HARVESTING - RESISTANT ESTIMATES OF AIR POLLUTION EFFECTS ON MORTALITY

Scott L. Zeger, Francesca Dominici, and Jonathan Samet

Running title: Harvesting-Resistant Estimates

Word count: 2748

Full names and affiliations: Scott L. Zeger, Department of Biostatistics, School of Hygiene and Public Health, Johns Hopkins University; Francesca Dominici, Department of Biostatistics, School of Hygiene and Public Health, Johns Hopkins University; Jonathan M. Samet, Department of Epidemiology, School of Hygiene and Public Health, Johns Hopkins University.

Address for correspondence: Prof. Scott L. Zeger, Department of Biostatistics, The Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205-3179. USA

Acknowledgments: Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the Environmental Protection Institute (EPA R824835) and automotive manufacturers. The contents of this article do not necessarily reflect the views and policies of HEI, nor do they necessarily reflect the views and policies of EPA, or motor vehicles or engine manufacturers. We are grateful to Julia Kelsall, co-developer of the software used to perform frequency domain regression.

Abstract

A number of studies have recently shown an association between particle concentrations in outdoor air and daily mortality counts in urban locations. In the public health interpretation of this evidence, a key issue is whether the increased mortality associated with higher pollution levels is restricted to very frail persons for whom life expectancy is short in the absence of pollution. This possibility has been termed the "harvesting hypothesis". We present an approach to estimating the association between pollution and mortality from times series data that is resistant to short-term harvesting. The method is based in the concept that harvesting alone creates associations only at shorter time scales. We use frequency-domain log-linear regression to decompose the information about the pollution-mortality association into distinct time scales and we then create harvesting-resistant estimates by excluding the short-term information that is affected by harvesting. We illustrate the methods with total suspended particles and mortality counts from Philadelphia for 1974-1988. We show that the total suspended particles-mortality association in Philadelphia is inconsistent with the harvesting only hypothesis and that the harvesting resistant estimates of the total suspended particles relative risk are actually larger, not smaller than the ordinary estimates.

Key words: mortality displacement, frequency domain log-linear regression, frailty models, air pollution.

The acute effects on morbidity and mortality of extreme episodes of particulate air pollution have been well documented by the 1952 London fog and other air pollution disasters (1). In more recent times, acute health effects have been associated with fluctuations in particulate air pollution well within the Environmental Protection Agency standards for the United States (2,3). Substantial evidence has accumulated over the last decade in support of associations of daily levels of air pollution with mortality counts and with measures of morbidity. The American Thoracic Society (4,5), and Dockery and Pope (6) provide overviews of this new literature. Samet et al. (7) have conducted re-analyses and have critically evaluated the pioneering work by Schwartz and Dockery (8) and largely confirm an acute association between mortality and particulate air pollution.

Nevertheless, uncertainty remains regarding the public health implications of these findings. First, controversy remains about whether a single constituent of air pollution is responsible for the increased mortality and morbidity or whether the adverse health effects are caused by combined actions of multiple pollutants (9). Even if a single constituent of the mix of pollutants in urban air, for example small particles, is largely responsible for increasing morbidity and mortality, a second question arises: is the increase in mortality only among extremely frail individuals whose remaining life expectancy would be short, in the absence of pollution? That is, are only a small number of total days of life lost from pollution, or are individuals dying who would otherwise have survived for substantial periods? The possibility that only extremely frail individuals die from exposure to air pollution has been termed the "harvesting" hypothesis (10), a phenomenon also referred to as mortality displacement".

One approach to investigating the possibility that only frail individuals are affected by air pollution uses a compartmental model (11). In the simplest frailty model, the death process is

assumed to have two steps. First, an individual moves from a relatively healthier population into a very frail subgroup from which all mortality occurs. In the second transition, persons in this frail pool die; the risk of dying increases at higher pollution levels. If the size of the pool of the frail persons is small, then the mean residency time in the frail condition is short, regardless of the exposure to pollution. In this case, harvesting can occur because persons in the frail pool have a short life expectancy in the absence of pollution. On the other hand, if the size of the frail population is large, the mean residency time in the frail state is relatively long, so that pollution-caused deaths substantially shorten life. More realistic extensions of this simple two-compartment model, for example to include three states: healthy, diseased and highly frail, can be developed but they would capture the harvesting principle in a similar way.

The evidence available in mortality time series data to assess whether harvesting occurs is in the pattern of mortality following days with a large number of deaths. If the frail subpopulation is small and if deaths can occur only from this pool, then the number of deaths on a day after a pollution event will be smaller than expected because the previous high mortality depletes the pool of at-risk frail individuals. Hence, we would expect a negative auto-correlation between the number of deaths on a day after a pollution episode and a day before that episode (12).

A few investigators have developed statistical models using the ideas above to estimate the size of the frail population and the expected days of life lost due to exposure to air pollution. Spix et al. (11) proposed a one-step Markov chain model, and demonstrated that the pollution relative risk estimates from Poisson regression are biased about 10%-30% by harvesting. Smith (13) used a two-compartment model as described above with the additional assumption that both the risk of becoming frail and the risk of death may depend on air pollution.

In this paper, we take a different approach to the harvesting issue. Rather than attempting to estimate directly the degree of harvesting, we propose a class of estimators of the pollution-mortality association that is resistant to shorter-term harvesting. That is, we propose an approach that ignores the information in the time-series data in which short-term harvesting would influence the mortality-pollution association. With this approach, pollution relative risks are close to unity if the association is due to harvesting alone. The method is based upon partitioning both the pollution and mortality time series data into components with variation occurring at different time scales and then relying on the longer-term components to estimate the effect of pollution on mortality.

First, we propose a simple, two-compartment model for harvesting and demonstrate that harvesting produces correlations between mortality and pollution data that are non-negligible only at short time scales. Then we review briefly an approach to time series modeling of the mortality-air pollution association that gives separate estimates of the pollution effect at different time scales. This "frequency domain log-linear regression" is described in detail by Kelsall et al. (14). We then propose a harvesting-resistant estimator that sets aside the short-term associations that are subject to the influence of harvesting. Further, we apply this methodology to the analysis of particulate air pollution and mortality for data from Philadelphia for the period 1974-1988, which were previously analyzed by several investigators (9,15).

A Simple Model for Harvesting

The simplest model that captures the harvesting phenomenon is based upon the assumption that individuals in the general population transition into a very frail subgroup, and that deaths only occur from among the frail subpopulation. This idea can be implemented in the

following difference equation $N_t = N_{t-1} - D_{t-1} + I_t$ where N_t *is* the size of the frail population; I_t is the number of new persons from the general population who become frail at the start of day *t;* D_t is the number of deaths, and x_t is the pollutant value for day *t.* This equation simply states that the frail population on day *t* comprises those frail individuals less the number of deaths from the previous day plus the newly frail. Hence on days following a large number of deaths, the very frail population at high risk of pollution-induced death is reduced and fewer deaths can result.

Given N_t frail persons on day *t*, we assume that each person independently experiences a daily hazard of death l_t, where the log-odds of death depends linearly on the pollution value x_t. Finally we assume that there is an effectively infinite total population and that, the number of persons I_t that enter the frail subpopulation at the starts of day *t* follows a Poisson distribution with mean $m_I\left(\frac{I^*_{t-1}}{m_I}\right)^{a}$, where I^*_{t-1} is the number of entrants from the previous day or a small positive constant if $I_{t-1} = 0$. This model for I_t allows the number of persons entering the frail state on a given day to depend on the number which entered on the previous day, introducing some positive auto-correlation (when $\alpha > 0$) that might reflect the influence of such events as influenza epidemics or stretches of bad weather (16). We assume that the long-term average of persons entering the frail subgroup $E[I_t]$ and the long-term average of the frail persons dying per day $E[D_t]$ are equal so the population neither grows nor shrinks in the long-run. Finally, to initiate the mortality D_t and frailty N_t time series, we assume that the number of persons at risk on the first day follows a Poisson distribution with mean equal to the long term average $E[N_t]$.

To demonstrate that harvesting induces associations between mortality and pollution only at short time scales, we have generated three long (N_t, D_t) series, which represent different degrees of harvesting. We chose the parameter values in Table 1 so that the mean number of

deaths per day is 50, similar to the Philadelphia data, and the mean residence times (MRT) in the frail state are 3, 30, or 300 days. The log-relative risk of mortality associated with particles was .05 for all scenarios. These data were generated using for x_t the total suspended particles series from Philadelphia for 608 days during the period 1980 to 1994 with 100 repetitions to create the simulated series of length 60800.

Figure 1 displays the association, as measured by the squared correlation (coherence), between the number of deaths (square root transformed) $\sqrt{D_t}$ and particle measurements x_t as a function of time-scale, for the three harvesting scenarios. These plots were generated by cross-spectral analysis (17). Note that in all three situations the correlation becomes non-negligible only at time scales that are less than about twice the mean residence time (MIRT). Hence, we see that harvesting induces association between mortality and pollution only at shorter time scales.

In the next two sections, we review our approach to estimating the mortality-pollution association separately at multiple time scales and then adopt the approach to produce harvesting-resistant estimates of pollution relative risks that ignore the short-term information that can be influenced by harvesting.

Frequency Domain Log-Linear Regression

The main idea of FDLLR is to decompose both the air pollution series x_t and the mortality series (here after referred to as y_t) into distinct component series x_{kt} and y_{kt}, one pair for each of many distinct time scales k and to calculate the association separately between x_{kt} and y_t for each time scale.

Figure 2 shows such a decomposition into five time scales -- i.e. roughly year, season, month, week, day -- for the total suspended particles and mortality series from Philadelphia for

the period 1974-1988. Note the top series comprise only the longest-term fluctuations, while the bottom series represents the shortest-term variations. The actual value of x_t (or y_t) on day t is obtained by summing the values of the five component series on that day. This type of decomposition can be obtained by smoothing with successively shorter running averages.

FDLLR estimates a separate coefficient $\hat{b}_k$ by regressing each y_t component on its corresponding component x_t giving a sequence of regression coefficients, e.g. $\hat{b}_1$, $\hat{b}_2$, $\hat{b}_3$, $\hat{b}_4$, $\hat{b}_5$ in this five component illustration.

The actual implementation of FDLLR uses a Fourier series decomposition of the *xt* and *yt* series, and produces a smooth pollution-mortality log relative risk function $\hat{b}_k$ of the time scale or equivalently the frequency: *k* cycles in the total period of observations, rather than estimates at only five or some other small number of time scales. Kelsall et al. (14) give a detailed description of FDLLR.

Harvesting-resistant estimates of mortality-pollution relative risk

As we demonstrated above, harvesting will only affect $\hat{b}_k$ s for large *k,* say $k \geq K$. Under the hypothesis that harvesting is the only cause of the pollution-mortality association, we would expect $\hat{b}_k$ to be near zero at low frequencies *(k* small) and to increase in absolute value towards higher frequencies (shorter time scales). A plot of $\hat{b}_k$ versus *k* is therefore informative with regard to the harvesting hypothesis. We can also calculate a single harvesting-resistant estimator

of b by taking an appropriately weighted average of the $\hat{b}_k$ *s* for $1 \leq k \leq K$, ignoring information at higher frequencies. The estimator is specifically defined as:

$$\hat{b}_k = \left(\Sigma_{k \leq K} w_k \bar{x}_k x_k\right)^{-1} \left(\Sigma_{k \leq K} w_k \bar{x}_k z_k\right) \qquad (1)$$

where x_k and z_k are the discrete Fourier transforms of x_t and of the linearized response z_t used in generalized linear models (16), $w_k^{-1} = Var(z_k)$ and $\bar{x}_t$ is the complex conjugate of *x*.

Many pollution-mortality analyses use daily time-series data (15). With series of *n* days, the possible values of *k* range from one to *n/2* complete cycles in *n* days corresponding to frequencies: one cycle in *n* days to one cycle in two days. We seek a harvesting-resistant estimator that ignores $\hat{b}_k$ *s* corresponding to periods shorter that about twice the mean residence time in the frail state. We showed above, that under a simple model, the pollution-mortality associations are negligible at the remaining time scales when harvesting is the only source of this association.

When our harvesting-resistant estimator is applied to the simulated data with *K* corresponding to twice the mean residence time, the estimated pollution effect is close to 0 in all three cases.

Table 2 shows the fraction of the total frequencies corresponding to the shorter-time scales that must be ignored to protect against harvesting for various values of the mean residence time in the frail state given daily time series data. A large fraction of the available information is affected by harvesting. Nevertheless, the remaining information at the longer frequencies can be used to estimate the pollution-mortality association without being biased by short-term harvesting as illustrated for Philadelphia below.

Application to Philadelphia data

To illustrate our method we consider mortality and air pollution in Philadelphia for the years 1974-1988. The data set is the same as that used by Samet et al. (15). In addressing the harvesting question we implement our FDLLR while adjusting for temperature and dew-point, and longer-term trends associated with factors such as changes in medical practice, demographics and influenza epidemics. The adjustment is actually implemented using smoothing splines (18) with 6 degrees of freedom for temperature and dew point and with 90 degrees of freedom for time, respectively. The adjustment is similar to but not identical to those by Samet et al. (18) who used 120 degree of freedom for time and included day of the week as well.

Our goal is to calculate estimates of the pollution-mortality association in time series data that sets aside (ignores) information affected by harvesting. We first apply the frequency domain regression methods that estimate the pollutant/mortality association separately at each time scale.

On the left of Figure 3 we show the time-scale-specific estimates of the mortality relative risk associated with total suspended particles by time scales. The horizontal axis is the time scale in days at which the association is measured. The solid and dotted lines are the estimated log relative risks plus or minus two estimated standard errors, respectively, at each time scale. The plot is scaled to show the expected change in mortality corresponding to a change of one interquartile range (lQR mg/m^3) in total suspended particles. On the right of Figure 3, we show the estimated harvesting-resistant effects of total suspended particles on mortality as successively more of the shorter-term information is removed. The estimate at a particular time scale is calculated disregarding information in the left hand panel at all shorter time scales.

Note that the pattern in the left panel is the opposite of the expectation under the harvesting hypotheses. The pollution relative rate is substantially different from zero at low

frequencies, and in fact, decreases rather than increases toward shorter-term frequencies. Hence the "harvesting" only hypotheses is inconsistent with the Philadelphia data. The harvesting-resistant estimators corresponding to mean residence times of 2 and 4 days are .022 (95 % Cl .012 to .032) and .024 (95 % CI .015 to .033), respectively indicating that the association between total suspended particles and mortality reflects factors other than harvesting alone.

Discussion

The method we present can be approximated by using filtering techniques to split the pollutant and mortality time series into components having variations on distinct time-scales as was done in Figure 2. A log-linear regression of the component mortality on the corresponding component pollution series could be performed to obtain a separate relative-risk estimates for each component pair. Our method has the advantage of giving relative-risk estimates that are continuous functions of time-scale rather than providing only a few discrete values. It also provides valid confidence intervals that are not directly available from log-linear regression programs.

In the current implementation of our harvesting-resistant estimator, we fix the time lag between the pollution-exposure and mortality, rather than estimate it from the data. Hence it is advisable to consider multiple, reasonable lags. For Philadelphia, the results were qualitatively the same when the total suspended particles was lagged 0, 1, 2 or 3 days.

In gauging the public health significance of the evidence from the daily time-series studies, we find that the extent of harvesting has been a major point of controversy. With little evidence of significant life shortening, the findings from the time-series studies may not warrant a regulating response. In the recent standard-setting process for particulate matter, two cohort studies have

figured prominently because their findings indicate longer-term effects (19,20). Using a new analytic approach, our reassessment of the Philadelphia data indicates that the previously reported associations between air pollution indicators and mortality cannot be attributed solely to harvesting. This analytic approach should be extended to additional data sets to assess the consistency of our findings in Philadelphia.

Figure Captions

Figure 1. Estimated squared coherency between each pair of $\sqrt{D_t}$ and x_t.

Figure 2. Decomposition into five component series for the mortality and total suspended particles ($\mu g/m^3$) series (on squared root scale) from Philadelphia for the period 1974-1988. Each plot represents the residual time series respect to the previous component.

Figure 3. Time-scale (frequency)-specific and cumulative estimates of the log-relative risk for mortality associated with current day total suspended particles versus time scales in days estimated from daily time-series data from Philadelphia for 1974-1988. The solid and dotted lines are the estimates plus or minus two estimated standard errors respectively. To the left of the cumulative-estimates plot there are harvesting-resistant estimators corresponding to a mean residence time (MRT) of 2 days and 4 days respectively. The cumulative estimates at a certain time scale *K* (no. of days per cycle) is calculated disregarding time scales shorter than *K*.

Table 1: Parameter Values for the Simulated Example

Simulated Example			
Parameters	MRT=3	MRT=30	MRT=300
$E[N_t]$	150	1500	15000
$E[I_t]$	50	50	50
$E[D_t]$	50	50	50
00	-.69	-3.36	-5.70
01	.05	.05	.05
a	.80	.80	.80
C	1	1	1

Table 2: Fraction of frequencies whose corresponding periods are shorter than twice the mean residence time (MRT) in the frail state

MRT (days)	% of frequencies to ignore
2	50
4	75
8	87.5
16	93.75

Reference List

1. Beaver, H. Interim Report (on London air pollution incident), Committee on Air Pollution: CMD 9011. London. Her Majesty's Stationary Office. 1953.

2. US Environmental Protection Agency (EPA). Proposed guidelines for neurotoxicity risk assessment. 1995; 60(192), p.52031.

3. US Environmental Protection Agency (EPA) and Office of Air Quality Planning and Standards. Review of the National Ambient Air Quality Standards for Particulate Matter: Policy Assessment of Scientific and Technical Information. OAQPS Staff Paper. Research Triangle Park, North Carolina. U.S. Government Printing Office. 1996; EPA-452\R-96-013.

4. American Thoracic Society, Committee of the Environmental and Occupational Health Assembly, Bascom R, Bromberg PA, Costa DA, Devlin R, Dockery DW, Frampton MW, Lambert W, Samet JM, et al. Health effects of outdoor air pollution. Part 1. Am J Resp Crit Care Med 1996;153:3-50.

5. American Thoracic Society, Committee of the Environmental and Occupational Health Assembly, Bascom R, Bromberg PA, Costa DA, Devlin R, Dockery DW, Frampton MW, Lambert W, Samet JM, et al. Health effects of outdoor air pollution. Part 2. Am J Resp Crit Care Med 1996;153:477-98.

6. Dockery DW, Pope CA, III. Acute respiratory effects of particulate air pollution. Annu Rev Public Health 1994;15:107-32.

7. Samet, J.M., Zeger, S.L., and Berhane, K. The Association of Mortality and Particulate Air Pollution. Cambridge, Massachusetts. Health Effects Institute. 1995; p.1 Particulate Air Pollution and Daily Morality: Replication and Validation of Selected Studies.

8. Schwartz J, Dockery DW. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am Rev Respir Dis 1992;145(3):600-4.

9. Samet JM, Zeger SL, Kelsall JE, et al. Air pollution, weather, and mortality in Philadelphia 1973-1988. Health Effects Institute Report, editor. Particulate air pollution and daily mortality: analyses of the effects of weather and multiple pollutants. 1997.

10. Schimmel H, Murawski TJ. Proceedings: the relation of air pollution to mortality. J Occup Med 1976;18(5):316-33.

11. Spix C. Daily time-series of mortality counts: estimating the harvesting effects. Stat Med 1997;(submitted).

12. Lipfert FW, Wyzga RE. Air pollution and mortality: issues and uncertainties. Journal of the Air & Waste Management Association 1995;45(12):949-66.

13. Smith L. ASA Proceedings: assessing the human health risk of atmospheric particle. Env Stat 1997.

14. Kelsall J, Zeger S, Samet J. Frequency domain log-linear models; air pollution and mortality. J Royal Stat Soc 1997.

15. Kelsall JE, Samet JM, Zeger SL, Xu J. Air pollution and mortality in Philadelphia, 1974-1988. Am J Epidemiol 1997;146(9):750-62.

16. Zeger SL, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. Biometrics 1988;44(4):1019-31.

17. Bloomfield P. Fourier Analysis of Time Series: An Introduction. New York, New York: John Wiley & Sons, Inc.; 1976.

18. Hastie TJ; Tibshirani RJ. Generalized additive models. New York: Chapman and Hall; 1990.

19. Dockery DW, Pope CA, III, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Jr., Speizer FE. An association between air pollution and mortality in six U.S. cities. N Engl J Med 1993;329(24):1753-9.

20. Pope CA, III, Dockery DW, Schwartz J. Review of epidemiological evidence of health effects of particulate air pollution. Inhal Toxicol 1995;7(1):1-18.