SPATIAL ANALYSIS & MORE

Thomas A. Louis, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
www.biostat.jhsph.edu/~tlouis/
tlouis@jhu.edu

Outline

- Prediction vs Inference
- FE vs RE
  - Teacher expectancy study
- Why include correlations or random effects?
- Analysis of Spatial data
  - General issues and approaches
  - Lip cancer example
- Diagnostics: Replace O–E by quantiles
- Analyzing spatially mis-aligned information

Prediction vs Inference

Prediction

- Build a prediction model that has very good, out of sample performance using all available information and modeling approaches
  - No explicit attention to "telling a story and coefficients" or "adjustments"
  - Possibly, stabilize via "regulation" via Bayes, lasso, ...
- Example: Modeling to support individual patient choice of hospital

Inference

- Care is needed in model form and components to ensure that the inference(s) {slopes(s), effects, ...} answer the intended question
  - Respect causal goals
  - Don't include variables "on the pathway"
- Example: Hospital profiling wherein you don't include hospital attributes in a risk adjustment; they are reserved for the hospital effect through which hospitals are compared

The Teacher Expectancy Study[1]

A research synthesis

$Y_k =$ the "expectancy effect score" for the k^{th} study

$\sigma_k =$ $\text{SE}_k = SE(Y_k)$, the within-study standard error

$\text{weeks}_k =$ the number of weeks of teacher-student contact before the experiment

$b_k =$ the "study effect" (unexplained variation in excess of σ_k)

$$\begin{aligned} b_1, \ldots, b_K &\sim N(0, \tau^2) \\ r_k &\sim N(0, \sigma_k^2) \end{aligned}$$

[1] Raudenbush and Bryk (1985). Empirical Bayes Meta-analysis. *J. Educational Statistics* **10**: 75–98.

The Teacher Expectancy Study[1]

A research synthesis

$Y_k =$ the "expectancy effect score" for the k^{th} study

$\sigma_k =$ $\text{SE}_k = SE(Y_k)$, the within-study standard error

$\text{weeks}_k =$ the number of weeks of teacher-student contact before the experiment

$b_k =$ the "study effect" (unexplained variation in excess of σ_k)

$$\begin{aligned} b_1, \ldots, b_K &\sim N(0, \tau^2) \\ r_k &\sim N(0, \sigma_k^2) \end{aligned}$$

$$\begin{aligned} [Y_k \mid b_k, \alpha, \text{weeks}_k, r_k] &= \mu + b_k + \alpha \cdot \text{weeks}_k + r_k \\ E[Y_k \mid b_k, \alpha, \text{weeks}_k] &= \mu + b_k + \alpha \cdot \text{weeks}_k \\ V[Y_k \mid b_k, \text{weeks}_k] &= \sigma_k^2 \\ E[Y_k \mid \text{weeks}_k] &= \mu + \alpha \cdot \text{weeks}_k \\ V[Y_k \mid \text{weeks}_k] &= \tau^2 + \sigma_k^2 \end{aligned}$$

[1] Raudenbush and Bryk (1985). Empirical Bayes Meta-analysis. *J. Educational Statistics* **10**: 75–98.

Estimates (via BUGS)

Model	E(μ) E(α)	SD	P-value	$\hat{\tau}$	-2LL
FE	0.060	0.036	0.098	0	70.7
RE	0.080	0.049	0.103	0.120	70.1
RE + weeks	0.088	0.044	0.046	0.086	62.4
$\hat{\alpha}$	-0.015	0.005	0.003		

- Including the between-study variation (the RE model) increases the SE of the estimated intervention effect to accommodate the broader inference to "studies like these" rather than "these studues"
  - SD: 0.036 $\uparrow$ 0.049
- Including the covariate (weeks) reduces unexplained variability
  - $\hat{\tau}$ 0.120 $\downarrow$ 0.086
- And, reduces the SD, but not to the FE level
  - SD: 0.049 $\downarrow$ 0.044 $>$ 0.036

Why include correlation and random effects?

Why include correlation?

- To improve efficiency
  - Only occasionally a worthy payoff and model can be fragile in that fixed-effects specification can be more demanding than for a working independence model
  - For example, a valid longitudinal analysis may require more than a valid cross-sectional model
- Produce a more valid likelihood and so,
  - Report more "honest" SEs, etc.
  - Under MAR come closer to "ignorability"
- Scientific interest
- Prediction!

Why include random effects (heterogeneity, longitudinal, spatial)?

- Surrogates for unmeasured or poorly measured covariates
  - Covariates $\leftrightarrow$ Covariance
- To broaden the inference space
- To induce correlations (see above)
- To "personalize" the model
- To support stabilization

Spatial data: Issues and Goals

- Tradeoff of geographic resolution and estimation stability
- Tradeoff of variance and bias
- Use spatial correlation and general covariates to accomplish the tradeoff

Why adjust and stabilize?

- There may be region and time-specific adjustment factors
  - age, gender and race distributions
  - differential exposures
- Observed rates may be very

Spatial data: Issues and Goals

- Tradeoff of geographic resolution and estimation stability
- Tradeoff of variance and bias
- Use spatial correlation and general covariates to accomplish the tradeoff

Why adjust and stabilize?

- There may be region and time-specific adjustment factors
  - age, gender and race distributions
  - differential exposures
- Observed rates may be very

N T B E
U S A L

Issues and Goals

- In estimating location-specific rates, need to tradeoff off geographic focus and statistical stability
- Another example of the variance/bias tradeoff

Overall vs Local Shrinkage

- An independent RE (Bayesian) model shrinks individual estimates to the fixed-effects model
- For spatially aligned data, one may want to shrink towards a region-specific focus and also shrink towards the fixed-effects model
- A spatial correlation structure accomplishes this goal
    - It can be considered a surrogate for unmeasured (or poorly measured) spatially aligned covariates

Spatial Correlation

- Specify a correlation/covariance matrix
  - d = a "distance" metameter
  - $\text{corr}(d) = e^{-\gamma d}$
  - Matérn or other flexible options
- Use a Conditional Autoregressive (CAR) model
  - Conditional on all other region-specific parameters, the target parameter has mean that is a weighted average of the other parameters and a variance that depends on the weights
  - The weights depend on distance and some values are illegal in that there isn't a joint distribution that would induce the conditional distribution
  - Monte Carlo methods are needed to fit the model
  - Indeed, the Gibbs sampler (Stuart and Don Geman, 1984) was motived by this kind of problem

Effect of Spatial Correlation

(a positive correlation decreasing with distance)

- Instead of shrinking to the overall mean, shrinkage is to a "local" mean
- Then, this local mean is shrunken towards the fixed-effects model
  - Generally, less than if there had been no local shrinkage
- This occurs for all locations and it takes a computer to sort it out

Consequence

- A collection of elevated, but unstable estimates in subregions of a region will remain elevated, due to local borrowing of information
- Without spatial correlation each subregion estimate would be shrunken a great deal towards the overall mean

A typical spatial model

- The data model is Poisson with expectation $m_k \psi_k$ for location k
- Internal or external standardization is used to estimate the null-hypothesis expectations m_k
- The ψ_k are relative risks with prior distribution

$$\log(\psi_k) = \mathbf{X}_k \boldsymbol{\alpha} + \theta_k + \phi_k$$

- The θ_k are independent random effects that produce extra-Poisson variation
- The ϕ_k are spatially correlated random effects
- Without repeats over time, θ and ϕ are partially confounded but the estimates of ψ are still available

Poisson Spatial Model

$$
\begin{aligned}
\boldsymbol{\eta} &\sim h(\boldsymbol{\eta}) \\
\boldsymbol{\psi} &\sim g(\boldsymbol{\psi} \mid \mathbf{X},\ \boldsymbol{\eta}) \\
Y_k \mid \psi_k &\sim Poi(m_k \psi_k) \\
f(y_k \mid \psi_k) &= \frac{1}{y_k!}(m_k \psi_k)^{y_k} e^{-m_k \psi_k}
\end{aligned}
$$

$$
\log(\psi_k) = \mathbf{X}_k \boldsymbol{\alpha} + \theta_k + \phi_k
$$

- The m_k are expecteds (may result from adjusting for some covariates)
- The θ_k are independent region effects and the ϕ_k are spatially correlated region effects
- They can be considered "model lack of fit" or "region-specific effects"
- Focus is on $\boldsymbol{\alpha}$ and on the adjusted relative risk: $\rho_{\mathbf{k}} = \mathbf{e}^{\theta_{\mathbf{k}} + \phi_{\mathbf{k}}}$

Region Effects

Independent

$$\begin{aligned} \log(\psi_k) &= \mathbf{X}_k\boldsymbol{\alpha} + \theta_k \\ \theta_1, \ldots, \theta_K &\text{ iid } N(0, \tau^2) \end{aligned}$$

- The θs shrink towards 0 and the MLE, region-specific estimates are moved toward the regression surface

Conditional Autoregressive (CAR)

- For a set of weights (w_{kj}) depending on the distance between regions k and j (e.g., 1/0 adjacency)

$$\begin{aligned} \log(\psi_k) &= \mathbf{X}_k\boldsymbol{\alpha} + \phi_k \\ \phi_k|\phi_{j\neq k} &\sim N(\bar{\phi}_k, \tau^2_{\phi_k}) \\ \bar{\phi}_k &= \frac{\sum_{j\neq k} w_{kj}\phi_j}{\sum_{j\neq k} w_{kj}} \\ \tau^2_{\phi_k} &= \frac{1}{\lambda \sum_{j\neq k} w_{kj}} \end{aligned}$$

Incidence of Lip Cancer in Scotland[2]

- County-specific information for Scotland's 56 counties, pooled over the six years 1970-1980
  - Y_k, the observed lip cancer cases in males
  - Expected lip cancer cases are computed from the male population and person-years at risk using internal standardization
  - $X_k = \text{AFF}_k$, the fraction of the male population engaged in agriculture, fishing and forestry
- CAR using adjacency, exchangeable, and combined,

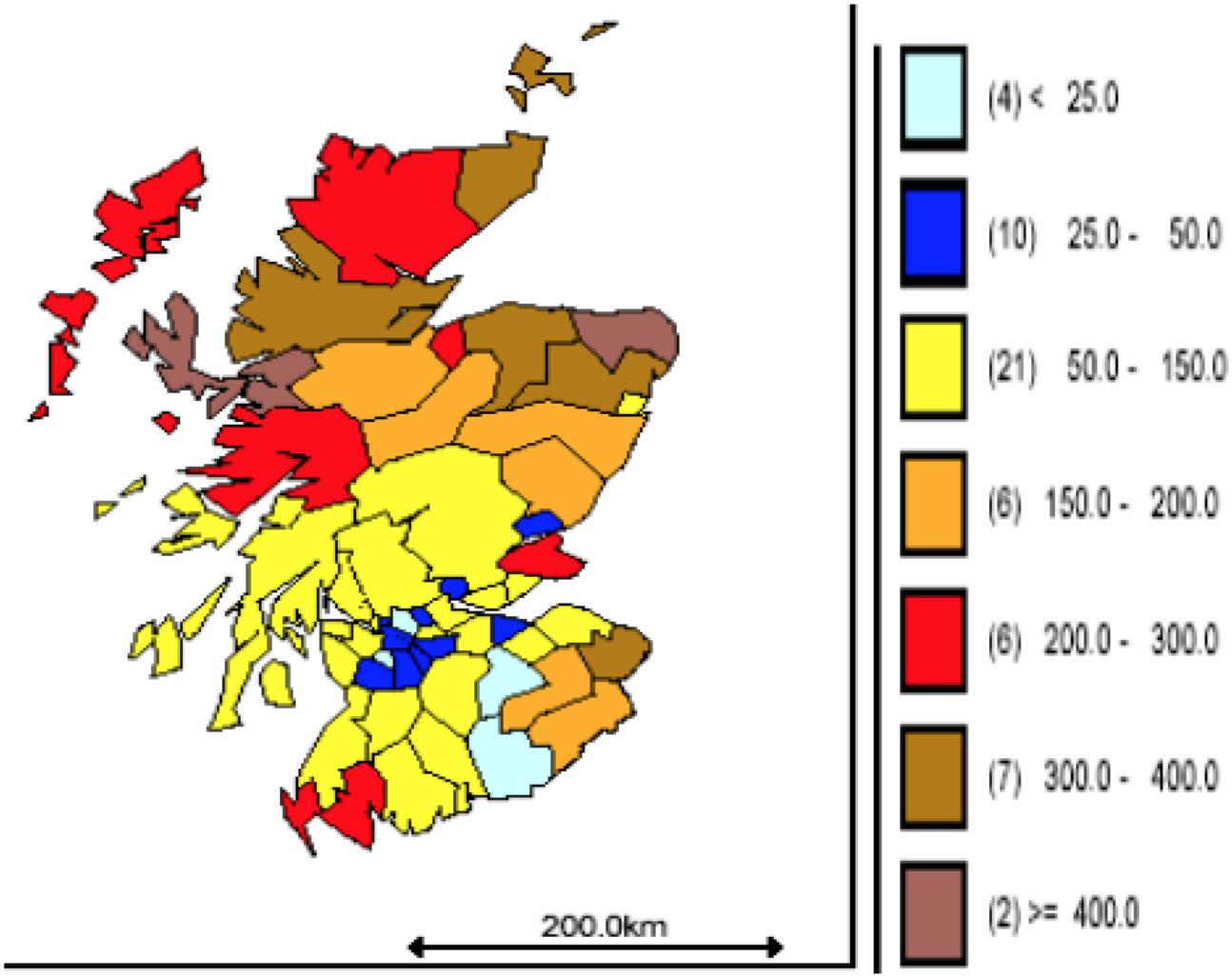
$$\begin{aligned} \log(\psi_k) &= \alpha \text{AFF}_k + \phi_k \\ \log(\psi_k) &= \alpha \text{AFF}_k + \theta_k \\ \log(\psi_k) &= \alpha \text{AFF}_k + \phi_k + \theta_k \end{aligned}$$

- With relative risk, $\rho_k = e^{\phi_k + \theta_k}$, deleting either ϕ or θ for a sub-model

[2] Clayton & Kaldor, 1987 *Biometrics*

CRUDE, COUNTY-SPECIFIC RELATIVE RISKS

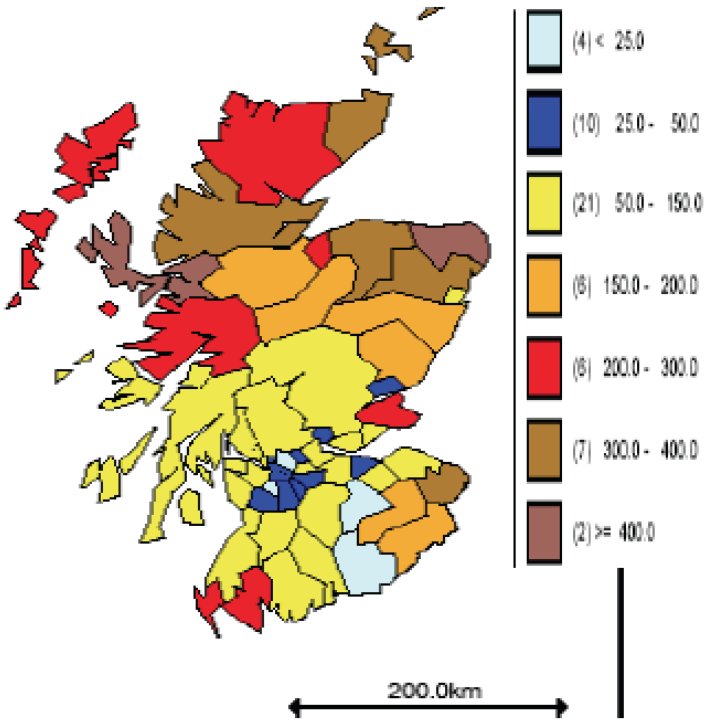
Rates appear to cluster, with a noticeable grouping of counties with SMR> 200 in the North



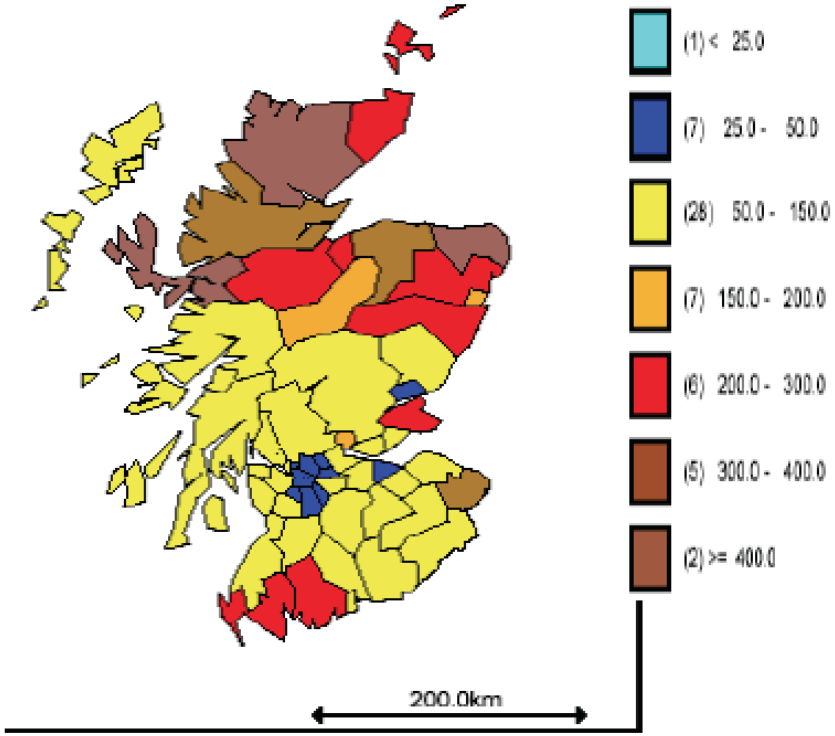
THE CAR MODEL

Local Smoothing

Crude SMR

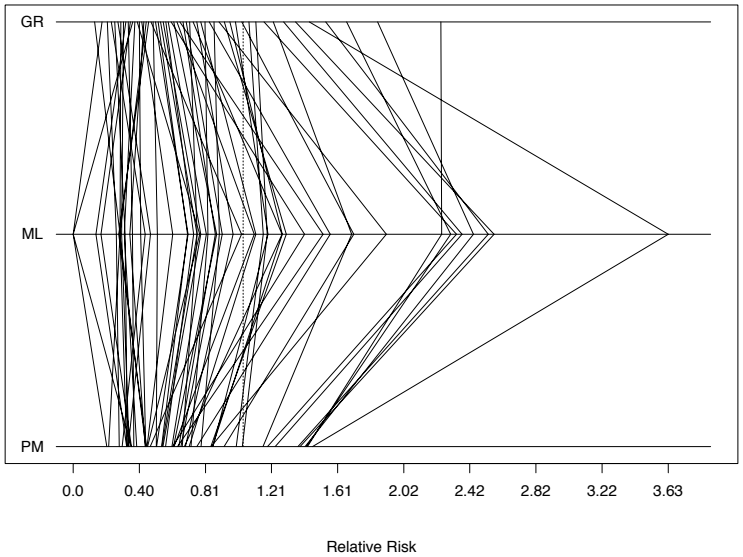


Smoothed SMR

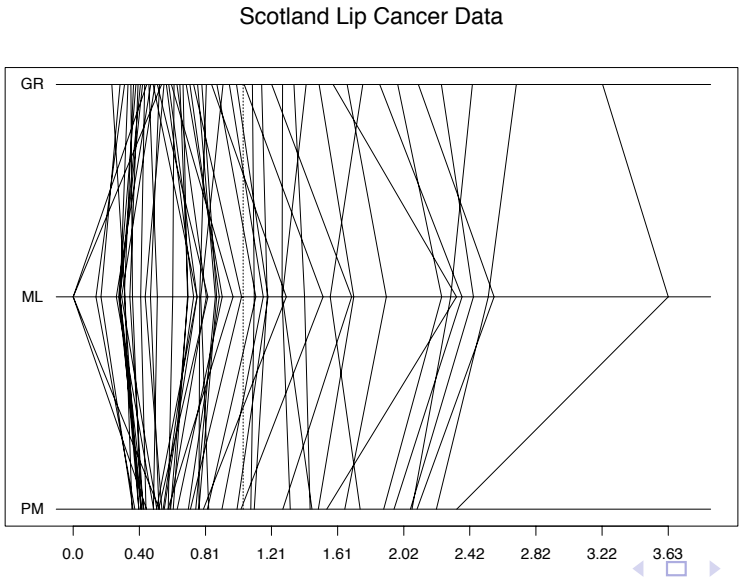


Shrinkage Plots

Exchangeable



CAR



Estimates: 2000 samples after 1000 burn-in

Exchangeable model

$\alpha_0 \sim N(0, 10^6)$, $\alpha \sim N(0, 10^8)$

$\tau^{-2} \sim$ Gamma, E $= 1$, V $= 1000$

Parameter	Posterior Mean	Posterior SD
α_0	-0.51	0.16
α	6.95	1.33
τ	0.62	*

- RR ≈ 2.0 for a 0.1 change in AFF

CAR model

$\alpha \sim$ N$(0, 10^8)$

$\tau^{-2} \sim$ Gamma; E $= 0.25$ V $= 1000$

Parameter	Posterior Mean	Posterior SD
α	4.04	1.13
τ	0.63	*

- RR ≈ 1.5 for a 0.1 change in AFF

BUGS Programming

$w_{kj} =$ 1 or 0 according as county j is or is not adjacent to county k

$c_k =$ the number of neighbors of region k

$$
\begin{aligned}
Y_k &\sim \text{Poisson}(m_k \psi_k) \\
\log(\psi_k) &\leftarrow \alpha \text{AFF}_k + \phi_k \\
\alpha &\sim \text{Normal}(0,\ 10^{-8}) \\
\phi_k \mid \boldsymbol{\phi}_{j \neq k} &\sim \text{Normal}(\bar{\phi}_k,\ \text{prec}_k) \\
\bar{\phi}_k &\leftarrow \frac{1}{c_k} \sum_{j \in neighbors(k)} \phi_j \\
\text{prec}_k &\leftarrow \lambda c_k \\
\lambda &\sim \text{Gamma}(0.25,\ 1000) \\
\psi_k^{ml} &\leftarrow Y_k / m_k
\end{aligned}
$$

The full predictive distribution

Binomial Example

$$\begin{aligned} [Y_k \mid P_k] &\sim \text{Binomial}(n_k, P_k) \\ \hat{P}_k &= \frac{Y_k}{n_k} \end{aligned}$$

- Have $P_k^{(\nu)}$ MCMC draws, $\nu = 1, \ldots, M$.
- For each $P_k^{(\nu)}$ generate

$$Y_k^{(\nu)} \sim \frac{1}{n_k} \text{Binomial}\left(n_k, P_k^{(\nu)}\right).$$

- The $\left(Y_k^{(\nu)}, P_k^{(\nu)}\right), \nu = 1, \ldots, M$ give the joint distribution

Moments of the Predictive Distribution

- The usual "mantras"

$$
\begin{aligned}
E_k &= E(Y_k) = Y_k^{(\bullet)} = E\{(E(Y_k \mid P_k)\} \approx P_k^{(\bullet)} \\
V_k &= V(Y_k) = \frac{1}{M}\sum_{\nu}\left\{Y_k^{(\nu)} - Y_k^{(\bullet)}\right\}^2 = E\{V(Y_k \mid P_k)\} + V\{E(Y_k \mid P_k)\} \\
&\approx \frac{1}{M}\sum_{\nu}\frac{1}{n_k}P_k^{(\nu)}\left(1 - P_k^{(\nu)}\right) + \frac{1}{M}\sum_{\nu}\left\{P_k^{(\nu)} - P_k^{(\bullet)}\right\}^2 \\
&\qquad \left\{\begin{matrix}\text{Binomial}\\ \text{Variance}\end{matrix}\right\} \quad + \quad \left\{\begin{matrix}\text{Model}\\ \text{Uncertainty}\end{matrix}\right\} \\
SD_k &= V_k^{\frac{1}{2}}
\end{aligned}
$$

- If a large number of $Y_k^{(\nu)}$ are generated for each $P_k^{(\nu)}$

  (specifically, $Y_k^{(\ell,\nu)}, \ell = 1, \ldots,$ "large"),

  then " $\approx$ " can be replaced by " $=$ ".

Residuals

- Standardized (Observed - Expected) residual

$$R_k^* = \frac{\hat{P}_k - E_k}{SD_k}$$

- These are fine for the Gaussian, but not so good for small P binomial
- Better is to find the percentile location of $\hat{P}_k$ amongst the $\left\{ Y_k^{(\nu)} \right\}$
- Denote it by ζ_k and for the residual use,

$$R_{kt}^{\ddagger} = \Phi^{-1}(\zeta_k)$$

- If the predictive distribution is exactly Gaussian, these will be identical to the R_k^* and in general are less dependent on the Gaussian assumption
- For example, here are comparisons of R^* and $R^{\ddagger}$ when $n = 25$, the direct estimate is 0 and there is only Binomial uncertainty (no model uncertainty)

$$R^* = -\left(\frac{nP}{1-P}\right)^{\frac{1}{2}} \qquad R^{\ddagger} = \Phi^{-1}\left\{(1-P)^n\right\}$$

P	.01	.05	.10	.50
R^*	-0.50	-1.15	-1.67	-5.00
$R^{\ddagger}$	+0.76	-0.59	-1.46	-5.42

Integrating mis-aligned information

Exposure assessment at the Fernald, OH superfund site[3,4]

- In the years 1951-1988 the former Feed Materials Production Center (FMPC) processed uranium for weapons production
- The Dosimetry Reconstruction Project sponsored by the CDC, indicated that during production years the FMPC released radioactive materials
- The primary exposure to residents of the surrounding community resulted from breathing radon decay products
- The risk assessment required estimates of the number of individuals at risk using block-group, age/sex population counts, and exposure as dictated by wind direction, distance from the plant and building density

[3] Mugglin and Carlin (1998). Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones. *JABES*, 3: 111-130.

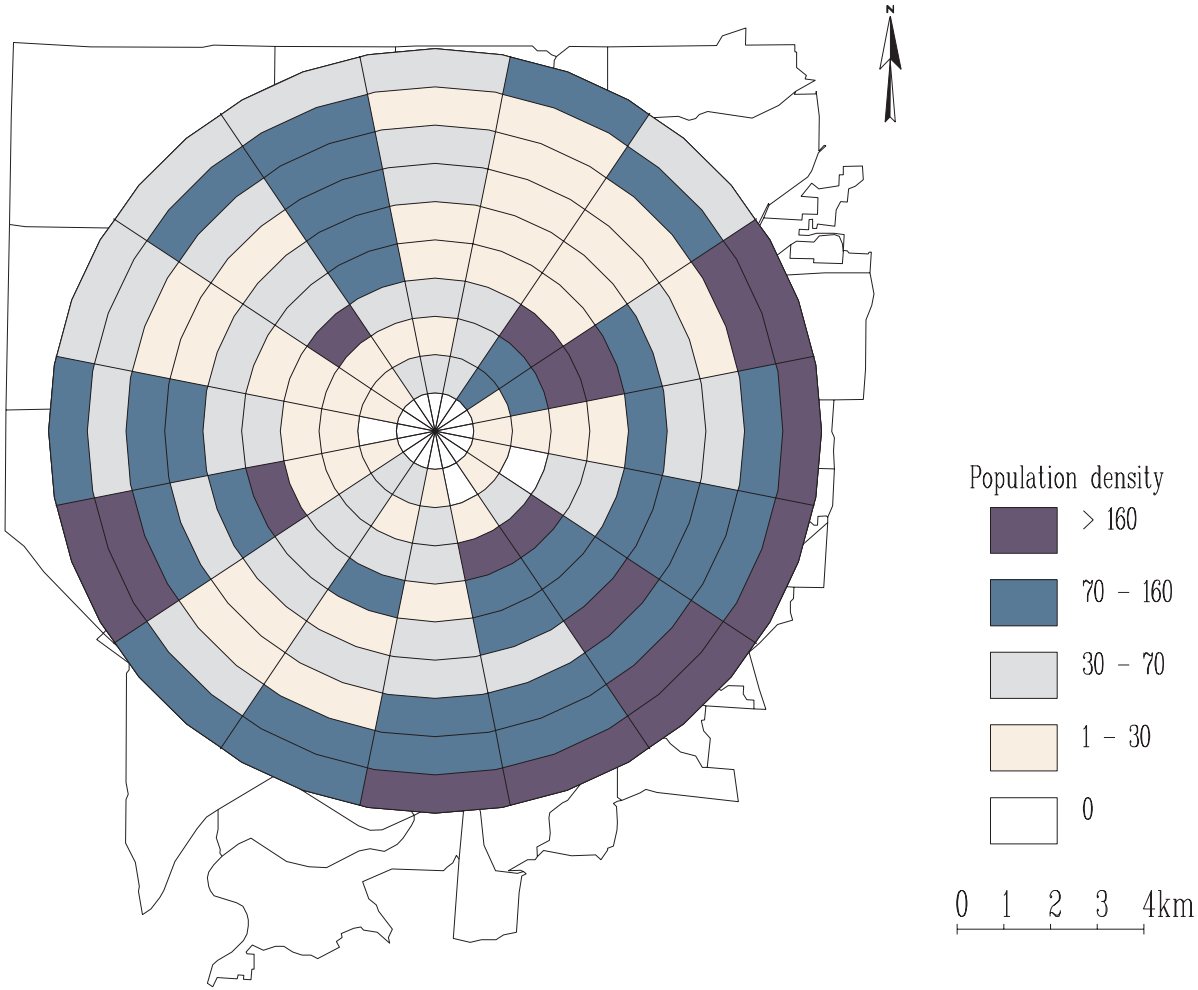
[4] For a more modern approach, see Bradley, Wikle, Holan(2015b) Spatio-temporal change of support with application to American Community Survey multi-year period estimates. *Stat*, 4: 255–270.

Estimating Health Effects

Need to estimate:

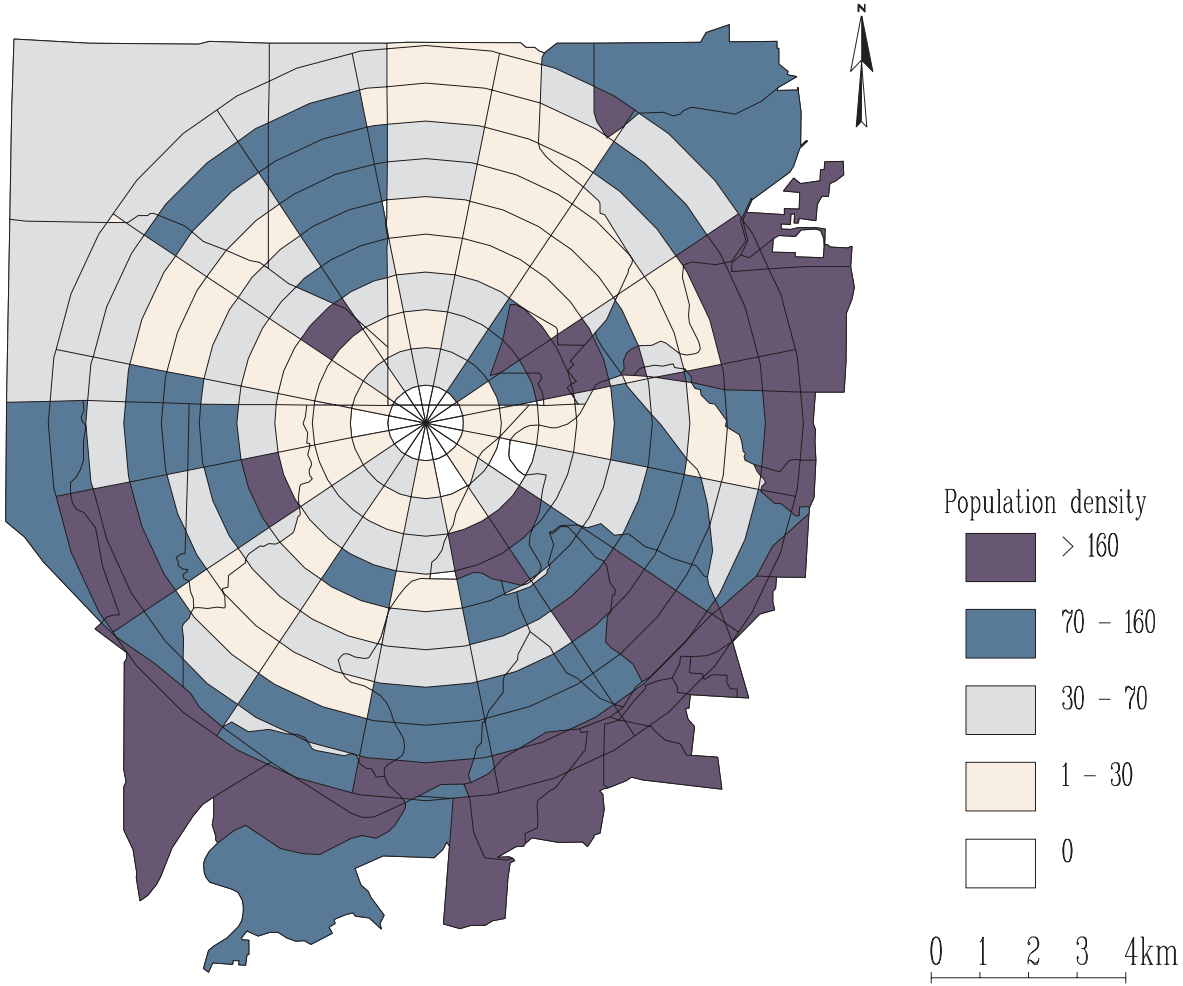
- The number of individuals at risk using block-group population counts, broken down by age and sex
- Exposure with the areas of interest dictated by direction and distance from the plant
- The following figures display exposure "windrose," consisting of 10 concentric circular bands at 1-kilometer radial increments divided into 16 compass sectors
    1. Population counts
    2. These overlayed on USGS maps
    3. These with counts of the number of structures (residential buildings, office buildings, industrial building complexes, warehouses, barns, and garages) within each cell
        - The hatching pattern in indicates the areal density (structures per square kilometer) in each cell

1. Population density & wind direction



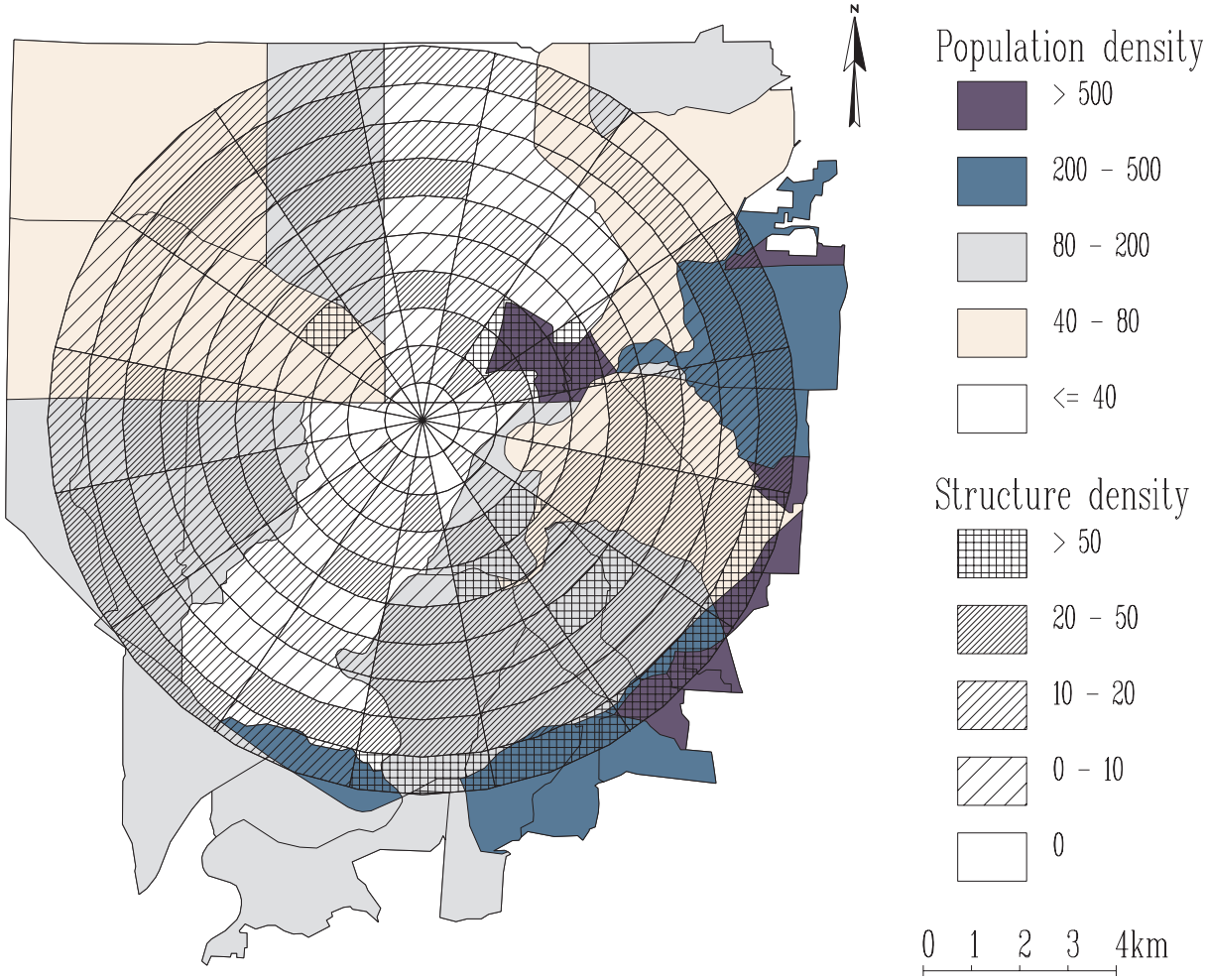
- Population density intersected with census units and wind direction centered around the exposure source

2. Population density, USGS map, . . .



- Population density intersected with census units and wind direction centered around the exposure source, overlay on USGS map

3. Population density, structure density, . . .



- Population density intersected, structure density, census units and wind direction centered around the exposure source

Integration and Risk Assessment

- It is necessary to interpolate subgroup-specific population counts to the windrose exposure cells
- These numbers of persons at risk can then be combined with cell-specific dose estimates and estimates of the cancer risk per unit dose to obtain expected numbers of excess cancer cases by cell
- The Bayesian formalism is necessary to combine and smooth the misaligned information, thereby producing a complex posterior distribution of population counts, exposures, etc. that supports the risk assessment
- The approach depends on constructing a Rosetta Stone linking the data sources and letting Markov Chain Monte-Carlo do the hard work