Supplementary material to "Spatial misalignment in time series studies of air pollution and health data"

Roger D. Peng[1]   Michelle L. Bell[2]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD, 21205, USA

[2]School of Forestry and Environmental Studies, Yale University, New Haven CT, 06511, USA

1 Simulation Study

We designed a simulation study to assess the properties of the regression calibration and two-stage Bayesian methods. In this study we created a hypothetical grid of monitors over a large region and created a separate "county" within that region. For each monitor we simulated a time series of pollutant concentrations of length 500, which is roughly equal to what we observe in our dataset. The grid of monitors is dense so that we can reasonably approximate the true ambient average concentration by taking a simple average across the grid of monitors.

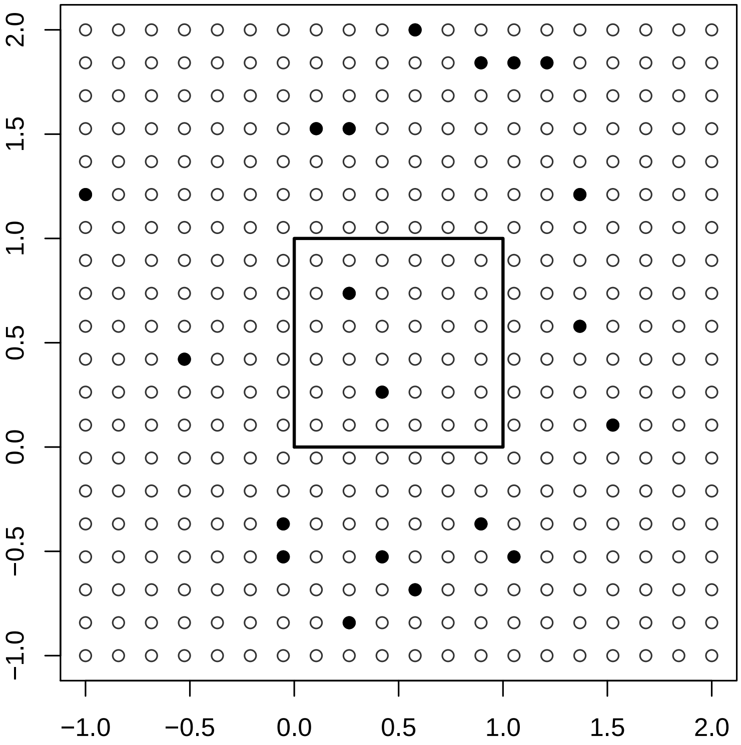

Figure 1: Design of simulation study. The interior box represents the county boundary and the dark solid circles represent observed monitor locations. The open circles represent unobserved monitors from which we can simulate the true (unobserved) ambient average exposure.

Figure 1 shows the setup for the monitor locations and the "county" boundaries. The open circles represent the dense grid of monitors that we use to estimate the true ambient average in the region. However, not all of the monitors are taken to be observed—the full grid of monitors is only used to calculate the true ambient average (what we previously called x_t). The solid black circles are the observed monitors that we use to create our simulated dataset. The solid square box within the region is our “county” and the solid black circles within that box are the observed monitors within that county (in this case there are only two monitors).

In the first step of the simulation we simulate spatially correlated time series at every monitor with varying degrees of spatial correlation. Once we have simulated data for all monitors we can compute the daily true ambient average concentration in the county by taking the mean of all monitors (i.e. all open circles) inside the county boundary. This gives us x_t. The health data are simulated using a Poisson model with the true county-wide ambient average as the predictor,

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t) \\ \log \mu_t &= \alpha + \theta x_t + \gamma z_t \end{aligned}$$

where α, θ, and γ are fixed beforehand and z_t is a confounder variable correlated with x_t. To simulate z_t, we used the linear model $z_t = 0.5x_t + \varepsilon_t$ where $\varepsilon_t \sim \mathcal{N}(0, \sigma_x^2/4)$ and σ_x^2 is the temporal variance of x_t. To create the dataset for analysis, we take the time series of health data along with the time series at each of the observed monitors, both inside and outside of the county (all solid black circles in Figure 1).

For simulating the spatially correlated time series, we used a Matérn model with $\kappa = 1$ and $\sigma = 0.73$, which was estimated using data from Cook County and was in the middle of the range of values estimated across chemical components. For ϕ we chose values 0.1, 0.5, and 1 which range from very rough to very smooth on the scale of the design in Figure 1. The three correlation functions used in the simulation are shown in Figure 2. We did not create any dependence in the temporal domain. Finally, we used $\alpha = 4$, which produced a mean outcome that was typical for a large US city and $\theta = 0.03$, which was a typical effect size for a chemical component, and $\gamma = 0.1$.

To estimate θ, we employed four methods: (1) a Poisson generalized linear model using the true county-wide ambient average exposure x_t; (2) the regression calibration approach; (3) the two-stage Bayesian model; and (4) the naive approach using $\bar{w}_t$ as the exposure estimate. We simulated 500 datasets and estimated θ using the four methods on each dataset. In each of the four approaches, z_t was also included into the model.

In the first step of the analysis, the spatial-temporal model is fitted to the data using all of the observed monitors (solid black circles) in the region. Then the fitted model is used to estimate the county-wide average for just the square box in the middle. For each day, the naive approach uses the mean of the values at the observed monitors as its estimate of the county-wide average. This gives us what we previously called $\bar{w}_t$. Figure 3 shows an example of a simulated dataset with the true county-wide average and the monitor average under the rough and smooth scenarios. For the "smooth" scenario the monitor average is a reasonable approximation of the truth, but for the "rough" scenario we see that the monitor average overestimates the temporal variation in the data. Figure 3 also shows estimates provided by the two-stage model for the county-wide average time series. In this case, the model estimates appear to match the truth reasonably well under both scenarios, with the estimate under the rough scenario appearing slightly underdispersed.

Figure 4 shows the distribution of the estimates of θ under the four different methods and the three levels of smoothness for the exposure data. We can see that for the rough pollutant data scenario, there is a clear bias/variance tradeoff between the regression calibration and two-stage model compared to the naive method which just uses the raw mean of the within-county monitors. Both the regression calibration and two-stage models appear reasonably unbiased across the simulations but estimate θ with much greater variability. The naive method is quite precise but is biased towards the null. Under the moderate smoothness scenario, all three methods do reasonably well with some bias incurred by the naive method. For the smooth scenario, all methods appear to perform equally well, as we expected. We also ran the simulations for value of σ that was ten times the value used to produce Figure 4. The conclusions were generally similar (Figure 5).

Table 1 shows the coverage performance of the 95% interval estimates for each method as well as the relative bias and the root mean squared errors. For the two-stage Bayesian model we use the 95% posterior interval. We can see that when $\phi = 0.1$ (the rough scenario), 95% confidence intervals for the naive method exhibit very poor coverage, only including the true estimate in its interval in 34% of the simulations. Under the smoother scenarios, the empirical coverage probabilities increase but do not reach the nominal 95% level. Similarly, the naive method is heavily biased under the rough scenario with estimates on average 87% smaller than the truth. The naive method makes up for its large bias by having a relatively small variance and hence a RMSE that is comparable to using the true county-wide ambient average. Both the regression calibration and two-stage Bayesian model performed similarly across scenarios and performance metrics. The methods were on average substantially less biased than the naive approach although their larger estimation variance contributed to a larger RMSE.

	ϕ	Truth	Reg Cal	Bayes	Naive
95% CI Coverage	0.1	0.953	0.953	0.935	0.343
	0.5	0.946	0.940	0.938	0.801
	1	0.921	0.933	0.931	0.885
Relative Bias	0.1	-0.101	-0.278	-0.311	-0.873
	0.5	0.019	-0.089	-0.093	-0.361
	1	-0.011	-0.038	-0.041	-0.201
RMSE	0.1	0.036	0.069	0.068	0.029
	0.5	0.015	0.016	0.016	0.017
	1	0.014	0.014	0.014	0.014

Table 1: Average coverage of 95% confidence/posterior intervals (CI), relative bias, and root mean squared error for the four estimation methods.

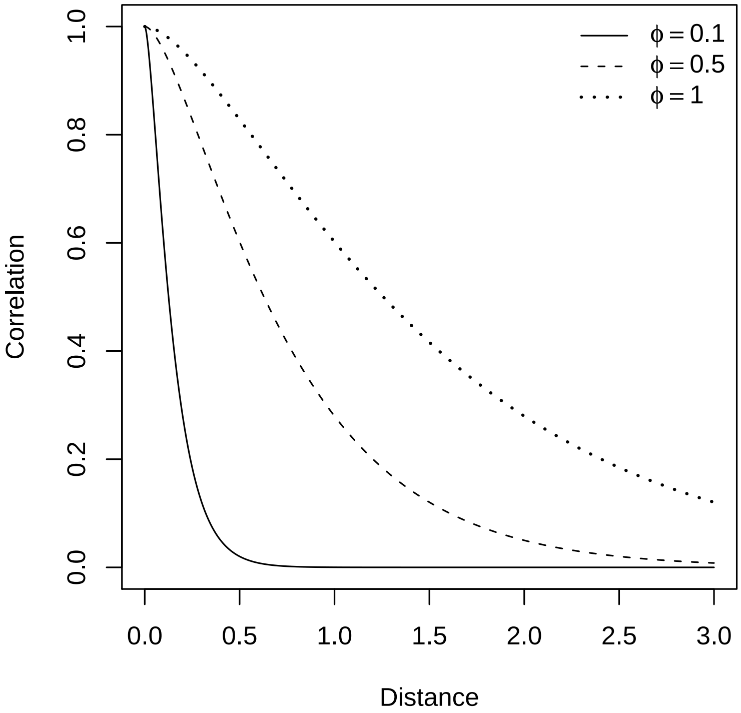

Figure 2: Matérn correlation functions used for simulation study; $\kappa = 1$ for all functions.

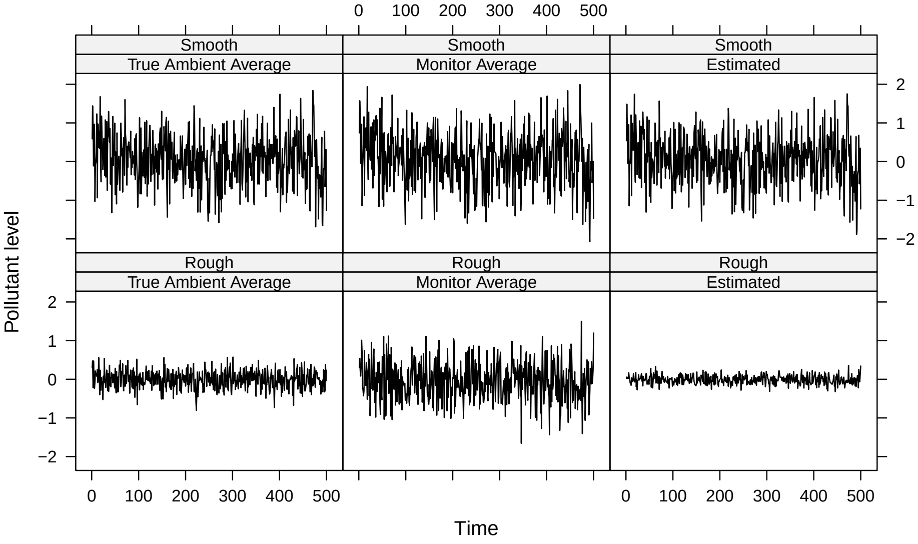

Figure 3: Simulated true county-wide average, observed monitor average, and estimated county-wide average using the two-stage Bayesian model under the rough and smooth scenarios.

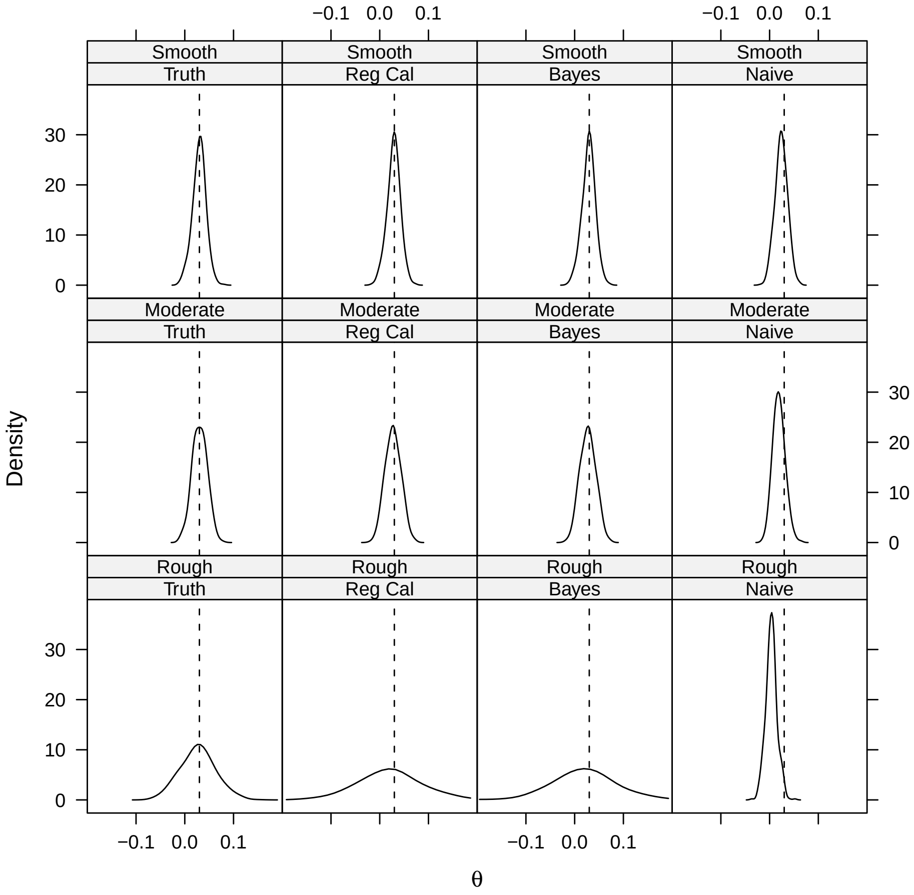

Figure 4: Distributions of $\hat{\theta}$ across all simulations using the regression calibration (Reg Cal) and two-stage Bayesian model (the true value is $\theta = 0.03$) when $\sigma = 0.73$.

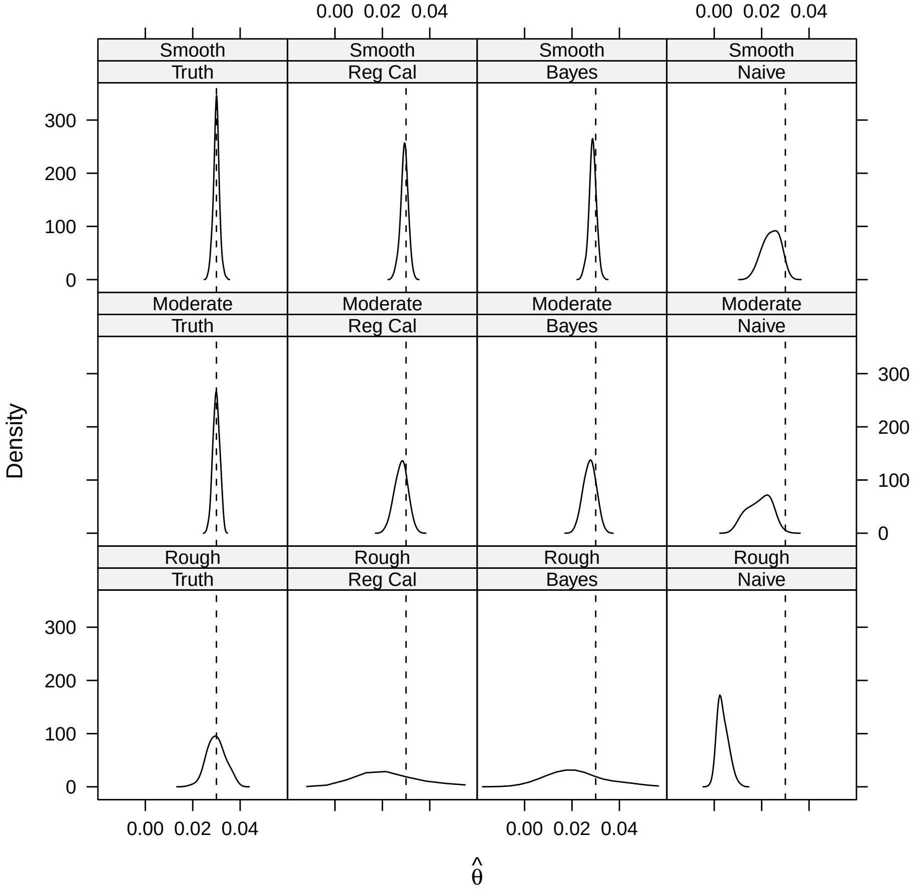

Figure 5: Distributions of $\hat{\theta}$ across all simulations using the regression calibration (Reg Cal) and two-stage Bayesian model (the true value is $\theta = 0.03$) when $\sigma = 7.3$

2 Supplementary Figures

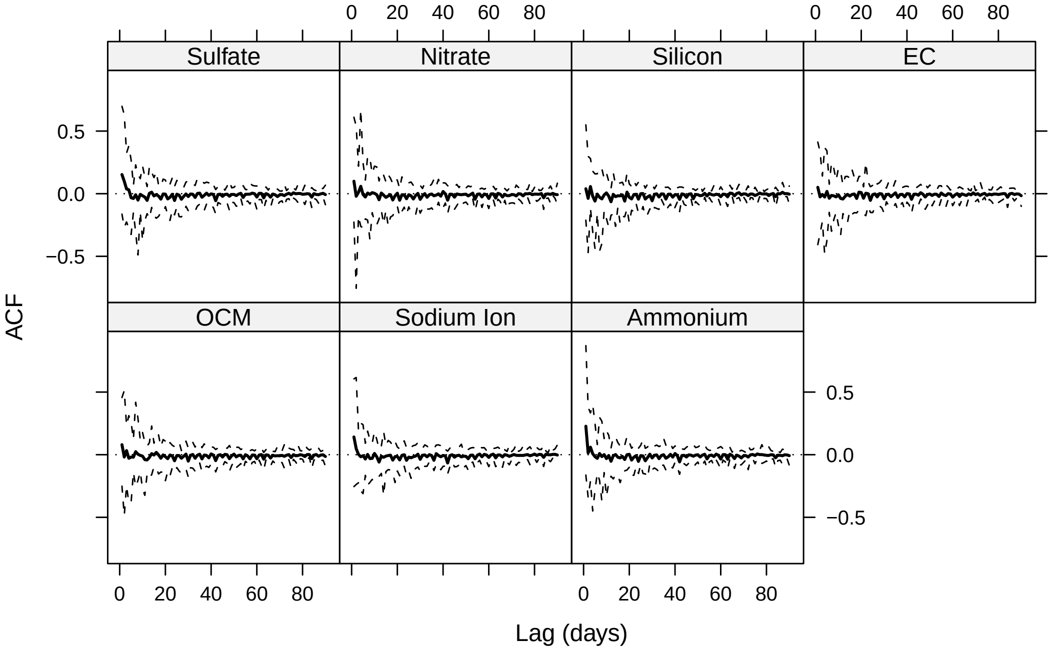

Figure 6: Autocorrelation functions (lags 1–90 days) for each chemical component after detrending, averaged across all monitors; dashed lines indicate lower 5% and upper 95% limits of the distribution across monitors.

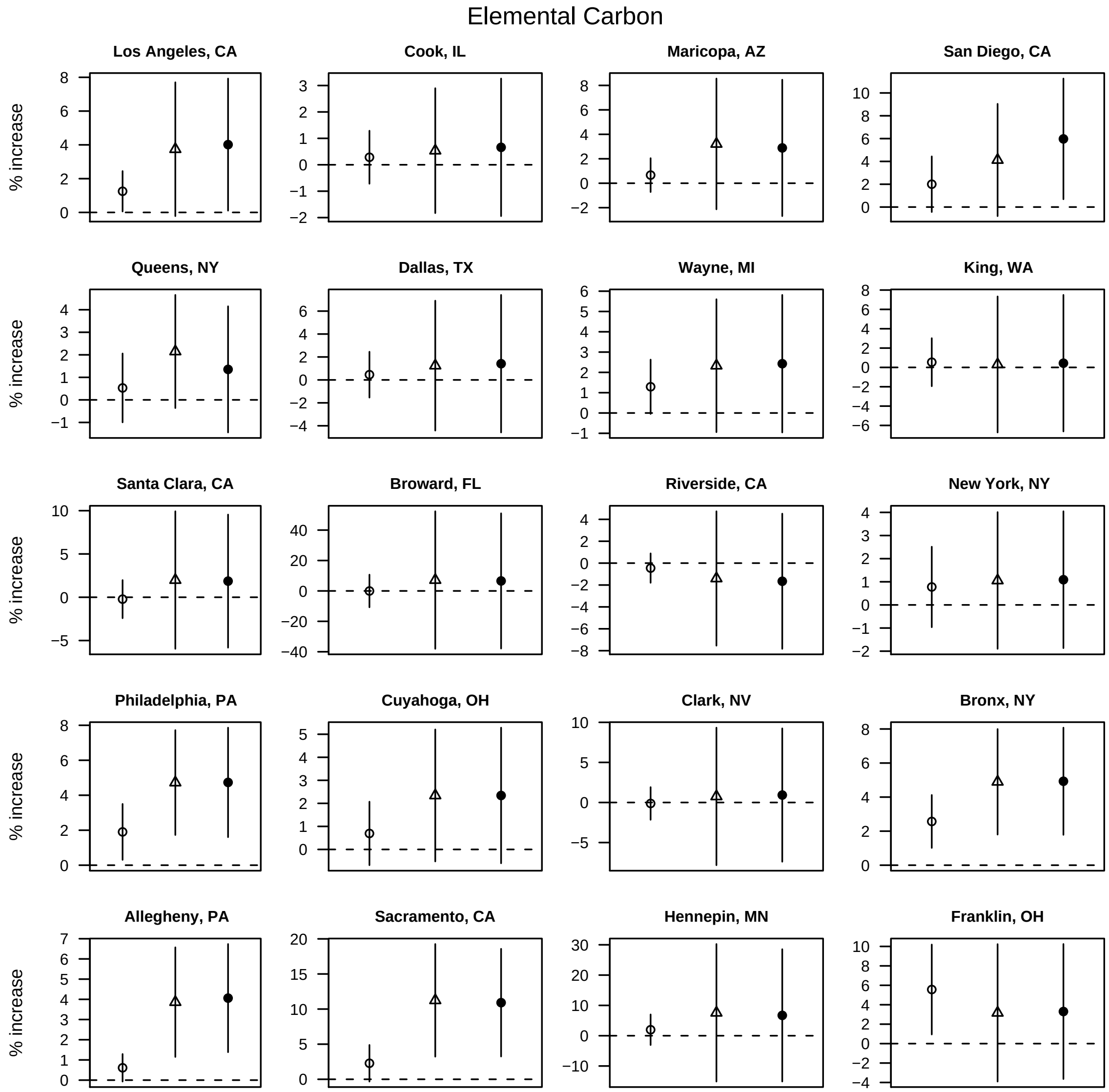

Figure 7: Percent increase in cardiovascular hospital admissions per 1 interquartile range (0.40 μg/m^3) increase in elemental carbon estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

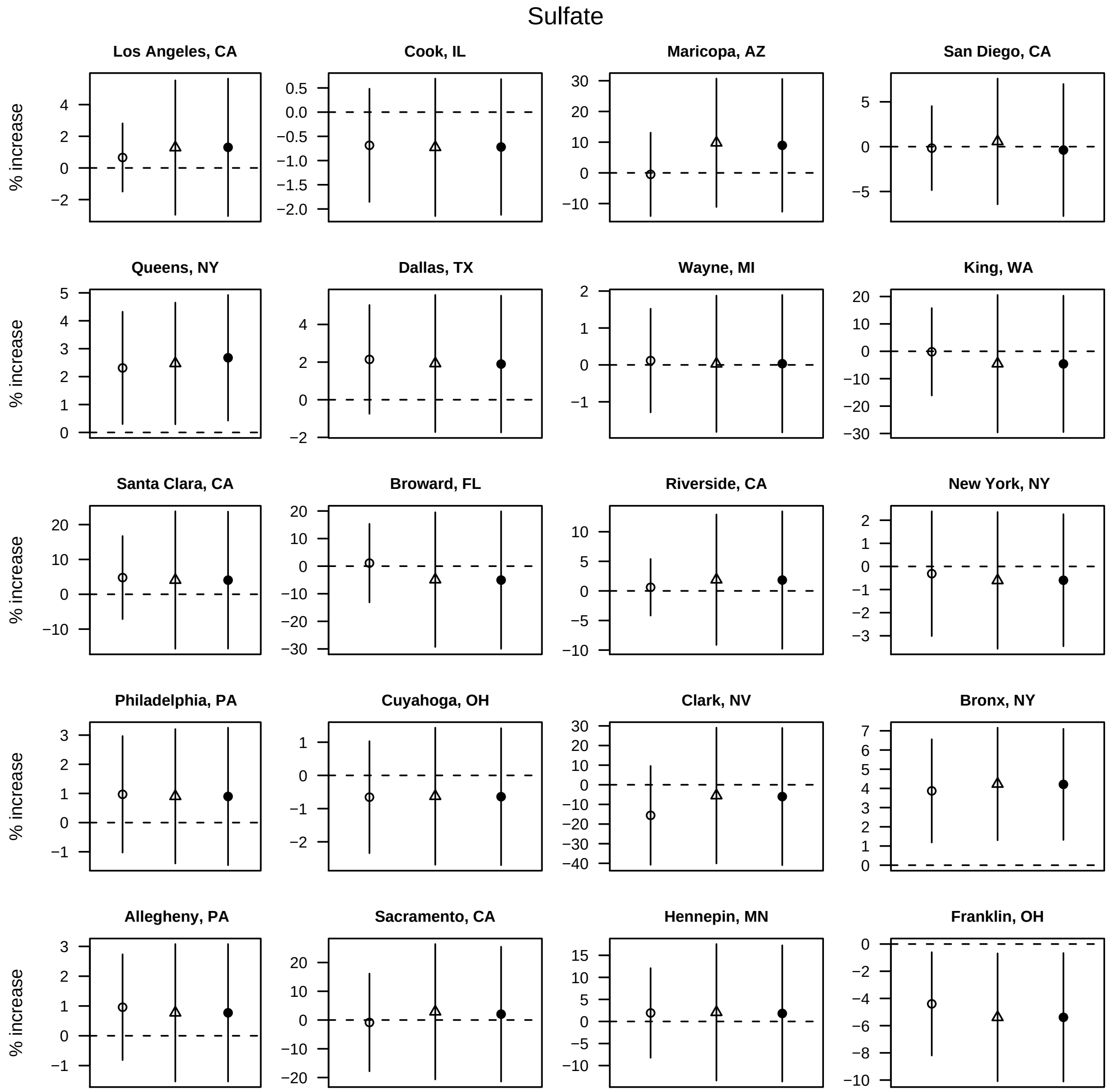

Figure 8: Percent increase in cardiovascular hospital admissions per 1 interquartile range (3.06 μg/m^3) increase in elemental carbon estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

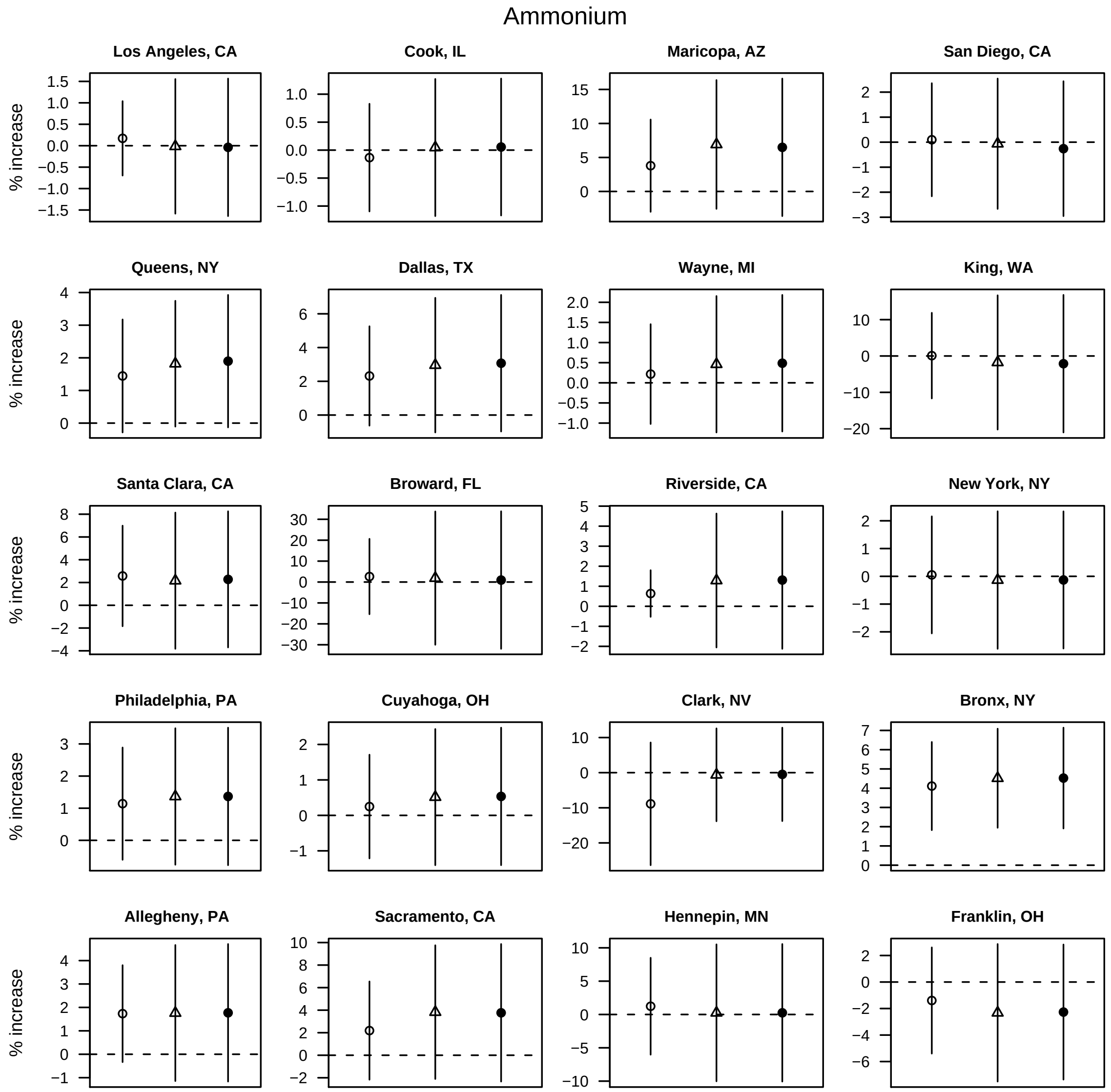

Figure 9: Percent increase in cardiovascular hospital admissions per 1 interquartile range (1.35 μg/m^3) increase in ammonium estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

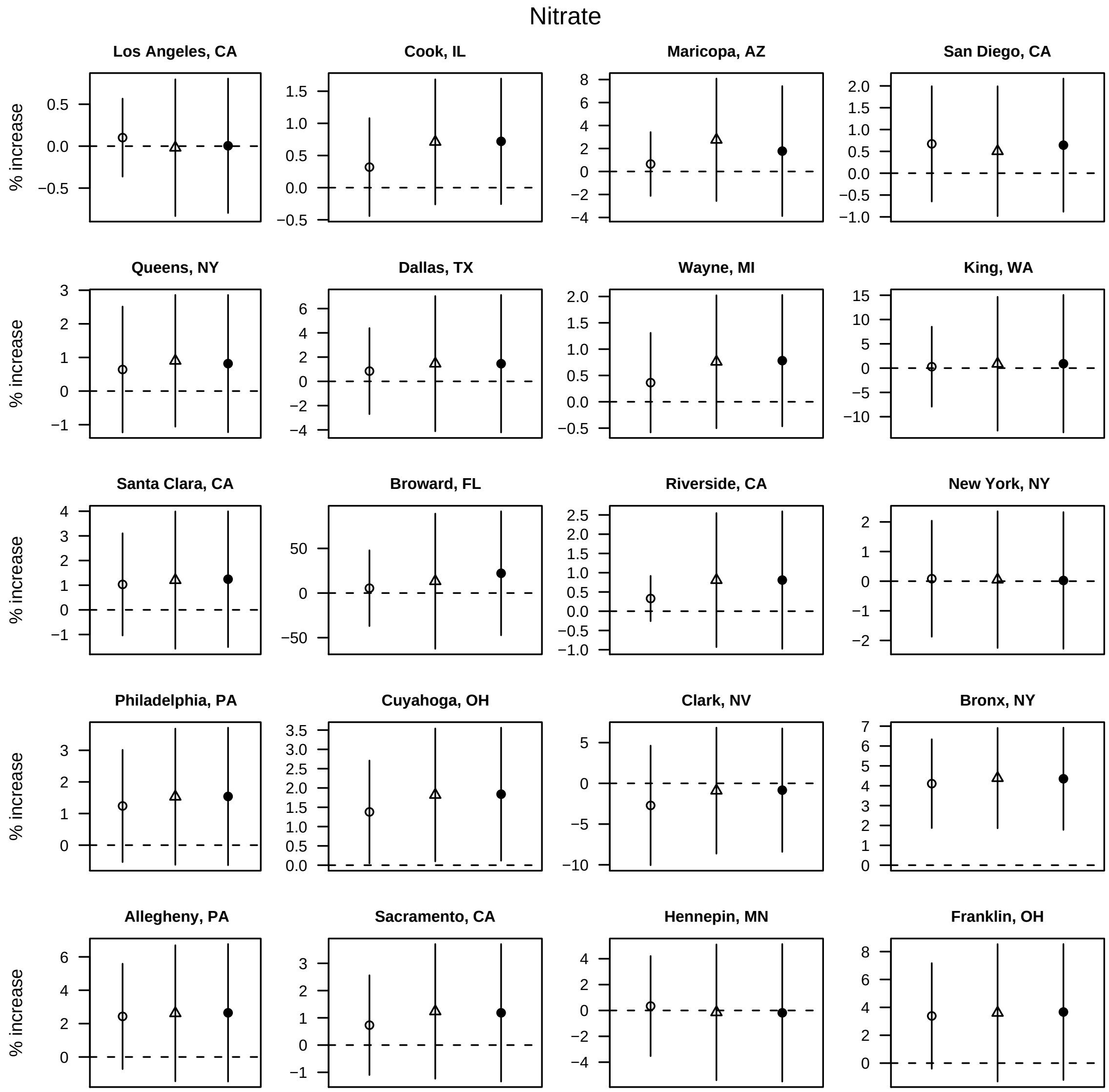

Figure 10: Percent increase in cardiovascular hospital admissions per 1 interquartile range (1.64 μg/m^3) increase in nitrate estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

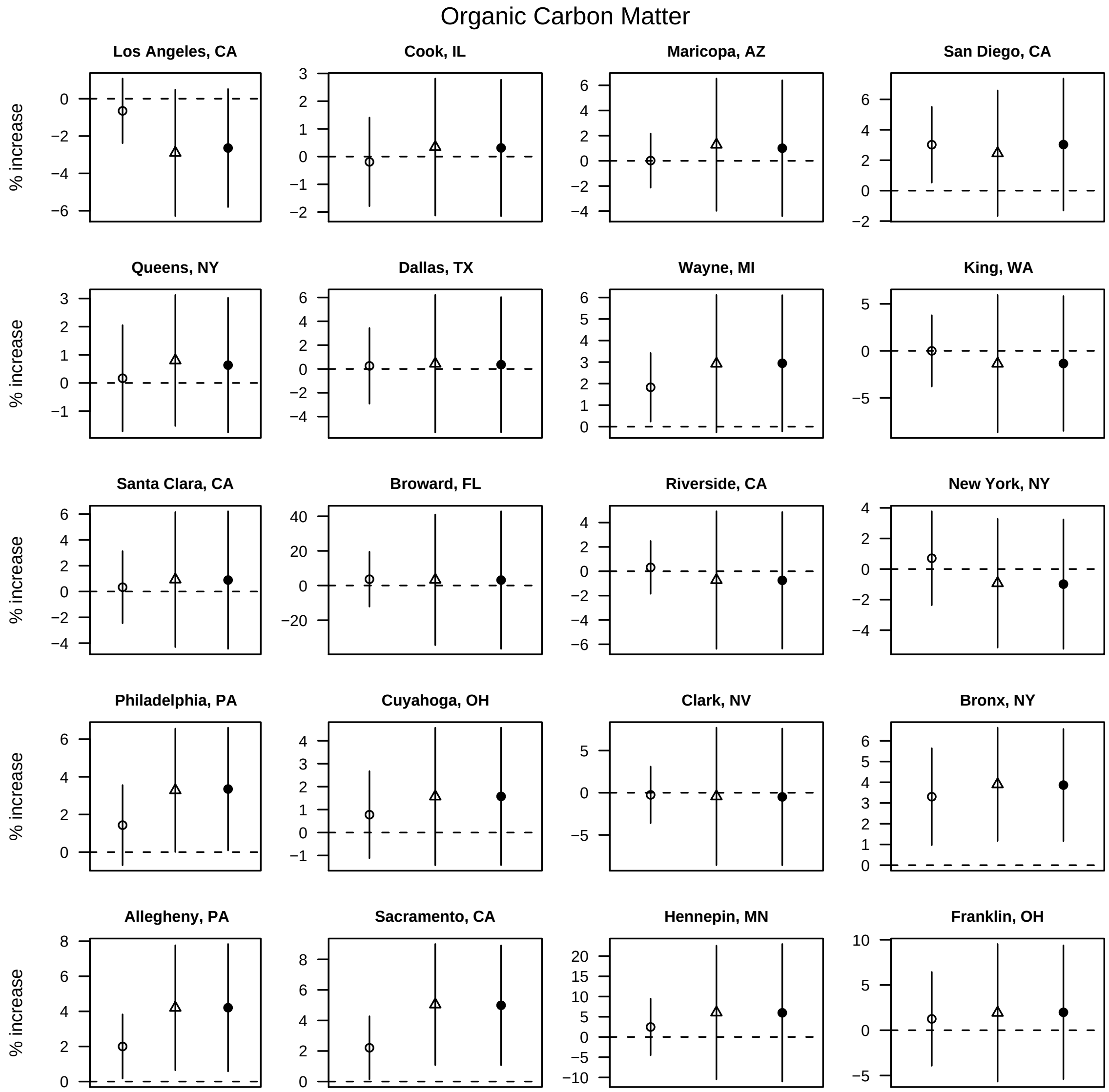

Figure 11: Percent increase in cardiovascular hospital admissions per 1 interquartile range (3.18 μg/m^3) increase in organic carbon matter estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

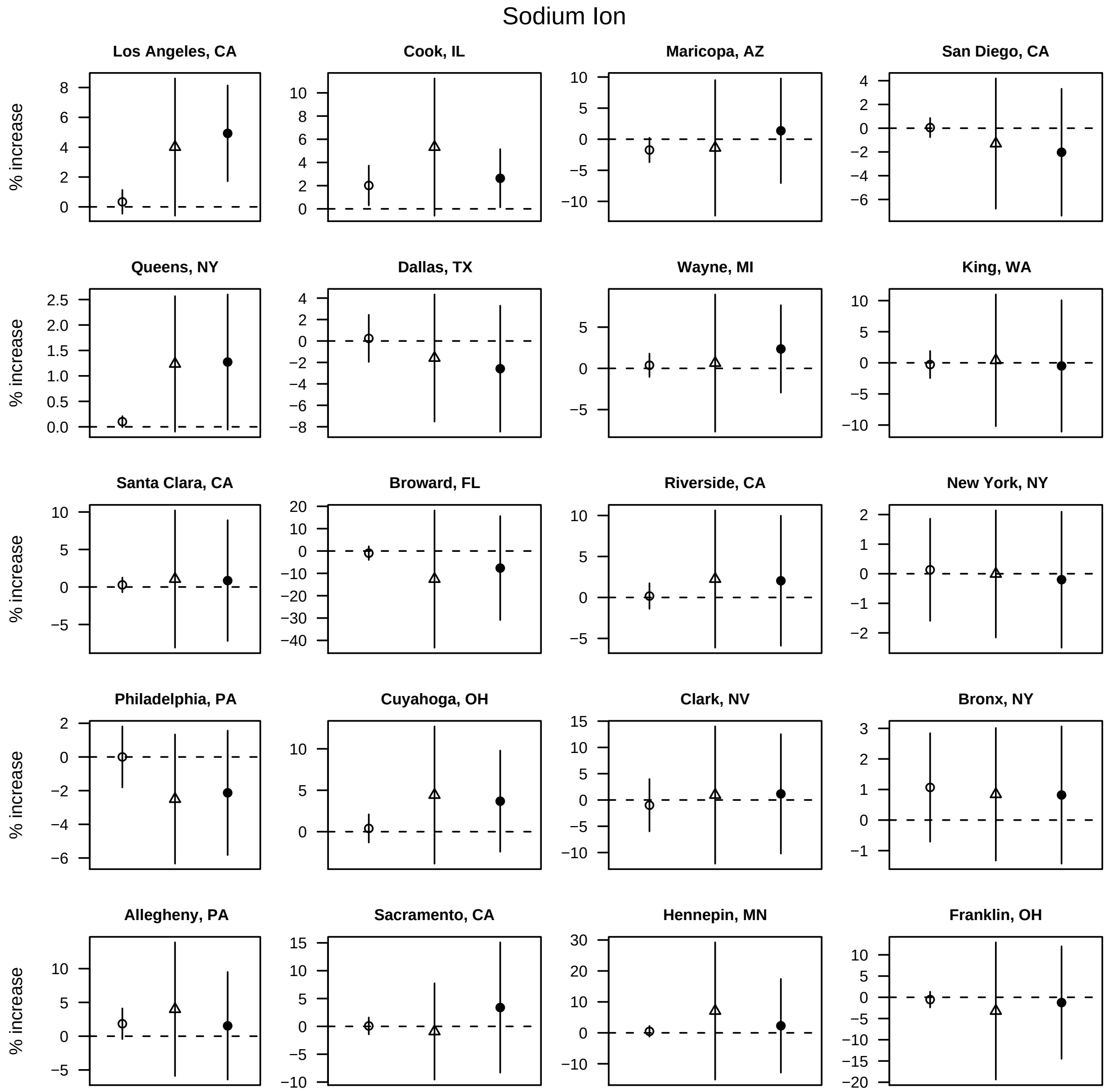

Figure 12: Percent increase in cardiovascular hospital admissions per 1 interquartile range (0.11 μg/m^3) increase in sodium ion estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

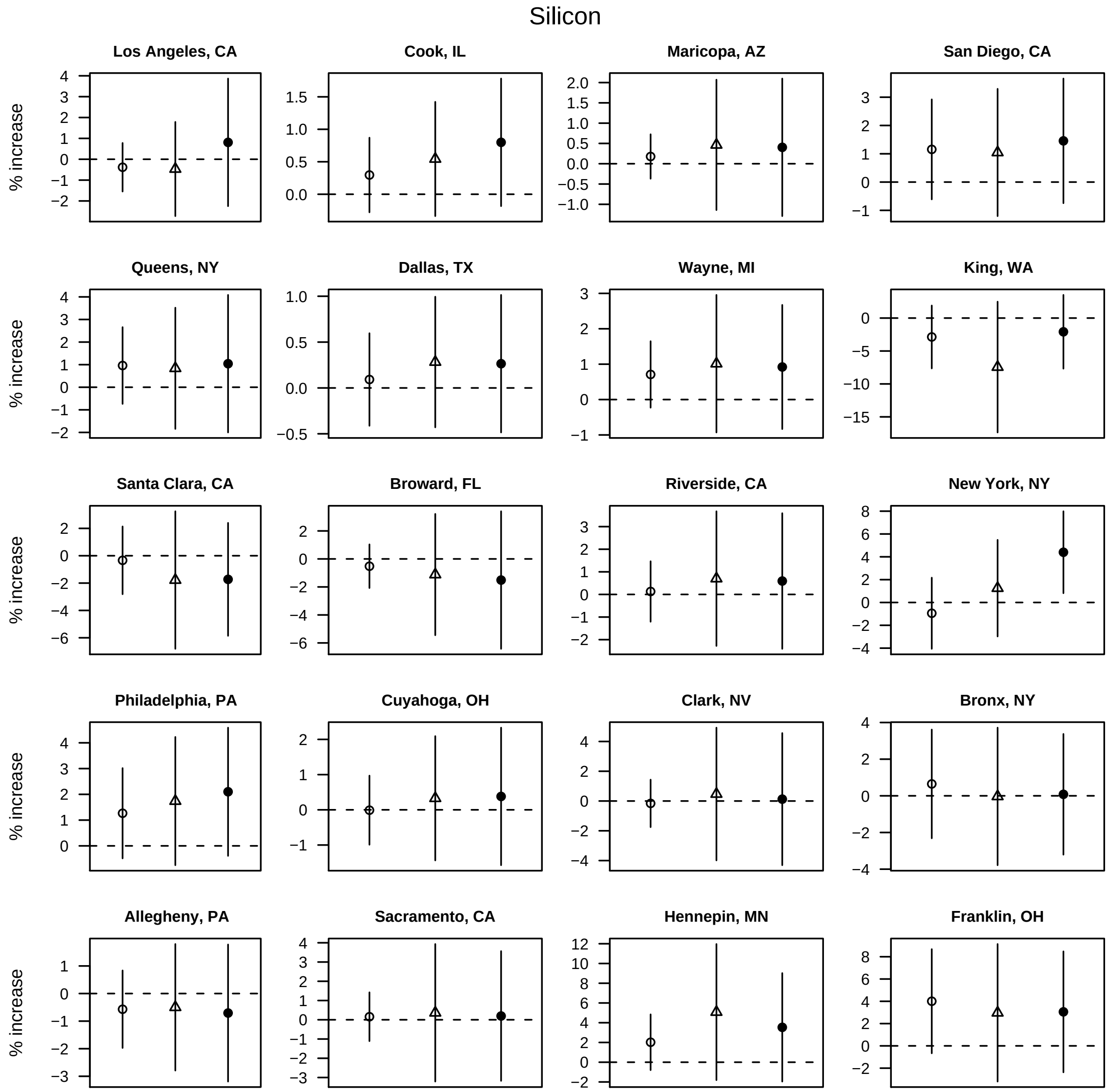

Figure 13: Percent increase in cardiovascular hospital admissions per 1 interquartile range (0.07 μg/m^3) increase in silicon estimated using maximum likelihood (open circle), regression calibration (triangle), and the two-stage Bayesian model (filled circle).

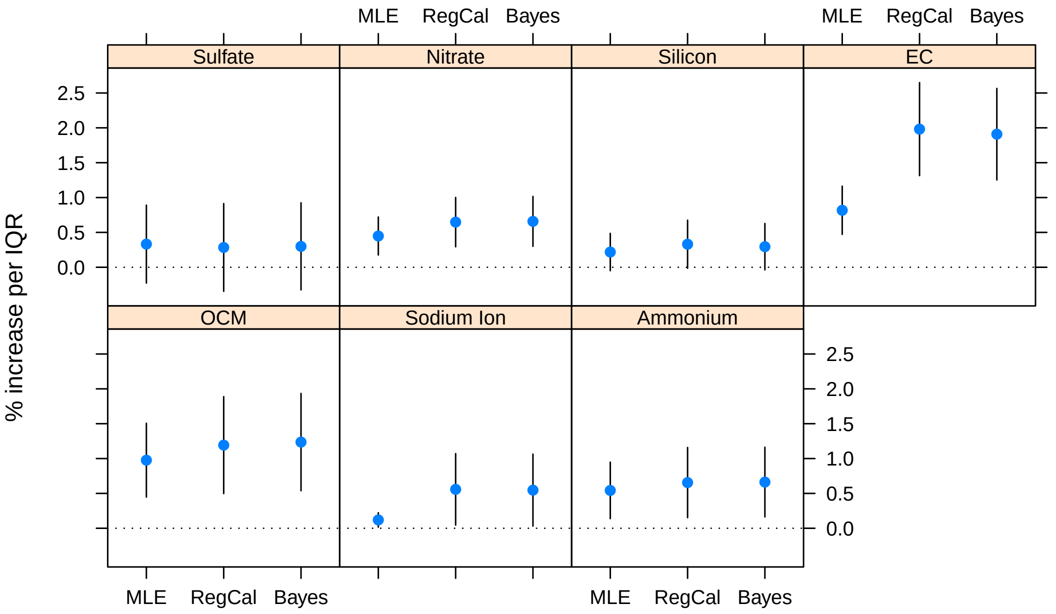

Figure 14: Inverse-variance weighted average (across 20 counties) percent increase in hospital admissions per 1 interquartile range increase in chemical component estimated using maximum likelihood, regression calibration, and the two-stage Bayesian model.