California case study: Spatial linkage and analysis of point data

- Eric Roberts, MD PhD
- Paul English, PhD MPH
- Craig Wolf, MS-Eng
- Makinde Falade, MS-GIS
- Svetlana Smorodinsky, MPH

Themes

1. Health outcome surveillance using point data
2. Hazard metric validity
3. Exposure validity
4. Analytic validity

Alameda County Demonstration Project

- Health outcomes
  - Vital records--preterm birth and term low birthweight
  - Asthma (multiple outcomes)
- Traffic exposure metrics
  - Measures based on traffic counts
  - Measures based on modeled NO_2
    - Land-use regression
    - Modeled NO_2 using ADMS-Urban

Focus on spatial data

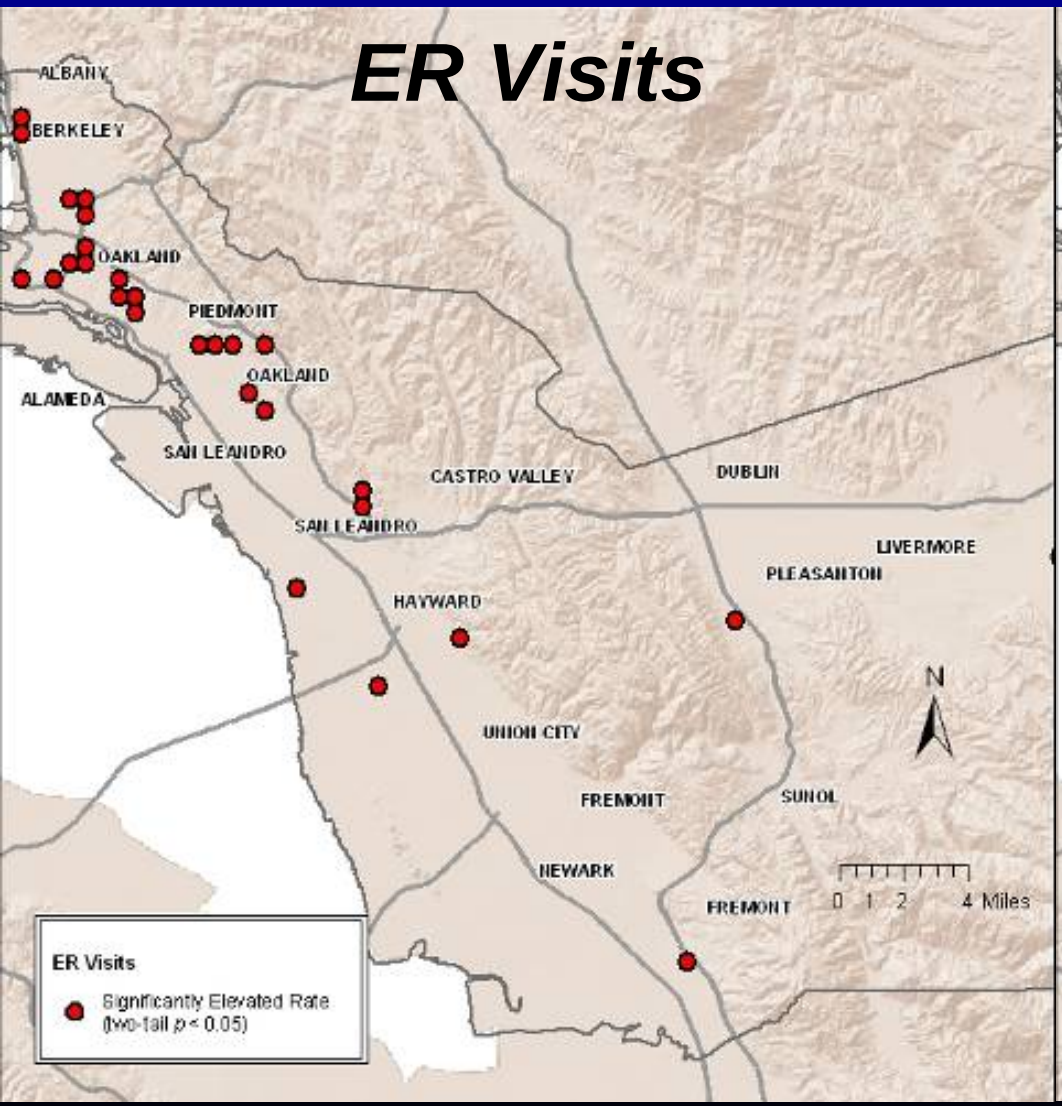
- Advantages
  - Conviction that residential space is important component of disparities, both health and social
  - Communities and populations often self-identify based on their spatial locations
  - Spatial presentation facilitates communication, educational objectives
- Disadvantages
  - Maximizes confounding for any associations
  - Complicates causal inference

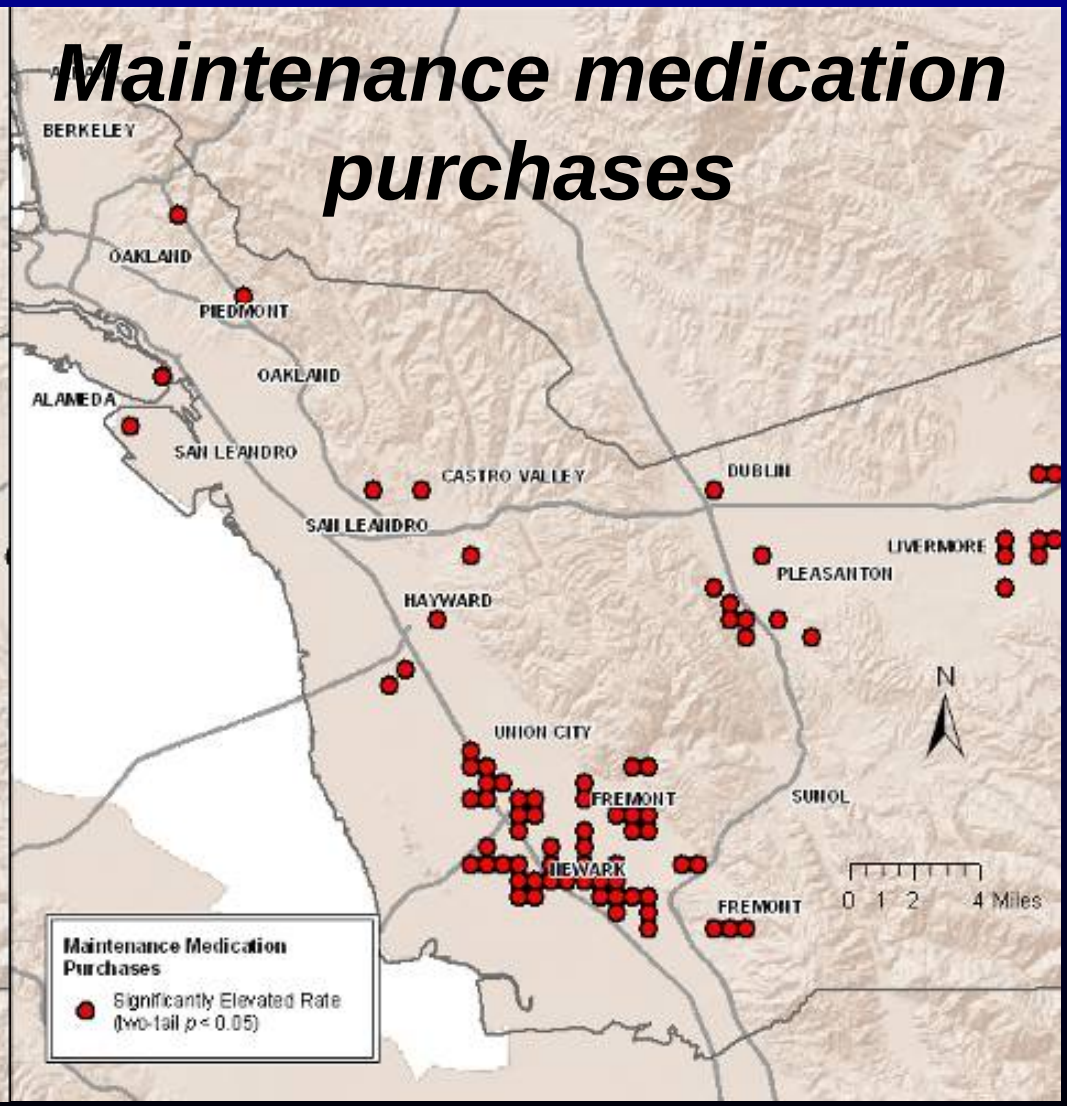
Health outcomes: Data sources

- Birth outcomes
  - Vital records
  - California Center for Health Statistics
  - 100% population sample
- Asthma outcomes
  - Administrative and billing records
  - Special research-oriented arrangement with Medi-Cal and Kaiser Permanente of Northern California
  - Approximately 1 of 3 county residents represented; data believed to have reasonable generalizability

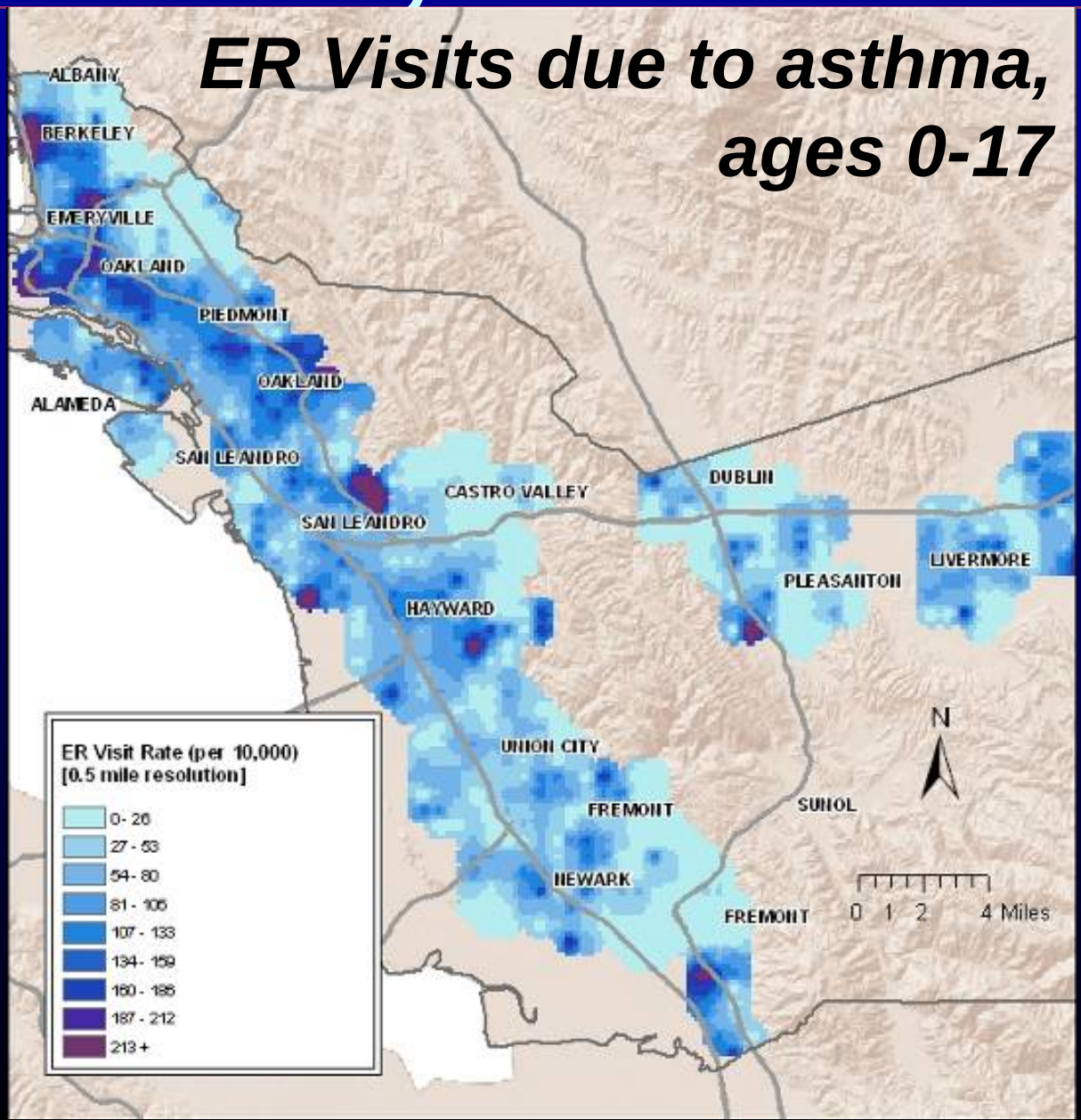
Point data for health outcomes: Primary illustrations

Local elevations in asthma event rates ($p \leq 0.05$), ages 0-17, Alameda County, 2001

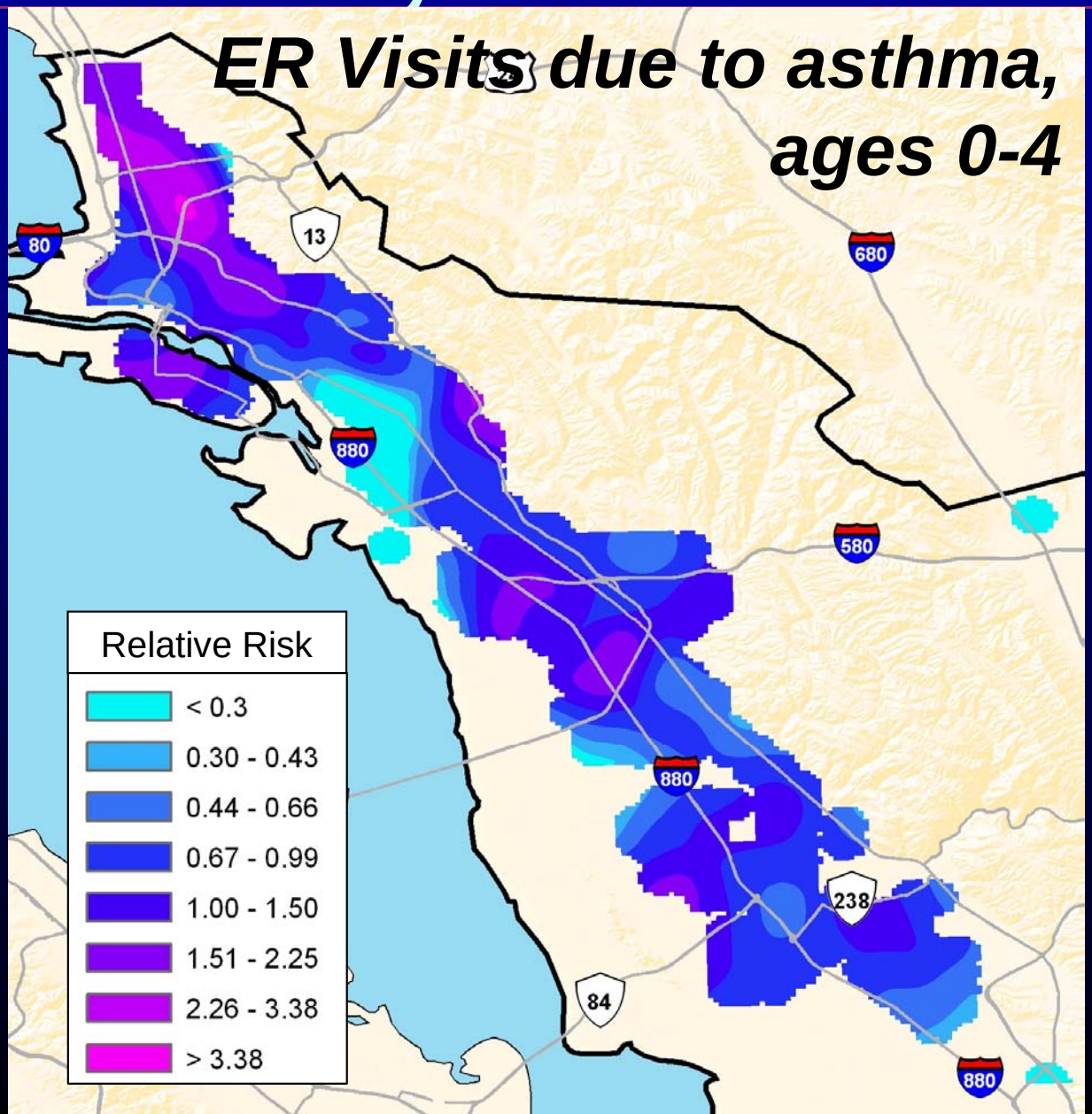




Point data for health outcomes: Primary illustrations

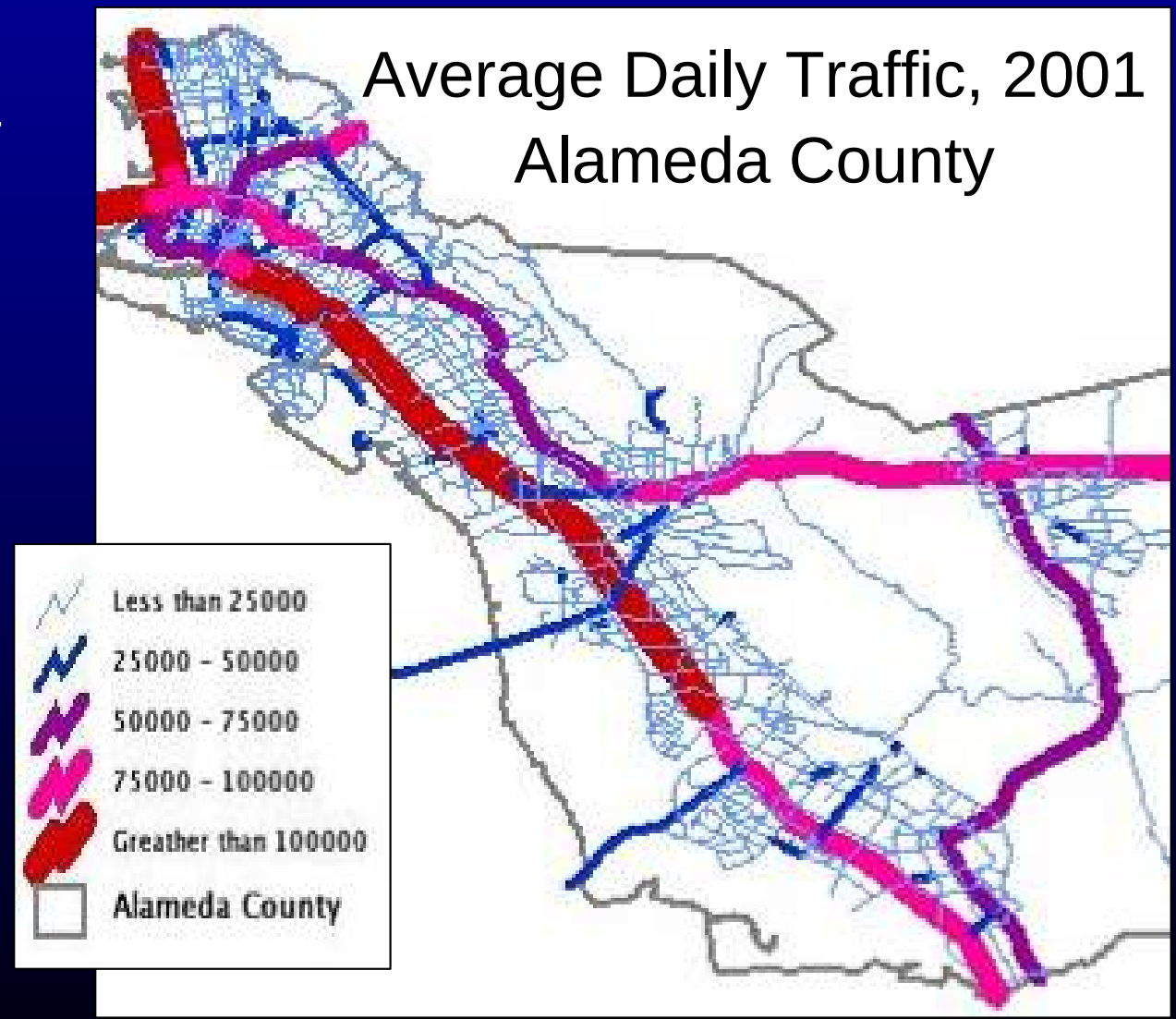


Point data for health outcomes: Primary illustrations



Traffic Count data

- Source: California Department of Transportation
- Lack of consistent collection schedules and protocols
- Low-volume roadways missing data



Traffic Count data

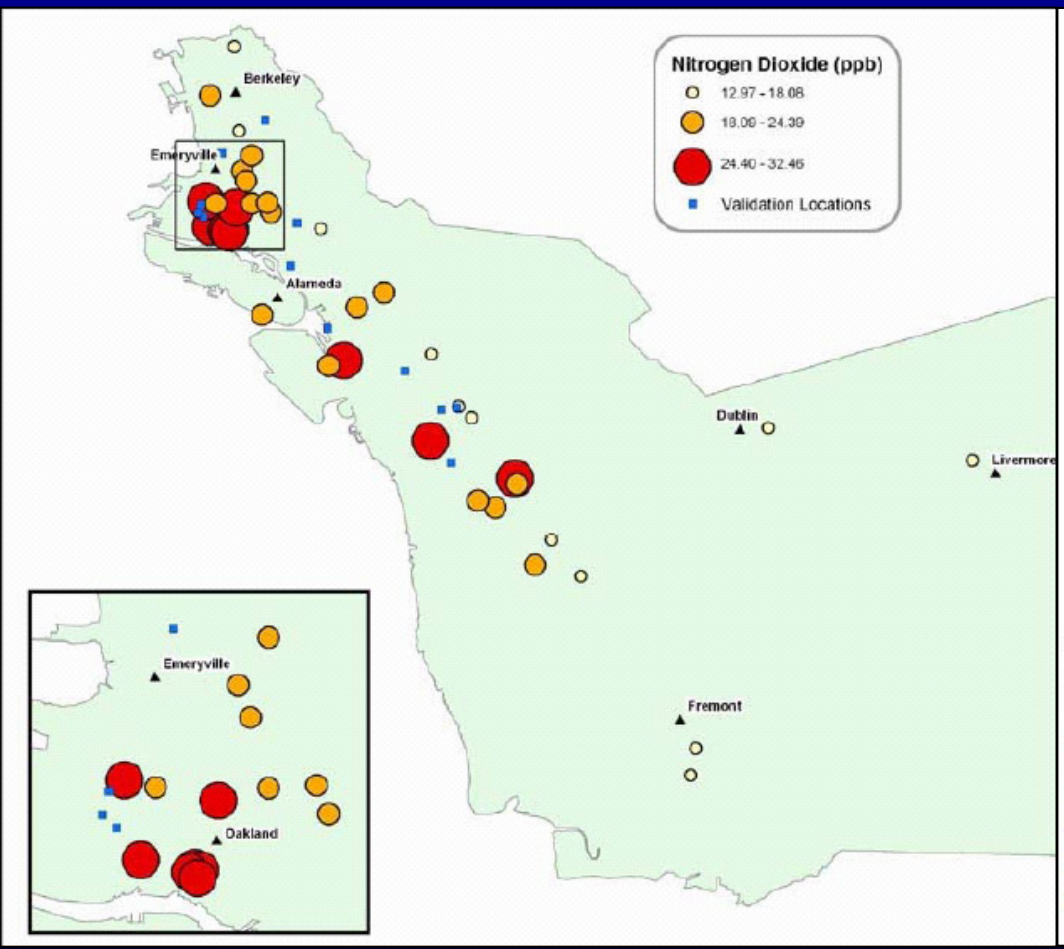
- Processing: automated interface developed by CEHTP (available on web)
- Metrics chosen:
  - Sum of AADT of all roadways within 300 m (R_{300})
  - Sum of AADT adjusted for lengths of segments within 300 m (L_{300})
  - Sum of AADT within 300 adjusted assuming a Gaussian dispersion based on distance: (G_{300})

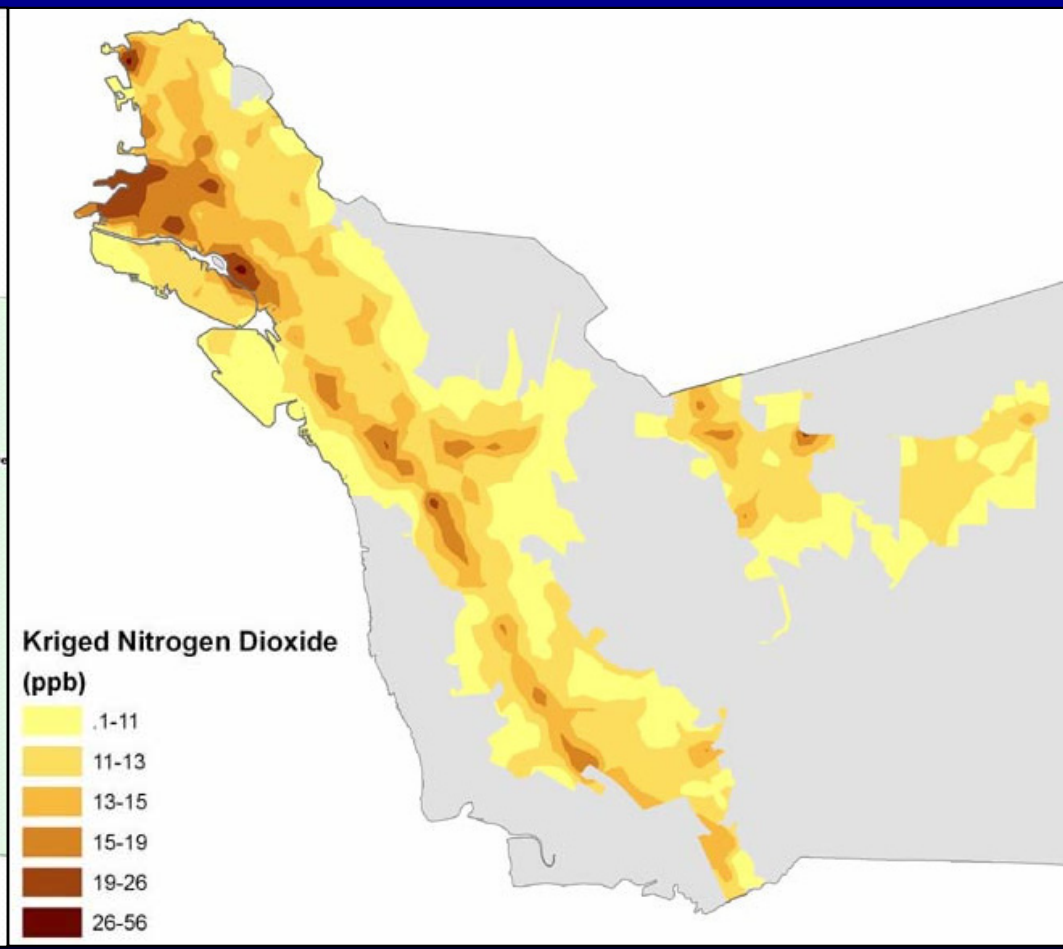
$$Y = \left(\frac{1}{0.4\sqrt{2\pi}}\right) e^{-\left(\frac{}{(0.4)^2}\right)}$$

Land-use regression

- 47 passive diffusion tubes placed around county during 2-week period in Spring 2005
- Analyzed by ion chromatography
- Predictors of log-transformed concentrations (R^2=0.69, $|\text{residual}|_{\text{mean}}$=1.85 ppb)
  - Total AADT within 40 m radius
  - Total AADT between 40 and 500 m radii
  - Total area of Port of Oakland within 1,000 m

Land-use regression

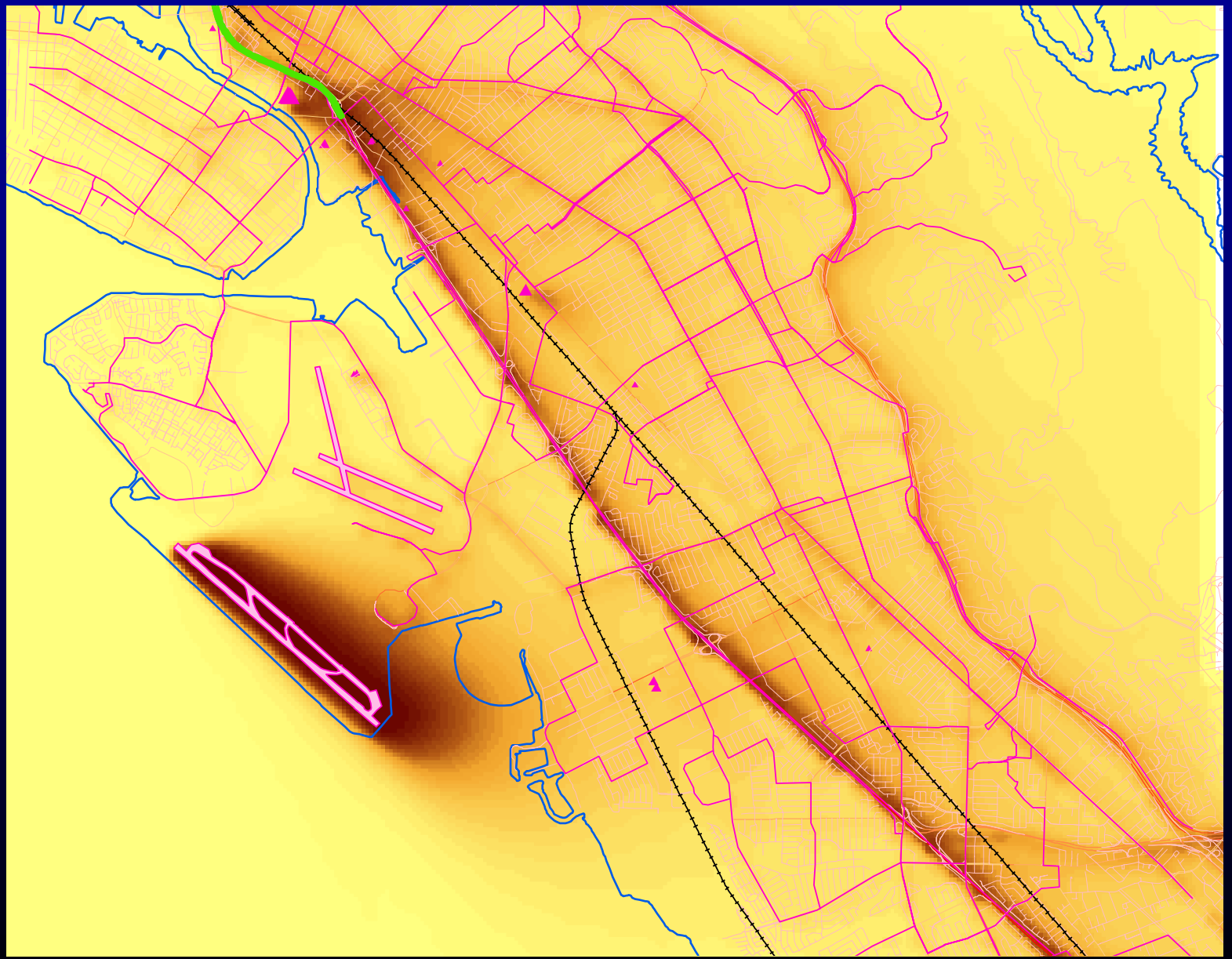




ADMS-Urban

- Incorporation of road data, point sources, meteorology, atmospheric chemistry
- Questions about NO_x/NO_2 assumptions: is NO_2 *as modeled* still best indicator of traffic exposure?
- Sensitive to small distance changes (e.g. road offset used for geocoded health data)

ADMS-Urban



Traffic metrics and measured NO_2

Indicator	Spearman *r*	*p*
Sum of all volumes in 300 m buffer	0.69	<0.0001
Sum of all volumes between 40 and 300 m	0.68	<0.0001
Maximum volume in 300 m buffer	0.57	<0.001
Gaussian adjusted maximum traffic in 1000 m buffer	0.48	0.002

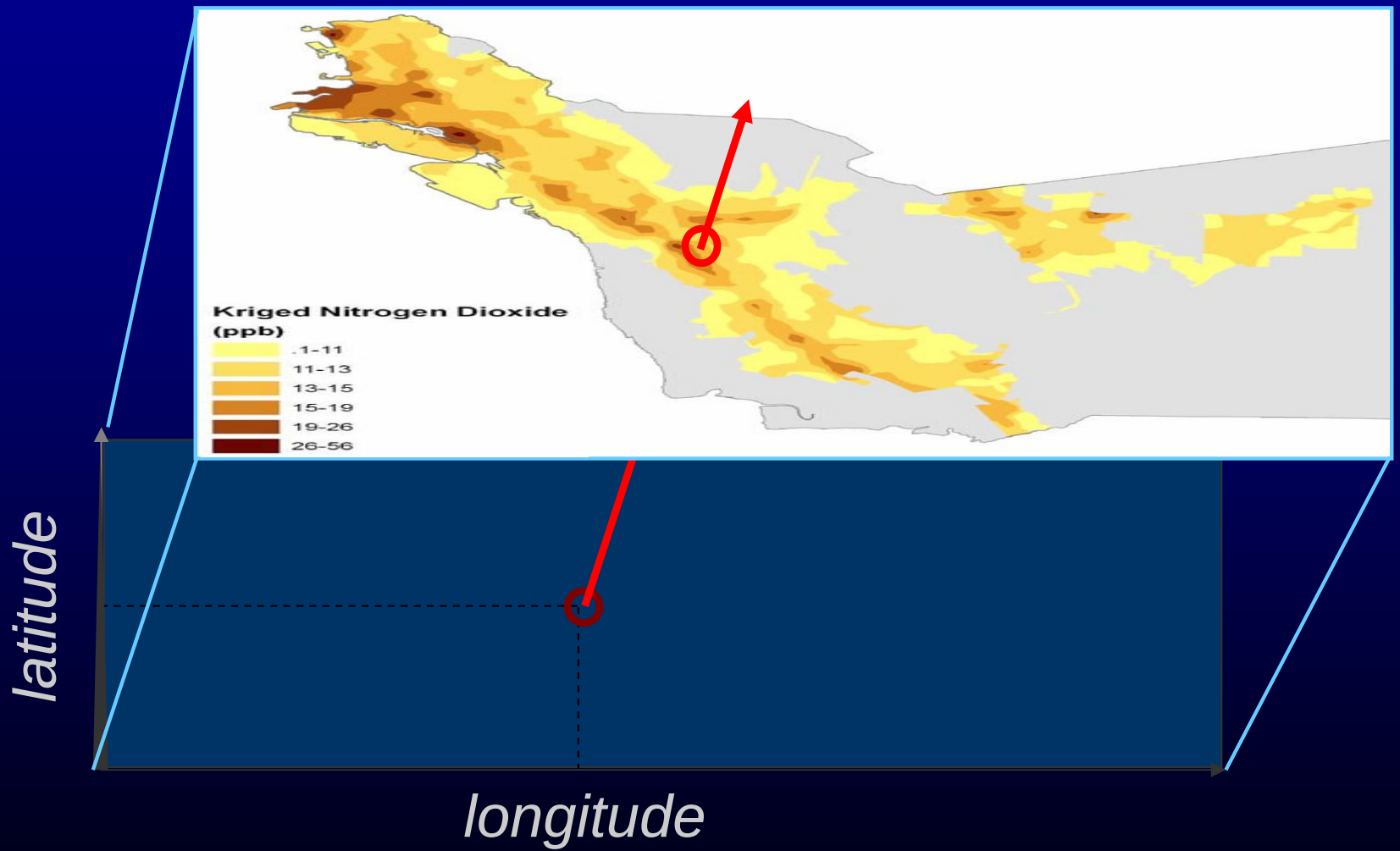
(Smorodinsky, et al. unpublished data)

Performance of NO_2 models

	ADMS-Urban (n=38) (all sources plus background)	Land Use Regression (n=12)
R^2	0.60	0.79
Fraction of 2	100%	100%
Fractional Bias	17.8%	11.9%
% within 5 ppb	68.4%	100%

(Smorodinsky, et al. unpublished data)

Linkage of point data: point-to-polygon intersection



Analytic problems

$$f(ER) = \beta_0 + \beta_{\exp} x_{\exp} + \beta_{\text{cov}} x_{\text{cov}} + \sum \varepsilon_i$$

- Assumes residuals ε_i have constant mean over study space
- This is equivalent to saying all spatial structure is accounted for by

  $\beta_{exp} x_{exp} + \beta_{cov} x_{cov}$

- Solution: Allow for the spatial structure of your residuals in your regression model

$$f(ER) = \beta_0 + \beta_{\exp} x_{\exp} + \beta_{\text{cov}} x_{\text{cov}} + Sp(x, y)$$

Analytic problems

- *Sp(x,y)* Could be a description of how the covariance between neighboring points decreases with distance (Kriging, SAR, CAR)
- Other options are non-parametric functions such as *locally weighted estimation* (loess) or *splines*
  - Currently this approach is the only one developed for point data
- Gotway and Waller: No method necessarily superior, but ***some*** allowance for spatially structured residuals is required

Questions about semi-parametric models

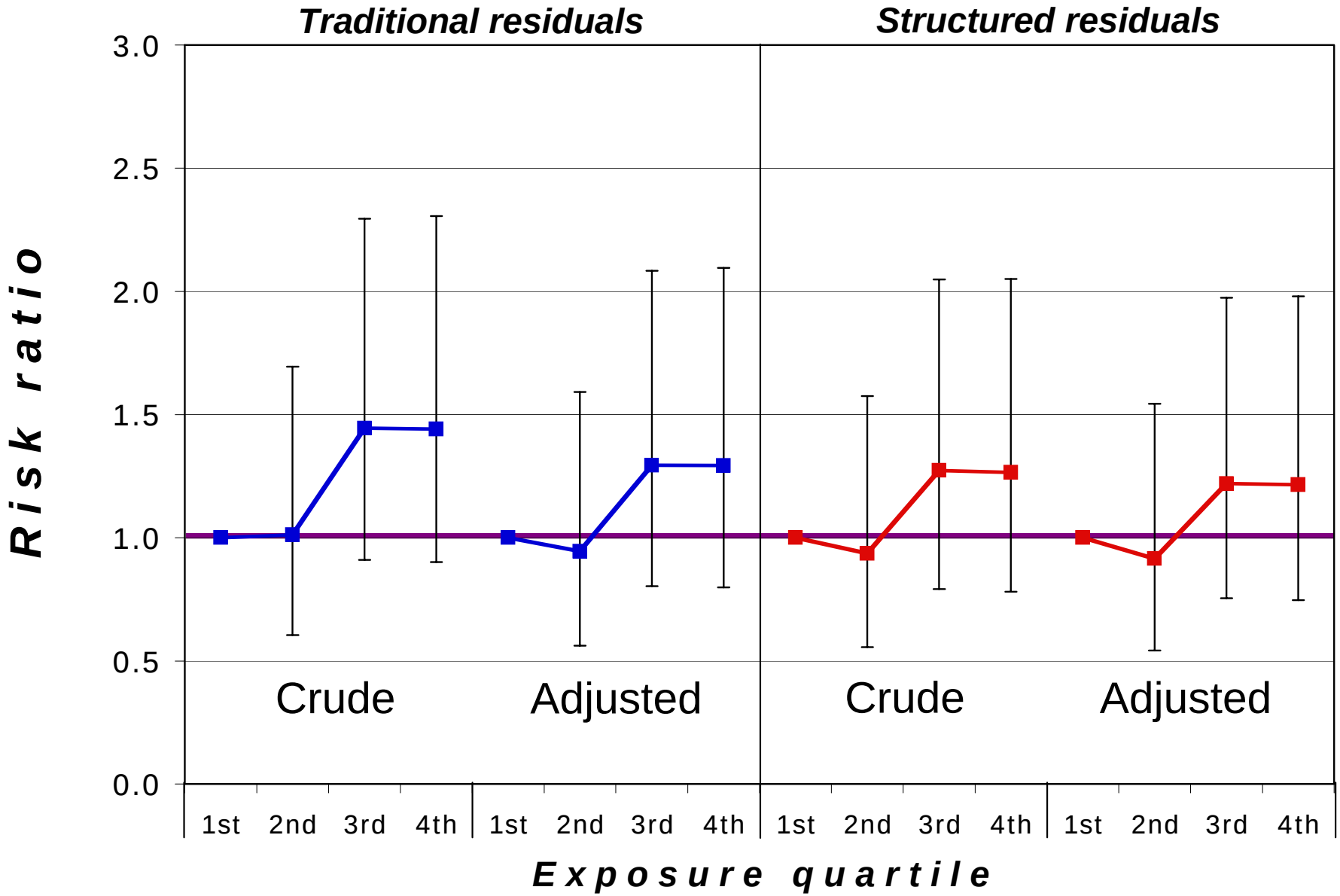
- As currently used, packages in S-plus and R may under-estimate standard errors for β
- If *lo(x,y)* adapts to fit whatever is left out of the model (*residual* structure), will β_{exp} change depending on whether we include our covariates anymore? (Answer: sometimes)

$$f(ER) = \beta_0 + \beta_{\exp} x_{\exp} + \beta_{\text{cov}} x_{\text{cov}} + lo(x, y)$$

$$f(ER) = \beta_0 + \beta_{\exp} x_{\exp} + lo(x, y)$$

- If β_{exp} can be relied upon to be independent of our choice of covariates, is this a solution to our problems with spatial confounding?

Example: AADT in 300 m radius and ER visits for asthma, ages 5-17

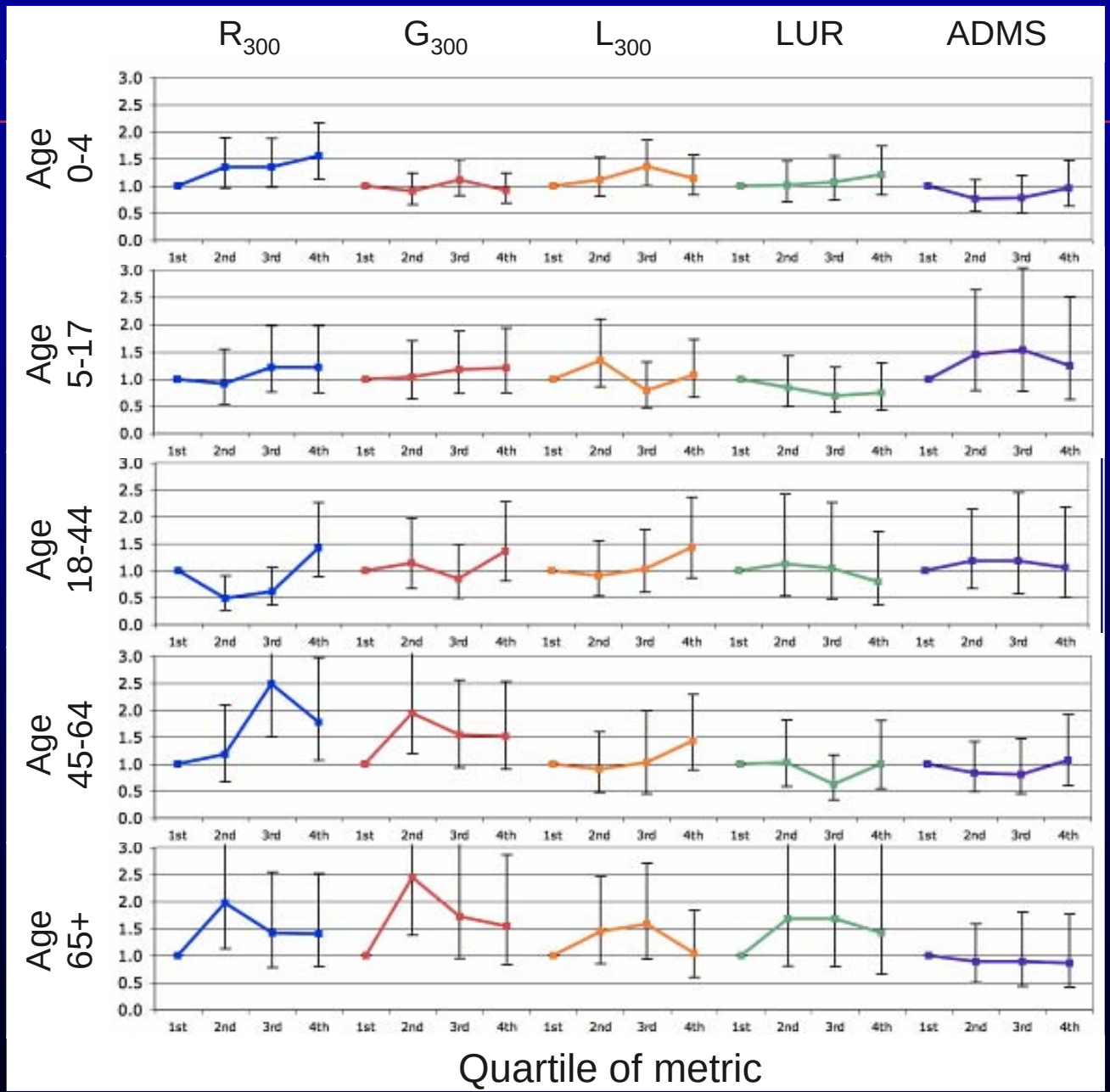


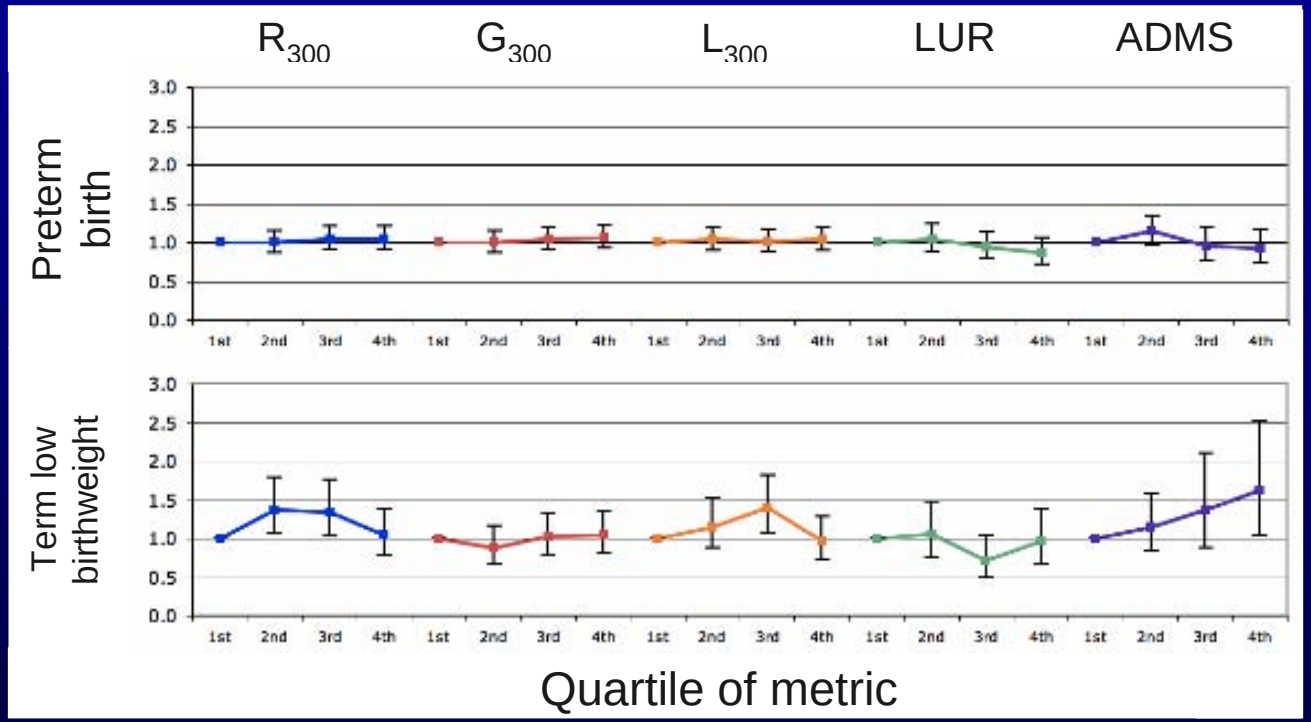
Does linkage of traffic metrics to health outcomes “work?”

- This depends what we mean when we say “work?” (What is the association we ***should*** find?)
- Answer may depend on analytic methods and choices of covariates as much as on health and exposure metrics

Does linkage of traffic metrics to health outcomes “work?”

- This attempt:
  - ER visits for asthma
    - Poisson regression
    - Covariates of median family income in census tract and Medicaid status included
  - Birth outcomes
    - Logistic regression
    - Covariates for maternal race/ethnicity included
  - Both analyses: Loess smoothing term to account for spatial structure of residuals (note this may under-estimate standard errors)


R_{300}
G_{300}
L_{300}
LUR
ADMS
Age 0-4
Age 5-17
Age 18-44
Age 45-64
Age 65+
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1st 2nd 3rd 4th
Quartile of metric


R300
G300
L300
LUR
ADMS
Preterm birth
Term low birthweight
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1st
2nd
3rd
4th
Quartile of metric

Summary

- ER visits due to asthma:
  - Metric with most consistent association: R_{300}
  - Age strata with strongest correlations: 0-4 and 45-64
- Birth outcomes:
  - Metric most consistently associated with preterm birth: none
  - Metric most consistently associated with term low birthweight: ADMS

Summary

- Lack of associations with birth outcomes may be due to pollution levels in Alameda County too low for an effect (e.g. compared to LA)

- In any case, ***associations are inconsistent enough so that we feel like we are cherry-picking the ones we like***

Limitations and next steps

- Hazard validity
  - Refinement of hazard metrics certainly possible, including use of vehicle profiles (truck counts) and pollutants besides NO_2
  - Still have to choose between source ("emissions") and pollutant modeling
  - In the absence of real gains in understanding of specific components of pollution responsible for health effects, "correct" focus is unknown
  - Temporal analysis is more likely to help than spatial analysis in this regard

Limitations and next steps

- Exposure validity
  - Time activity patterns appear to be the next major refinement, but is this likely to lead to increased linkability of Tracking systems?
  - When considering health effects of chronic exposure, residential history may be the more important variable for incorporation into linkage systems

Limitations and next steps

- Analytic validity
  - Not necessarily considered part of linkage, but lack of valid analytic approach will always be a roadblock to making spatial linkage useful
  - Need to address spatial confounding, since can never include ***all*** important covariates in model--is this more an analytic problem than a data collection one?
  - Spatial analytic methods ***designed for point data*** need further development

Thank you!

The CDC Environmental Public Health Tracking Program

Lance A. Waller, PhD*

**For invaluable comments and advice; any errors and misconceptions are my fault, not his!*