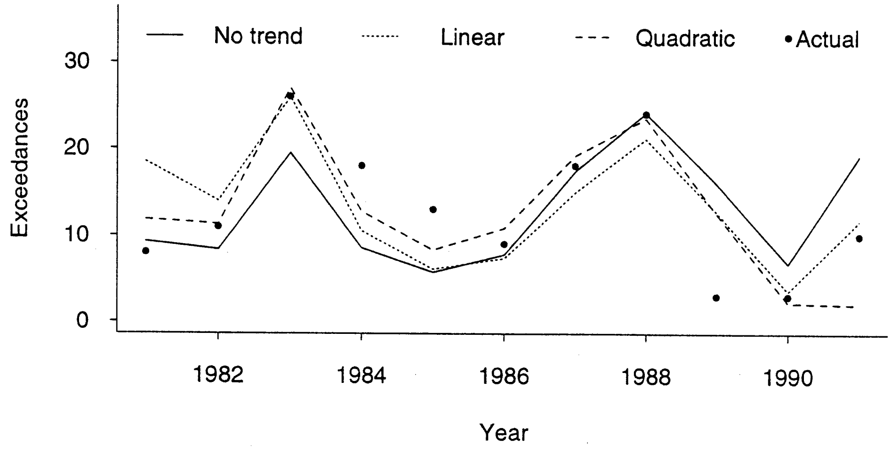
Table 7: Brier scores and test statistics

No trend, no PDAY	94.82	(0.52)
No trend, include PDAY	92.54	(0.52)
Linear trend, no Pday	93.28	(0.40)
Linear trend, include PDAY	91.63	(0.40)
Quadratic trend, no PDAY	90.81	(0.75)
Quadratic trend, include PDAY	89.67	(0.86)

These results confirm the pattern that the model improves with increasing order of trend, and is better including PDAY than omitting it.

In Figure 1, we show the observed numbers of exceedances together with the expected exceedances under each of the models with no trend, linear trend and quadratic trend. It can be seen that the model with no trend systematically underpredicts in the first part of the data, and overpredicts in the second – a sure sign that a trend is present – while the predictions for quadratic trend form the closest overall fit to the data.

Fig. 1: Exceedances and projections, network maxima



3.3 Extrapolation to higher thresholds

There are a number of reasons why one might want to extend this analysis to thresholds above 120 ppb. One argument is that, although the threshold is set for regulatory purposes at 120, it is generally acknowledged that it is only at rather higher levels – perhaps 160 or 180 – that real measurable health effects are seen. Therefore, to draw meaningful conclusions about human health effects, it might be necessary to look at a higher threshold than 120. Another reason for looking at different thresholds is that sometimes one draws different conclusions (regarding the existence of a trend, for example) at different threshold levels – Smith (1989) provided an early example of that, using data from Houston in an earlier time period. The third reason in trying different thresholds is the general interest in building and testing models for extreme values, in a situation where there is enough data around to test the validity of the resulting extrapolations.

A natural approach to the extrapolation problem is to divide the modelling exercise into two components: first, fit a model to the distribution of exceedances over a fixed threshold, which we will again take to be the standard 120, and then fit a model to the *excesses* over that threshold, i.e. the conditional distribution of the amount by which the threshold is exceeded, given that it is exceeded. One distribution which has become widely used in this context is the *generalized Pareto distribution* (henceforth GPD) whose distribution function is given by

$$G(y;\sigma,\xi) = 1 - \left(1 + \xi\frac{y}{\sigma}\right)_{+}^{-1/\xi}, \quad y > 0, \tag{2}$$

where $\sigma > 0$, ξ is any real number, and $x_+ = \max(x,0)$. Thus the range of y is $0 < y < \infty$ for $\xi \geq 0$ and $0 < y < -\sigma/\xi$ if $\xi < 0$. The exponential distribution, $1 - e^{-y/\sigma}$, appears as a limiting case when $\xi \to 0$. This distribution was originally motivated by the limit theorems of extreme value theory, and has been found to be a good fit in very many situations of fitting data over high thresholds, e.g. Smith (1989), Davison and Smith (1990). See also Smith and Shively (1995), for a parallel analysis of current data from Houston.

In the context of trying to fit meteorological covariates, it is logical to extend the model to one in which the excess (if there is one) on day i is representated by the $G(\cdot;\sigma_i,\xi_i)$ with σ_i and ξ_i depending on covariates. In practice it is usually adequate (and a lot simpler) to assume ξ constant, while the interpretation of σ_i as a scale parameter suggests naturally that a logarithmic link function would be appropriate. Thus we are lead to consider models of the form

$$\log \sigma_i = \sum_j x_{ij}\gamma_j, \quad \xi_i = \xi \tag{3}$$

in terms of new coefficients $\{\gamma_j,\ j = 1,2,...\}$. There is no reason why the significant covariates should be the same as in the binary analysis of section 3.1, so in general we would expect to repeat the variable-selection procedure based on the excesses over the threshold. For the time being, we assume the daily values are independent.

Once these models have been fitted, it is natural to define a "residual" $Y_i/\widehat{\sigma}_i$ with respect to tyhe i'th excess Y_i and the associated estimated scale parameter $\widehat{\sigma}_i$. The residuals are arranged in order and plotted against expected order statistics. If the model fits the data, then the residuals should be tightly clustered around a straight line of unit slope through the origin.

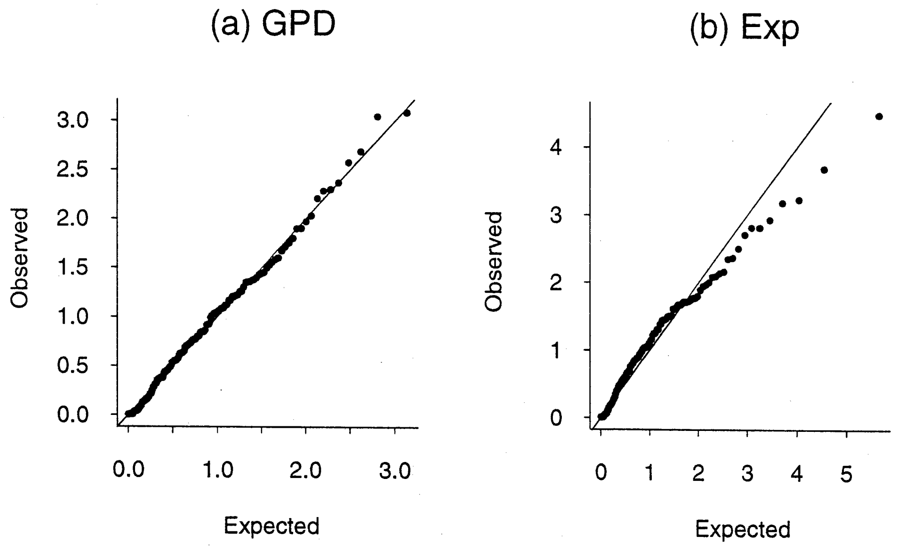
Table 8 shows the results of this analysis for the network maxima. In this case a model selection carried out for a number of individual stations showed that the covariates T, WSPD and T.WSPD were significant for a range of data sets; in the present analysis it looks as though the last two variables could be dropped but they were kept in the analysis for comparison with individual stations. Also note that ξ is significantly different from 0 and negative, a result which indicates an upper tail which is shorter than exponential. Finally, the coefficient for YEAR, though not statistically significant, is negative, indicating that the downward trend observed for threshold 120 remains at higher thresholds.

Table 8: GPD fitted to excesses

Variable	Estimate	Stand. error	t ratio
CONST	-.8527	1.685	-.51
YEAR	-.04	.03098	-1.29
T	.05795	.02026	2.86
WSPD	-.1005	.1307	-.77
T.WSPD	-.004424	.005415	-.82
ξ	-.2333	.08884	-2.63

Fig. 2(a) shows a probability plot of residuals from the model of Table 8, which appears very close to the theoretical straight line of unit slope through the origin. In contrast, the same model with $\xi = 0$ (the exponential model) shows a definite deviation in the upper tail (Fig. 2(b)). This is of some importance because earlier ozone studies (including Smith (1989)) have suggested that the exponential distribution fits high levels of ozone data very well; the present study would seem to indicate that this may not be the case if due account is taken of meteorological covariates.

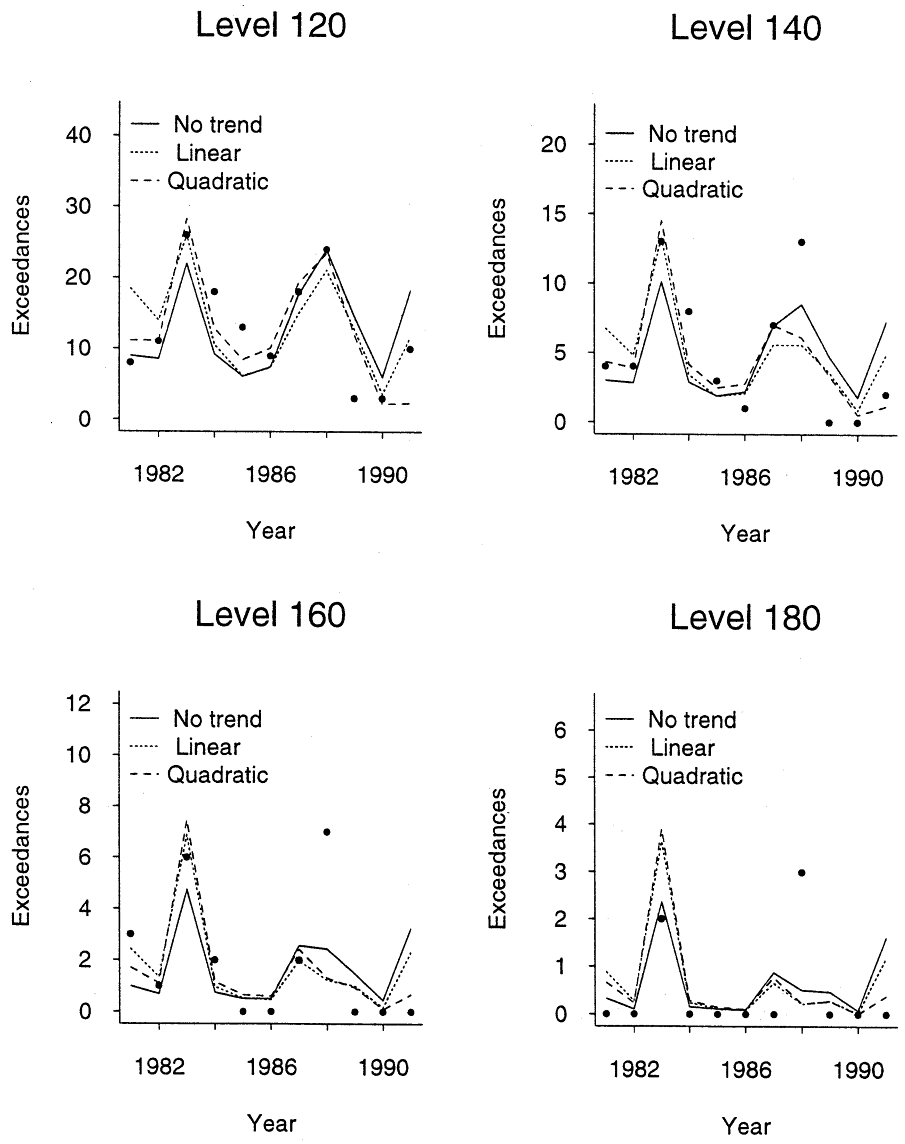
Fig. 2: Network maxima excesses



Now let us consider predictive diagnostics for these models. If we combine the models used earlier for exceedance probabilities of 120 with the GPD model for excesses fitted here, and then compute predictive diagnostics for each of the levels 120, 140, 160 and 180 (generalising the calculation of Fig. 1), we obtain the plots of Fig. 3. We can see from the plots that at the higher threshold levels, the models noticably fail to predict the high levels observed in 1988.

It is possible to construct χ^2 tests from these plots in the same way as was done earlier for the exceedances of level 120. We conpute a table similar to Table 6, including the z values in the last column, and form $\sum z^2$ as an approximate χ^2 statistic. However, for this purpose it is desirable that the e values in the fourth column be large enough for the validity of a χ^2 approximation (analogous to the $e \geq 5$ criterion often adopted for the standard Pearson χ^2 test) and for this purpose it is desirable to group years with low exceedances together. Thus we formed six groups based on the years 1981–2, 1983, 1984–6, 1987, 1988 and 1989–91, from which we calculate $\sum z^2 = 13.36$, significant at the level $p = 0.038$ against the nominal χ^2_6 distribution. Combined with the visual impression gained from Fig. 3, this suggests that the model is not an adequate fit. Variants on the model, e.g. using different combinations of the trend parameters and meteorological covariates, failed to resolve this difficulty.

Fig. 3: Exceedances over several levels



Level 120

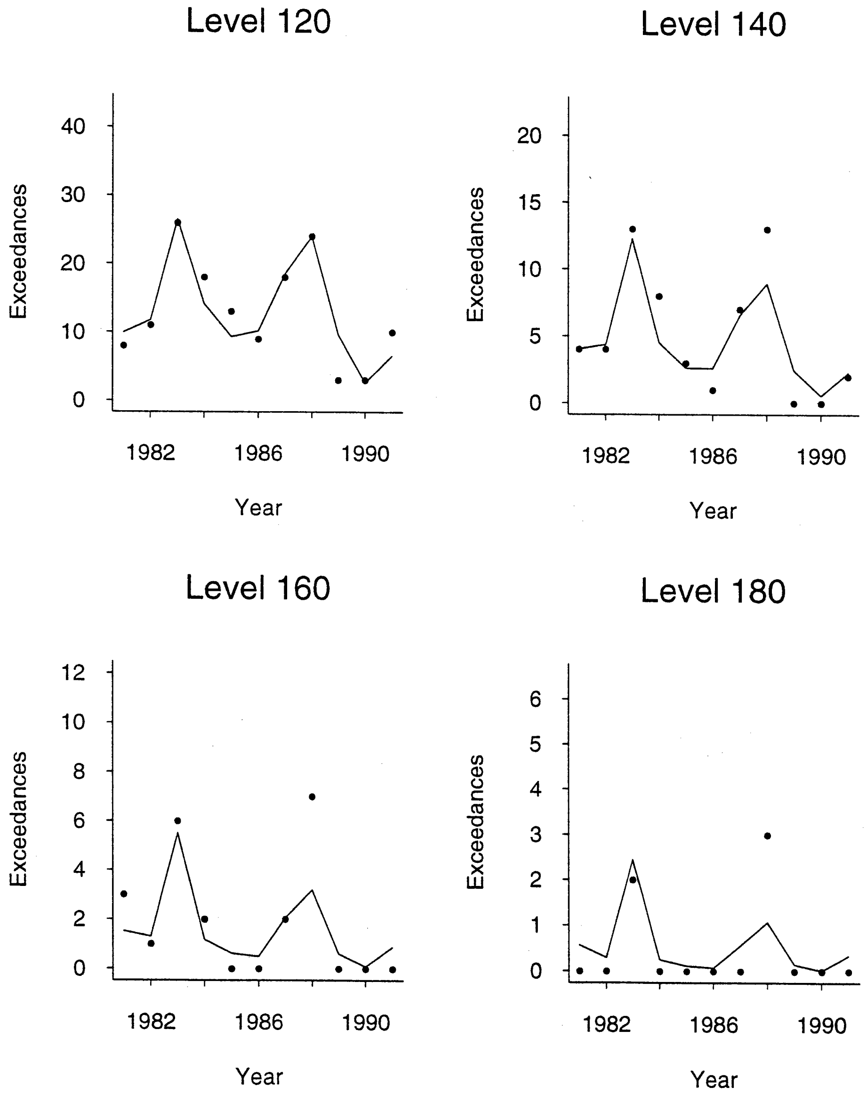
Level 140

Level 160

Level 180

A detailed description of the manner in which this lack of fit was resolved would take us too far from our main theme in this paper, but the essence of it was to model the process as a first-order Markov chain, using bivariate extreme value distributions to model the joint dependence. Models of this nature have been developed by Smith, Tawn and Coles (1993) and by Ledford and Tawn (1994), and are reviewed by Smith (1994). The trend for exceedances over 120 was taken to be quadratic in this case. Recomputing the plots of Fig. 3 for this model led to the plots of Fig. 4, and while they still underpredict the observed numbers of exceedances for 1988, the *overall* fit is much better — χ^2 statistics computed for both of the levels 140 and 160 produced insignificant results.

Fig. 4: Exceedances over several levels, dependent model



To summarise: the predictive model diagnostics proved to be a sensitive test of whether the extreme value model fitted the data. The original model, based on independent exceedances conditioned on the meteorological variables, did not provide an adequate fit. However, when an alternative model incorporating serial dependence in the extreme values was tried, the resulting model seemed to fit the data much better.

4. Atmospheric particles and human mortality

My second example is much less clear-cut than the first; indeed it is extremely confusing. However it does address an issue of considerable public interest at the present time: the notion that there is a direct relationship, far stronger than had previously been supposed, between human mortality and the presence in the atmosphere of small particles. Most of the studies have concentrated on pm10, which is defined by the EPA to be "particulate matter with an aerodynamic diameter less than or equal to a nominal 10 micrometers". The studies have led to claims that as many as 60,000 deaths per year in the US, and 10,000 deaths per year in the UK, may be due to particulates in the atmosphere (sources: *The New York Times*, July 19 1993; *The New Scientist*, March 12 1994). The background to such claims lies in a series of epidemiological studies relating deaths from nonaccidental causes in the elderly population to pm10 counts, taking into account additional relevant meteorological covariates. The paper by Schwartz (1993) is representative of the kinds of statistical analyses used, and Schwartz (1994) has given a parallel analysis using hospital admissions as the dependent variable. More recent studies (e.g. Seaton *et al.* 1995, Pope *et al.* 1995) seem to have left little doubt that there is an association of some sort, but there remains much controversy over whether an observed association can be interpreted as a causal relationship, and especially over the magnitude of the effect as reflected in the numbers just quoted. The following discussion, based on the paper by Styer *et al.* (1995) and some follow-up work of my own, is intended to show just how much the conclusions depend on seemingly arbitrary decisions related to model selection.

For this study, data were collected from Cook County, Illinois (the county which includes the city of Chicago). Mortality data consist of daily deaths of residents of the county, for the period January 1985 to December 1990. Deaths of individuals under 65 years of age, or from accidental causes, or those which took place outside Cook County, are excluded. There remain an average of about 80 deaths per day. Air pollution data consist of daily average pm10 readings of about twelve monitoring stations. Most of the stations record only every six days, but for each day in the study, averages across all available stations were taken. Preliminary analysis (Styer *et al.* 1995) had shown that three-day averages are the best predictor of deaths, i.e. for each day we take the average of that day, the previous day and the previous day but one. There are still a substantial number of missing days, particularly in the early part of the period of observation, but data are available for 1,963 of the 2,191 days of the sampling period.

Units of measurement for pm10 are micrograms per cubic metre – in this data set a typical daily value is about 40, while the current EPA standard is 150.

Meteorological data were based on the same sources as used for the ozone study, but different variables, including some lagged variables. Altogether twelve meteorological variables were considered, as follows:

Table 9: Meteorological variables for pm10 study

Variable	Meaning
press	Daily mean pressure in millibars
plag1	Pressure lagged by one day
plag2	Pressure lagged by two days
tmean	Daily mean temperature in ^{o}C
tlag1	Temperature lagged by one day
tlag2	Temperature lagged by two days
qmean	Daily mean specific humidity
qlag1	Humidity lagged by one day
qlag2	Humidity lagged by two days
wchill	Windchill factor
disind	Discomfort index (an indicator of hot and humid conditions)
gsum	Total solar radiation

Similar data were also available for Salt Lake County (Utah). In this case the period covered is from June 1, 1985, to December 31, 1990 (2,040 days), of which pm10 readings are available for all but 12 days. The average number of deaths per day here is 6.7, reflecting the much lower population. The meteorological variables are the same as for Cook County with the exception of the windchill and solar radiation variables.

The main method is based on normal regression, using log deaths as the dependent variable. Most of the studies by Schwartz and his co-authors have been based on a form of Poisson regression: for the data reported here, this was actually an inferior fit to the lognormal regression, though this conclusion may be specific to the Cook County data. The initial model fits are based on an assumption uncorrelated errors from one day to the next, though as with the ozone study, this is the subject of some discussion later on.

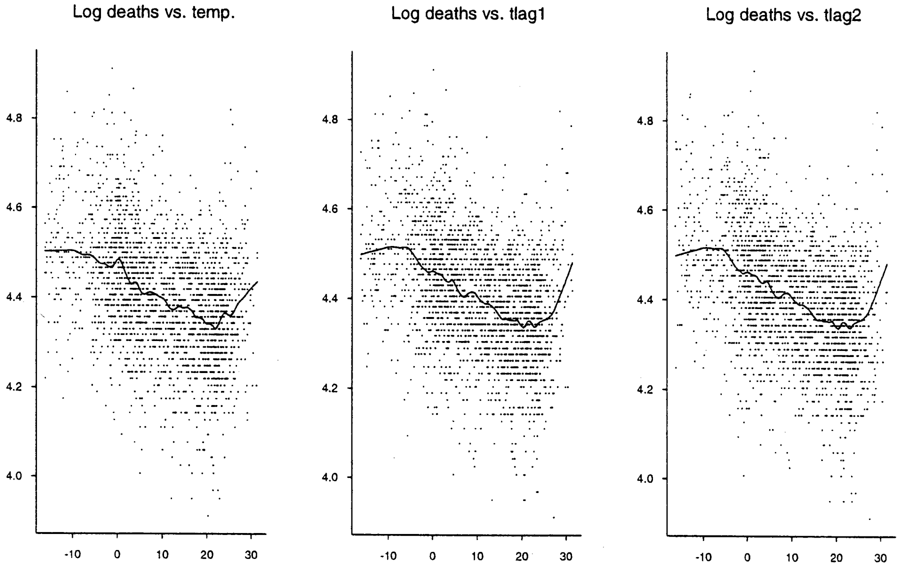
A scatterplot of log deaths against lagged and unlagged temperatures is shown in Fig. 5, together with a smooth curve through the data points (computed using the `lowess` command in Splus). This shows a changepoint at around 22°C, in all three curves, and suggests the addition of a variable which we call t22, defined by

$$\mathrm{t22} = (\mathrm{tmean} - 22)_+,$$

with corresponding lagged values t22lag1 and t22lag2. Similar scatterplots were drawn against the other meteorological variables but did not yield any conclusions of particular interest.

Further preliminary analysis indicated a definite variability from year to year and so suggested adding five indicator variables to the data, e.g. the variable I(year=86) takes on the value 1 if the year is 1986, 0 otherwise. Study of squared and cross-product terms led to the addition of four variables formed in this way. When all these variables were taken into account (but not, at this stage, pm10) the model listed in Table 10 was fitted to the data.

Fig. 5: Deaths vs. unlagged and lagged temperatures



If now the variable pm10 is added to the model in Table 10, we are led to a coefficient of .00080, with a standard error of .00021 — consistent with several results by Schwartz and co-workers, who have typically claimed a coefficient (in these units) of about .001.

However, there are a number of reasons why the simple model just described does not lead to an adequate fit to the data. Here are some of the main points which have arisen in further analysis of these data:

1. There is a strong *seasonal effect* over and above anything that can be explained by meteorology. Particularly striking is a tendency for the residuals from the above model to rise around Christmas, with smaller but still significant variations the rest of the year.

2. It also appears that there is a significant season × year *interaction*, i.e. the seasonal effect varies from year to year.

3. Styer *et al.* (1995) found a significant season × pm10 interaction. Specifically, they found that if the pm10 coefficient was estimated separately for each of the four seasons of the year, that it was strongest in the autumn, with a weaker but still significant effect in the spring, and no effect at all in the winter and summer (in fact their coefficient for

"summer pm10" was negative, though not statistically significant). However, there is no known explanation for such seasonal variation.

Table 10: Parameter estimation for lognormal model

Parameter	Coefficient	S.E.	*t* Ratio
constant	4.47373	.01365	327.71000
press–1000	–.00127	.00071	-1.79757
tmean	–.00104	.00078	-1.32730
tlag1	–.00088	.00102	–.86631
tlag2	–.00289	.00092	-3.13354
qlag2	–.01005	.00340	-2.95841
gsum	–.00577	.00175	-3.29315
t22	.02711	.00653	4.15318
t22lag1	–.01011	.00695	–1.45451
I(year=86)	.03047	.00887	3.43323
I(year=87)	.01990	.00890	2.23609
I(year=88)	.02680	.00908	2.95329
I(year=89)	.01651	.00885	1.86515
I(year=90)	.01596	.00889	1.79461
$t22^2$	–.00260	.00095	-2.73869
$t22lag1^2$	.00304	.00095	3.18820
$(press–1000)^2$	.00008	.00004	2.02903
tlag2×qlag2	.00024	.00011	2.12990

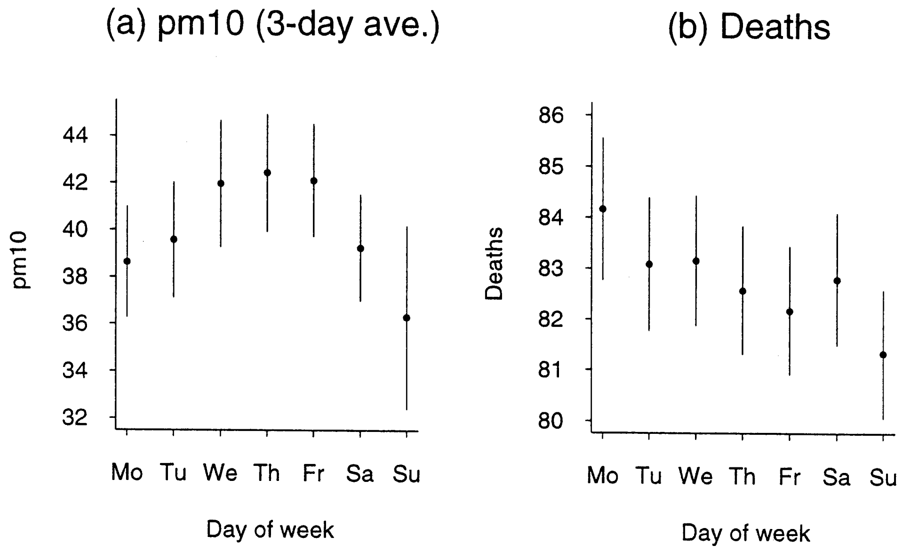
4. Styer *et al.* also found significant variability among different groups of the population, with males being more strongly affected than females, though with no significant difference between whites and blacks. They also performed a more detailed breakdown by "cause of death", finding that (as expected) deaths from respiratory causes were associated with the largest coefficient, but also finding a significant effect on deaths from cancer, which is harder to explain.

5. There is also a strong *day of week* effect (Fig. 6) – it appears that deaths are highest on a Monday and lowest on Sunday (one can speculate on a number of quite different reasons for this). The explanation for a weekday effect on pm10 is easier to explain, because the effect is consistent with traffic patterns (remember that our pm10 variable is based on three-day running averages so we would expect the highest levels to be on Wednesday, Thursday and Friday). However, when weekday is added as a variable to the model, the estimated pm10 coefficient does not change much.

6. Another difficulty is the presence of significant serial correlations in all simple models. There is no logical reason for this, but we are not talking about an infectious

disease, and the most obvious explanation for a spurious correlation – that of lagged effects – has already been taken into account in the specification of the model.

Fig. 6: Weekday effects and 95% confidence intervals

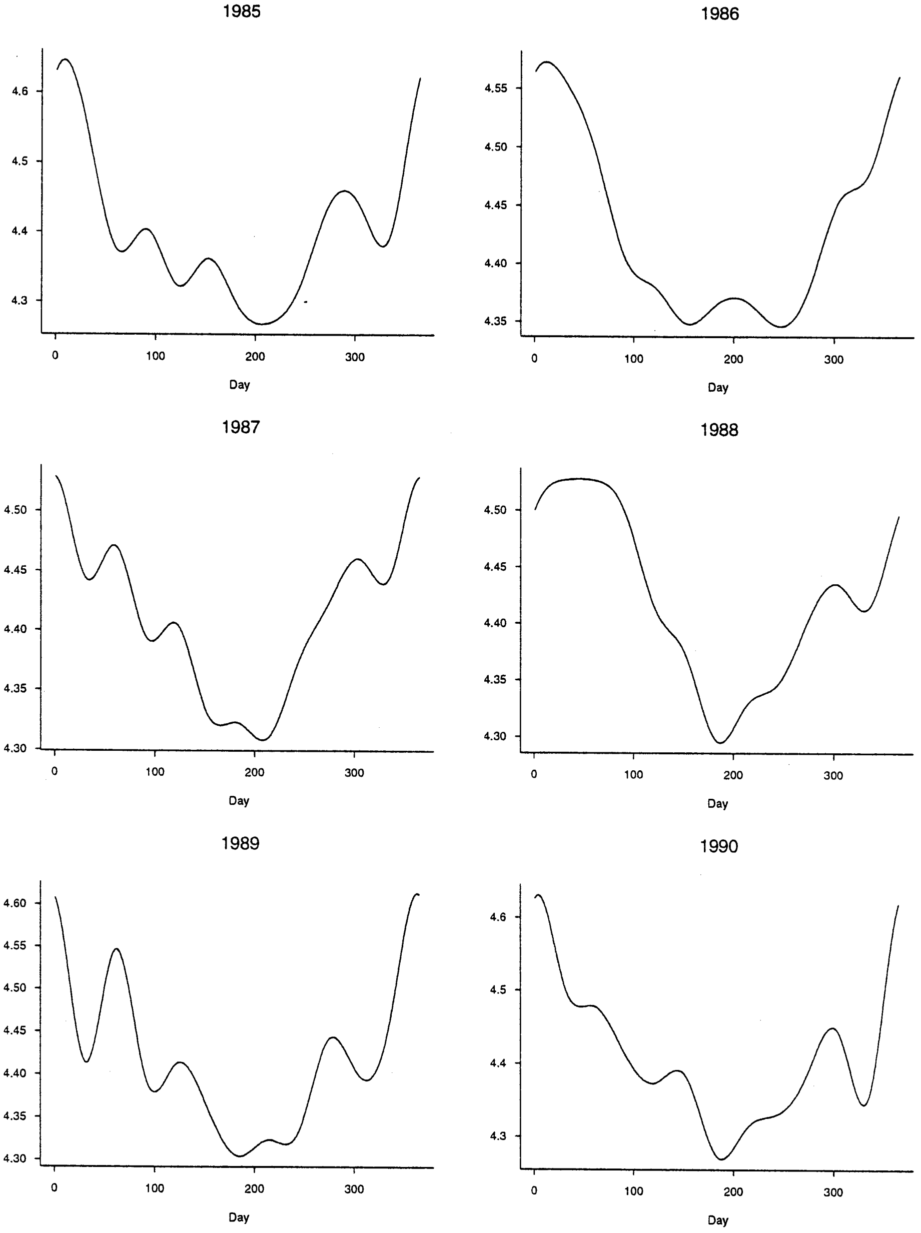


After all these considerations have been taken into account, the best model that I have been able to find uses a separate 12-knot spline to model the seasonal effect in each year. This model takes care of the season × year interaction and is the only model to remove all traces of serial correlation. Fig. 7 shows the estimated seasonal effects.

However, with this model the estimated pm10 coefficient is only .00036, less than half the value originally estimated and well below the effect claimed by Schwartz and his co-workers, and the standard error of this is .00021, raising the question of whether it is significant at all. Studies of similar models using the Salt Lake County data (where the total population base is much smaller) failed to show a significant effect in any version of the model.

To summarise, although this analysis does not disprove the existence of a pm10 effect, it does show that the questions surrounding it are very complicated. There does not appear to be any clear-cut choice of the best model, yet the model selection has a crucial effect on the results obtained.

Fig. 7: Seasonal trends within each year



REFERENCES

Bloomfield, P. (1992), Trends in global temperature. *Climatic Change* **21**, 1-16.

Bloomfield, P. and Nychka, D. (1992), Climate spectra and detecting climate change. *Climatic Change* **21**, 275-287.

Bloomfield, P., Royle, A. & Yang, Q. (1993), Accounting for meteorological effects in measuring urban ozone levels and trends. National Institute of Statistical Sciences Technical Report #1.

Collett, D. (1991), *Modelling Binary Data.* Chapman and Hall, London.

Cox, D.R. and Isham, V. (1988), A simple spatial-temporal model of rainfall. *Proc. Roy. Soc. Lond. A* **415**, 317-328.

Davison, A.C. and Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). *J.R. Statist. Soc.*, **52**, 393-442.

Dawid, A.P. (1982), The well-calibrated Bayesian (with discussion). *J. Amer. Statist. Assoc* **77**, 605-613.

Dawid, A.P. (1984), Statistical theory: the prequential approach (with discussion). *J.R. Statist. Soc. B* **147**, 278-292.

Dawid, A.P. (1986), Probability forecasting. In *Encyclopedia of Statistical Sciences*, Vol. 7, eds. S. Kotz, N.L. Johnson and C.B. Read. New York: Wiley-Interscience, 210-218.

Fowler, D. *et al.* (1993), Ozone in the United Kingdom. United Kingdom Photochemical Oxidants Review Group. Published by the Air Quality Division of the Department of the Environment.

Handcock, M.S. and Wallis, J.R. (1994), An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Statist. Assoc.* **89**, 368–390.

Ledford, A. and Tawn, J. (1994), Statistics for near independence in multivariate extreme values. Submitted for publication.

Pope, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E. and Heath, C.W. (1995), Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *Am. J. Respir. Crit. Care Med.* **151**, 669–674.

Rodriguez-Iturbe, I., Cox, D.R. and Isham, V. (1987), Some models for rainfall based on stochastic point processes. *Proc. Roy. Soc. Lond. A* **410**, 269-288.

Rodriguez-Iturbe, I., Cox, D.R. and Isham, V. (1988), A point process model for rainfall: further developments. *Proc. Roy. Soc. Lond. A* **417**, 283-298.

Schwartz, J. (1993), Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology* **137**, 1136–1147.

Schwartz, J. (1994), Air pollution and hospital admissions for the elderly in Birmingham, Alabama. *American Journal of Epidemiology* **139**, 589–598.

Seaton, A., MacNee, W., Donaldson, K. and Godden, D. (1995), Particulate air pollution and acute health effects. *The Lancet* **345**, 176–178 (January 21 1995).

Seillier-Moiseiwitsch, F. (1995), Predictive assessment of logistic models. To appear, *Statistics in Medicine.*

Seillier-Moiseiwitsch, F. & Dawid, A.P. (1993), On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc* **88**, 355-359.

Shively, T.S. (1991), An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment* **25B**, 387-396.

Smith, R.L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science* **4**, 367-393.

Smith, R.L. (1993), Long-range dependence and global warming. In *Statistics for the Environment* (V. Barnett and F. Turkman, editors). John Wiley, Chichester, 141-161.

Smith, R.L. (1994), Multivariate threshold methods. In *Extreme Value Theory and Applications*, eds. J. Galambos, J. Lechner and E. Simiu. Kluwer Academic Publishers, Dordrecht, pp. 225–248.

Smith, R.L. and Huang, L.-S. (1994), Modeling high threshold exceedances of urban ozone. National Institute of Statistical Sciences Technical Report #6.

Smith, R.L., Huang, L.-S. and Seillier-Moiseiwitsch, F. (1995), Predictive diagnostics for high threshold exceedances of urban ozone. In preparation.

Smith, R.L., Tawn, J. and Coles, S. (1993), Markov chain models for threshold exceedances. Submitted for publication.

Smith, R.L. and Shively, T.S. (1995), A point process approach to modeling trends in tropospheric ozone. To appear, *Atmospheric Environment*.

Smith, R.L. and Robinson, P.J. (1995), A Bayesian approach to modelling spatial-temporal precipitaion data. Proceedings of the Sixth International Meeting on Statistical Climatology, Galway, Ireland.

Styer, P., McMillan, N., Gao, F., Davis, J. and Sacks, J. (1995), The effect of outdoor airborne particulate matter on daily death counts. *Environmental Health Perspectives*, May 1995.

Young, P.C. and Lees, M. (1993), The active mixing volume: a new concept in modelling environmental systems. In *Statistics for the Environment* (V. Barnett and F. Turkman, editors). John Wiley, Chichester, 3–43.