APPENDIX AVAILABLE ON THE HEI WEB SITE

Research Report 183

Development of Statistical Methods for Multipollutant Research

Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants

E.S. Park et al.

Appendix F. Application to the Harris County Mortality Data from Multiple Subregions

Note: Appendices available only on the Web have been reviewed solely for spelling, grammar, and cross-references to the main text. They have not been formatted or fully edited by HEI.

Correspondence may be addressed to Dr. Eun Sug Park, Texas A&M Transportation Institute, The Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135; e-park@tamu.edu.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83467701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee.

Appendix F: Application to the Harris County Mortality Data from Multiple Subregions

In this section, we illustrate the use of estimated source contributions from spatially-enhanced multivariate receptor models in the health effects modeling. We defined “subregions” within Harris County by dividing the entire county into M (= 9) contiguous subregions containing the 9 monitoring sites of Canister VOCs. The following rule was applied to assign each census tract in Harris County to one of those 9 subregions. If a monitor is in a census tract that tract was assigned to its monitor. Otherwise, the closest monitor to the center of a census tract determined the census tract assignment to each subregion. Each subregion plays a role of a 10-mile buffer defined earlier (i.e., in the analysis near Clinton Drive). The mortality data were obtained for each subregion, and aggregated into daily counts. The time series of estimated source contributions for each subregion can then be obtained using the enhanced multivariate receptor models presented in the previous section.

For the health model, the average health outcome in day t and subregion m $\left(m=1,\cdots,M\right)$ may be represented as

$$\begin{aligned}\log E\left(y_t^m\right) &= \alpha + \sum_{k=1}^{q}\beta_k\gamma_{t-l,k}^m + \eta_w workday + \eta_{ET} ExtremeTemp \\ &\quad + \eta_{time} ns\left(time, df_{time}\right) + \eta_T ns\left(\mathrm{T}_t^m, df_T\right) + \eta_D ns\left(D_t^m, df_D\right) + \log\left(N_t^m\right)\end{aligned}$$

where $\gamma_{t-l,k}^m$ is the lag l source contribution for day t in subregion m, $ns\left(time, df_{time}\right)$ represents a natural cubic spline of time with df_{time} degrees of freedom, $ns\left(\mathrm{T}_t^m, df_T\right)$ represents a natural cubic spline of temperature with 2 days lag with df_{T} degrees of freedom, $ns\left(D_t^m, df_D\right)$ represents a natural cubic spline of dew point temperature with df_{D} degrees of freedom, and N_t^m is the size of the population at risk in subregion m on day t.

We fitted the above model to the daily counts of cause-specific mortality data (Cardiovascular, IHD, Acute MI, Heart Failure, Stroke, Respiratory, COPD, and Pneumonia), temperature and dew point temperature data, and the time series of estimated source contributions for each subregion obtained by the enhanced multivariate receptor models. We used 24 *df* (or 4 *df*/yr) for the smoothing of calendar time, and 4 *df* for the smoothing of temperature and 3 *df* for dew point temperature, with 2-day lags for temperature and 0-day lags for dew point, respectively. Lag days 0–2 were explored for each cause-specific mortality time series. Due to a large number of models that needed to be fitted, we used GLM procedure in R (R Development Core Team, 2006) in this illustration rather than using MCMC. Table F.1 presents source-specific effects on various cause-specific daily mortality counts associated with an IQR increase of estimated VOCs source contribution. The effects of Refinery on Cardiovascular mortality (lag 1), Acute MI (lag 0), Respiratory (lag 0 and lag 2), and COPD (lag 0) were all significant at $\alpha = 0.05$. The effects of Petrochemical production on Cardiovascular mortality and Pneumonia were both significant at lag 1. The effects of unburned gasoline on Acute MI (lag 2) and Heart failure (lag 1) as well as the effect of Vehicle Exhaust on Heart Failure (lag 0) were also significant. The effects of Natural Gas were significant for Cardiovascular mortality (lag 1), IHD (lag 1), Acute MI (lag 1), Heart Failure (lag 2), and Stroke (lag 1), but strangely with negative coefficients.

Table F.1. VOC source-specific effects on cause-specific mortality for residents ≥ 65 years at the time of death residing in Harris County

Cause of Mortality		**Source 1 (Refinery)**	**Source 2 (Petrochem)**	**Source 3 (Gasoline)**	**Source 4 (Natural Gas)**	**Source 5 (Vehicle Exhaust)**
Cardiovascular	β (lag 0)	0.026	0.009	0.018	–0.013	0.018
	β (lag 1)	**0.046**	**0.033**	–0.006	**–0.052**	–0.005
	β (lag 2)	0.016	0.013	–0.004	–0.023	0.015
IHD	β (lag 0)	0.020	0.031	0.003	–0.007	0.002
	β (lag 1)	0.020	0.033*	–0.031	**–0.041**	0.004
	β (lag 2)	–0.001	0.025	0.012	–0.008	–0.007
Acute MI	β (lag 0)	**0.060**	–0.010	0.004	0.020	0.004
	β (lag 1)	0.044	0.043*	–0.032	**–0.072**	–0.016
	β (lag 2)	0.001	–0.013	**0.065**	–0.007	–0.022
Heart Failure	β (lag 0)	0.038	0.009	0.047	–0.049	**0.125**
	β (lag 1)	0.048	–0.006	**0.117**	–0.074	–0.026
	β (lag 2)	–0.057	–0.018	–0.001	**–0.107**	0.073
Stroke	β (lag 0)	0.023	0.007	0.021	0.013	0.016
	β (lag 1)	0.061	0.030	–0.012	**–0.103**	–0.031
	β (lag 2)	0.029	–0.002	–0.024	–0.016	0.018
Respiratory	β (lag 0)	**0.077**	–0.017	0.016	–0.026	–0.019
	β (lag 1)	–0.056	0.046	0.051	–0.001	0.017
	β (lag 2)	**0.063***	0.040	0.001	–0.005	–0.031
COPD	β (lag 0)	**0.105**	0.007	–0.008	–0.060	0.011
	β (lag 1)	–0.082	0.005	0.048	0.030	0.025
	β (lag 2)	0.044	–0.013	–0.041	0.011	–0.008
Pneumonia	β (lag 0)	0.017	–0.042	0.021	0.021	–0.049
	β (lag 1)	–0.057	**0.125**	0.046	–0.017	0.057
	β (lag 2)	–0.054	0.053	0.087	0.046	–0.034

Notes: 1. The β coefficient of VOCs contributions from each source type represents the estimated log-relative risk per IQR increase of estimated VOCs source contribution; 2. Statistically significant effects (at α = 0.05) are denoted in bold; 3. * indicates that the statistical significance changed (from insignificant to significant or vice versa) when *df* for time changed from 4/year to 7/year.

We conducted a sensitivity analysis with varying degrees of freedom for natural splines for calendar time (4 *df*/year and 7 *df*/year) and weather variables (temperature: 2–6 *df*, dew point temperature 2–6 *df*). In general, no significant differences were observed. The statistical significance of β coefficients in Table F.1 mostly stayed the same except for a very few cases. When 7 *df*/year (instead of 4 *df*/year) was used for calendar time, β coefficient for petrochemical production for IHD (lag 1) and Acute MI (lag 1) changed to 0.041 to 0.0703 and became statistically significant while β for refinery on respiratory mortality (lag 2) changed to 0.0564 and lost its statistical significance (*P* value = 0.06).