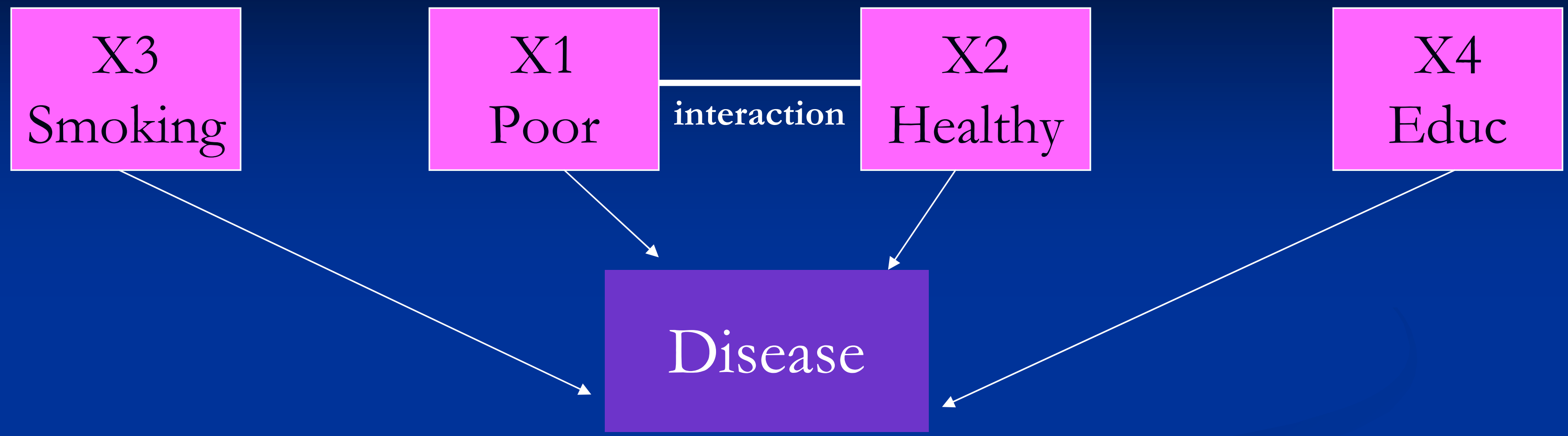
Spatial Mapping of Multivariate Profiles

John Molitor
Imperial College, London

Aug 26, 2010

Motivation- deal with correlated data



D= β_0 ...

β_1X1+ β_2X2 +β_3X3+ β_4X4 +…. Xp+

+β_5*X1*X2 +…+β_{10}X3*X4 + +…. Xp-1 * Xp+

...or

P=20

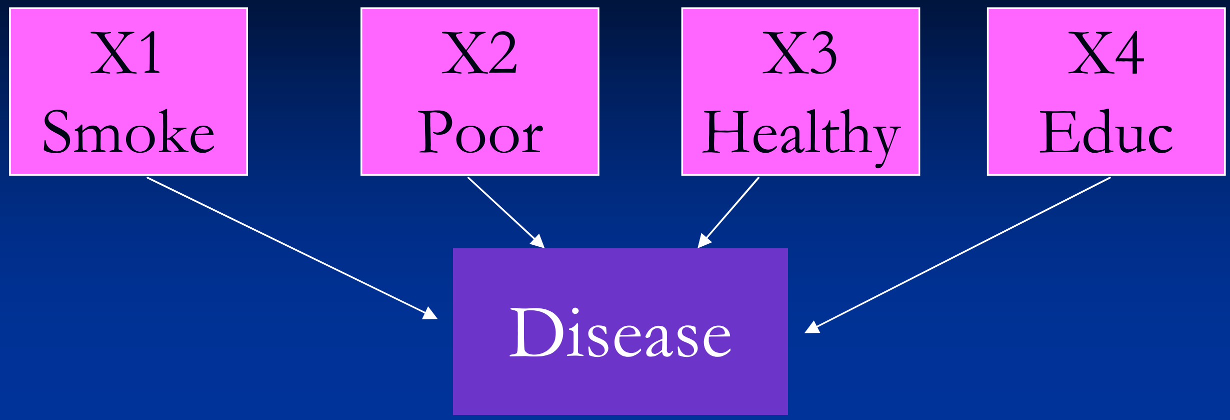
2-ways interaction: 190

3-ways interaction: 1140

-Muticollinearity

- Pattern of interaction effects may be illusive

Individual Covariates versus Profiles



Use a sequence of covariates values to form **different profiles**

profile 1: 1, 1, 0, 0 (Smoke, Poor)

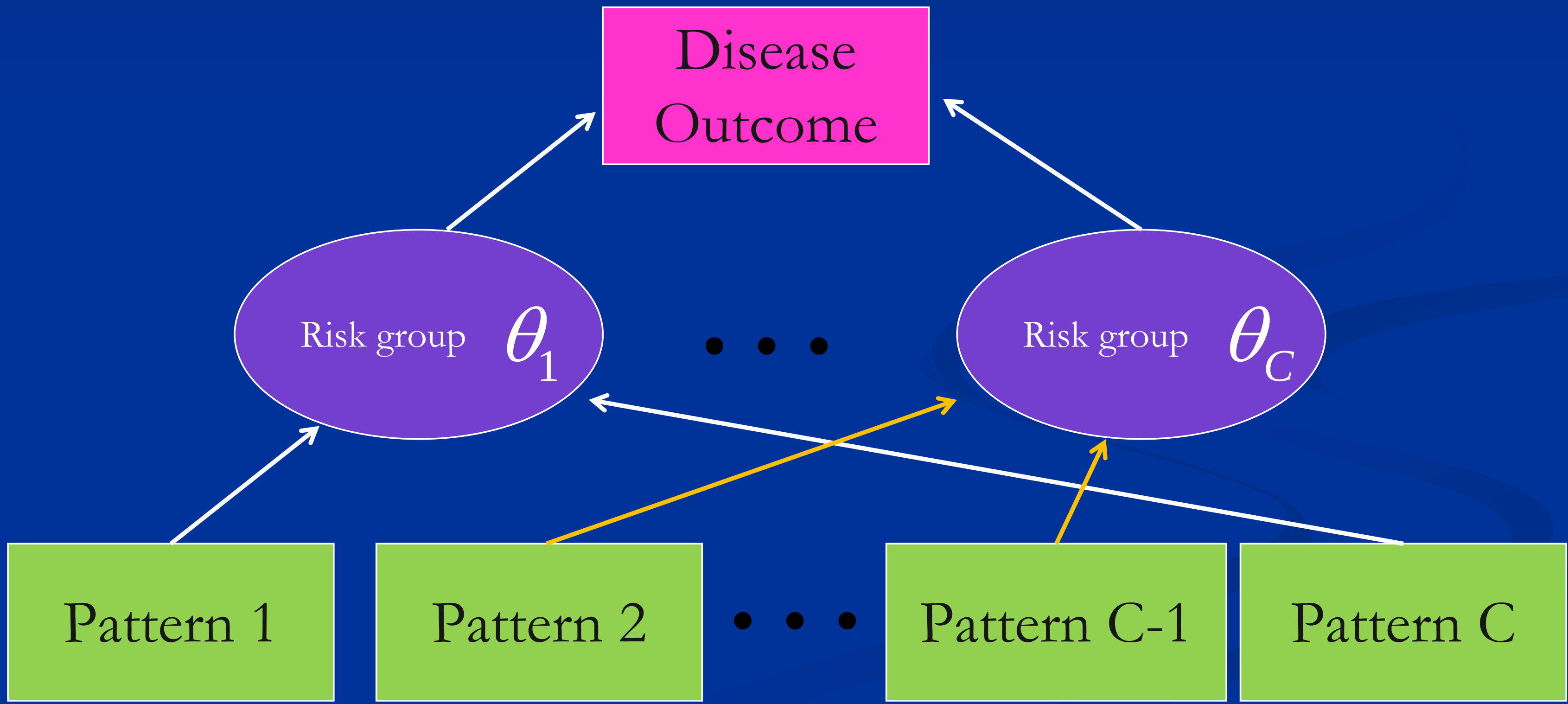
profile 2: 1, 0, 0, 1 (Smoke, Educ)

…

profile N: 0, 0, 1, 1 (Healthy, Educ)

Profile Regression

- Idea : Use pattern as basic unit of inference. Cluster these patterns into a relative small numbers of risk groups and use these risk groups to predict an outcome of interests.



Profile Regression- modeling framework

- **Assignment Model:**

Model the probability that an individual is assigned to particular cluster.

$$f(\mathbf{x}_i) = \sum_{c=1}^{C} \psi_c f(\mathbf{x}_i \mid \theta_c)$$

- **Disease Model:**

Model the risk associated with a individual pattern group.

$$\text{logit}(y_i) = \theta_{z_i} + \beta W_i, \quad z_i = c$$

Or, alternatively,

$$\text{logit}(\text{y}_i) = \alpha + \theta^*_{z_i} + \beta W_i, \sum_{c=1}^{C} \theta^*_c = 0$$

How to decide the number of clusters?

$$f(\mathbf{x}_i) = \sum_{c=1}^{C} \psi_c f(\mathbf{x}_i \mid \theta_c)$$

- Reversible Jump - complicated split/merge moves

- Flexible Approach - finite number of clusters
  - Truncated Dirichlet Process
  - Choose more clusters than needed. (Clusters allowed to be empty.)
  - Chose the enough clusters to avoid estimating a large number of unnecessary cluster parameters.

Stick-breaking prior cluster probabilities

- Determines prior probabilities for cluster allocations
- Prior probability assigned to first cluster is obtain by breaking stick of length one.
- Subsequent probabilities obtained by breaking “left over” part of stick.

Truncated Dirichlet Process

When specified the number of clusters

Infinite	Finite
Dirichlet	**Truncated Dirichlet**

$$f(x_i) = \sum_{c=1}^{\infty} \psi_c f(x_i \mid \theta_c) \qquad \approx \qquad \sum_{c=1}^{C} \psi_c f(x_i \mid \theta_c)$$

Markov Chain Monte Carlo (MCMC) Parameter Estimation

- Fits model as a unit.
- Both outcome (y’s) and covariates (x’s) influence cluster membership
- Flexible (e.g. easy to change form of disease model)
- Implemented in WinBugs (could use JAGS or custom code)

Model Averaging through Post-Processing

- Estimating the risk of a new profile
- Examination of Average Clustering
- Estimate the partition of interest.
- Deal with typical clustering algorithm problems such as label-switching.

Estimating the Risk of a New Profile – A Model Averaging Approach

1. Probabilistically assign the profile to the appropriate cluster

$$\Pr\left(z_{new} \mid x_{new}\right) \propto \Pr\left(x_{new} \mid z_{new}\right) \Pr\left(z_{new}\right)$$

2. Profile risk is equal to the risk of cluster to which pattern is assigned

$$\text{profile risk} = \theta_{z_{new}}$$

3. Average over varying number of clusters used at each iteration of MCMC sampler

Examination of Average Clustering (invariant to label switching)

- At every iteration of MCMC sampler, we have a partition of individuals:

  $\mathbf{z}_1$=(2, 2, 2, 5, 5, 5, 7, 7, 7, 5)
  $\mathbf{z}_2$=(2, 2, 2, 5, 5, 5, 5, 7, 7, 7)
  $\mathbf{z}_3$=(2, 2, 2, 5, 5, 5, 5, 7, 5, 7)
  $\mathbf{z}_4$=(2, 2, 2, 5, 5, 7, 5, 7, 7, 5)
  …

- Find the best partition, $\mathbf{z}_{\text{best}}$. Represents as average way in which the algorithm groups individuals into clusters.

  e.g. $\mathbf{z}_{\text{best}}$ = (2, 2, 2, 5, 5, 5, 5, 7, 7, 7)
  $\mathbf{z}_{\text{best}}$ = (a, a, a, b, b, b, b, c, c, c)

Best Partition Z_{best}

- Construct the score matrix (S_Z)
  - Record 1 if individual i and j are in the same cluster and record 0 otherwise (repeating for each iteration)
  - Averaging the score matrices obtained at each iteration
  - Define S_{ij} as empirical prob. which individual i and j in the same cluster
- Finding $\mathbf{z}_{best}$: Use the following "least squares" formula (Dahl 2006)

$$Z_{best} = \arg\min_{z \in Z} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{N} \left(S_{z,ij} - S_{ij} \right)^2 \right\}$$

Accounting for uncertainty when finding the best partition using model averaging

- Individuals in each single group of $\mathbf{z}_{best}$ may appear in the different cluster at each iteration.
- Variability from cluster is used to access the uncertainty related to group defined by the $\mathbf{z}_{best}$
- At each iteration of MCMC sampler, we find average risk for all individuals in each subgroup of best partition, $\mathbf{z}_{best}$.
  (Same procedure for covariate probabilities)
- Important to properly assess uncertainty as all datasets will have "best" grouping.

Subgroup Assignment at Each Iteration of MCMC Sampler

$$\theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_4 \quad \theta_5 \quad \theta_6 \quad \theta_7 \quad \theta_8 \quad \theta_9 \quad \theta_{10}$$

Subgroup 1

$$\bar{\theta}_a = \frac{\theta_2 + \theta_2 + \theta_2}{3}$$

Subgroup 2

$$\bar{\theta}_b = \frac{\theta_5 + \theta_5 + \theta_6 + \theta_6}{4}$$

Subgroup 3

$$\bar{\theta}_c = \frac{\theta_6 + \theta_7 + \theta_7}{3}$$

Cluster Risks					
$\theta_1 = 0.2$	$\theta_2 = 0.4$	$\theta_3 = 0.6$	$\theta_4 = 0.1$	$\theta_5 = 0.7$	$\theta_6 = 0.8$

Partition Sub-Groups		
1,8,5	2,6,4	7,3

Individual Cluster Assignment								Sub-Group Risk $\bar{\theta}$		
1	2	3	4	5	6	7	8	1	2	3
1	3	5	3	2	3	5	1	$\underline{z} = (1,1,2)$ $\bar{\theta} = 0.27$	$\underline{z} = (3,3,3)$ $\bar{\theta} = 0.6$	$\underline{z} = (5,5)$ $\bar{\theta} = 0.7$
1	3	3	3	4	3	5	1	$\underline{z} = (1,1,4)$ $\bar{\theta} = 0.17$	$\underline{z} = (3,3,3)$ $\bar{\theta} = 0.6$	$\underline{z} = (5,3)$ $\bar{\theta} = 0.65$

Mean: (0.2+0.2+0.4)/3=0.27

Applications: Los Angeles Data: Air Pollution and Deprivation

- The multi-pollutant profile approach developed will be applied to estimates of air pollution concentrations for NO_2 (ppb), $PM_{2.5}$ ($ug\ m^{-3}$) , Ambient Diesel on-road and Diesel off-road concentrations ($ug\ m^{-3}$) exposures obtained using a recently published paper (Su, Morello-Frosch et al. 2009) for Census Tracts (CT) in Los Angeles County.

- Outcome: Deprivation: Number of deprived individual within each CT.

Example: Vulnerable Populations in Los Angeles

1. Assignment Model

$$f\ (x_i) = \prod_{c=1}^{C} \psi_c f(x_i \mid \mu_c, \Sigma_c)$$

2. Disease Model

$$y_i \sim Bin\left(n_i, p_i\right)$$

$$logit[p_i] = \alpha + \theta^*_{z_i} + \varepsilon_i, \quad \sum \theta^*_c = 0$$

Pure Model Averaging (No best clustering) Percentage of Variance Explained

- Percentage of poverty variation explained by air pollution.

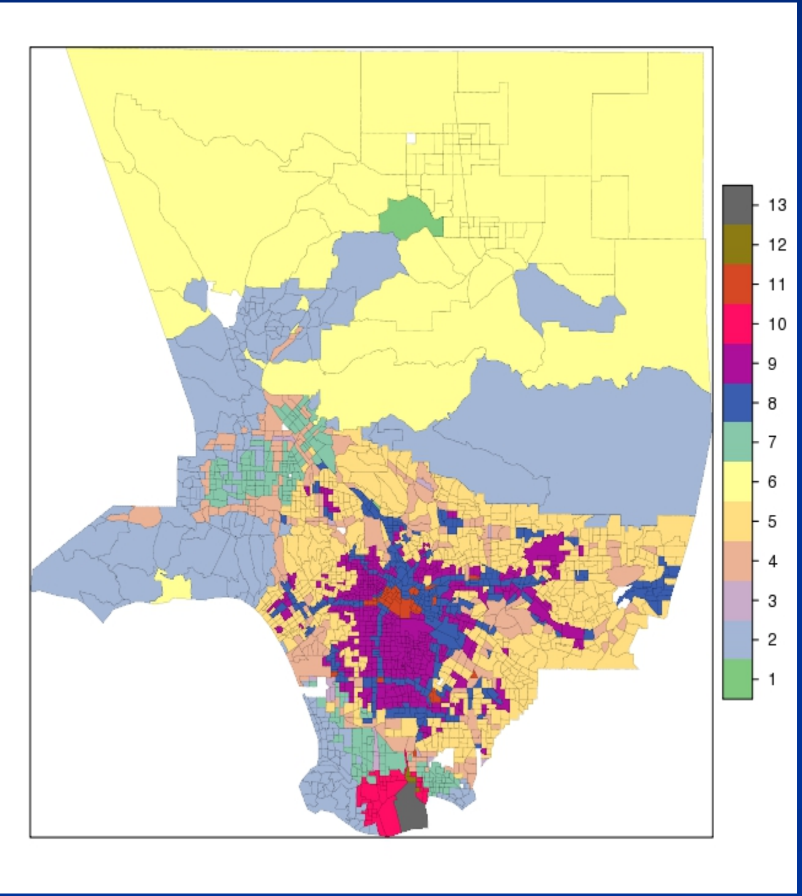
$$y_i \sim Bin(n_i, p_i)$$

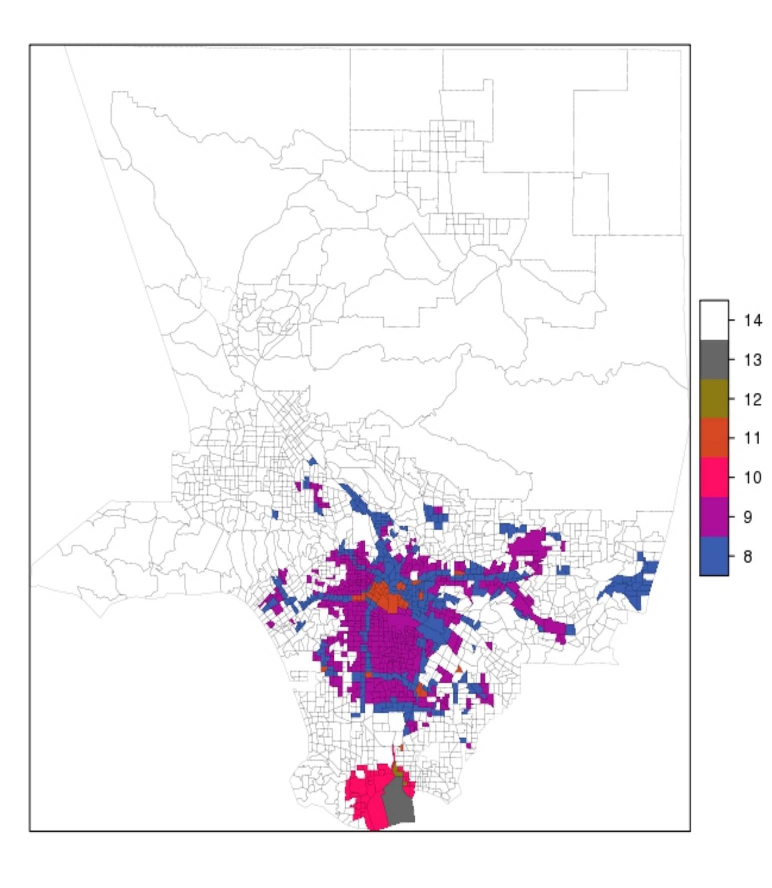
$$logit[p_i] = \alpha + \theta^*_{z_i} + \varepsilon_i$$

$$\rho = \frac{Var(\theta^*_{z_i})}{Var(\theta^*_{z_i}) + Var(\varepsilon_i)}$$

Air pollution/Poverty clusters

Poverty/ Air pollution clusters with statistically significant association with poverty in positive direction.





Air Pollution / Poverty Clusters

Cluster	NO2	PM2.5	Diesel (road)	Diesel (off-road)	Percent Pov	AP Effect
8	26.67 (26.25, 27.11)	21.67 (21.54, 21.80)	1.20 (1.14, 1.25)	1.29 (1.25, 1.33)	0.26 (0.256,0.258)	0.55 (0.47, 0.62)
9	24.20 (23.92, 24.48)	21.70 (21.63, 21.78)	0.72 (0.70, 0.74)	1.43 (1.40, 1.46)	0.29 (0.289,0.291)	0.66 (0.61, 0.71)
10	20.44 (19.37, 21.48)	16.60 (16.08, 17.15)	0.81 (0.67, 0.99)	7.95 (6.45, 9.36)	0.28 (0.281,0.287)	0.76 (0.53, 0.97)
11	32.32 (30.17, 34.41)	21.96 (21.62, 22.31)	2.44 (2.05, 2.84)	1.77 (1.57, 1.99)	0.36 (0.355,0.363)	1.10 (0.90, 1.30)
12	23.60 (14.54, 33.28)	17.45 (12.87, 22.75)	1.99 (0.77, 3.21)	6.91 (4.29, 9.41)	0.54 (0.509,0.574)	1.73 (0.80, 2.52)
13	17.91 (-12.23, 46.97)	17.48 (2.45, 34.90)	0.61 (-2.29, 3.58)	6.70 (-1.87, 13.65)	0.99 (0.991,0.998)	6.91 (5.43, 8.56)

Air Pollution / Poverty Clusters

Percentage of Variation explained by Deprivation ϱ= 0.59 (0.57, 0.62)

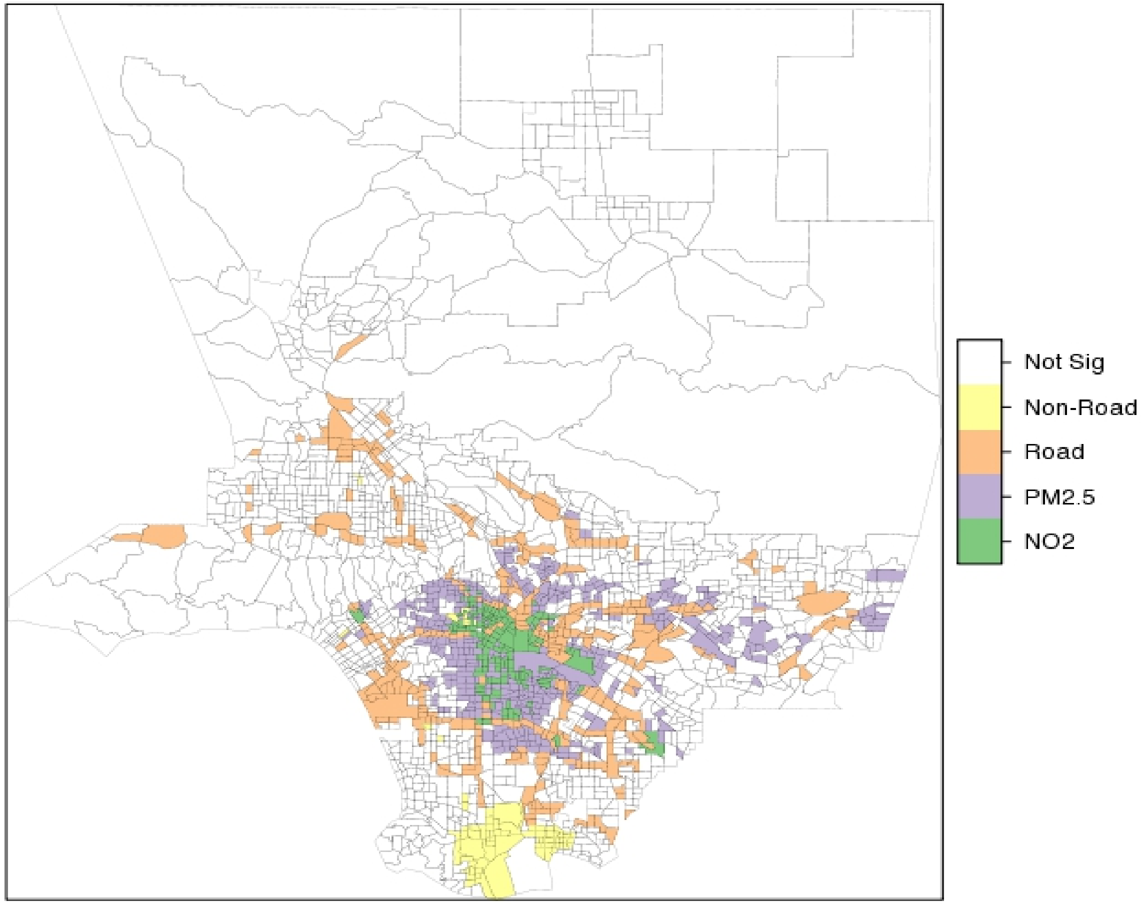
Cluster	NO2	PM2.5	Diesel (road)	Diesel (off-road)	Percent Pov	AP Effect
1	18.67 (1.74, 36.48)	16.69 (6.70, 27.59)	0.44 (-0.93, 1.84)	0.95 (-3.26, 4.52)	0.00 (0.001,0.002)	-4.97 (-6.34, -3.24)
2	15.50 (14.96, 16.08)	17.10 (16.71, 17.48)	0.45 (0.43, 0.48)	1.13 (1.06, 1.21)	0.04 (0.042,0.043)	-1.28 (-1.39, -1.17)
3	22.98 (20.99, 24.88)	19.64 (18.68, 20.52)	1.50 (1.14, 1.87)	1.66 (1.25, 2.31)	0.07 (0.069,0.073)	-0.63 (-1.05, -0.25)
4	22.05 (21.27, 22.75)	20.14 (19.85, 20.40)	0.95 (0.90, 1.01)	1.09 (1.03, 1.16)	0.09 (0.091,0.092)	-0.45 (-0.55, -0.35)
5	21.84 (21.59, 22.11)	21.23 (21.11, 21.34)	0.60 (0.59, 0.62)	1.08 (1.06, 1.11)	0.10 (0.099,0.099)	-0.39 (-0.45, -0.34)
6	16.64 (15.42, 17.80)	12.15 (11.03, 13.47)	0.33 (0.29, 0.38)	0.62 (0.53, 0.74)	0.16 (0.159,0.162)	-0.13 (-0.30, 0.02)
7	19.98 (19.35, 20.61)	18.47 (18.14, 18.75)	0.60 (0.56, 0.64)	1.52 (1.41, 1.64)	0.20 (0.201,0.203)	0.11 (0.00, 0.21)
8	26.67 (26.25, 27.11)	21.67 (21.54, 21.80)	1.20 (1.14, 1.25)	1.29 (1.25, 1.33)	0.26 (0.256,0.258)	0.55 (0.47, 0.62)
9	24.20 (23.92, 24.48)	21.70 (21.63, 21.78)	0.72 (0.70, 0.74)	1.43 (1.40, 1.46)	0.29 (0.289,0.291)	0.66 (0.61, 0.71)
10	20.44 (19.37, 21.48)	16.60 (16.08, 17.15)	0.81 (0.67, 0.99)	7.95 (6.45, 9.36)	0.28 (0.281,0.287)	0.76 (0.53, 0.97)
11	32.32 (30.17, 34.41)	21.96 (21.62, 22.31)	2.44 (2.05, 2.84)	1.77 (1.57, 1.99)	0.36 (0.355,0.363)	1.10 (0.90, 1.30)
12	23.60 (14.54, 33.28)	17.45 (12.87, 22.75)	1.99 (0.77, 3.21)	6.91 (4.29, 9.41)	0.54 (0.509,0.574)	1.73 (0.80, 2.52)
13	17.91 (-12.23, 46.97)	17.48 (2.45, 34.90)	0.61 (-2.29, 3.58)	6.70 (-1.87, 13.65)	0.99 (0.991,0.998)	6.91 (5.43, 8.56)

Pure Model Averaging (No best clustering) Calculating Dominant Pollutant

$$
\begin{aligned}
p_{NO_2} &= \Pr(\mu_{NO_2} > \bar{\mu}_{NO_2}) \\
p_{PM_{2.5}} &= \Pr(\mu_{PM_{2.5}} > \bar{\mu}_{PM_{2.5}}) \\
p_{Diesel} &= \Pr(\mu_{Diesel} > \bar{\mu}_{Diesel}) \\
p_{Non-Diesel} &= \Pr(\mu_{Non-Diesel} > \bar{\mu}_{Non-Diesel})
\end{aligned}
$$

$$
p_{Dominant} = \max(p_{NO_2}, p_{PM_{2.5}}, p_{Diesel}, p_{Non-Diesel})
$$

Statistically Significant Dominant Pollutant – Model



The End

Thanks to
Sylvia Richardson, NuooTing Molitor,
Mike Jerrett