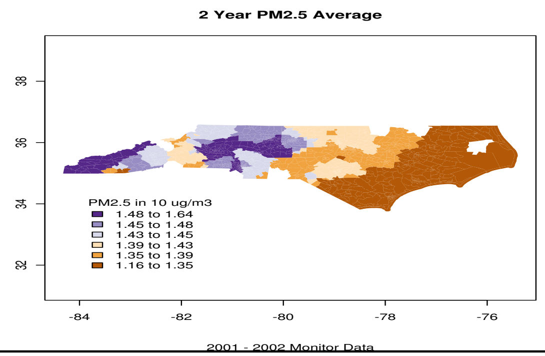



# Spatial-temporal impact of air pollution on human health

**Montse Fuentes**

**fuentes@stat.ncsu.edu**

**In Collaboration with**

**E. Kalendra (NCSU), M.L. Miranda (Duke), and B. Strauss (Duke).**

## Motivation

Recent studies have linked exposure to fine particulate matter ($PM_{2.5}$) and ozone ($O_3$) to mortality counts (county level).

This study uses a unique spatial data architecture consisting of geocoded North Carolina mortality data for 2001-2002, combined with U.S. Census 2000 data.

In our analysis we work with different levels of aggregation for the mortality data, and different metrics and sources of information for the pollution.

We also take into account distances to roadways and other important covariates.

## Our contribution

There is an increased interest in modelling association between mortality counts and pollution monitoring data.

Modelling the exposure surface and estimating exposure might lead to bias in the estimated health effect.

We introduce a model that is easy to implement that can adjust for this bias, without making a distributional assumption for the exposure.

Of considerable interest is potential non-additivity of effects of important co-pollutants.

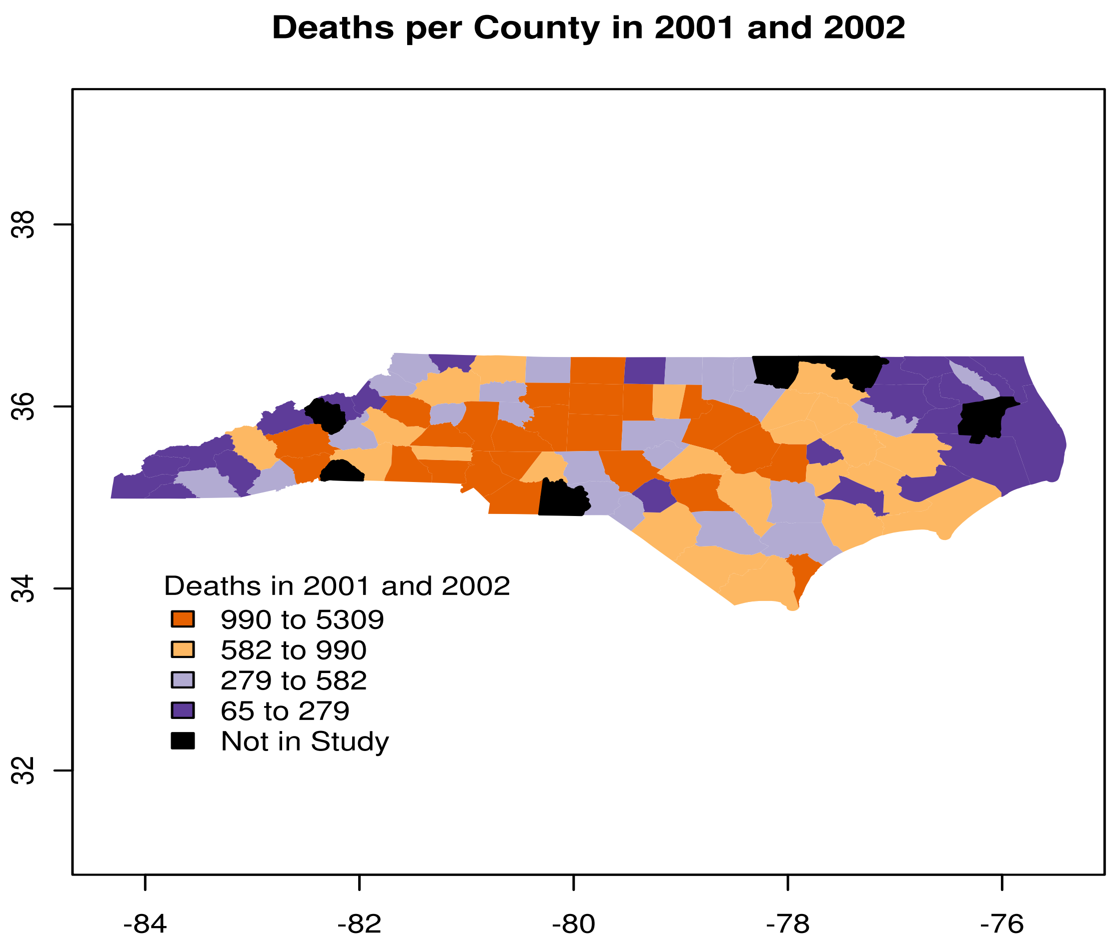
To investigate complex interactions, we introduce an alternative parameterization for $\mathrm{PM}_{2.5}$ and $O_3$ effects that allows a flexible, spatially-varying bivariate surface to characterize joint effects of $O_3$ and $\mathrm{PM}_{2.5}$.

We apply the nonadditive models to other co-variates, i.e. $\mathrm{PM}_{2.5}$ and distance to roadways.

Population data:

- Geocoded (lon/lat) mortality data in North Carolina for years 2001-2002.
- Natural deaths.
- Population: $> 65$ years-old.
- U.S. 2000 Census data.

Figure 1: Number of deaths in NC per county in 2001-2002 (for $> 65$ years old).



Weather (from weather stations):

- Daily average temperature.
- Daily precipitation.
- Location based daily pressure.
- Daily Dewpoint.

Exposure data:

- Different metrics:
  - Monitoring data for daily 8-hour max ozone and daily average of $PM_{2.5}$.
  - Output of air quality model (CMAQ) at 12 km resolution.
  - EPA fused data (combining CMAQ with monitoring data).
- Using GIS we obtain distances to primary (interstate and highways) and secondary (state) roads.

Figure 2: Monitoring stations for $PM_{2.5}$.

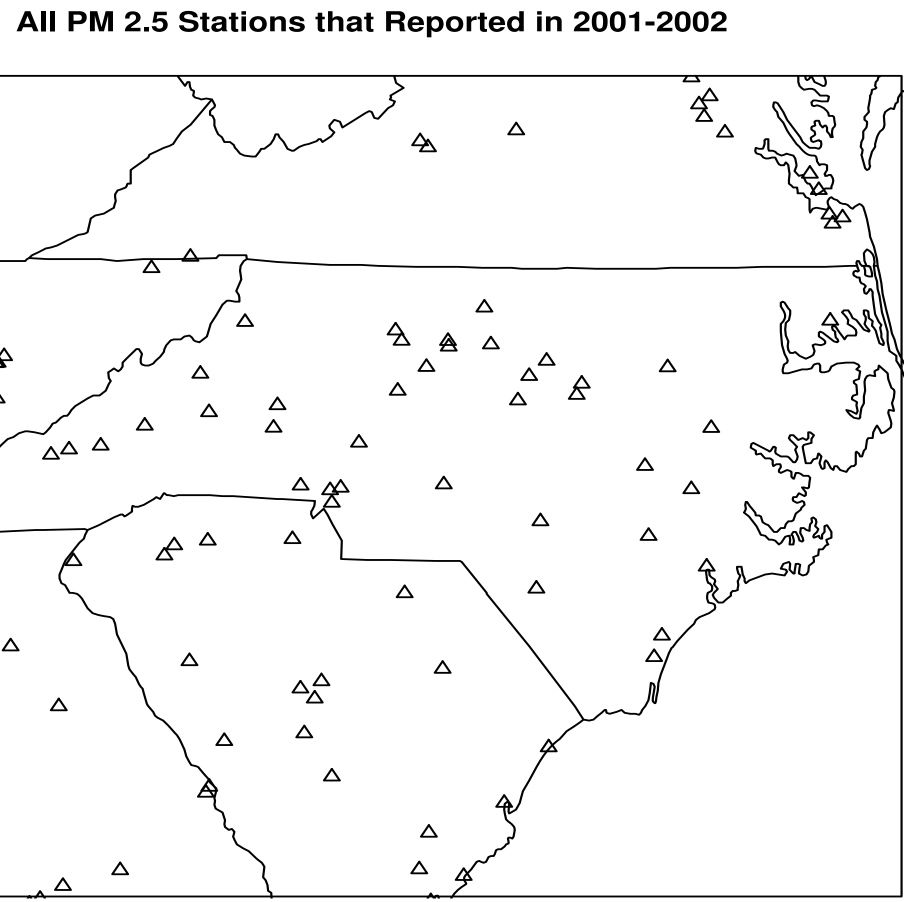


Figure 3: Monitoring stations for ozone.

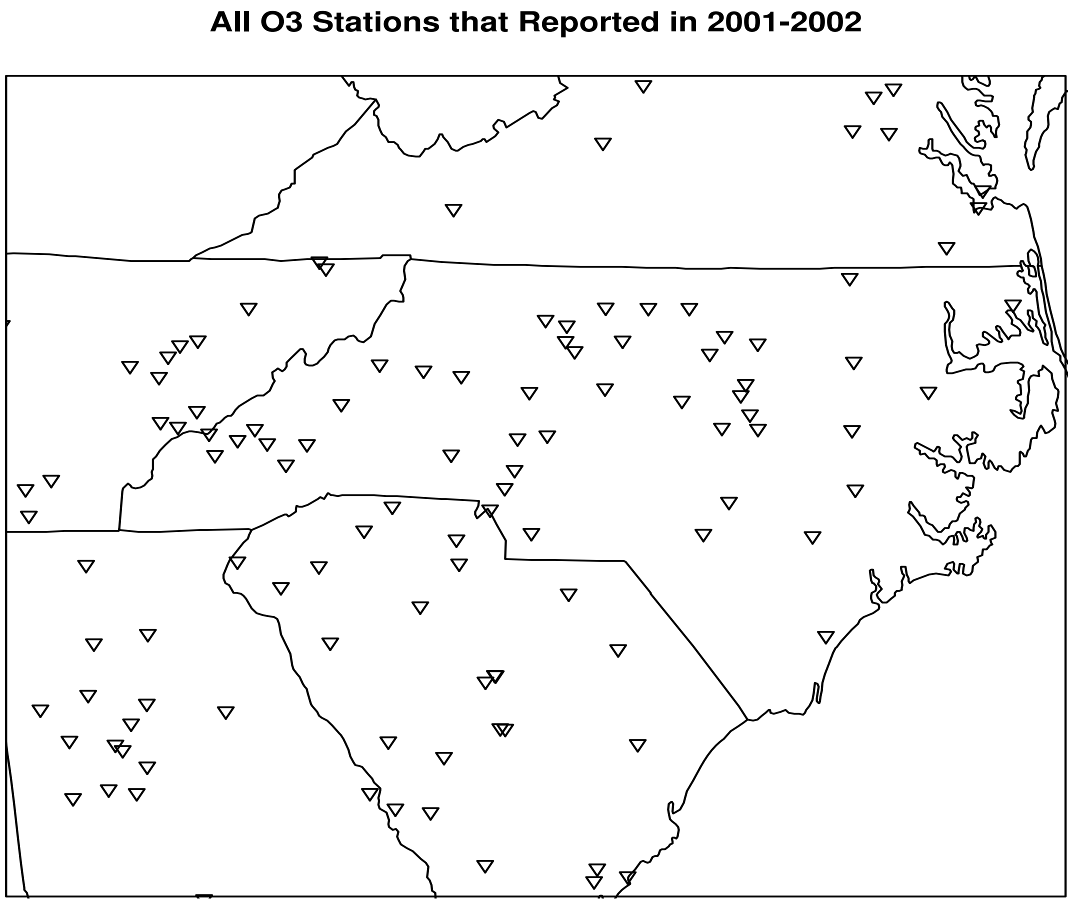


Figure 4: Average of daily $PM_{2.5}$ concentrations in 2001-2002 in NC. Spatial surface based on closest monitoring station to tract centroid.

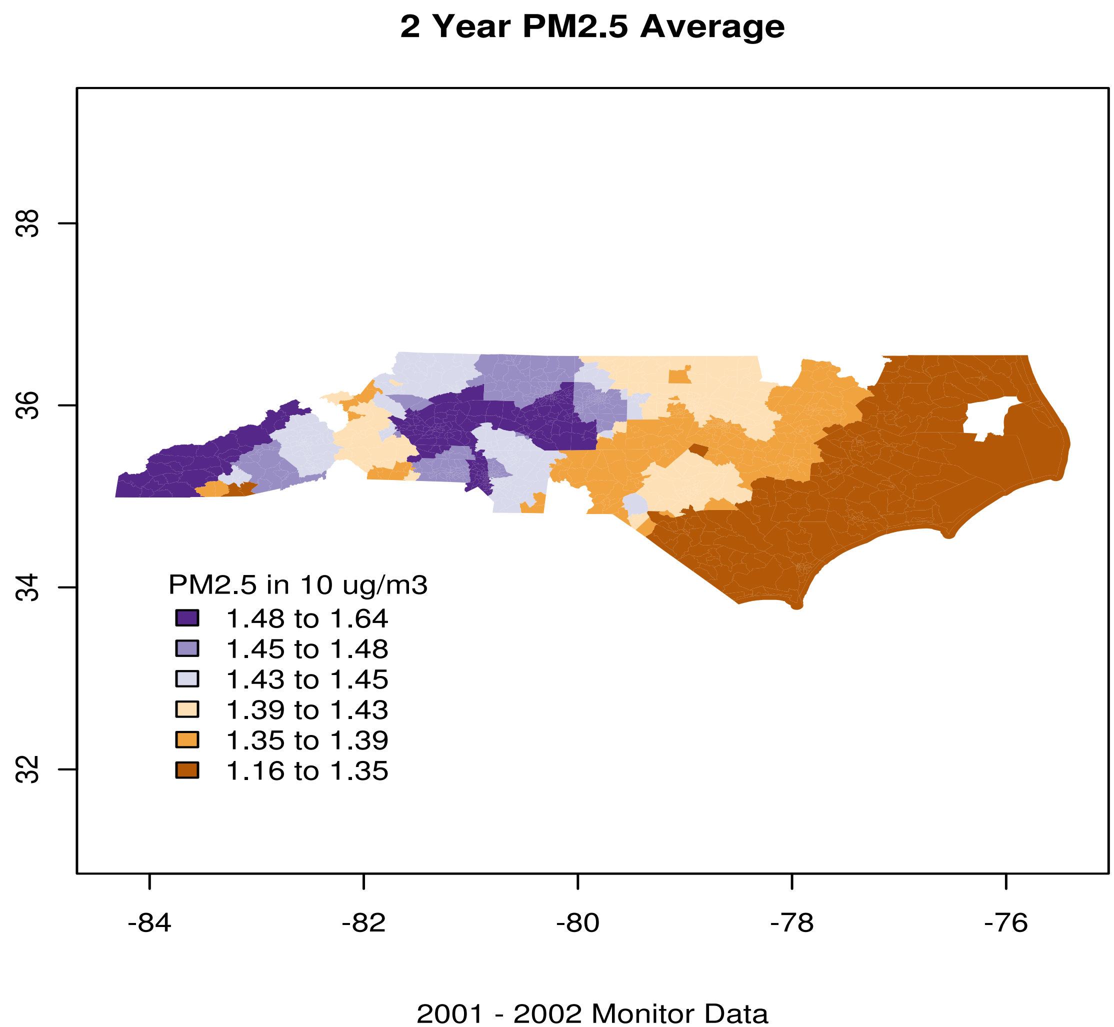


Figure 5: Average of daily 8-hour max. ozone concentrations in 2001-2002 in NC. Spatial surface based on closest monitoring station to tract centroid.

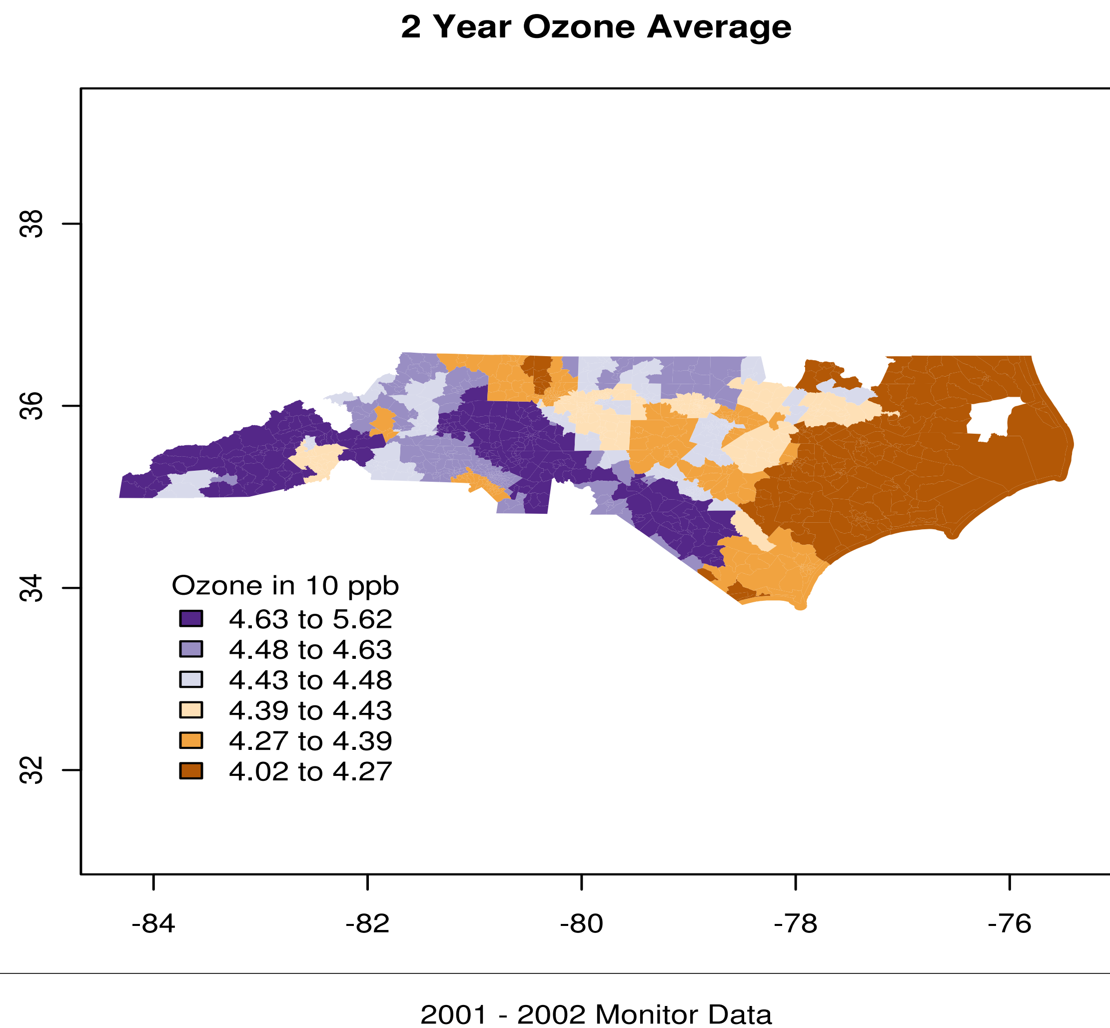


Figure 6: Distance to closest primary roads (km) from tract centroids.

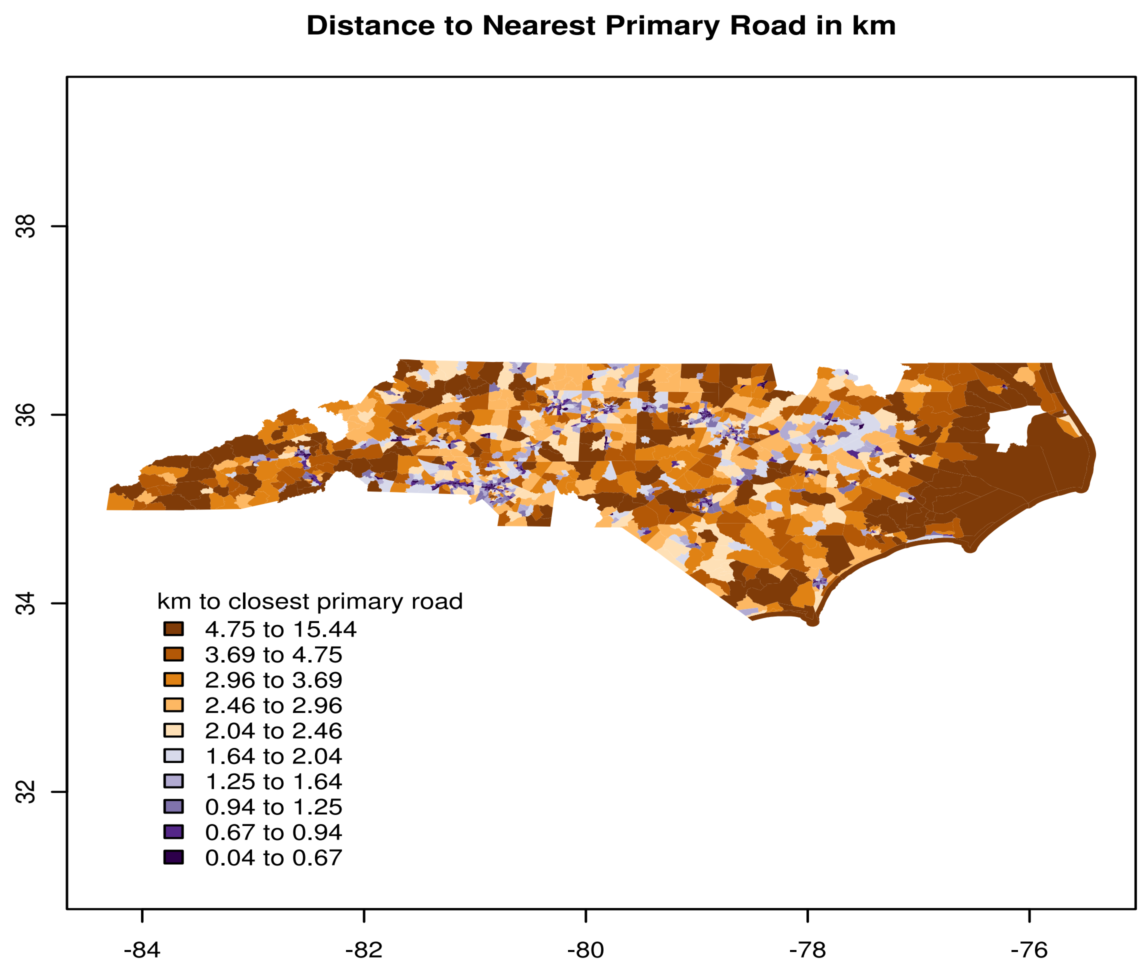
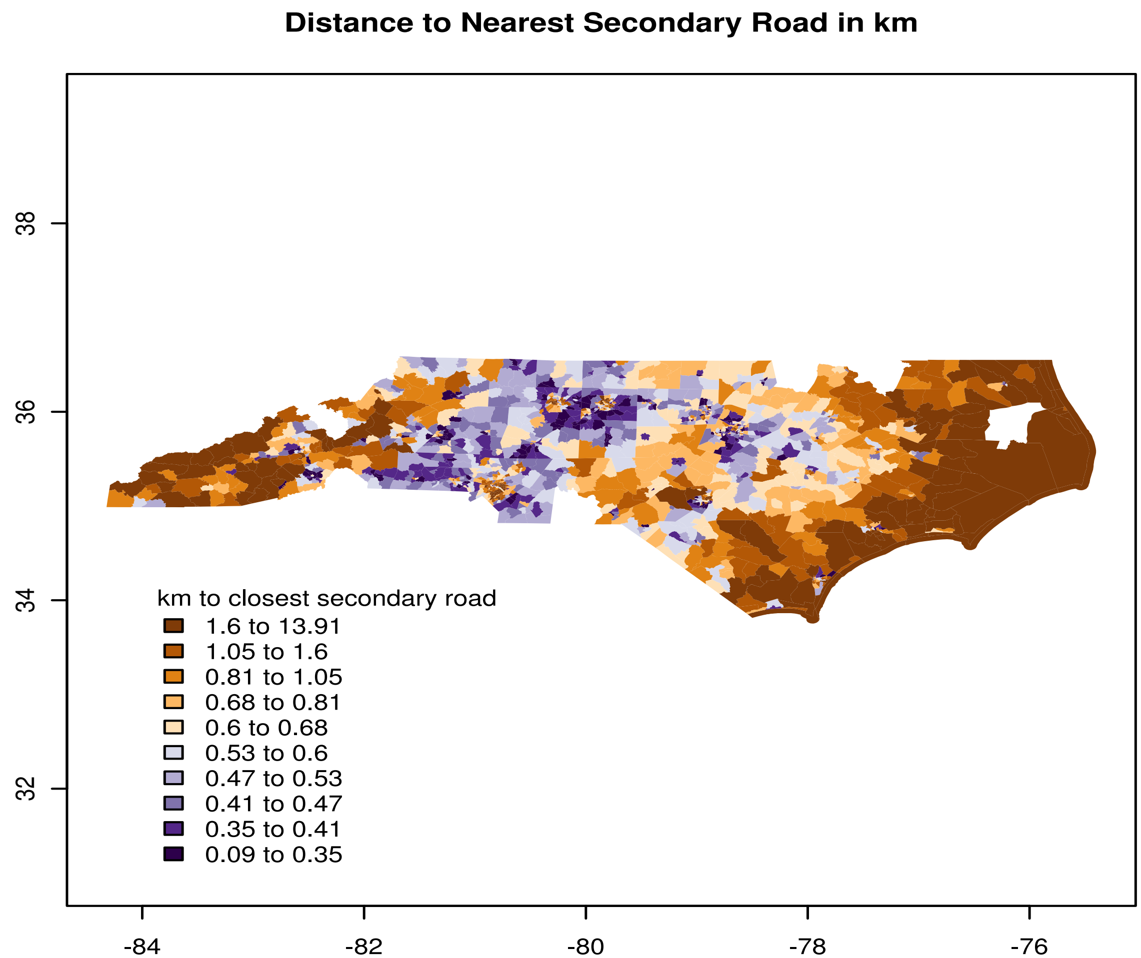


Figure 7: Distance to closest secondary roads (km) from tract centroids.



# Effect of aggregating data

Our geocoding of mortality data, allows us to investigate the impact of aggregation, going from models at the individual level to modelling counts of mortality (tract, county, region levels).

Model at individual level

$Y_{itk}$: whether an individual $i$ in region $k$ died day $t$.

$x_{ikt}$: exposure data at location $s_i$ (residence of individual $i$),

$z_{ikt}$: other covariates (e.g. weather).

$$Y_{itk}|x_{ikt}, z_{ikt}, \beta \sim Bernoulli(p(x_{ikt}, z_{ikt}, \beta)) \tag{1}$$

where the probability of death for individual $i$ at time $t$ in region $k$ is

$$p(x_{ikt}, z_{ikt}, \beta) = \exp(f(x_{ikt}, z_{ikt}), \beta), \tag{2}$$

$p$ is usually very small (rare event).

## Modelling counts

$Y_{kt}$ mortality counts in region $k$.

$N_{kt}$ population data.

We model mortality counts in region $k$:

$$E\left[Y_{kt}|x_{kt}, z_{kt}, \beta\right] = N_{kt} q_{kt} \tag{3}$$

where

$$q_{kt} = \frac{1}{N_{kt}} \sum_{i=1}^{N_{kt}} \exp(f(x_{ikt}, z_{ikt}), \beta) \tag{4}$$

Because $N_{kt}$ is typically large and $p(x_{ikt}, z_{ikt}, \beta)$ is small, the binomial can be approximated with a Poisson distribution.

Using $N_{kt}q_{kt}$ gives the convolution model (Wakefield, 2006)

$$Y_{kt}|x_{kt}, z_{kt}, \beta \sim_{\text{ind}} \text{Poisson}\left\{\sum_{i=1}^{N_{kt}} \exp(f(x_{ikt}, z_{ikt}), \beta)\right\} \tag{5}$$

The convolution model requires exposure information for every individual in region $k$ at time $t$.

Approximations of $N_{kt}q_{kt}$ can lead to bias of the estimated parameters. The standard Poisson model for mortality counts using exposure areal averages leads to ecological bias.

Standard Poisson model for mortality counts:

$$Y_{kt} \sim \text{Poisson}\left\{N_{kt} \exp(f(x_{kt}, z_{zt}), \beta)\right\}, \tag{6}$$

where $x_{kt}$ is the average exposure in region $k$.

## Our model for aggregated data

A method to approximate $N_{kt} q_{kt}$ is to use the Taylor expansion.

$$\sum_{i=1}^{N_{kt}} \exp(f(x_{ikt}, z_{ikt}), \beta) = e^{\beta_0} \sum_{i=1}^{N_{kt}} \exp(\beta_1 x_{ikt}). \tag{7}$$

Instead of making a distributional assumption, we expand the term using a Taylor series approximation.

$$\sum_{i=1}^{N_{kt}} \exp(\beta_1 x_{ikt}) \approx (N_{kt}) \left(1 + \beta_1 \frac{\sum_i x_{ikt}}{N_{kt}} + \frac{1}{2} \beta_1^2 \frac{\sum_i x_{ikt}^2}{N_{kt}}\right). \tag{8}$$

This leads to the first order approximation

$$\sum_{i=1}^{N_{kt}} \exp(\beta_1 x_{ikt}) \approx (N_{kt}) \exp\left(\beta_1 \frac{\sum_i x_{ikt}}{N_{kt}}\right). \tag{9}$$

And second order approximation

$$(N_{kt}) \exp\left(\beta_1 \frac{\sum_i x_{ikt}}{N_{kt}} + \frac{1}{2}\beta_1^2 \left(\frac{\sum_i x_{ikt}^2}{N_{kt}} - \left[\frac{\sum_i x_{ikt}}{N_{kt}}\right]^2\right)\right). \tag{10}$$

Instead of computing the computationally expensive sum, we only need to store $\frac{\sum_i x_{ikt}}{N_{kt}}$ and $\frac{\sum_i x_{ikt}^2}{N_{kt}}$, which is only an increase of one variable compared to the typical Poisson regression.

The estimate of $\beta$ is computed using a restricted Poisson regression model.

The ecological bias increases with the *population-weighted* sample variance for the exposure (term multiplying $\beta_1^2$). This ecological bias is more of an issue for exposure variables with spatial heterogeneous variability.

## Standard Poisson model

Standard Poisson model for mortality counts, where we are interested in estimating the effect of pollution.

$$Y_{kt} \sim Poisson\left\{N_{kt}\exp(f(x_{zt}, z_{kt}))\right\} \tag{11}$$

$$\begin{aligned} f(x_{zk}, z_{kt}) &= \beta_0 + \mathrm{h(t)} + \mathrm{ns(TAVG)} + \mathrm{ns(PRES)} + \mathrm{ns(DPTP)} \\ &+ \mathrm{ns(Dist.\ Pri)} + \mathrm{ns(Dist.\ Sec)} + \mathrm{g}_k(\mathrm{PM}_{2.5}, O_3) \end{aligned}$$

where h(t) is a temporal trend (4 Fourier components), ns() is natural splines (5 d.f.), TAVG is average daily temperature, PRCP daily precipitation, PRES location based pressure, and DPTP dewpoint. Dist. is the distance to the nearest roadway, with Pri. meaning primary and Sec meaning secondary.

$g_k()$ is generally additive: $g_k(\mathrm{PM}_{2.5}, O_3, ) = \beta_1 \mathrm{PM}_{2.5} + \beta_2 O_3$

## Non-additive models

Additive model:

$$g_k(Pollution) = \beta_1 X_1(k) + \beta_2 X_2(k)$$

Nonadditive model we propose:

$$g_k(Pollution) = g_k(X_1(k), X_2(k))$$

where

$$g_k(a, b) = \sum_{m=1}^{M} w_m(k) b_m(a, b)$$

$$b = \{b_m\}_{m=1}^{M}$$

are the two-dimensional basis functions (e.g., thin plate splines, polynomials), and the $w_m(k)$ are spatially-varying coefficients.

# Results to compare metrics and levels of aggregation

Comparing the effect of using different metrics for $PM_{2.5}$ and different levels of aggregation, using the standard Poisson model previously presented.

| | NC | Region | County | Tract |
|---|---|---|---|---|
| Monitor | $0.008_{(.007)}$[1.2] | $0.009_{(.006)}$[1.5] | $0.008_{(.006)}$ [1.4] | $0.005_{(.006)}$[0.9] |
| CMAQ | $0.007_{(.008)}$[1.0] | $0.013_{(.006)}$[2.0] | $0.013_{(.006)}$[2.2] | $0.014_{(.006)}$[2.5] |
| Fusion | $0.010_{(.007)}$[1.5] | $0.015_{(.006)}$[2.5] | $0.022_{(.005)}$ [4.1] | $0.024_{(.005)}$[4.5] |

Table 1: Estimated $\beta_{SD}$[z-value]. $\beta$: percent increase of mortality per increase of 10 units of $PM_{2.5}$.

# Results to study distance to roadways

Analysis at the tract level, using monitoring data. We include distance to nearest roadways, as an additive effect.

| | |
|---|---|
| $PM_{2.5}$ | $0.005_{(.006)}$ [0.9] |
| Dist Pri | $-0.062_{(.011)}$ [-5.5] |
| Dist Sec | $-0.044_{(.006)}$ [-7.2] |
| $O_3$ | $0.004_{(.003)}$ [1.4] |
| Dist Pri | $-0.063_{(.011)}$ [-5.5] |
| Dist Sec | $-0.044_{(.006)}$ [-7.2] |

Table 2: Tract Level. Estimated $\beta_{SD}$[z-value].

- The health effects due to the two pollutants ($PM_{2.5}$ and $O_3$), do not seem to change by adding in the model the distance to primary and secondary roads.

- The distance to primary and secondary roadways seem to be more relevant in explaining mortality, than the monitoring data.

- Possible explanation: Monitored concentrations might not represent near roadway concentrations.

## Results for our bias-adjustment framework

We use monitoring data at the county level, to study the impact of our population-based exposure averaging approach (using the first and second order appr.), rather than using the standard linear method with exposure areal averages.

We apply our bias-adjustment method to the $PM_{2.5}$ and dist. to secondary road variables.

We introduce analysis at the tract level as reference (bias is more negligible at that level).

| | Est | SD | Z |
|---|---|---|---|
| Tract level | 0.008 | .006 | 1.4 |
| County level (standard areal aggregation) | 0.010 | .006 | 1.7 |
| County level (pop.-based averages, first order appr.) | 0.009 | .006 | 1.5 |
| County level (pop.-based averages, second order appr.) | 0.009 | .006 | 1.5 |

Table 3: Estimated $\beta$ for $PM_{2.5}$ (% increase in mortality per 10 units increase of $PM_{2.5}$).

| | Est | SD | Z |
|---|---|---|---|
| Tract level | -0.104 | 0.013 | -8.2 |
| County level (standard areal aggregation) | -0.054 | .012 | -4.2 |
| County level (pop.-based averages, first order appr.) | -0.095 | .018 | -5.3 |
| County level (pop.-based averages, second order appr.) | -0.098 | .019 | -5.2 |

Table 4: Estimated $\beta$ for distance to secondary roads (% increase in mortality per 1 km increase of the dist. to second. road).

- The results from our bias-adjusted model are more similar to the results at the tract level for the distance to secondary ($\beta \sim -.1$). The potential bias with the standard model is about 1/2 the magnitude of this health effect.

- The impact of this bias-adjustment framework is more negligible with $PM_{2.5}$ than with distance to roadways, because this variable is less significant and there is less spatially heterogeneity in the $PM_{2.5}$ variance.

Figure 8: **Results for the non-additive model.** Gradient vector of the risk of mortality due to joint exposure to ozone and $PM_{2.5}$ (lag 1). The circles represent the data. The changes in the background contourplot represent .01 change in the actual health effect.

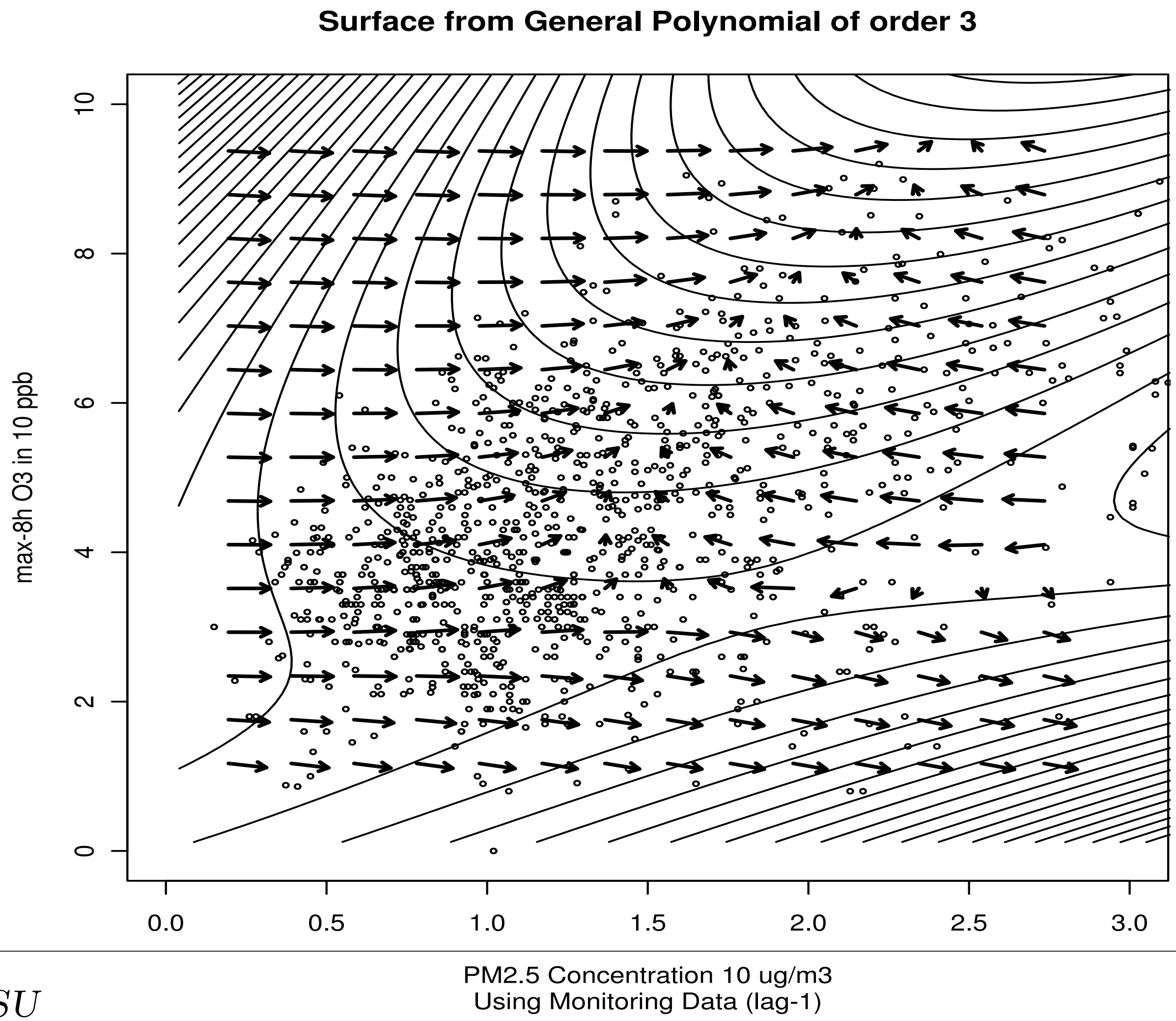
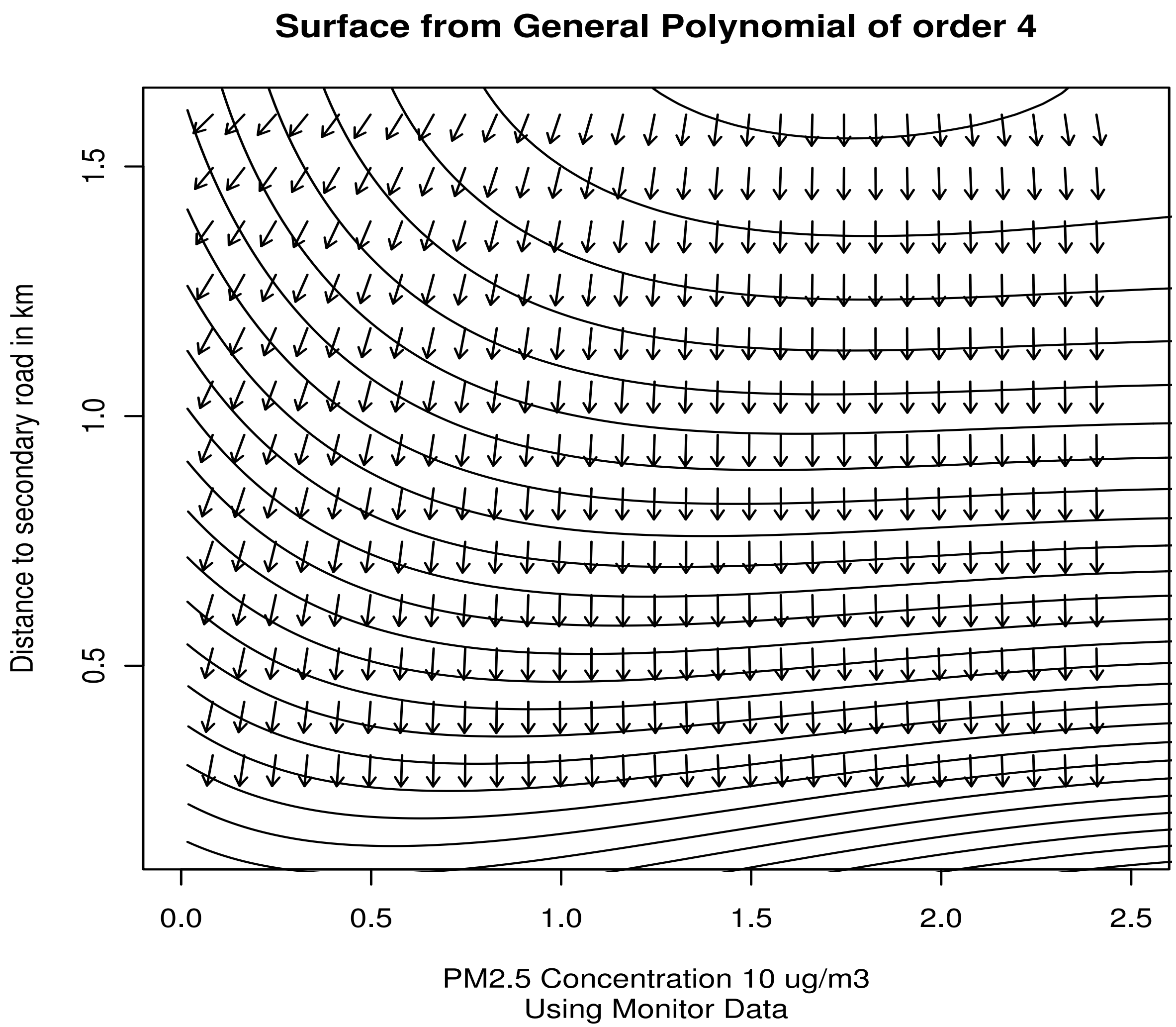


Figure 9: Gradient of the risk of mortality due to joint exposure to $PM_{2.5}$ and distance to closest secondary road.



# Conclusions

The following results from our study, could have a significant impact in air quality regulation, managing and policy:

- There is a significant risk of mortality associated to fine particulate matter and ozone.
- The spatial scale at which the analysis are done matters a lot. Different results at different scales.
- The EPA fused data product (combining CMAQ and monitoring data) gives more power to characterize the risk of mortality due to pollution.
- Monitored PM concentrations might not represent near roadway concentrations. Thus, other variables, such as distance to roadways, might be a better indicator of near roadway exposure.

- Most of the associations between pollution and mortality are done using areal exposure data and mortality counts. It is important to use population-based aggregation methods, like the one presented here, to avoid bias in the estimated health effect.
- Co-Pollutants health effects seem to be non-additive. Additive methods could result in misleading results, due to interactions.