Discussion of, Combining Information from Survey and non-Survey Data Sources: Challenges & Opportunities by, Sharon Lohr & Trivellore Raghunathan

Thomas A. Louis, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
tlouis@jhu.edu

Disclaimer

Outline

- Comments on Lohr & Raghunathan
- A (non-probability) sample of Census projects
- Integrating misaligned information:
  Risk assessment at the Fernald OH superfund site
- Coda

Comments on Lohr & Raghunathan

- (Proper) combining confers benefits, but care is needed
  - Explicit mapping of assumed relations
  - Calibration, (probabilistic) record linkage, disclosure limitation, operating characteristic evaluation, ...
  - Multiple imputation and other "honest" assessments of uncertainties

Comments on Lohr & Raghunathan

- (Proper) combining confers benefits, but care is needed
  - Explicit mapping of assumed relations
  - Calibration, (probabilistic) record linkage, disclosure limitation, operating characteristic evaluation, ...
  - Multiple imputation and other "honest" assessments of uncertainties
- Use of administrative data as augmenters, benchmarks/calibrators is relatively straightforward and safe
  - Use these to supplement/complement high-quality surveys
- Sensors and other objective data-generators have high potential
  - Traffic monitors likely should replace the ACS leave for work and commuting time questions
- Combining survey with non-curated information has great potential, but is risky and we need to develop principled approaches

Comments on Lohr & Raghunathan

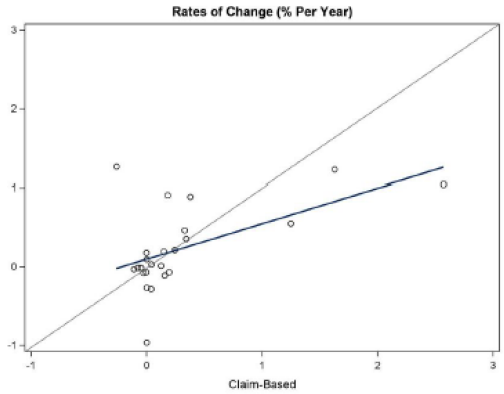
- (Proper) combining confers benefits, but care is needed
  - Explicit mapping of assumed relations
  - Calibration, (probabilistic) record linkage, disclosure limitation, operating characteristic evaluation, ...
  - Multiple imputation and other "honest" assessments of uncertainties
- Use of administrative data as augmenters, benchmarks/calibrators is relatively straightforward and safe
  - Use these to supplement/complement high-quality surveys
- Sensors and other objective data-generators have high potential
  - Traffic monitors likely should replace the ACS leave for work and commuting time questions
- Combining survey with non-curated information has great potential, but is risky and we need to develop principled approaches
- What may appear to be small selection effects in self-selected, web-based "surveys" can induce considerable bias and substantially increase MSE

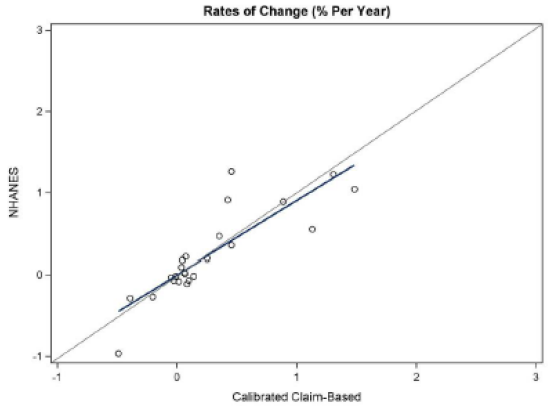
  See Meng's discussion of Keiding & Louis (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussion and response). *JRSS-A*, 179: 319–376.

Comments on Lohr & Raghunathan

(Continued)

- Complex linking and calibration
  - NHANES vs Self-report prevalence Calibration successfully lines them up

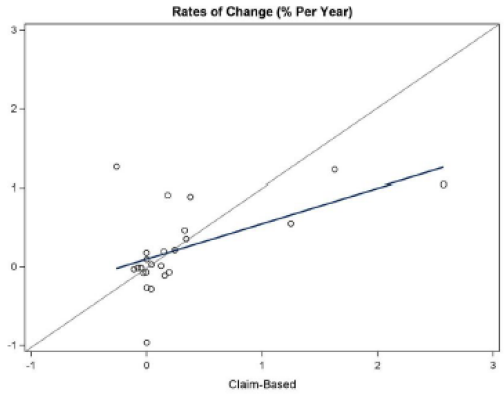


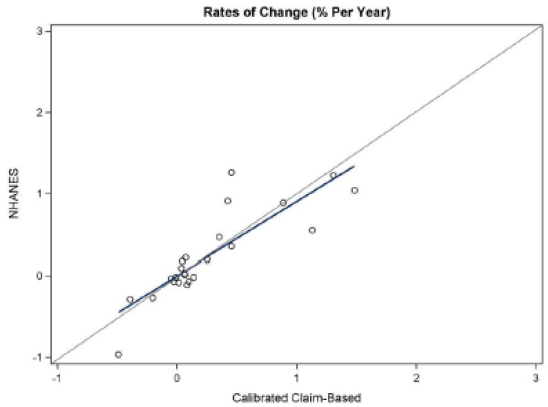


Comments on Lohr & Raghunathan

(Continued)

- Complex linking and calibration
  - NHANES vs Self-report prevalence Calibration successfully lines them up



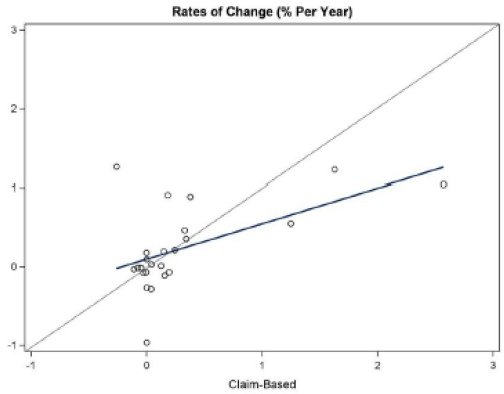


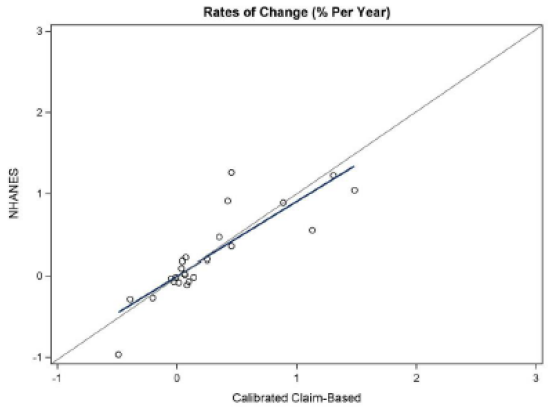
- Hierarchical models/small domain estimation:
  Give the small sample size domains a break
  Jiang, Nguyen, Rao (2011). Best Predictive Small Area Estimation. *JASA*, 106: 732–745.
- Combining data sources is a missing data problem

Comments on Lohr & Raghunathan

(Continued)

- Complex linking and calibration
  - NHANES vs Self-report prevalence Calibration successfully lines them up





- Hierarchical models/small domain estimation:
  Give the small sample size domains a break
  Jiang, Nguyen, Rao (2011). Best Predictive Small Area Estimation. *JASA*, 106: 732–745.
- Combining data sources is a missing data problem
  - All of statistical inference is a missing data problem
- **Complex modeling challenge**
  Quantify the relative influence of data and models on final results

CENSUS: LEHD[1]

Longitudinal Employer-Household Dynamics

Merging administrative and survey data at Census: The LEHD program

An innovative federal statistical program, collecting existing data and ...

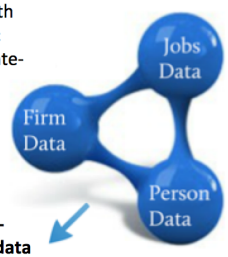
..linking it together to provide new information sources at low cost

New linked national jobs data for the U.S.

United States Census Bureau | U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

The LEHD data:

Firm data: *survey* data on firms from both Census economic programs and state-provided QCEW.



Jobs data: *administrative* data on worker earnings at jobs from state unemployment insurance systems. Key record for linking firm and person data.

Linked employer-employee microdata

Person data: demographics of workers sourced from mix of Census *survey* and SSA and IRS *administrative* data.

New public use statistics tabulated from linked frame

United States Census Bureau | U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU census.gov

[1] Thanks to Erika McEntarfer

CENSUS: LEHD

Bringing data together

How the data are brought together:

Unique state/federal data sharing partnership

- 49 states (& DC) share unemployment insurance wage records and QCEW data with Census
- LEHD program links administrative data from state and federal sources with survey data from Census and BLS economic and demographic programs.

Advantages:

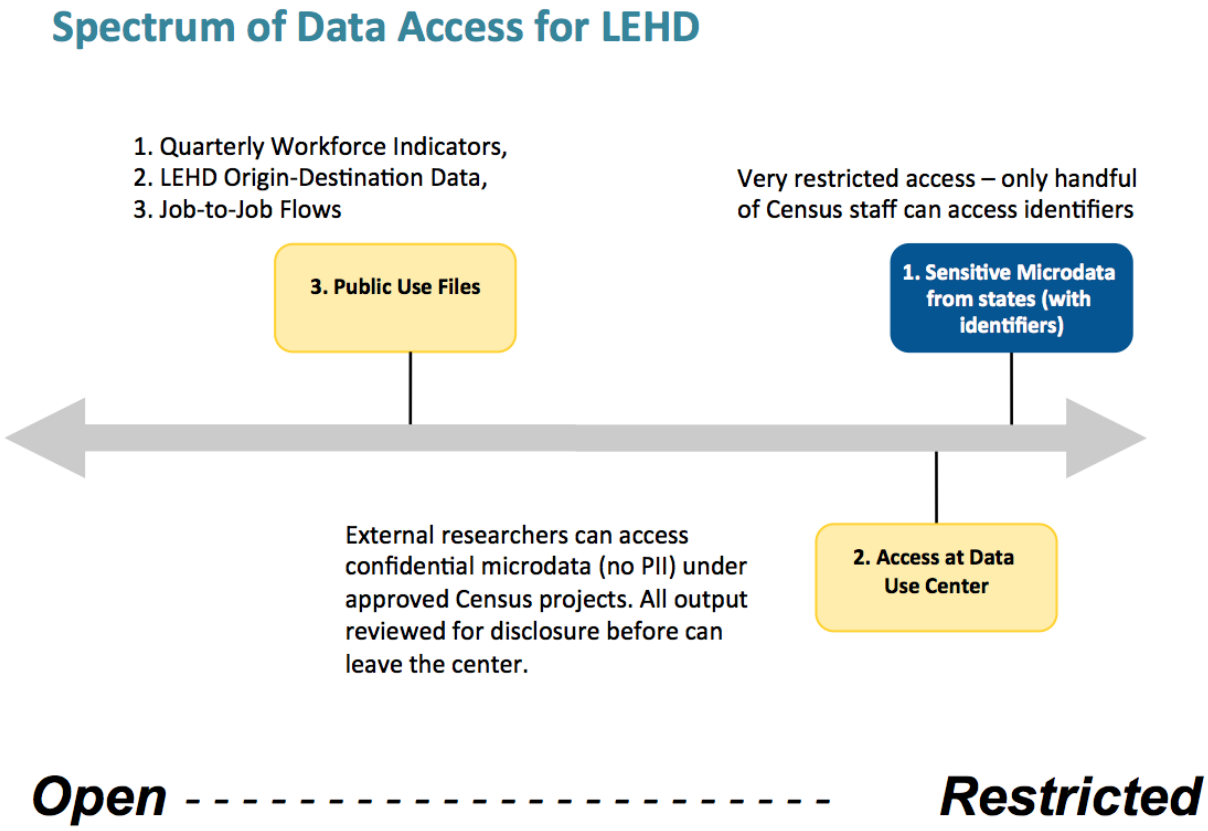
- Leverage existing survey data with administrative data allows production of new statistics at low cost with no additional respondent burden

Challenges:

- Cleaning and linking data across various sources
- Partnerships with other agencies can end (WY recently withdrew), threatening the viability of the statistical program

CENSUS: LEHD

Data Access



CENSUS: Innovation Measurement Initiative (IMI)

University admin data, Census ACS, Business register adrecs, web scraped info

Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients
Nikolas Zolas, Nathan Goldschlag, Ron Jarmin, Paula Stephan, Jason Owen- Smith, Rebecca F. Rosen, Barbara McFadden Allen, Bruce A. Weinberg and Julia I. Lane (December 10, 2015)
Science Translational Medicine **350** (6266), 1367-1371. [doi: 10.1126/science.aac5949]

Editor's Summary

Tracking the knowledge economy

Although the U.S investment in scientific research can be documented readily, its output is harder to track. Zolas *et al*. combined data obtained from eight universities on their doctorate recipients with data from business registries and the U.S. Census Bureau. This allowed them to link Ph.D. recipients to all their subsequent employers. Doctoral recipients tended to stay in academia or join large companies with high salaries. Roughly 20% stayed in the state in which they received their degree. In the year after receiving a Ph.D., mathematicians and computer scientists received the highest salaries, and biologists received the lowest.

Science, this issue p. 1367

IMI: Continued

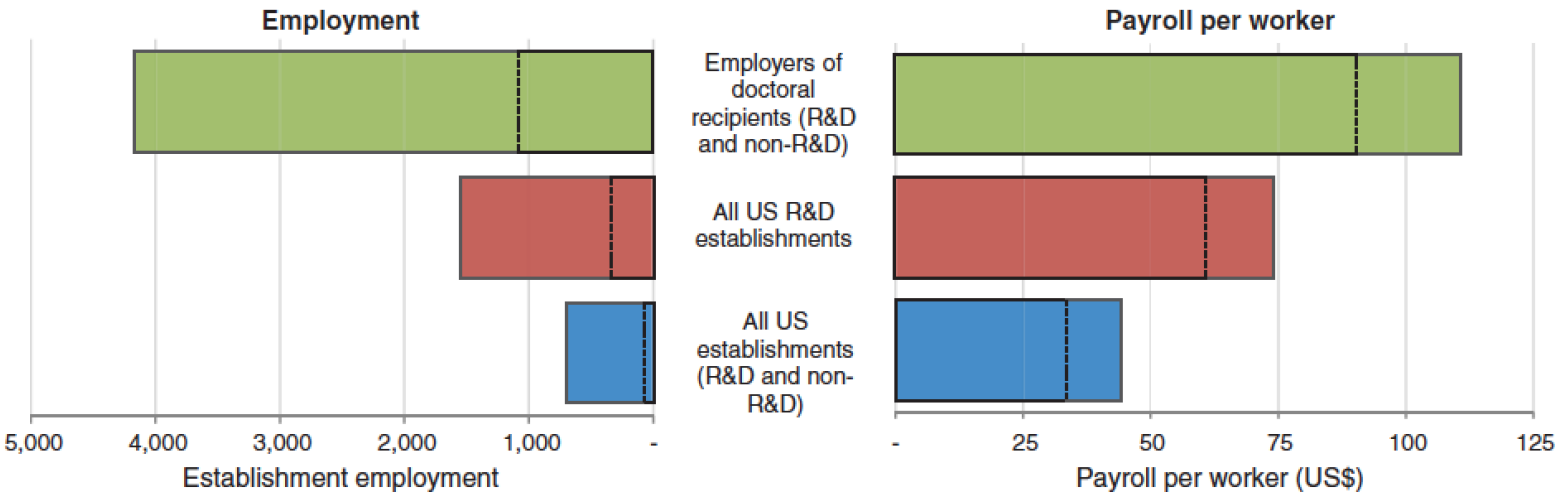
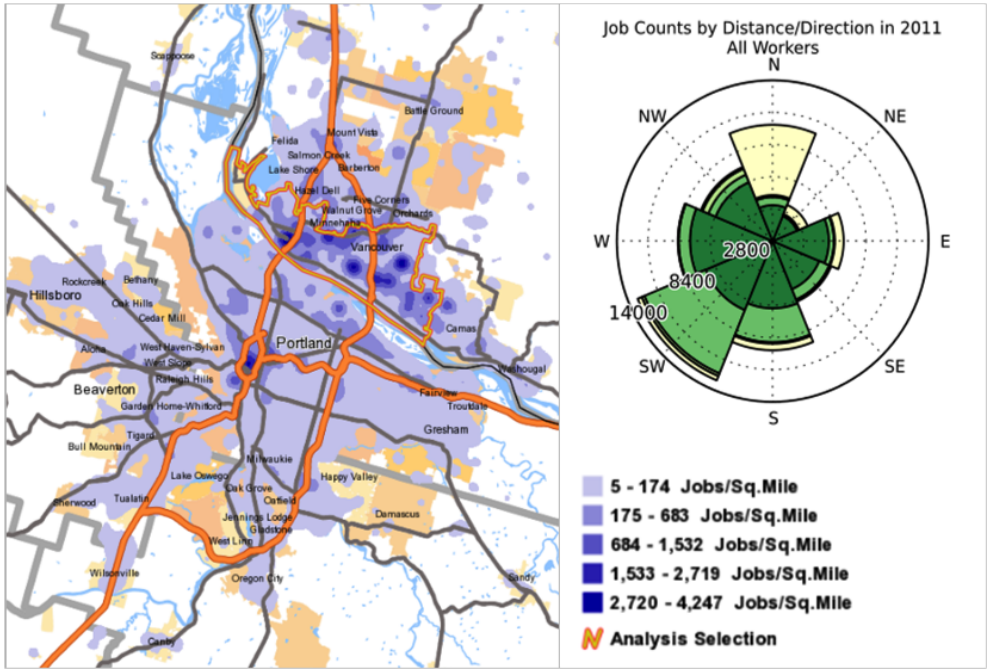


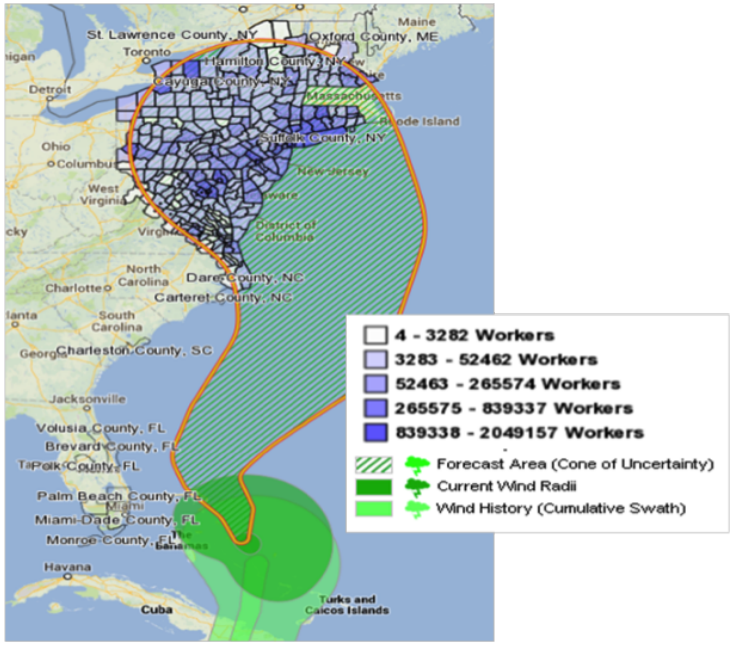
Fig. 1. UMETRICS doctoral recipients are placed at establishments that are larger and have higher payrolls per worker. Medians are dashed inner lines, and means are solid outer lines. The standard deviations in employment at establishments that employed UMETRICS doctoral recipients, at all U.S. establishments owned by R&D performing firms, and all U.S. establishments are 6407, 3661, and 2362, respectively; the standard deviations in annual payroll per worker are $120,199; $56,252; and $44,327, respectively; the differences in employment size and payroll per worker are statistically significant. Annual payroll per worker is the average payroll (the total payroll divided by the number of employees) across all employees at the three types of establishments—all U.S. establishments, all U.S. establishments owned by firms that perform R&D, and the establishments that employed UMETRICS doctoral recipients (regardless of whether they are owned by firms that perform R&D). National and R&D establishments are weighted by total establishment employment, whereas doctoral recipient establishments are weighted by the number of doctoral recipients employed. Values for annual payroll per worker are U.S.$1 ×1000.

Census: Partially Synthetic data allow users to select custom geographies in "OnTheMap"

Commuting Patterns, Portland OR



Hurricane Sandy



Combining Estimates from Related Surveys via Bivariate Models

(Application: using ACS estimates to improve estimates from smaller U.S. surveys)

William R. Bell and Carolina Franco, U.S. Census Bureau

2016 Ross-Royall Symposium

February 26, 2016

Application I: 2010 Disability Rates for U.S. States: SIPP borrowing from ACS

y_{1i} = SIPP disability estimate, y_{2i} = ACS disability estimate

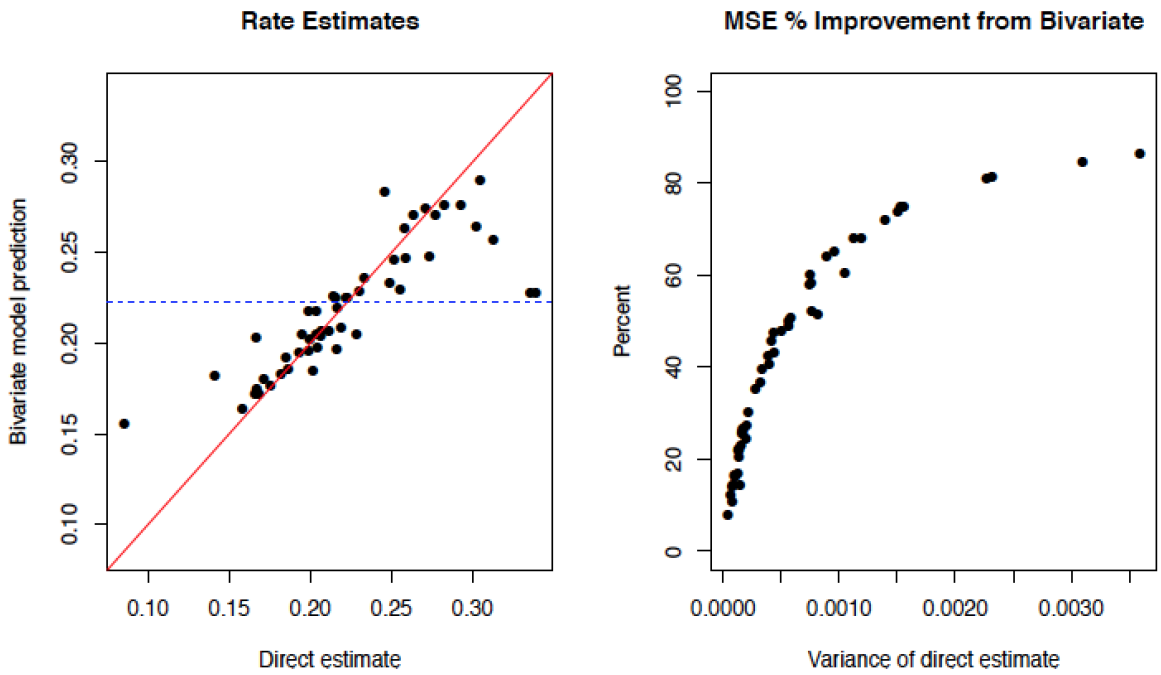
Smoothing of SIPP direct sampling variance estimates is applied.

$\hat{\rho}$= **.82**

- Univariate shrinkage yields an MSE decrease of $2\% - 67\%$ from direct, with a median of 19%
- The MSE decrease from bivariate vs. univariate model is $6\% - 59\%$ with a median of 29%
- The MSE decrease from bivariate vs. direct is $\mathbf{8 - 86\%}$, **with a median decrease of 43%**

Disability Rates for U.S. States, 2014

Bivariate model for SIPP and ACS estimates



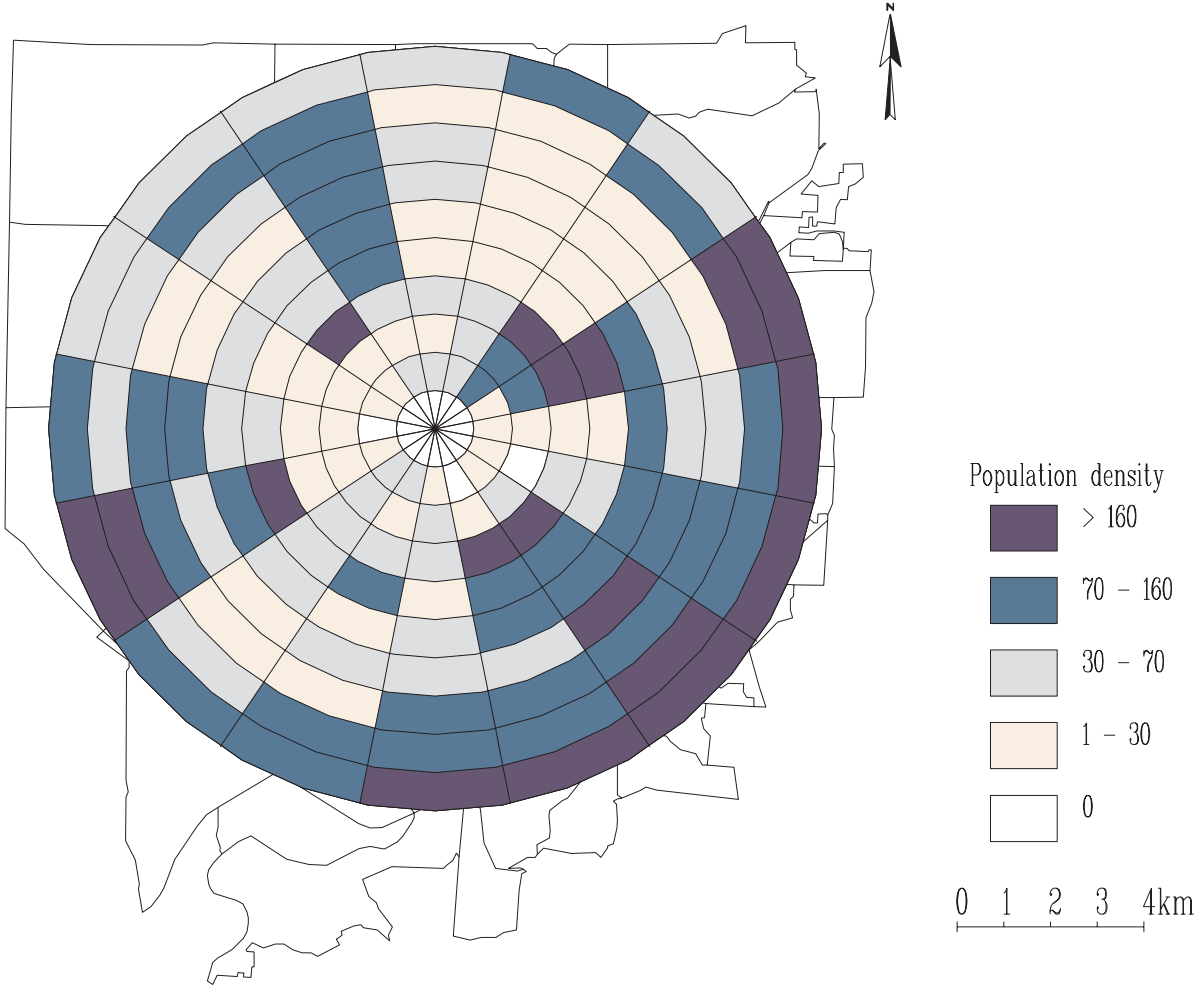
Integrating mis-aligned information

Exposure from the Fernald, OH superfund site[1]

- 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

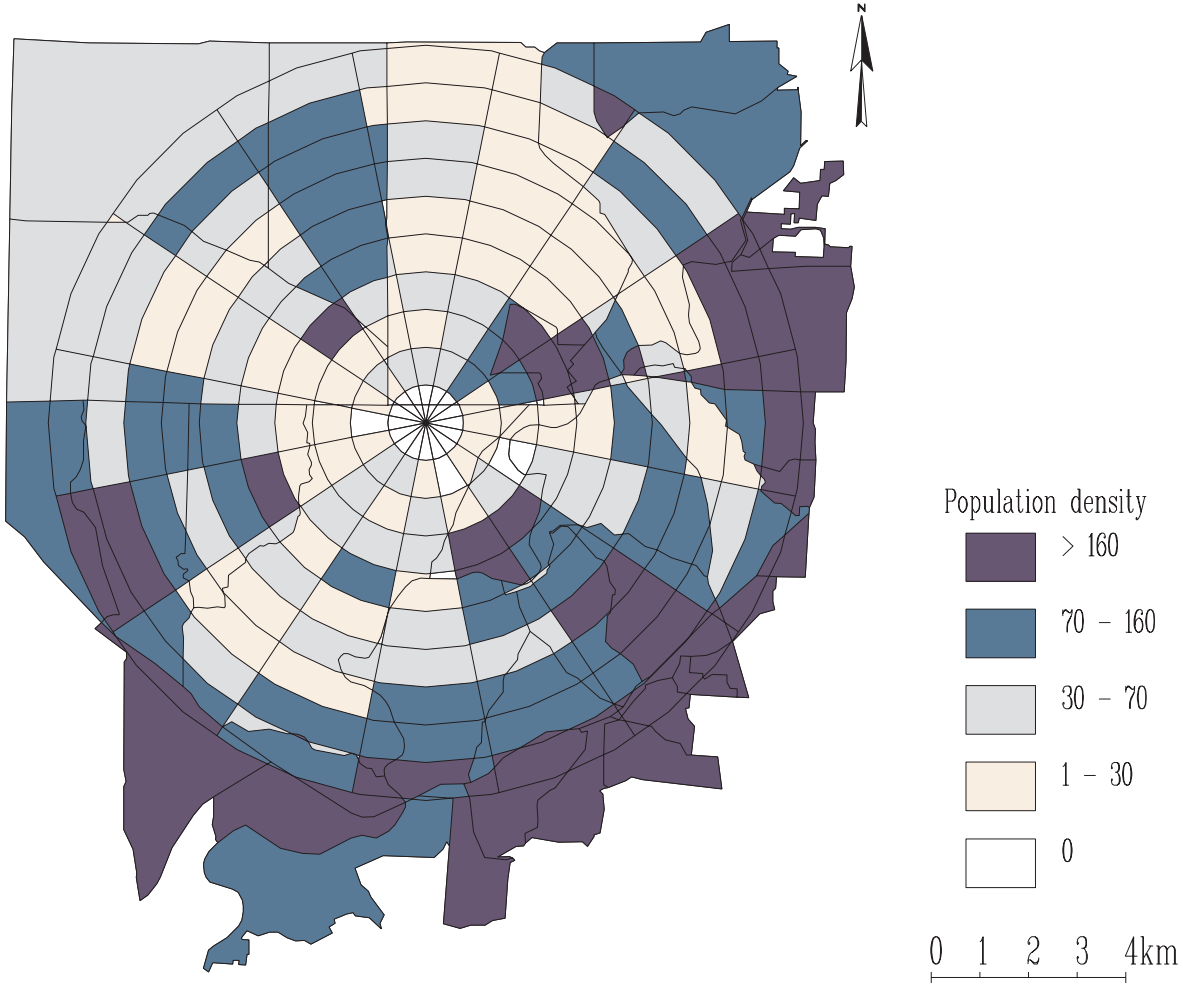
[1] Mugglin and Carlin (1998). Hierarchical modeling in Geographic Information Systems: population interpolation over incompatible zones. *J. of Agricultural, Biological, and Environmental Statistics*, 3: 111-130.

1. Population density & wind direction



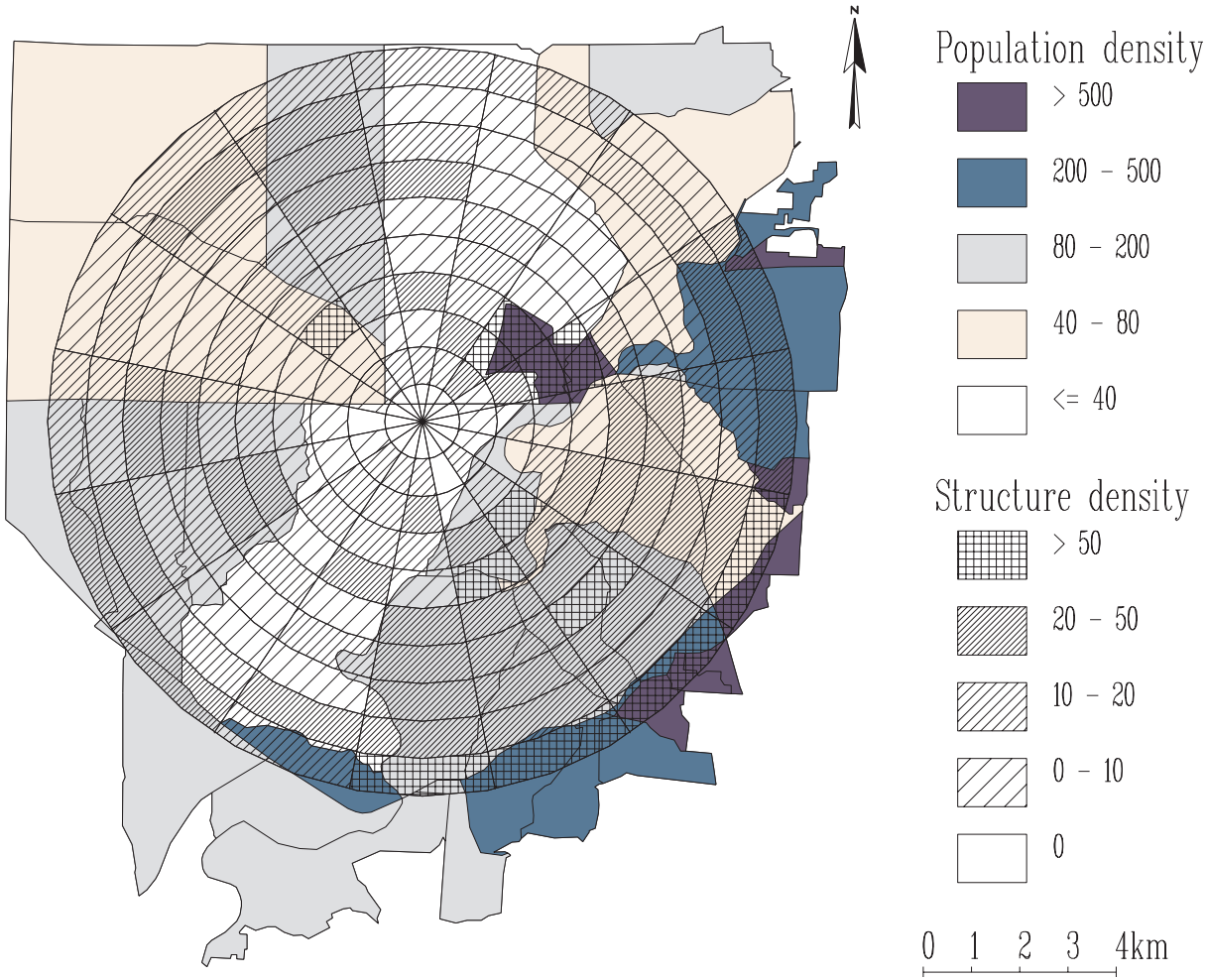
- Population density intersected with census units and wind direction centered around the exposure source

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 absolutely 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

Coda

- In the complex situations addressed by Lohr, Ragunathan, Census, and Mugglin, the Bayesian formalism is essential
- As are sensitivity analyses and ensuring reproducible research
- Model assessment is challenging and essentially a frequentist act

Coda

- In the complex situations addressed by Lohr, Ragunathan, Census, and Mugglin, the Bayesian formalism is essential
- As are sensitivity analyses and ensuring reproducible research
- Model assessment is challenging and essentially a frequentist act

  "Bayesians get the glory, but frequentists do the hard work"
  Brad Efron said that

- The quality of results depends on the expertise of the investigative team and the quality input information

Coda

- In the complex situations addressed by Lohr, Ragunathan, Census, and Mugglin, the Bayesian formalism is essential
- As are sensitivity analyses and ensuring reproducible research
- Model assessment is challenging and essentially a frequentist act

  "Bayesians get the glory, but frequentists do the hard work"
  Brad Efron said that

- The quality of results depends on the expertise of the investigative team and the quality input information

  "Space-age procedures will not rescue stone-age data"
  I said that

Coda

- In the complex situations addressed by Lohr, Ragunathan, Census, and Mugglin, the Bayesian formalism is essential
- As are sensitivity analyses and ensuring reproducible research
- Model assessment is challenging and essentially a frequentist act

  "Bayesians get the glory, but frequentists do the hard work"
  Brad Efron said that

- The quality of results depends on the expertise of the investigative team and the quality input information

  "Space-age procedures will not rescue stone-age data"
  I said that

#thankyou