Small area estimation of cancer rates: A case study of lung cancer in Florida, 2000-2010

Lan Hu, Yongwan Chun, Daniel A. Griffith

Geospatial Information Sciences program
University of Texas at Dallas
Email: lan.hu@utdallas.edu

Introduction

- Small area estimation (SAE) furnishes a cost-effective method for obtaining reliable small area data
  - SAE has been widely applied in many research areas (e.g., economics, demography, epidemiology)
  - One major problem of SAE is how to generate reliable estimates of interest for small areas with data only available at a coarser geographic resolution
- Lung cancer, the leading cause of death for both man and woman, has received substantial research attention
  - The literature notes that spatial autocorrelation frequently is detected in lung cancer incidence rate
  - Spatial SAE models have been proposed to address spatial autocorrelation [e.g., Spatial Empirical Best Linear Unbiased Predictor (Spatial EBLUP)]

Research Objectives

- Estimate lung cancer incidence rates at the census tract resolution in Florida
  - A synthetic method, a Poisson regression method, and a Poisson eigenvector spatial filtering method
  - Estimates are compared with actual observed cancer rates
- Compare the estimated results of these methods
- Examine whether or not accounting for spatial autocorrelation can lead to a better estimation

Literature Review

- Small areas can be defined as geographic areas (e.g., census units such tracts, communities), or socio-demographic groups

- SAE is an important endeavor in global health, epidemiology
  - Reveal disparities of disease at local small areas (e.g., Krieger et al., 2002; Mobely et al., 2012)
  - Provide a fundamental basis for understanding the complex interaction between human and environmental systems (e.g., Langford et al., 2008)
  - Allow investigation into the factors responsible for geographic disparities in health (Ruther et al., 2017)
  - Identify priority areas for action, and optimize the use of limited resources (Zhang et al., 2013)

Literature Review (cont'd)

- SAE indirect estimator
  - Synthetic estimator
    - Compute a designed unbiased estimator across all the areas, and then apply it for every small area (Pfeffermann 2013)
    - Assume homogeneity that may yield a large bias (e.g., Jia et al., 2004)
  - Regression model (e.g., logistic or Poisson regression)
    - Incorporate areal level covariates to consider disparities at different geographic resolutions
    - Further extended to a mixed model, including random effects to improve model fits (e.g., Li et al., 2009; Zhang et al., 2014)
  - SAE spatial model (e.g., spatial EBLUP)
    - Consider to incorporate spatial autocorrelation in an estimation model specification (e.g., Pratesi and Salvati 2008)

Our Dataset

- Lung cancer incident points in the State of Florida
  - Florida Cancer Registry
  - An 11-year period from 2000 to 2010
  - A total of 172,460 cases after data cleaning
  - Age, sex, race/ethnicity

Variables	Categories
Age	< 20, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, >85
Sex	Male, Female
Race/ethnicity	White, Black, Asian, Hispanic, Other
County	67 counties

Methodology

- Eigenvector Spatial Filtering (Griffith 2003)
  - Utilizes eigenvectors from a transformed spatial weights matrix

$$(\mathbf{I} - \mathbf{1}\mathbf{1}^T/\mathrm{n})\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/\mathrm{n})$$

  - n: the number of spatial units; **C**: a spatial weights matrix
  - **I**: an identity matrix; **1**: a vector of ones; T: matrix transpose operator

- An ESF model specification for a Poisson random variable (Griffith 2002)

$$E(\mathbf{Y}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\gamma})$$

  - $g(\cdot)$ link function; $E(\cdot)$ expectation operator
  - **Y**: response variable; **X**: covariates; **E**: a set of eigenvectors
  - $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$: parameters to be estimated
  - A stepwise approach to select eigenvectors from a candidate set based on a criterion of maximizing model fit

Methodology (cont'd)

- **Synthetic method**
  - Construct demographic strata by age/sex/race (150 strata)
  - Compute cancer rate for each strata

$$strata.rate_i = \frac{cancer\ count\ for\ each\ strata\ i}{Populaiton\ for\ each\ strata\ i}$$

  - Compute cancer count for each census tract/block group

$$Count.bg_j = \sum_i pop_{ij} * strata.rate_i$$

  - Adjust cancer count for each spatial unit (ensure the total count for each county equal to the observed count)

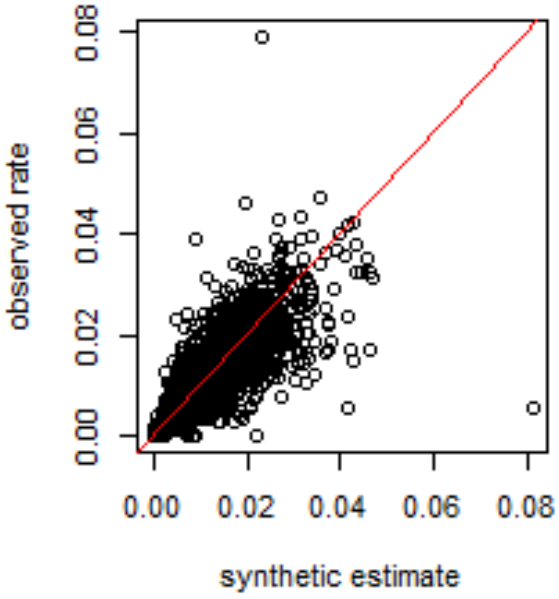
Methodology (cont'd)

- **Poisson and Poisson ESF methods**
  - Construct demographic-geographic strata age/sex/race/county (10,050 strata)
  - Model cancer counts for the strata with Poisson and Poisson ESF models (population per stratum as an offset variable)
  - Calculate a cancer rate for each stratum
  - Compute a cancer count for each census tract/block group
  - Adjust a cancer count for each spatial unit (ensure the total count for each county equal to the observed count)

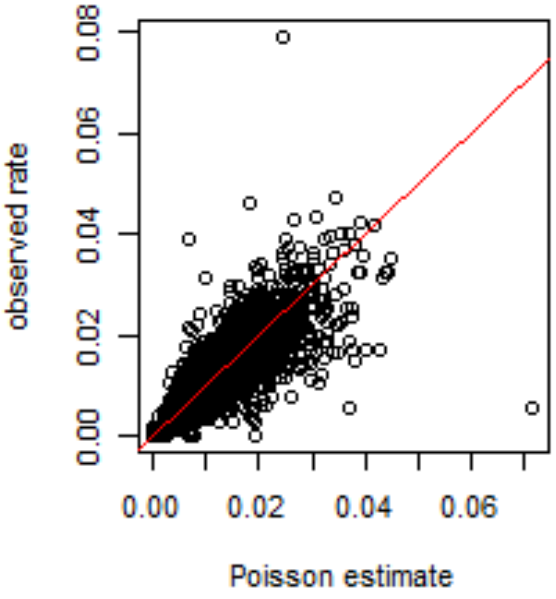
	Poisson	**Poisson ESF**
specifications	$E[\ln(\text{count})] = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 Race + \beta_4 \text{Poverty}$	$E[\ln(\text{count})] = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 Race + \beta_4 \text{Poverty} + \beta_5 ESF$

	Poisson		Poisson ESF	
(Intercept)	-13.123	0.310***	-13.073	0.308***
raceblack	0.791	0.060***	0.775	0.059***
racehispanic	0.355	0.060***	0.348	0.059***
raceother	-1.160	0.098***	-1.186	0.098***
racewhite	1.213	0.058***	1.201	0.057***
agea2024	1.271	0.430***	1.300	0.427**
agea2529	2.282	0.357***	2.290	0.355***
agea3034	2.956	0.334***	2.936	0.332***
agea3540	4.253	0.312***	4.194	0.310***
agea4045	5.353	0.306***	5.268	0.304***
agea4550	6.168	0.305***	6.052	0.303***
agea5055	6.742	0.304***	6.641	0.303***
agea5560	7.289	0.304***	7.236	0.302***
agea6065	7.669	0.304***	7.621	0.302***
agea6570	8.126	0.304***	8.083	0.302***
agea7075	8.451	0.304***	8.421	0.302***
agea7580	8.650	0.304***	8.583	0.302***
agea8085	8.575	0.304***	8.467	0.302***
agea85	8.286	0.304***	8.242	0.302***
sexmale	0.340	0.008***	0.362	0.011***
poverty	1.280	0.154***	1.280	0.153***
AIC	38,014		37,681	
Moran’s I (areal effects)			-0.04 (-0.61)	
selected eigenvector			5/48	

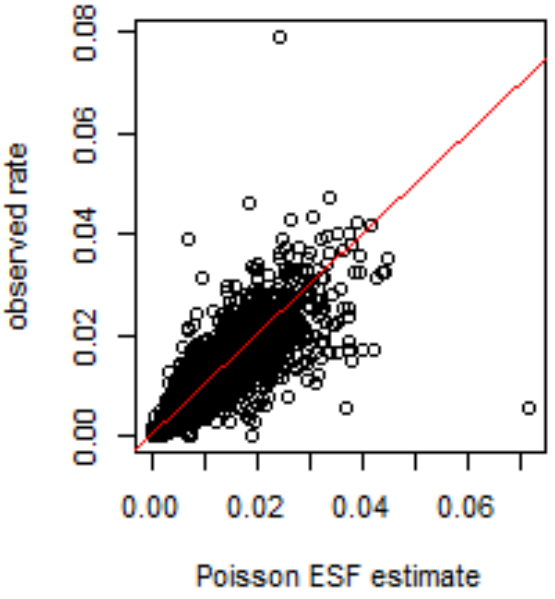
Results



Cor : 0.78

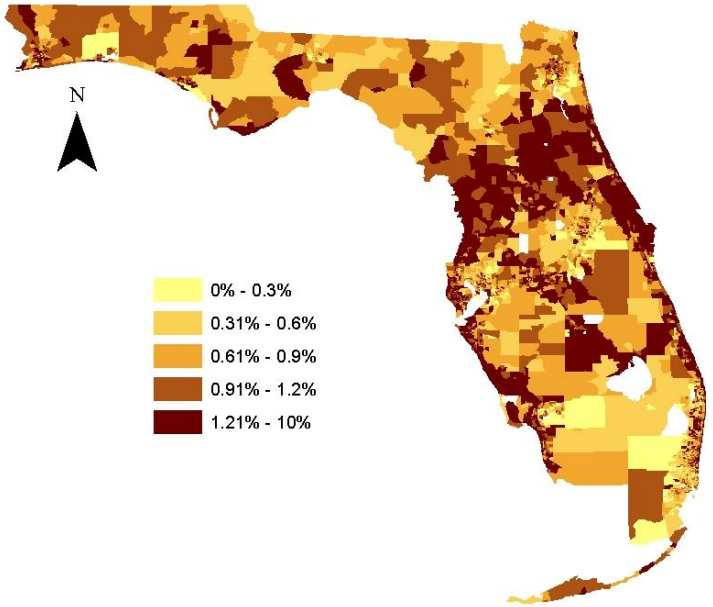


Cor : 0.81

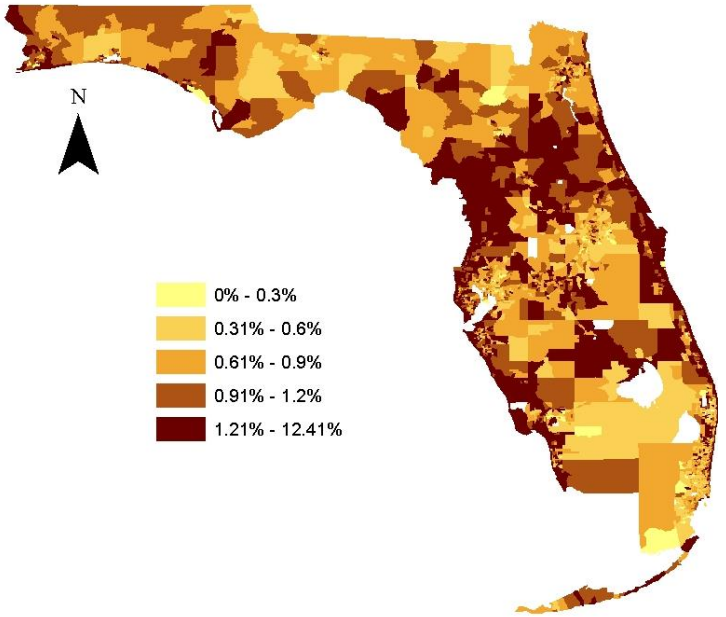


Cor : 0.81

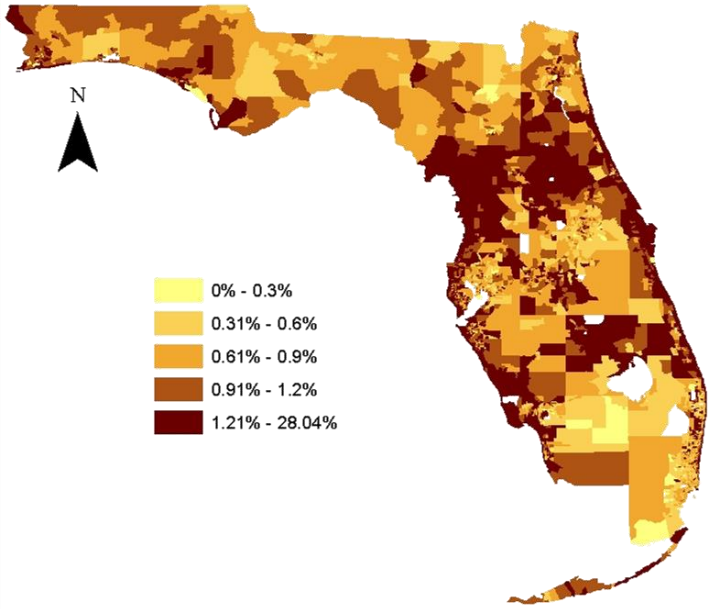
Observed rates



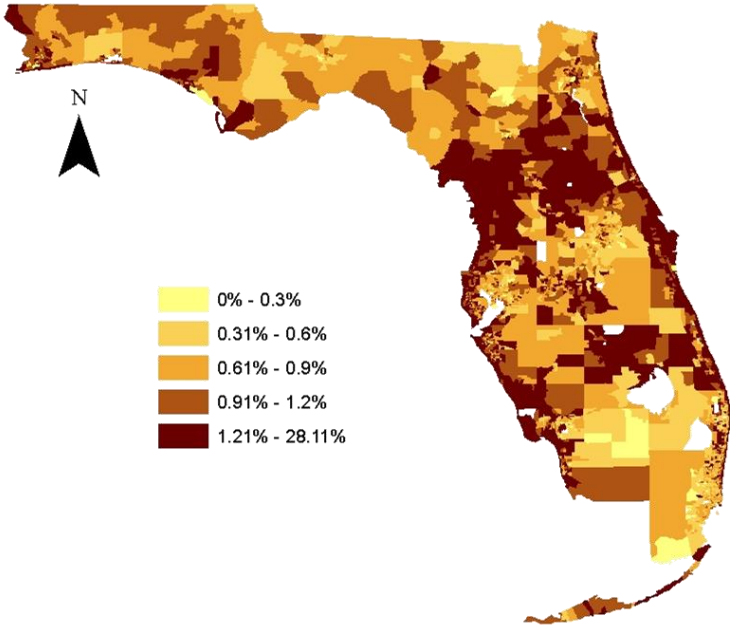
Synthetic estimates



Poisson estimates



Poisson ESF estimates



Summary and Discussions

- The estimation results are reasonably good
  - The correlations between estimates and observed cancer rates are high
  - age, race and sex are important factors to describe lung cancer rate
- Poisson models generate better estimates than synthetic method
  - The synthetic method assumes homogeneity while Poisson models consider local variations
  - However, the spatial model improves estimation marginally
- Estimates generated with the three methods capture the major spatial pattern of lung cancer rate (e.g., high/low cancer rate areas)
  - Over/under-estimations exist on all three maps
  - Estimations need to be further improved to reveal local disparities

Future work

- A better spatial model specification is desired
  - The current method that accounts for spatial autocorrelation only marginally improved estimates
- Similar estimates will be extended for the census block group resolution
  - to evaluate if these SAE methods produce comparable results at different geographic resolutions
  - to examine how good estimates are for finer level of spatial units