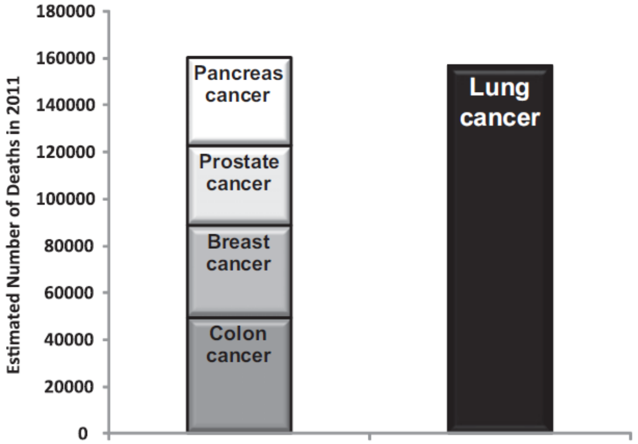
Estimating lung cancer mortality rates in U.S. counties using Bayesian spatial models

Melissa Jay, BA[1], Mary Charlton, PhD[2], and Jacob Oleson, PhD[1]

[1] Department of Biostatistics, University of Iowa ; [2] Department of Epidemiology, University of Iowa

Lung cancer in the United States

- Leading cause of cancer death
- 5-year survival rate (2007): 16%
- Incidence and mortality rates have been decreasing steadily for men but only recently started to decrease for women
- Known risk factors include:
  - Cigarette smoking
  - Secondhand smoke
  - Radon
  - Air pollution
  - Diesel exhaust
  - Family history



Source: Dela Cruz et al. (2011)

Why do we need rate estimates?

- For diseases with low death counts, there is a large amount of variability in raw mortality rates
- For example, say a county has a population of 800
  - Rate if there were 0 deaths: (0/800)*100,000 = 0
  - Rate if there was 1 death: (1/800)*100,000 = 125
  - Rate if there were 2 deaths: (2/800)*100,000 = 250
- The raw mortality rates are also spatially correlated (Moran's I = 0.150, p = 0.002)
- Using statistical models, we can produce more reliable lung cancer mortality rates by incorporating information from both the counties themselves and their neighbors
- Trustworthy estimates can help health departments allocate resources and promote prevention/ intervention efforts

Project goals

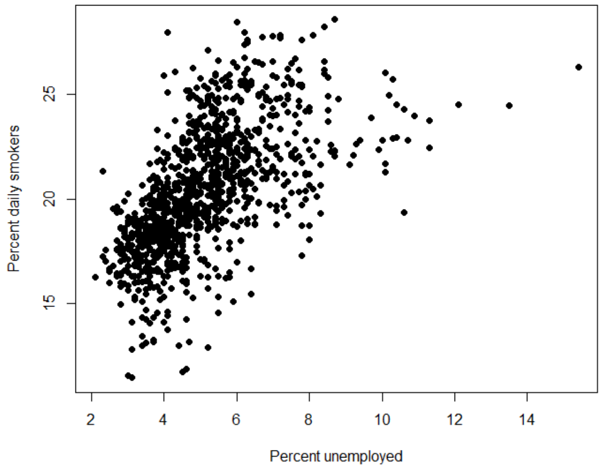
- Estimate cancer mortality rates for U.S. counties
- Determine which county-level variables explain cancer mortality rates
- Identify geographic regions with high and higher than expected cancer mortality rates
- In this presentation, we will focus on lung cancer mortality rates in the Midwest

Data collection

- Annual lung cancer death counts were derived from the National Center for Health Statistics Restricted-Use Vital Statistics data files and age-adjusted using the 2010 U.S. standard population
- County-level variables were collected from agencies including the EPA, CDC, USDA, and Institute for Health Metrics and Evaluation (IHME)
- Variables included in analysis:
  - Behavioral: prevalence of daily smokers, high alcohol consumption, and obesity
  - Environmental: PM2.5, proximity to active coal mine, radon zone, diesel emissions
  - Healthcare access: proximity to NCI cancer center, proportion uninsured
  - Geographic: recoded rural-urban continuum code
  - Demographic: proportions Hispanic, Black, Asian, and American Indian
  - Socioeconomic: median household income, percent unemployed, income inequality
- We will focus on all Midwest counties (n = 1,055) for two years (2007 and 2017)
- Many of the study variables were selected from O'Connor et al. (2018)

Correlated variables

- Many of the variables in the dataset are highly correlated
  - Examples:
    1. Smoking and percent unemployed (r = 0.594)
    2. Smoking and obesity (r = 0.586)
- By including each variable individually, it may be difficult to interpret how a variable explains lung cancer mortality rates
  - We know that smoking is related to lung cancer mortality, but if obesity came out as a significant variable would the pattern in mortality rates be related to obesity? Or is obesity simply further explaining a county's smoking habits?
- We can use a factor analysis to classify some of these highly correlated variables into a smaller set of latent variables, or factors

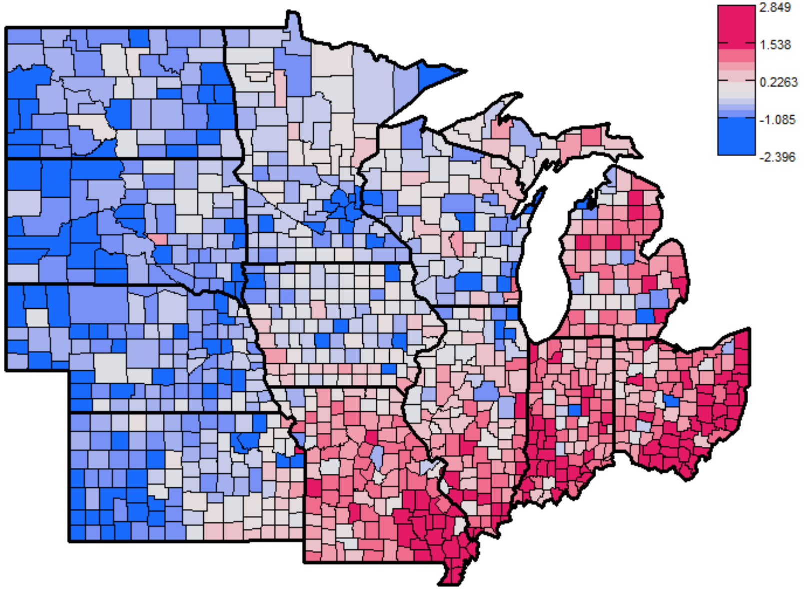


Factor variables - 2007

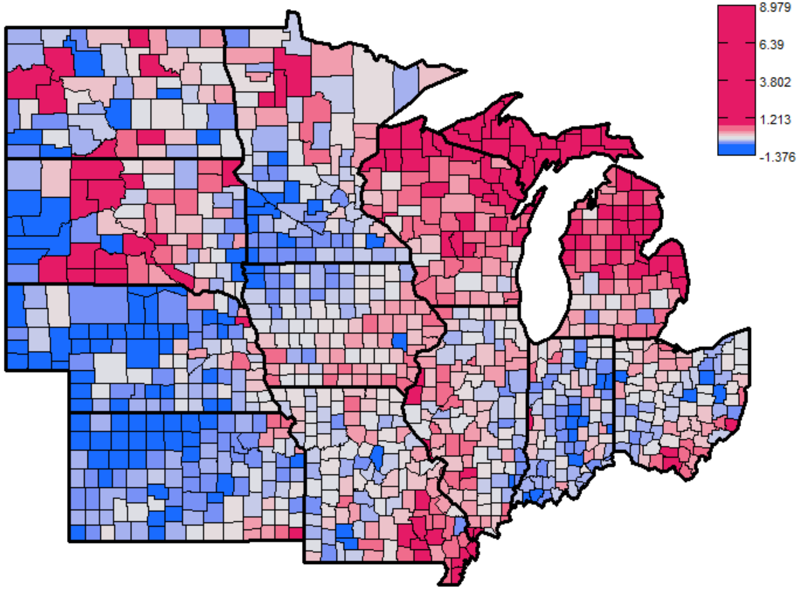
Factor	Variables Included	Loadings
1	Smoking	0.82
	Unemployment	0.62
	PM2.5	0.58
	Obesity	0.48
	Coal mining	0.37
2	Household income	0.95
	Urban	0.54
	Uninsured	-0.50
	Cancer center	0.43
	PM2.5	0.39
	Income inequality	-0.32
3	Proportion American Indian	0.79
	Heavy alcohol consumption	0.76
	Obesity	0.61
	Unemployment	0.49

Factor	Variables Included	Loadings
4	Proportion Black	0.69
	Proportion Asian	0.68
	Diesel exhaust	0.64
	Income inequality	0.42
	Cancer center	0.36
	Urban	0.34
5	Suburban	0.86
	Urban	-0.45
6	Obesity	0.56
	Proportion Hispanic	0.32

Selected 2007 factors



Factor 1: smoking, unemployment, PM2.5, obesity, coal mining



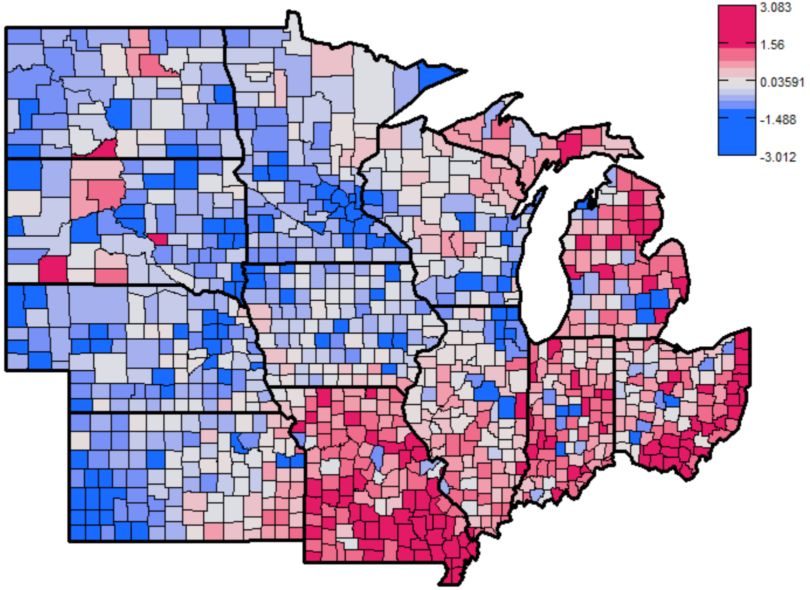
Factor 3: proportion American Indian, heavy alcohol consumption, obesity, unemployment

Factor variables - 2017

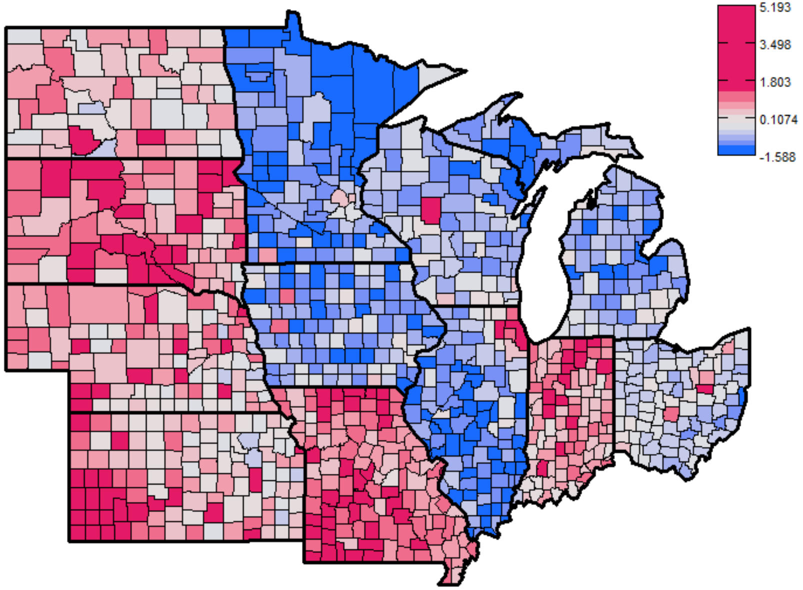
Factor	Variables Included	Loadings
1	Smoking Obesity Household income Unemployment Radon zone Proportion Asian	0.90 0.69 -0.64 0.63 0.37 -0.36
2	Diesel exhaust Proportion Black Urban Proportion Asian Cancer center Household income	0.91 0.63 0.55 0.54 0.52 0.34
3	Proportion American Indian Heavy alcohol consumption	0.98 0.76

Factor	Variables Included	Loadings
4	Proportion uninsured Proportion Hispanic	0.89 0.39
5	Urban Suburban	0.73 -0.61
6	Income inequality Household income Proportion Black	0.54 -0.43 0.41

Selected 2017 factors



Factor 1: smoking, obesity, -household income, unemployment, radon zone, -proportion Asian



Factor4: uninsured, proportion Hispanic

Statistical models

Bayesian hierarchical Poisson regression model

$$Y_i \sim Poisson(\theta_i)$$

$$\log(\theta_i) = \log(n_i) + \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta} + \gamma_i + \epsilon_i$$

$$R_i = \left(\frac{\theta_i}{n_i}\right) * 100000$$

Statistical models

Bayesian hierarchical Poisson regression model

$$Y_i \sim Poisson(\theta_i)$$

Models county i's age-adjusted lung cancer death count with a Poisson distribution with expected value θ_i

Statistical models

Bayesian hierarchical Poisson regression model

$$\log(\theta_i) = \underbrace{\log(n_i)}_{\text{Log of population size}} + \underbrace{\boldsymbol{x}_i'\boldsymbol{\beta}}_{\text{Factor variables multiplied by their coefficients}} + \underbrace{\gamma_i}_{\text{Spatial random effect that accounts for correlation using a conditional autoregressive model}} + \underbrace{\epsilon_i}_{\text{Additional error term}}$$

Statistical models

Bayesian hierarchical Poisson regression model

$$R_i = \left(\frac{\theta_i}{n_i}\right)*100000$$

Age-adjusted rate is calculated by dividing θ_i by the county population size and multiplying by 100,000 people

Used vague normal priors on the regression coefficients and vague inverse-gamma priors on the variance parameters

Factor effects - 2007

Factor	Variables Included	Estimate	95 % Credible Interval
1	Smoking, unemployment, PM2.5, obesity, coal mining	1.142	(1.110, 1.172)*
3	Proportion American Indian, heavy alcohol consumption, obesity, unemployment	1.083	(1.053, 1.113)*
6	Obesity, proportion Hispanic	0.978	(0.960, 0.995)*
2	Household income, urban, -uninsured, cancer center, PM2.5, -income inequality	1.012	(0.993, 1.032)
4	Proportion Black, proportion Asian, diesel exhaust, income inequality, cancer center, urban	1.010	(0.999, 1.021)
5	Suburban, -urban	0.998	(0.980, 1.017)

Multiplicative effects of a one-standard deviation increase in factor variable on lung cancer mortality rates

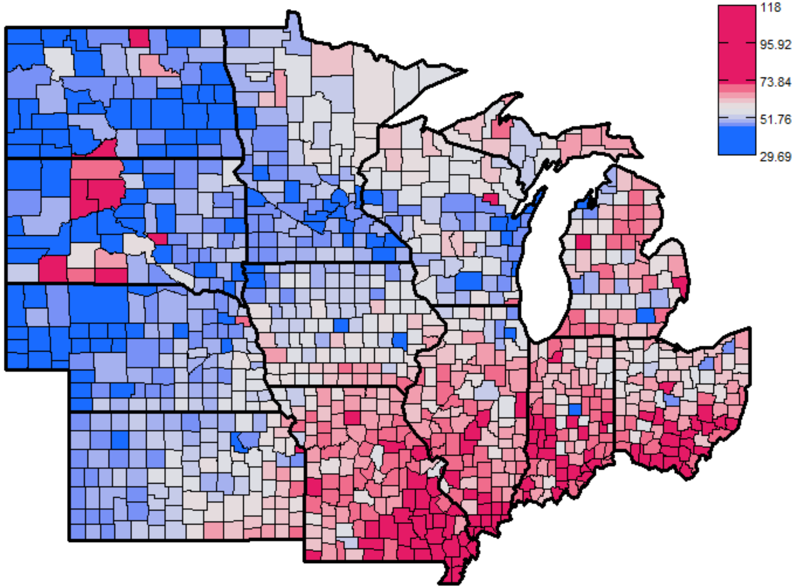
Factor effects - 2017

Factor	Variables Included	Estimate	95 % Credible Interval
1	Smoking, obesity, -household income, unemployment, radon zone, -proportion Asian	1.165	(1.146, 1.183)*
4	Proportion uninsured, proportion Hispanic	0.977	(0.957, 0.997)*
6	Income inequality, -household income, proportion Black	0.985	(0.970, 1.000)*
5	Urban, -suburban	0.987	(0.971, 1.003)
3	Proportion American Indian, heavy alcohol consumption	1.008	(0.975, 1.040)
2	Diesel exhaust, proportion Black, urban, proportion Asian, cancer center, household income	1.006	(0.991, 1.022)

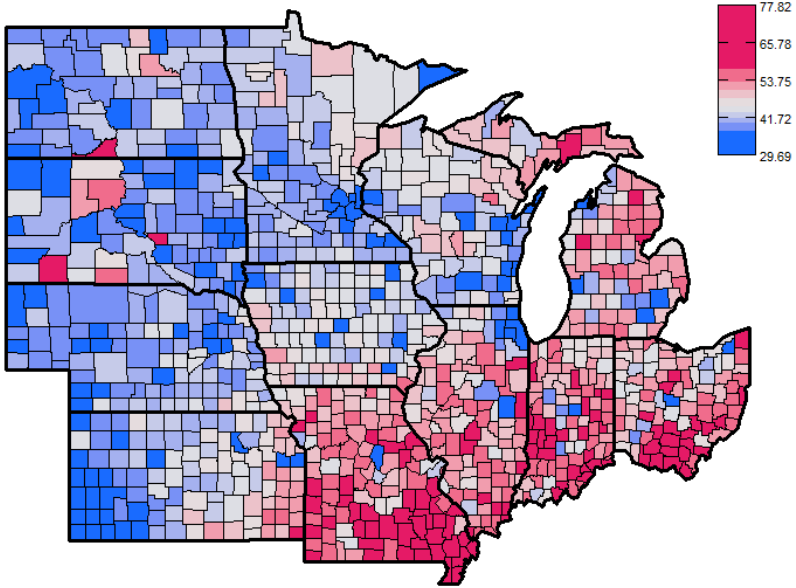
Multiplicative effects of a one-standard deviation increase in factor variable on lung cancer mortality rates

Estimated lung cancer mortality rates

2007



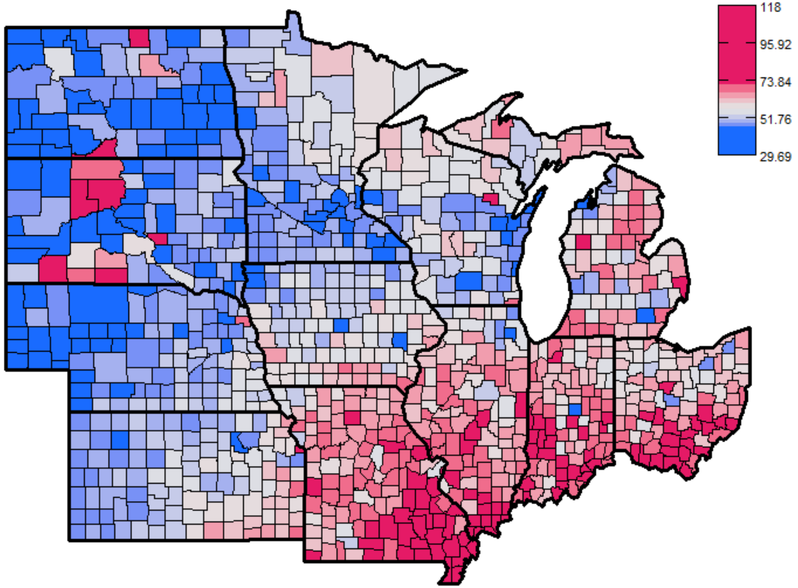
2017



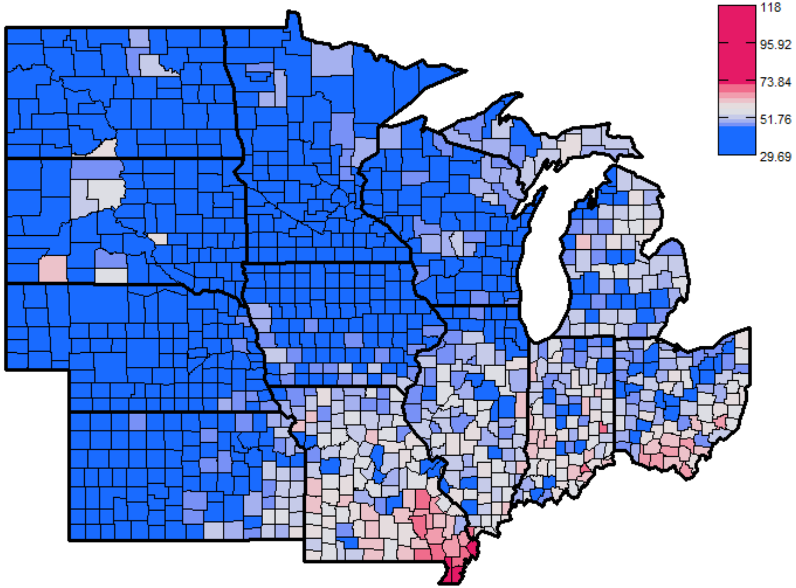
(Different scales)

Estimated lung cancer mortality rates

2007



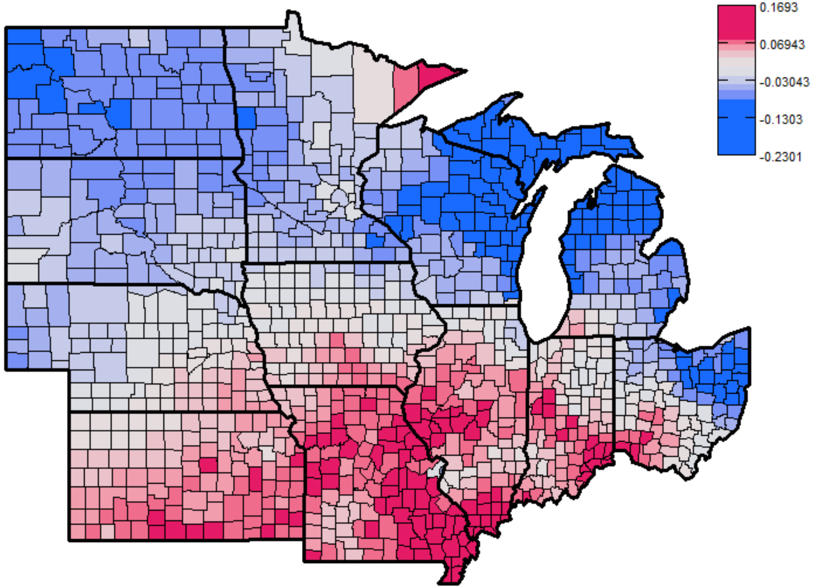
2017



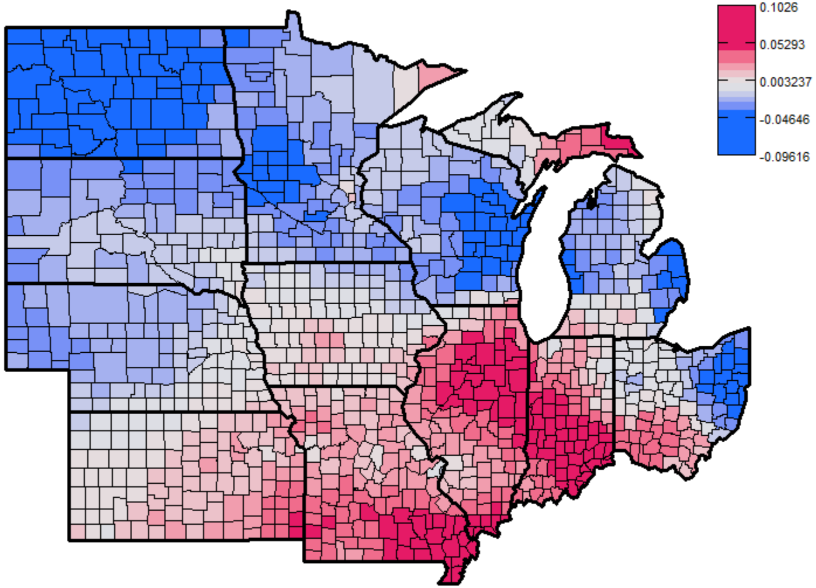
(Same scales)

Unexplained spatial variation

2007



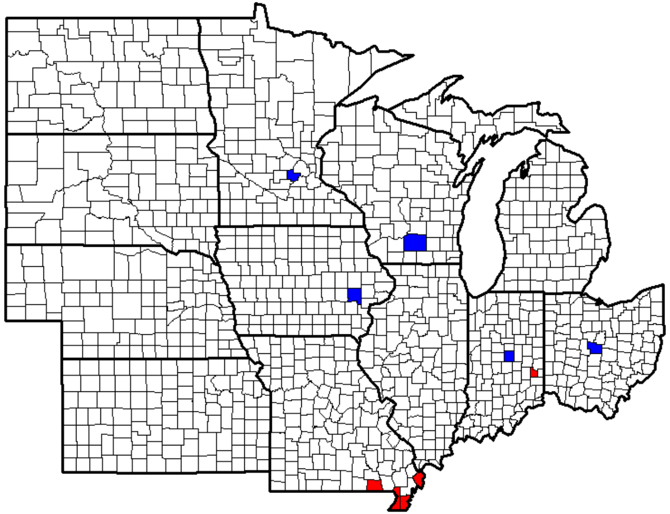
2017



Highest and lowest mortality rates - 2017

County	State	Rate
Dunklin	MO	77.82
Pemiscot	MO	74.55
Mississippi	MO	72.18
Fayette	IN	70.66
Ripley	MO	69.63

County	State	Rate
Hamilton	IN	29.69
Carver	MN	29.71
Dane	WI	30.03
Johnson	IA	30.37
Delaware	OH	30.52



Conclusions and limitations

- Conclusions
  - The combination of county-level smoking, obesity, and unemployment explained some of the patterns in mortality rates across both years
  - Patterns in mortality rates were consistent from 2007 to 2017, but rates have decreased over time
  - In both 2007 and 2017, southern Missouri, Illinois, Indiana, and Ohio, and select counties in South Dakota had the highest estimated lung cancer mortality rates
  - There still exists unexplained spatial variation in these models
- Limitations
  - Temporal correlation was not captured in this analysis
  - Age-adjustment occurred before the modeling process, rather than in the model to reduce the model run time

Future work

- Use all collected mortality data (1982-2017) to estimate mortality rates
  - Including each year's data into a single model will allow us to account for the correlation between mortality rates across years
- Create mortality rate estimates for eight cancer types
- Estimate cancer incidence rates in SEER-registry states using the same modeling framework

Acknowledgements

- Environmental Health Sciences Research Center Pilot Grant N. NIEHS/NIH P30 ES005605
- National Science Foundation Graduate Research Fellowship Program Grant No. 000390183
  - Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

environmental health sciences
research center

GRFP
NSF GRADUATE RESEARCH FELLOWSHIP PROGRAM

Data sources

- Bureau of Labor Statistics (BLS)
- Centers for Disease Control (CDC)
- CDC Wide-Ranging Online Data for Epidemiologic Research (WONDER)
- Environmental Protection Agency (EPA)
- Environmental Public Health Tracking Network
- EPA National Air Toxics Assessment (NATA)
- Institute for Health Metrics and Evaluation (IHME)
- National Center for Health Statistics (NCHS)
- National Cancer Institute (NCI)
- U.S. Census Bureau
- U.S. Department of Agriculture (USDA)
- U.S. Energy Information Administration (EIA)

References

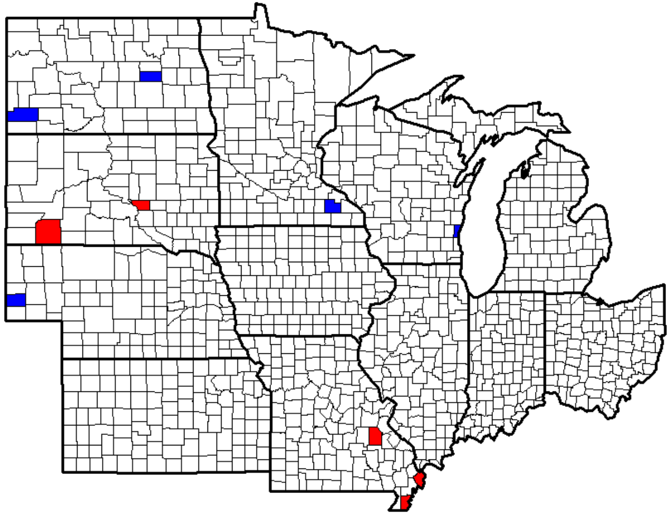
1. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. *Clin Chest Med*. 2011;32(4):605-644. doi:10.1016/j.ccm.2011.09.001.
2. Field RW, Steck DJ, Smith BJ, et al. Residential radon gas exposure and lung cancer: The Iowa radon lung cancer study. *Am J Epidemiol*. 2000;151(11):1091-1102. doi:10.1016/j.ccm.2011.09.001. 10.1093/oxfordjournals.aje.a010153.
3. Mokdad AH, Dwyer-Lindgren L, Fitzmaurice C, et al. Trends and patterns of disparities in cancer mortality among US counties, 1980-2014. *JAMA.* 2017;317(4):388-406. doi:10.1001/jama.2016.20324.
4. National Center for Health Statistics. *Detailed Mortality – All Counties 1989-2017*, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.
5. O'Conner JF, Sedghi T, Dhopapkar M, Kange MJ, Gross CP. Factors Associated with cancer disparities among low-, medium-, and high-income US counties. *JAMA Netw Open*. 2018;1(6):e183146. doi:10.1001/jamanetworkopen.2018.3146.
6. Puskin JS. Smoking as a confounder in ecologic correlations of cancer mortality rates with average county radon levels. *Health Phys.* 2003;84(4):526-532.

Thank you for listening! Questions?

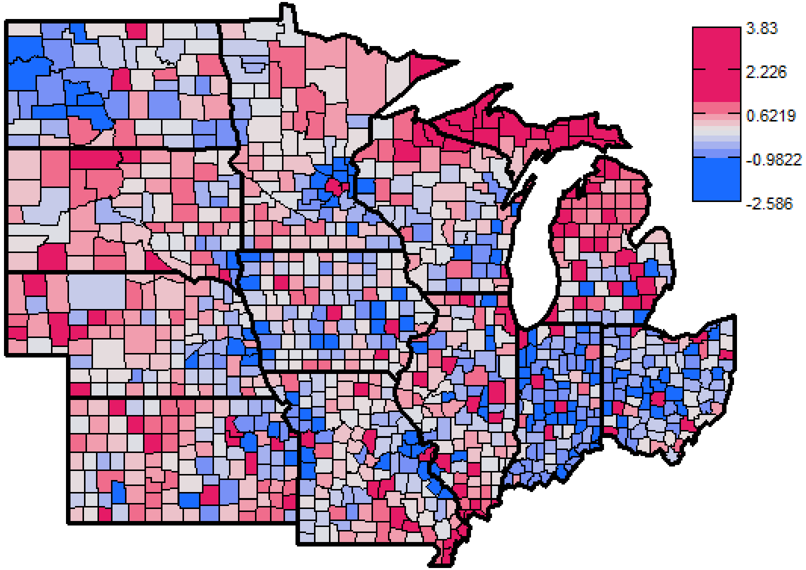
Highest and lowest mortality rates - 2007

County	State	Rate
Buffalo	SD	118.00
Pemiscot	MO	101.43
Washington	MO	101.43
Oglala Lakota	SD	99.14
Mississippi	MO	95.13

County	State	Rate
Olmsted	MN	39.61
Ozaukee	WI	40.01
Banner	NE	40.15
Slope	ND	40.43
Foster	ND	41.28



Factor 6 - 2017



Income inequality, -household income, proportion Black

Creating factor variables

- Factor analysis
  - Accounts for correlation between explanatory variables
  - Assumes there are latent variables that can be described by the variables we have collected
  - For example, several of the variables combined (presence of a cancer center, diesel emissions, air pollution, etc.) could more generally be describing an urban environment
- Conducted two separate factor analyses: one for 2007 and one for 2017
- Used six factors as variables in our regression models
  - Factors included explained 56% of the variability in the full dataset