Hierarchical Bayesian Modeling of Spatio-temporal Patterns of Lung Cancer Incidence Risk in Georgia, USA: 2000–2007

By: Ping Yin, Lan Mu, Marguerite Madden, John E. Vena

Yin, P., Mu, L. Madden, M., Vena, J.E. (2014). Hierarchical Bayesian Modeling of the Spatio-Temporal Patterns of Lung Cancer Incidence Risk in Georgia, U.S., 2000-2007. *Journal of Geographical Systems.* doi: 10.1007/s10109-014-0200-4

Abstract:

Lung cancer is the second most commonly diagnosed cancer in both men and women in Georgia, USA. However, the spatio-temporal patterns of lung cancer risk in Georgia have not been fully studied. Hierarchical Bayesian models are used here to explore the spatio-temporal patterns of lung cancer incidence risk by race and gender in Georgia for the period of 2000–2007. With the census tract level as the spatial scale and the 2-year period aggregation as the temporal scale, we compare a total of seven Bayesian spatio-temporal models including two under a separate modeling framework and five under a joint modeling framework. One joint model outperforms others based on the deviance information criterion. Results show that the northwest region of Georgia has consistently high lung cancer incidence risk for all population groups during the study period. In addition, there are inverse relationships between the socioeconomic status and the lung cancer incidence risk among all Georgian population groups, and the relationships in males are stronger than those in females. By mapping more reliable variations in lung cancer incidence risk at a relatively fine spatio-temporal scale for different Georgian population groups, our study aims to better support healthcare performance assessment, etiological hypothesis generation, and health policy making.

Keywords: Hierarchical Bayesian model | Spatio-temporal pattern | Lung cancer | Socioeconomic status | Georgia | GIS

Article:

1 Introduction

In Georgia, USA, during 2001–2005, the age-adjusted lung cancer incidence rate was 53 per 100,000 in females and 104 per 100,000 in males, the second highest cancer incidence rates after breast cancer in females and prostate cancer in males (Georgia Department of Public

Health *2008*). To the best of our knowledge, however, the spatio-temporal patterns of lung cancer risk in Georgia have not been reported, and most of the related research focuses on descriptive analyses at a coarse spatial scale (e.g., health district or county level) or temporal scale (e.g., 5-year period aggregation). Such results usually obscure detailed variations in lung cancer risk in space and time, leading to limited ability to formulate etiological hypotheses or limited support for accurate healthcare performance assessments and efficient health interventions.

One of the challenges for mapping lung cancer risk at a fine spatio-temporal scale is limited sample sizes. For rare diseases such as cancers, the total counts of cases could become very sparse at fine spatio-temporal scales. This situation could become more obvious when many demographic dimensions are considered simultaneously, such as gender, age, and race. With the sparseness of the counts, some traditional estimates of disease risk, such as the standardized incidence ratio (SIR), could become unreliable and may lead to a large misunderstanding of the true disease risk due to high sampling variability. Recently, hierarchical Bayesian models have been widely used to map disease risk spatially or spatio-temporally (Abellan et al. *2008*; Bernardinelli et al. *1995*; Best et al. *2005*; Knorr-Held *2000*; Lawson *2009*; Mollié *2001*; Wakefield et al. *2001*; Waller et al. *1997*; Xia and Carlin *1998*). For example, Fortunato et al. (*2011*) used Bayesian modeling to study the spatio-temporal patterns of bladder cancer incidence in Utah from 1973 to 2004. Richardson et al. (*2006*) conducted Bayesian modeling of joint patterns of male and female lung cancer risks in Yorkshire in UK. For sparse count data, the integration of both data fit and subjective prior information makes it possible for Bayesian models to mitigate the inferential biases of frequentist methods that totally depend on data fit. In addition, under the Bayesian framework, it is easy to develop model-based spatial and spatio-temporal smoothing methods, which not only consider the effects of disease risk factors, but also borrow strengths from neighboring areas and/or time periods.

In this study, we used hierarchical Bayesian models and GIS to explore the spatio-temporal patterns of lung cancer incidence risk in Georgia. We used the term of "hierarchical" to emphasize the hierarchical structure of Bayesian models instead of covariates at different organizational levels. The risk under the study was relative risk (RR), which was defined as the ratio of local risk in a spatio-temporal unit to the average risk across the whole study area over the entire time period. The analyses are conducted for four population groups stratified by gender and race at the census tract level over four 2-year periods from 2000 to 2007. A total of seven spatio-temporal models under two different modeling frameworks are proposed and compared. One framework is to model the RR of each population group separately, and the other framework is to jointly model the RR of each population group under the assumption that some common disease risk factors exist in all of the population groups. One of the seven models is finally chosen based on specific criterion, and its results are interpreted. The aim of the study is to obtain reliable spatio-temporal patterns of lung cancer incidence risk by gender and race in Georgia at a relatively fine aggregation scale. These patterns can help identify spatio-temporal hotspots of lung cancer risk among different population groups for further study and facilitate the related health policy decisions in Georgia. The effect of area-based socioeconomic status (SES) on the lung cancer incidence risk of population groups also is explored in the modeling.

2 Study area and data

Our study area is Georgia, USA, with 1,618 census tracts in 2000. Figure 1 shows the spatial distribution of population density by census tract in Georgia 2000. The 10 most populous cities in 2000 are also identified in this map. We can see that the population is mainly concentrated in the north region of Georgia, especially in the metropolitan Atlanta area that includes the cities of Atlanta, Sandy Springs, Roswell, and Marietta. All population and socioeconomic data were downloaded from the US Census Bureau.

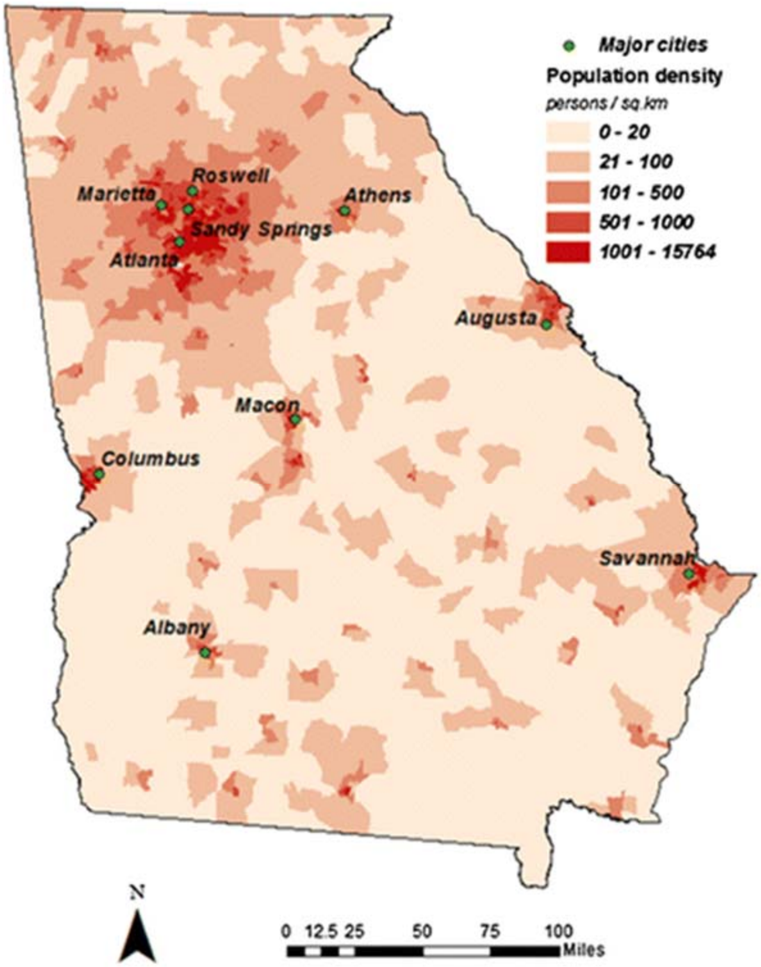

Fig. 1 Population density by census tract and the 10 most populous cities in Georgia 2000

The lung cancer data were extracted from the Georgia Comprehensive Cancer Registry (Georgia Department of Public Health *2011*). A total of 44,671 lung cancer cases were diagnosed in Georgia from 2000 to 2007. In this study, we only consider the cases among white and black individuals over 20 years old, which reduces the number of cases to 44,348. A total of 4,063 cases were excluded from the analyses because their recorded residences cannot be geocoded to the census tract or more detailed levels. A total of 40,285 cases were included in the final analyses. Among them, 34,347 cases (85.3 %) can be geocoded to the street level and the rest of them can meet the census tract level. Table 1 shows the distributions of cases by gender and race.

Table 1 Total numbers of cases of individuals over 20 years old and the included cases in the analyses

	White			**Black**		
	Total cases	**Included cases**	**Included cases (%)**	**Total cases**	**Included cases**	**Included cases (%)**
Male	20,547	18,614	90.59	5,557	4,991	89.81
Female	14,882	13,596	91.36	3,362	3,084	91.73

Compared to the zip code or county levels, census tracts are more homogenous and can provide more detailed information due to their smaller spatial sizes. Using finer aggregation levels on space (e.g., census block group level) and time (e.g., 1-year period) tends to decrease the precision of models due to an extremely high degree of data sparseness, to exclude more disease cases from the research due to a higher requirement on geocoding accuracy, and to increase the computational difficulty due to a significant increase in analytical unit amount. To balance the above considerations, therefore, the analyses aggregated the cases into the 1,618 census tracts and four 2-year periods, 2000–2001, 2002–2003, 2004–2005, and 2006–2007. The average number of cases per census tract per 2-year period was 2.88 for white males, 2.10 for white females, 0.77 for black males, and 0.47 for black females. Their average medians were 2, 1.5, 0, and 0, respectively. More detailed distribution of the observed numbers of cases is shown in the supplementary material.

3 Methods

3.1 Population estimation for intercensal years

The population at risk is important to the calculation of expected cases and the estimation of disease risk. However, the census population data at the tract level are only available at the census years (e.g., 2000 and 2010), and geographic boundaries of census tracts may change from time to time. For example, there were a total of 1,618 tracts in Georgia according to Census 2000 and that number became 1,969 in Census 2010. In this study, the boundaries of census tracts in 2000 were used as the standard geography for the whole study period. At the county level, the Census Bureau (Population Estimates Program *2011*) provides the estimates of population by race, gender, and age group for each intercensal year. With the population census data at the tract level and the estimates at the county level, we estimated the population by race, gender, and age group at the census tract level for each intercensal year.

First, we used the overlay function in the GIS, ArcGIS™ (ESRI, Inc.), and the areal weighting interpolation method (Goodchild and Lam *1980*) to estimate the population in 2010 using the geography of the 2000 census tracts. To improve the accuracy, we used the 2010 population data at the block level instead of the tract level. Then, we followed Best and Wakefield (*1999*)'s Model B to allocate the county population in each race–gender–age group to the census tracts in that county. Specifically, we assume the county population in a race–gender–age group, *N*, are multinomially distributed to the census tracts in that county with a vector of apportionment probabilities $p = (p_1,\ldots,p_I)^T$, where *I* denotes the number of census tracts in that county andp_i is the proportion of the population in census tract *i* in the population of the county *N*. The probabilities *p* for each intercensal year are estimated via a simple linear interpolation between the censuses (i.e., 2000 and 2010).

3.2 Relative risk (RR) and expected cases

We modeled the RR of lung cancer incidence for each gender–race population group in each spatio-temporal unit defined by the census tract and the 2-year period. The reference rate of each gender–race population group was defined as the indirect age-adjusted incidence rate of that population group across the whole state of Georgia over the entire time period 2000–2007. Following Vena (*1983*)'s study on lung cancer, ten age groups were considered in this research including age groups from 20 to 39 and 40 to 49, seven five-year age groups from 50 to 84, and one group from 85 and over. Using the US 2000 standard population for standardization, the direct age-adjusted (over 20 years old) lung cancer incidence annual rates (per 100,000 population) in Georgia from 2000 to 2007 were 132.7 for white males, 75.3 for white females, 135.2 for black males, and 54.5 for black females.

The expected number of cases, which reflects the reference rate, was needed in the modeling of RR. To calculate the expected number of cases for each gender–race population group in each spatio-temporal unit, we first calculated the age-specific incidence annual rate of each gender–race population group across Georgia over the entire time period and then calculated and summed up the expected number of cases in each age group for each spatio-temporal unit.

The standardized incidence ratio (SIR), defined as the ratio of the number of observed cases to the number of expected cases, is regarded as the best maximum likelihood estimate for RR in frequentist methods. In this study, we drew a comparison between the SIRs and our modeling results of RR.

3.3 Area-based SES measure

Due to the relative homogeneity, the area-based SES measure at the census tract level is a good surrogate of individual SES in a health study when individual SES is unavailable (Krieger *1992*). Detailed discussions of area-based SES measures can be found in the literature (Darden et al. *2009*; Carstairs *2001*; Krieger et al.*1997*, *2002*). Various single variable or composite measures can capture different aspects of socioeconomic characteristics. In this study, we used the modified Darden–Kamel Composite Index (Darden et al. *2009*) to measure the SES at the census tract level and evaluate its relationships with the lung cancer incidence risk by race and gender in Georgia. The modified Darden–Kamel Composite Index is an average Z-score of total nine socioeconomic variables in US census data (Table 2).

Table 2 Variables incorporated in the modified Darden–Kamel Composite Index

Modified Darden–Kamel Composite Index
1. Percentage of residents with university degrees
2. Median household income
3. Percentage of managerial and professional positions
4. Median value of dwelling
5. Median gross rent of dwelling
6. Percentage of homeownership
7. Percentage below poverty
8. Unemployment rate
9. Percentage of households with vehicle

Based on Census 2000 data, the modified Darden–Kamel Composite Indices for the census tracts in Georgia were calculated and the value range was from −31.05 to 24.77. A larger value means a higher SES. Based on the index, the census tracts in Georgia were divided into five SES groups using quintile classification so that each group had the same (or very close) number of census tracts. Group 1 had the highest SES, and group 5 had the lowest one. Figure 2 shows the spatial distribution of the SES by census tract. The top 20 % SES regions were mainly concentrated in the suburban areas in Georgia.

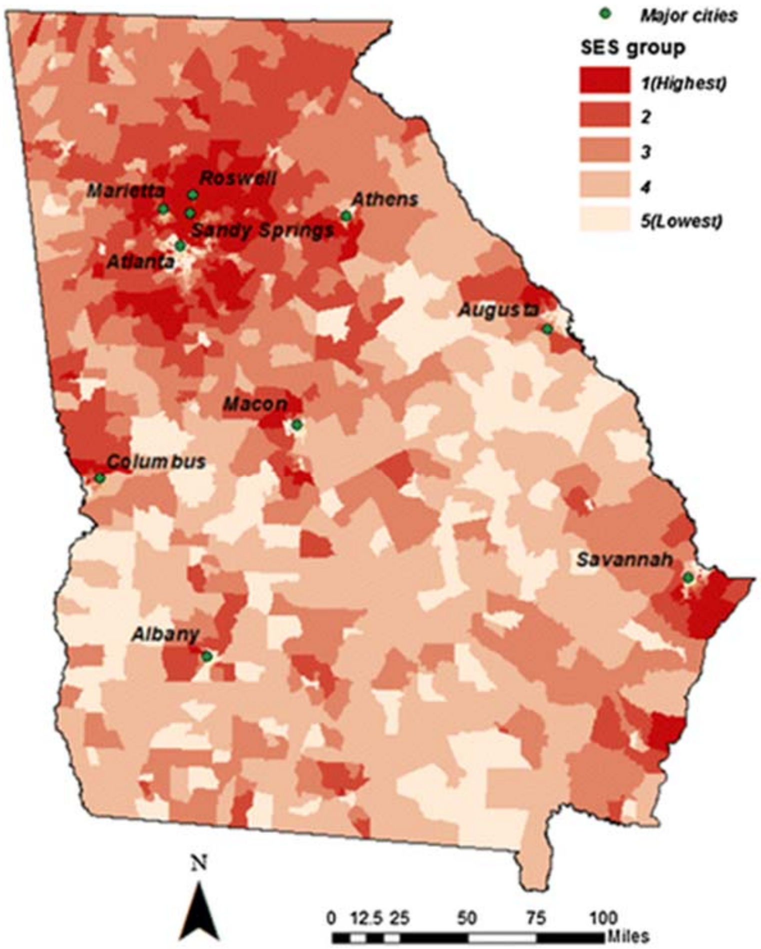

Fig. 2 Quintile map of SES in Georgia by census tract in 2000

3.4 Bayesian spatio-temporal models

Bayesian models naturally have hierarchical structures. At the first level, the number of observed cases $y_{itk}$for census tract $i = 1,\ldots,1{,}618$, time period $t = 1,\ldots,4$ (1: 2000–2001; 2: 2002–2003; 3: 2004–2005; 4: 2006–2007) and gender-race population group $k = 1,\ldots,4$ (1: white male; 2: white female; 3: black male; 4: black female) was assumed to follow a Poisson distribution with mean $E_{itk}\mathrm{RR}_{itk}$, where E_{itk} and RR_{itk} are, respectively, the estimated expected number of cases and the unknown RR in census tract *i*, time period *t,*and population group *k*. At the second level, the logarithms of RRs were decomposed into fixed effects for those measured risk factors, such as SES, and random effects for those unmeasured or unobserved risk factors. In Bayesian spatio-temporal models, three random effects are usually considered: spatial random main effect, temporal random main effect, and spatio-temporal interaction random effect. Both spatial and temporal random main effects would be further divided into a structured component and an unstructured component, which reflect the dependent and heterogeneous variations in risks in space and time, respectively. In the Bayesian paradigm, prior distributions were needed to be assigned to the model parameters and the random effects. Then, the inferences were made from the simulation-based posterior distributions of the parameters and random effects.

In this study, we modeled the RR of each population group individually under two modeling frameworks. The first framework used separate modeling where each population group had an independent set of random effects. The second framework used joint modeling where there were shared random effects representing some common unmeasured or unknown risk factors among all of the population groups. This joint modeling framework has been used to map one disease for multiple population groups or multiple diseases that have common risk factors (Richardson et al. *2006*; Knorr-Held and Best *2001*; Held et al. *2005*; Downing et al.*2008*; Tassone et al. *2009*; Wheeler et al. *2008*). We compared a total of seven models including two separate models and five joint models. Table 3 shows the components of the logarithm of RR in each model. These seven models are nested models where model 7 is the full model. In all joint models, the coefficients $\delta_{1,k}$ and $\delta_{2,k}$ allow gradients of the shared spatial and temporal components among all the population groups. For the two components $\varphi\ ik$ and $\theta\ tk$ in models 4–7, we set them equal to 0 for white male models ($k = 1$), so that these two components in other population group models ($k = 2$, 3 and 4) actually are the differentials of the spatial and temporal random main effects between that population group and white male group.

Table 3 Components of logarithms of RRs in the seven Bayesian spatio-temporal models

Framework	**Model #**	**Logarithms of RRs**
Separate	Model 1	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\varphi ik+\theta tk$
	Model 2	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\varphi ik+\theta tk+\omega itk$
Joint	Model 3	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\delta 1,k\lambda i+\delta 2,k\xi t+\omega itk$
	Model 4	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\delta 1,k\lambda i+\delta 2,k\xi t+\varphi ik+\theta tk$
	Model 5	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\delta 1,k\lambda i+\delta 2,k\xi t+\zeta it+\varphi ik+\theta tk$
	Model 6	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\delta 1,k\lambda i+\delta 2,k\xi t+\varphi ik+\theta tk+\omega itk$
	Model 7	$\log(\mathrm{RR}itk)=\alpha k+\beta Tkxi+\delta 1,k\lambda i+\delta 2,k\xi t+\zeta it+\varphi ik+\theta tk+\omega itk$
Fixed effects		
$\alpha\ k$—Overall log-RR for population group k across the whole study area over the whole study period		
$\beta\ k$—Coefficients associated with the SES group vector $x\ i$ for population group k		

Population group-specific random effects
$\varphi\ ik$—Spatial random main effect for population group *k* in census tract *i*
$\theta\ tk$—Temporal random main effect for population group *k* in time period *t*
$\omega\ itk$—Spatio-temporal interaction for population group *k* in census tract *i* and time period *t*
Population group-shared random effects
$\lambda\ i$—Shared spatial component in census tract *i*
$\xi\ t$—Shared temporal component in time period *t*
$\delta\ 1,k,\ \delta\ 2,k$—Coefficients of $\lambda\ i$ and $\xi\ t$ for population group *k*
$\varsigma\ it$—Shared spatio-temporal interaction in census tract *i* and time period *t*

In preliminary analyses, we tested models with different combinations of structured and unstructured components for spatial and temporal random main effects under both separate and shared modeling frameworks. The results showed that models involving both structured and unstructured components were generally more difficult to converge. Therefore, we only considered structured components in spatial and temporal random main effects for all of the models in Table 3. Specifically, the widely used Gaussian intrinsic conditional autoregression normal (CAR normal) prior proposed by Besag et al. (*1991*) was used to represent the dependent variations in RRs over space and time. For the population group-specific random effects $\varphi\ ikand$ $\theta\ tk$, the CAR priors were independent for each population group *k*. For population group-shared random effects $\lambda\ i$ and $\xi\ t$, the same CAR priors were applied across the different population groups. For a spatial random effect in an area, CAR normal specifies that its conditional distribution, given all other spatial effects, is a normal distribution with mean equal to the average spatial effects of its neighboring areas and variance inversely proportional to the number of these neighbors. In this study, the spatial neighbors were Queen neighbors, defined if they shared a border or a vertex. For a temporal random effect in a time period, CAR normal smoothes it toward the temporal effects of its applicable previous and next time periods.

Due to the lack of strong prior knowledge, vague prior distributions were used for other parameters in the models based on the current literature. We assigned a flat prior on the overall log-RR terms, α_k, and assigned independent normal (0, 10^{-5}) prior distributions to fixed effects $\beta\ k$. Independent normal (0, 5) prior distributions were assigned to the logarithms of scaling parameters $\delta_{1,}k$ and $\delta_{2,}k$, so that these scaling parameters have a 99 % possibility of lying between 0.36 and 2.82 with a mode at 1. This prior was also used by Richardson et al. (*2006*) and Downing et al. (*2008*). With respect to the spatio-temporal interaction random effects, independent normal prior distributions with means equal to 0 and precisions $\tau_{\omega k}$, $k = 1,\ldots,4$, were assigned to ω_{itk} in model 2 for each population group. Independent normal prior distributions with means equal to 0 and precisions τ_{ς} were assigned to ς_{it} in models 5 and 7, and a multivariate normal prior distribution with covariance matrix Σ was assigned to ω_{itk} in models 3, 6, and 7 to allow correlations among the population groups (Richardson et al. *2006*; Downing et al. *2008*). Following the previous studies (Best et al. *2005*; Downing et al. *2008*; Kelsall and Wakefield *1999*), independent conjugate hyperprior distribution Gamma (0.5, 0.0005) was assigned to all of the precision parameters in the normal priors for shared components τ_{λ}, τ_{ξ}, τ_{ς} and for population group-specific components $\tau_{\varphi k}$, $\tau_{\theta k}$, $\tau_{\omega k}$, $k = 1,\ldots,4$. The precision matrix (i.e., the matrix inverse of covariance matrix Σ) in the multivariate normal prior was assigned a Wishart (B, 4) distribution, where B is set to be a diagonal matrix with 0.01 s (Richardson et al. *2006*). This was a relative vague hyperprior making the expectation of precision matrix be a diagonal matrix with 400 s.

All of the models were fitted by the Markov chain Monte Carlo (MCMC) algorithms using WinBUGS software (Lunn et al. *2000*) (code for model 6 is given in the supplementary material). MCMC algorithms use iterative simulation of parameter values within a Markov chain to obtain the posterior distribution to which this chain converges. For each model, two independent chains were run for 60,000 iterations. The parameters we monitored included all fixed effects, scaling parameters, and standard deviations of all random effects. We also randomly selected varied numbers of overall RRs and random effects to monitor. Brooks–Gelman–Rubin diagnostics (Brooks and Gelman *1998*) and visual checks of tracing plots confirmed convergence for most monitored parameters by 50,000 iterations (selected tracing plots are shown in the supplementary material). Precision $\tau_{\omega k}$ and covariance matrix Σ for the prior distributions of spatio-temporal interaction component ω_{itk}did not converge well. However, their incapability of convergence did not affect the convergence of ω_{itk} that was achieved within 50,000 iterations. With the remaining 10,000 iterations (i.e., 20,000 samples with two chains), the Monte Carlo errors of the monitored parameters were <5 % of the sample standard deviation. Therefore, these samples were used for inference on all convergent parameters.

Similar to the joint mapping of lung cancer risk in males and females by Richardson et al. (*2006*), the scaling parameters $\delta_{2,k}$ were difficult to converge during the data fitting of models. This could be because only four time periods could not provide enough information to differentiate the shared and specific temporal patterns. Therefore, we fixed $\delta_{2,k} = 1$ for all joint models. Our preliminary analyses showed that fixing $\delta_{2,k}$ did not greatly affect the estimation of RRs in the joint models. The correlations of the estimated RRs were over 0.98 between the models with fixed $\delta_{2,k}$ and the corresponding ones with random $\delta_{2,k}$.

We used the deviance information criterion (DIC) to choose the best of the seven models. The DIC was proposed by Spiegelhalter et al. (*2002*) as the sum of $\bar{D}$ and pD, where $\bar{D}$ is the posterior mean of the deviance measuring the goodness of fit of a model, and pD is the number of effective model parameters measuring model complexity. The model with a smaller DIC was preferred due to its overall advantage in both data fit and model complexity.

4 Results

4.1 Model comparison

From Table 4, we can see that joint model 6 had the smallest DIC value of 64,155.6 among the seven models. Model 7 had the smallest $\bar{D}$ value indicating it had the best data fit, and model 4 had the smallestpD value indicating it was the simplest model. All of the joint models except for model 3 were better than the two separate models based on their DICs. Keeping parsimony in mind, in the following, we chose the results of model 6 ($\log(RR_{itk})=\alpha_k+\beta_k^T x_i+\delta_{1,k}\lambda_i+\delta_{2,k}\xi_t+\varphi_{ik}+\theta_{tk}+\omega_{itk}$) to interpret. In model 6, the shared components included spatial and temporal random main effects λ_i and ξ_t, and the specific components included spatial and temporal random main effects φ_{ik} and θ_{tk} as well as spatio-temporal interaction random effect ω_{itk}.

Table 4 DICs of the seven models

Framework	**Model #**	$\bar{D}$	***pD***	**DIC**
Separate	Model 1	63,349.2	962.636	64,311.8
	Model 2	63,029.5	1,264.91	64,294.4
Joint	Model 3	62,996.6	1,383.51	64,380.1
	Model 4	63,328.4	**869.157**	64,197.6
	Model 5	63,099.8	1,064.9	64,164.7
	Model 6	62,908.1	1,247.48	**64,155.6**
	Model 7	**62,904.5**	1,347.36	64,251.9

4.2 Spatio-temporal patterns of relative risk

For comparison, Fig. 3 shows the spatial patterns of the crude SIRs by race and gender in the first time period 2000–2001. These SIRs were calculated based on age-adjusted rates. In addition to the census tracts without residents or without lung cancer cases, we classified the census tracts into five categories based on their SIR values. The range of 0.95–1.05 represents the risk close to the reference rate, and other four categories represent the risk obviously higher or lower than the reference rate. Due to the uneven population distribution and possible missing in data collection, these SIR maps, especially those for black males and black females, show many census tracts with SIR value of zero due to zero cases observed in that tract and that time period. However, it was highly probable that lung cancer risk existed in these census tracts in reality. In addition, it is obvious that the SIR surfaces are not smooth across the whole area since most of the SIRs fall into either the very high or very low category. These facts indicated that SIR was not an appropriate estimate of the lung cancer risk in this study.

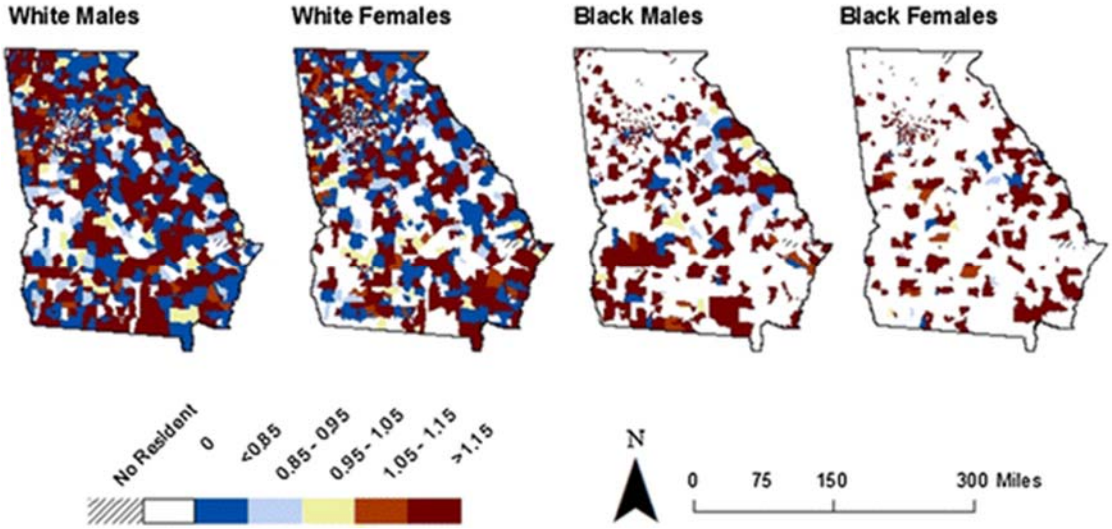

Fig. 3 Maps of crude SIRs by race and gender by census tract during 2000–2001

Figure 4 shows the maps of posterior median RR of model 6 in the four time periods for all of the gender–race population groups. In these maps, we used the classification breakpoints in Fig. 3 to classify the census tracts in terms of their modeled RR values. Compared to the crude SIRs in Fig. 3, the model-based RR shows a much smoother spatial pattern without RR equal to 0 in each population group. These maps show different spatial patterns of lung cancer risk exist among the four population groups. For white males and white females, the high RRs were mainly concentrated in the northwest, southeast, and middle regions of Georgia. For black males, the high RRs were mainly concentrated in the northwest, east, and south Georgia. The high RRs for black females were mainly concentrated in the northwest of Georgia. Comparing the maps of different time periods, for white males and black males, more census tracts with moderate and low RRs emerged and the number of census tracts with high RRs decreased over the time, while the situations reversed for white females and black females.

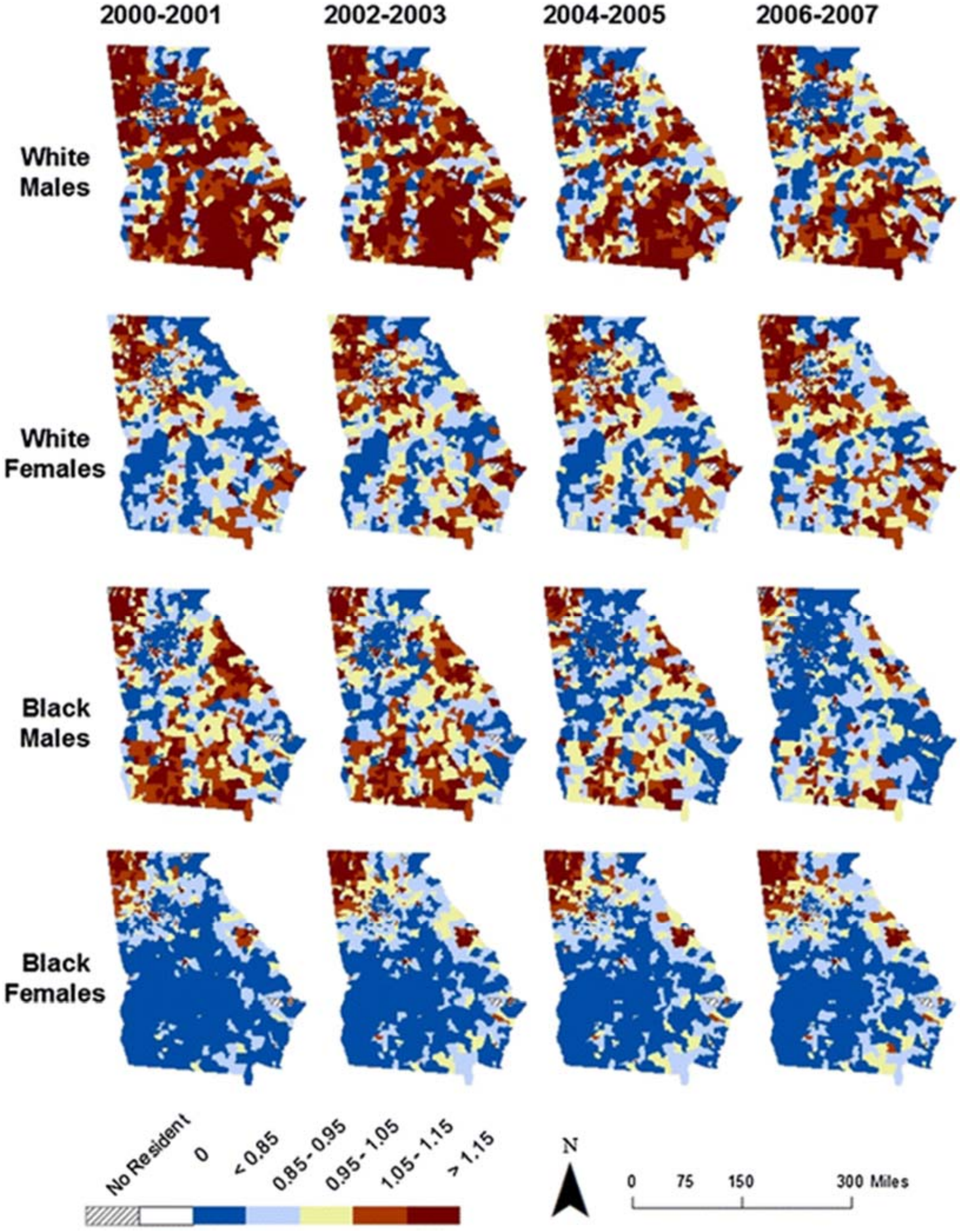

Fig. 4 Maps of the posterior median RR for all population groups in the four time periods

Those areas with consistently high disease risks over the time can only be identified using spatio-temporal analyses. Richardson et al. (*2004*) showed that Bayesian disease-mapping models are essentially conservative, with high specificity but low sensitivity if the elevated-risk areas have only a moderate (<twofold) excess. To obtain high specificity (around 95 %) and reasonable sensitivity to pick out areas where the true RR is moderately elevated (e.g., around twofold), they suggested using a cut-off rule of 0.8 on the posterior probability that an area had an estimated RR >1. Figure 5 shows the maps indicating how many times each census tract had an estimated elevated RR during the four time periods based on the rule of prob(RR > 1) > 0.8. The frequency associated with each census tract reflected the stability of elevated RR in that area over the entire time period. The northwest of Georgia and the area near Augusta had consistently high RRs for all population groups. The identification of those census tracts with consistently high RRs over the time could be helpful to generate some etiological hypotheses and support health policy making, such as the distribution of resources.

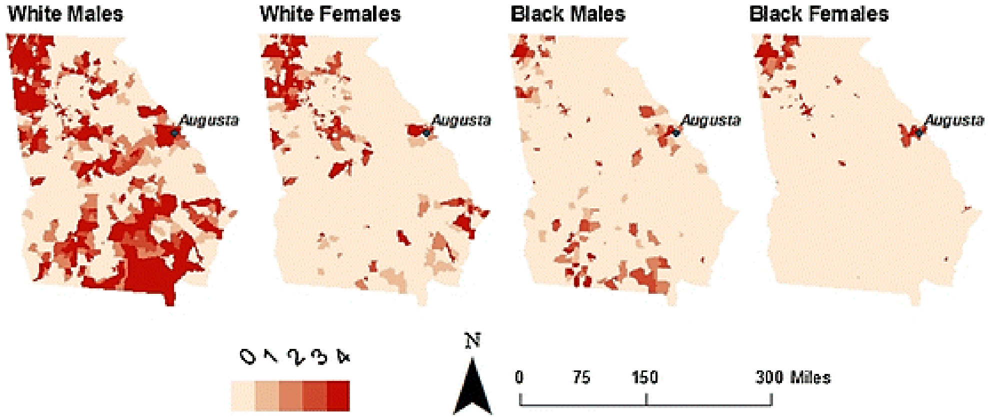

Fig. 5 Maps of elevated RR frequency (prob(RR > 1) > 0.8) by race and gender during 2000–2007

We studied the spatial patterns of lung cancer incidence RR among the population groups by looking at the shared and the population group-specific spatial components in model 6. The map of the shared spatial components in Fig. 6 captures the common spatial variation in RR among the four population groups. Taking the white male group as the reference with its scaling parameter equal to 1 for the shared spatial component, the posterior medians of the scaling parameters for white females, black males, and black females were 0.743 [95 % credible interval (CI) 0.606, 0.892], 0.538 [95 % CI 0.343, 0.761], and 0.571 [95 % CI 0.355, 0.818], respectively.

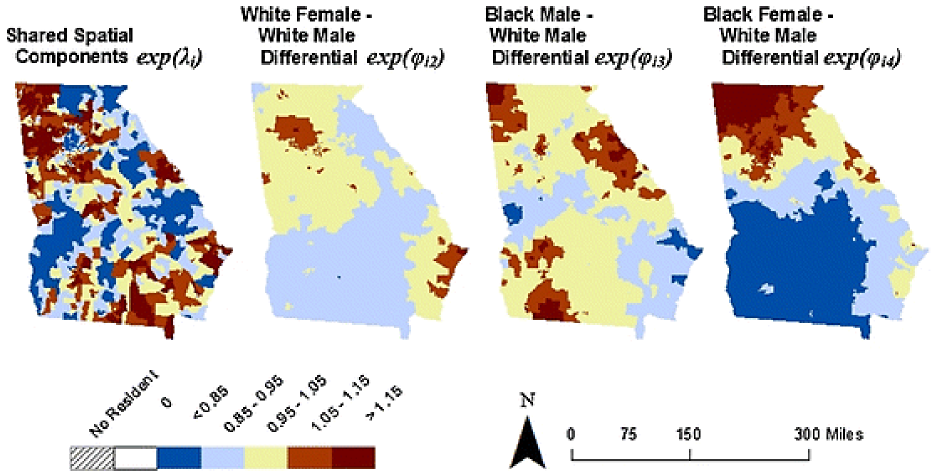
Shared Spatial Components exp(λi)
White Female - White Male Differential exp(φi2)
Black Male - White Male Differential exp(φi3)
Black Female - White Male Differential exp(φi4)
No Resident
0
<0.85
0.85-0.95
0.95-1.05
1.05-1.15
>1.15
N
0 75 150 300 Miles

Fig. 6 Maps of the posterior medians of the shared spatial component $exp(\lambda)$ and differential spatial components$exp(\varphi)$

The population group-specific spatial components reflect the deviation of spatial pattern in each population group from the common spatial pattern. This deviation could be caused by the population group-specific location-related risk factors. Since we took the white male group as the reference by setting its specific spatial component equal to 0, the shared spatial component totally represented the spatial effect among white males, and the specific spatial components in other three population groups reflected the differentials of spatial effect between that population group and white male group. From the differential maps (on exponential scale) of specific spatial components in Fig. 6, we can see that both of the white female–white male differential and the black male–white male differential show a major portion of the area with a value ranging from 0.85 to 1.15, which indicates that the pattern of the shared spatial component (i.e., the spatial pattern of RR in white males) captures well the variations in the spatial effects on RR for both white females and black males. Exceptions (i.e., high and low differential values) existed in the areas of metropolitan Atlanta and Savannah for white females, and existed in the northwest, northeast, and southwest in Georgia for black males. In the black female–white male differential map, a large southern area with a value <0.85 and a large northern area with a value larger than 1.15 reflect that there was an obvious difference in the spatial pattern of RR between white males and black females. The spatial pattern of RR in white males underestimated the RR in black females in northern Georgia and overestimated that in southern Georgia.

Table 5 shows the posterior medians and 95 % CIs of the shared temporal component and the differential temporal components. The shared temporal trend stayed flat in the first two periods and slightly decreased after 2004. The three differentials showed that the shared temporal trend captured well the temporal trend in the RR of black males, but was different from those of white females and black females.

Table 5 Posterior medians (95 % CIs) of the shared temporal components $exp(\xi)$ and differential temporal components $exp(\theta)$

Time period	**Shared temporal components$exp(\xi t)$**	**White female–white male differential $exp(\theta t_2)$**	**Black male–white male differential$exp(\theta t_3)$**	**Black female–white male differential $exp(\theta t_4)$**
2000–2001	1.04 (1.02, 1.07)	0.93 (0.90, 0.97)	1.01 (0.98, 1.06)	0.92 (0.86, 0.98)
2002–2003	1.04 (1.01, 1.06)	0.97 (0.94, 1.00)	1.00 (0.97, 1.04)	0.97 (0.92, 1.02)

2004–2005	0.98 (0.96, 1.00)	1.02 (0.99, 1.05)	1.00 (0.97, 1.04)	1.03 (0.98, 1.08)
2006–2007	0.95 (0.92, 097)	1.09 (1.05, 1.13)	0.98 (0.94, 1.02)	1.09 (1.03, 1.16)

4.3 Effect of SES

The posterior medians of the SES-related RR in Fig. 7 show the effect of SES on RR in each gender–race population group. The highest SES group was taken as the reference. The general trend among all population groups was that lower SES leads to a higher RR. However, the gradients of SES effects on RR in males were larger than those in females. The socioeconomic disparities in lung cancer RR were more obvious in males in Georgia.

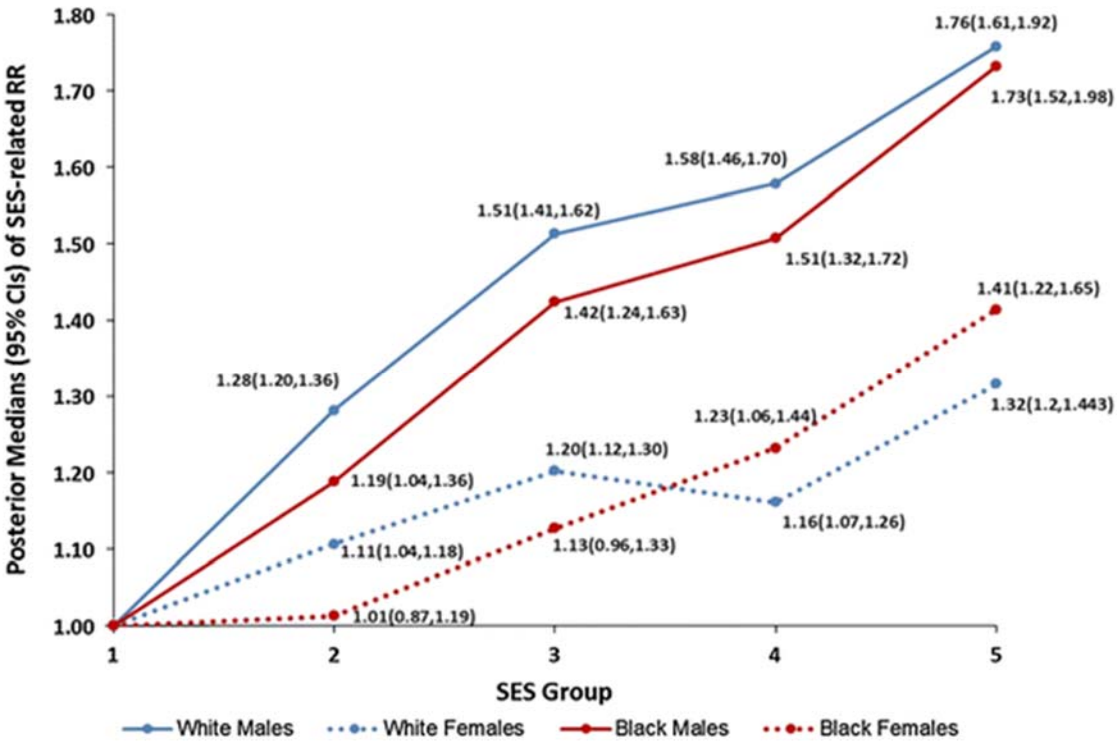

Fig. 7 Posterior medians (95 % CIs) of the RR for SES quintile *exp*(*β*) in the four population groups

4.4 Urban versus rural setting

An urban census tract was defined as a census tract with over 50 % of its area falling within the urbanized region defined by Census 2000. Following this definition, there were 827 urban census tracts and 791 rural census tracts. As shown in Fig. 8, males had higher average risk in rural areas while females had higher average risk in urban areas. The largest difference occurred in black females. One-way analysis of variance showed that all of the urban–rural differences were statistically significant ($p < 0.05$).

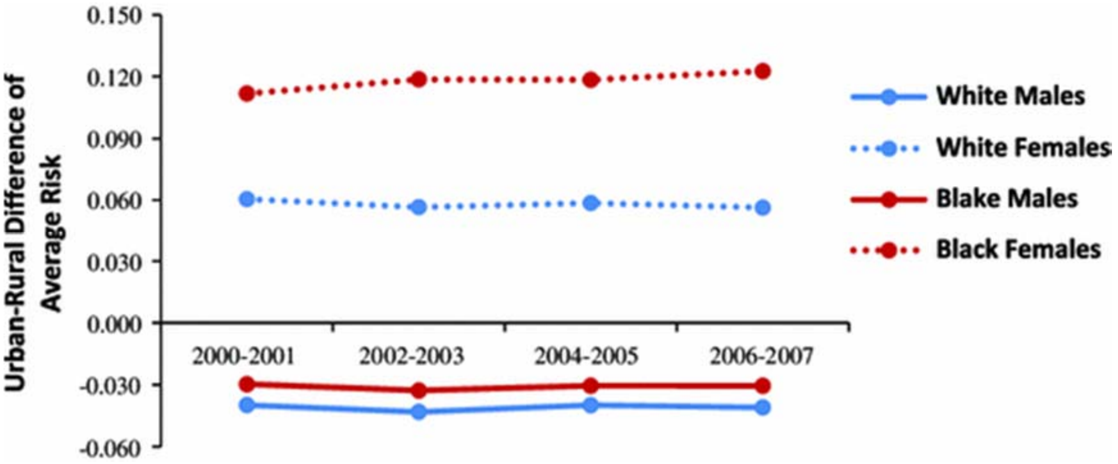

Fig. 8 Urban–rural difference of average risk

4.5 Sensitivity analysis

Bayesian modeling is sensitive to the choice of priors and hyperpriors. Following Downing et al's (*2008*) work, we performed a sensitivity analysis using an alternative hyperprior distribution Gamma (1,1) to replace Gamma (0.5, 0.0005) for the precision parameters in model 6. The Gamma (0.5, 0.0005) distribution made the variances (inverse of precision) have a 99 % probability of lying between 0.000151 and 6.25 with a mode at 0.00033. For the Gamma (1, 1) distribution, the 99 % probability range of the variances was from 0.217 to 100 and the mode was at 0.5. After 50,000 burn-in iterations with two independent chains that confirmed convergence, 20,000 samples were used for reference. The DIC of the model was 64,160.8 (D^{-} = 62,892.7,pD = 1,268.1), slightly larger than model 6. Table 6 shows the correlations between the posterior median RRs using model 6 with the two types of hyperpriors. The two groups of results showed a good concordance in general, but the correlations in black individuals were slightly lower than those in white individuals. These differences may be due to the different degrees of data sparseness between races.

Table 6 Correlations between the posterior median RRs using model 6 with two different types of hyperpriors

Time period	White males	White females	Black males	Black females
2000–2001	0.998	0.992	0.988	0.990
2002–2003	0.998	0.991	0.988	0.989
2004–2005	0.998	0.991	0.987	0.988

2006–2007	0.998	0.991	0.987	0.988

5 Discussion and limitations of the study

This study explored the spatio-temporal patterns of lung cancer incidence risk for four gender–race population groups at the census tract level within four 2-year periods 2000–2007. These patterns, which are usually smoothed out in spatial and/or temporal analyses at coarser scales, can answer such queries as who, where, and when the risk of lung cancer varies. For example, in addition to the northwestern Georgia where all population groups have stable elevated lung cancer risks over the study period, more attention is also needed to the higher rates among white males in many census tracts in the south of Georgia. With visual comparison or other explorative spatial analysis methods, these spatio-temporal patterns, especially the individual spatial and temporal components in the modeling, can aid the establishment of etiological hypotheses of lung cancer with regard to environmental- or lifestyle-related risk factors. These assumptions can then be validated with further models of cause and effect or biological experiments. This study showed a general inverse relationship between SES and lung cancer incidence risk for all population groups, and a larger gradient exists in males. This result was consistent with the findings of several previous studies (Mao et al. *2001*; van Loon et al. *1995*). To explain the socioeconomic disparities in the lung cancer risk, further study is needed, such as the exploration of occupational differences between males and females in Georgia.

It is well known that an individual's smoking behavior is an important risk factor for lung cancer. However, one of the limitations in this study was the lack of suitable smoking data at the fine spatial scale. The smoking data from the Behavioral Risk Factor Surveillance System (CDC *2013*) can be readily obtained. However, it is only available to 22 % counties in Georgia at the level of metropolitan statistical area since 2002. Recently, several studies show that the associations of SES with lung cancer may be attributable to incomplete adjustment for smoking (Matukala Nkosi et al. *2012*; Menvielle et al. *2009*). To some extent, the random effects in our hierarchical Bayesian spatio-temporal models can approximate the total effects of unmeasured or unknown risk factors including smoking. However, we believe that integrating appropriate smoking data into the models can greatly reduce the uncertainty of the models.

Although our study included about 90 % (i.e., a total of 40,285) Georgian lung cancer cases diagnosed during the study time period (see Table 1), it is important to note the potential bias introduced by the exclusion of cases. Figure 9 shows the spatial distribution of 3,039 excluded cases that can be geocoded at best to counties instead of census tracts. Most of the counties excluded less than 25 % cases. However, there were seven rural counties in central Georgia with large percentages (>50 %) for excluded cases. The high percentages of exclusion led to large uncertainty on the RR estimates in those areas, requiring more carefulness to use the study results. The percentages of the excluded cases for the four time periods are 12.1 % for 2000–2001, 8.9 % for 2002–2003, 8.0 % for 2004–2005, and 7.8 % for 2006–2007.

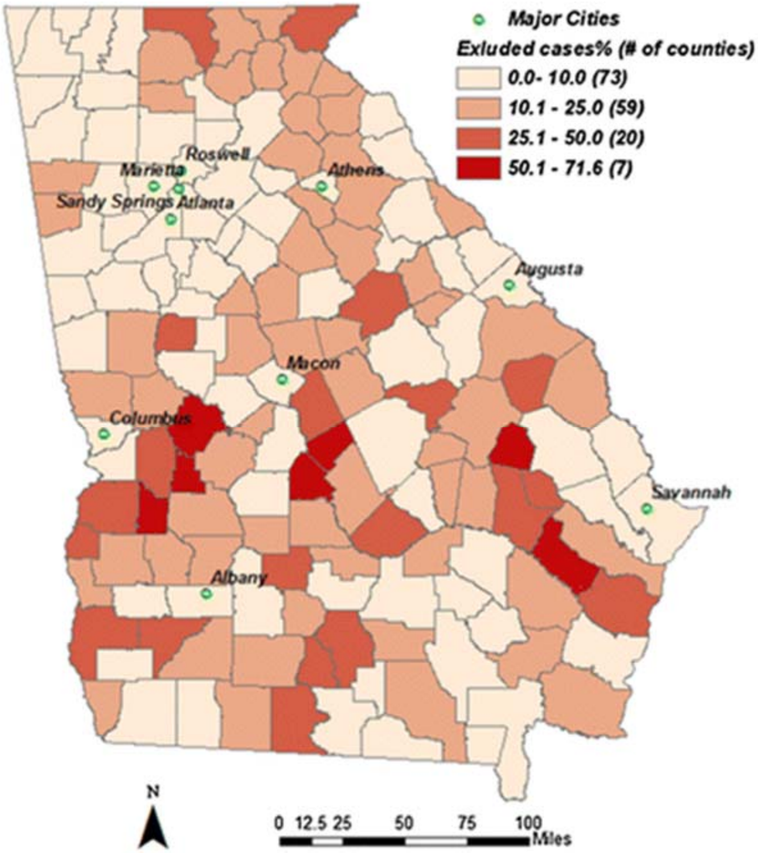

Fig. 9 Distribution of percentage of excluded cases by county

According to the 2009 American Community Survey (ACS *2013*), about 13.9 % of the population (1 year old and over) move each year in USA during our study period, about 8.3 % moving within the county, and about 2.7 % moving to another county within the state. Population mobility has been a big challenge in the studies of the diseases with a long latency period such as cancers (Wheeler et al. *2012*). In this study, we measure the area-based SES with Census 2000 data and assume they could reflect the individual SES during the long latency period. This assumption made this study suffer from measurement error. However, decennial census data at the census tract level was the best data we can get to approximate individual SES. Without detailed mobility data, integrating this census information can reduce the modeling uncertainty to some degree. In addition to the assumption about population mobility, the analysis of the relationship between disease RR and SES is subject to the modifiable area unit problem (Openshaw and Taylor *1981*) and ecological fallacy problem. The inferences based on the analyses at current scale and/or unit definition may not be generalized to other scales and/or unit definitions.

Estimation of population in small areas is a hot research topic in geography and statistics recently. In our study, we used an apportionment method to estimate the population by race, gender, and age in each census tract in each intercensal year. Improvement in population estimation models could greatly benefit the disease-mapping models.

6 Conclusion

We reported hierarchical Bayesian models to contribute to the literature about lung cancer studies in Georgia and to explore the spatio-temporal patterns of lung cancer incidence risk in Georgia for the time period 2000–2007. The study was conducted at the census tract level using 2-year time period as the temporal unit. Compared to the commonly used county level and 5-year time period, the finer spatial and temporal scales enabled our study to show more detailed variations in lung cancer incidence risk in space and time, which can better support healthcare performance assessment, etiological hypothesis generation, and health policy making. In this study, a total of seven Bayesian spatio-temporal models under the separate (each population group with an independent set of random effects) and joint (shared random effects representing some common unmeasured or unknown risk factors among all of the population groups) modeling frameworks were proposed and compared. The modeling results showed that the joint models generally produced better performance than the separate models using DIC as the criterion. Compared to the crude SIR, the estimated disease risk from Bayesian spatio-temporal models can be more reliable at a relatively fine spatio-temporal scale. The study also showed that there were strong inverse relationships between SES and lung cancer incidence risk in males and weak inverse relationships in females in Georgia. These relationships and the patterns of the spatial and temporal random effects in these Bayesian models may provide some implications on the underlying disease risk factors for further ecological studies.

Acknowledgments

All of the authors want to thank Dr. Sara Wagner for her help on the Georgia cancer data acquisition. Dr. John E. Vena's effort was sponsored in part by a Distinguished Cancer Scientist award from the Georgia Cancer Coalition (GCC).

Electronic supplementary material

Below is the link to the electronic supplementary material.

Supplementary material 1 (DOCX 255 kb)

References

Abellan JJ, Richardson S, Best N (2008) Use of space-time models to investigate the stability of patterns of disease. Environ Health Perspect 116(8):1111

ACS (2013) American Community Survey data on geographical mobility/migration. http://www.census.gov/hhes/migration/data/acs.html. Accessed 10 Oct 2013

Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M (1995) Bayesian analysis of space-time variation in disease risk. Stat Med 14(21–22):2433–2443

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43(1):1–20

Best N, Wakefield J (1999) Accounting for inaccuracies in population counts and case registration in cancer mapping studies. J R Stat Soc Ser A (Stat Soc) 162(3):363–382

Best N, Richardson S, Thomson A (2005) A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 14(1):35

Brooks SP, Gelman A (1998) Alternative methods for monitoring convergence of iterative simulations. J Comput Gr Stat 7:434–455

Carstairs V (2001) Socio-economic factors at areal level and their relationship with health. Spatial. Epidemiology 1(9):51–68

CDC (2013) Behavioral risk factor surveillance system. http://www.cdc.gov/brfss/index.htm. Accessed 10 Oct 2013

Darden J, Rahbar M, Jezierski L, Li M, Velie E (2009) The measurement of neighborhood socioeconomic characteristics and black and white residential segregation in metropolitan Detroit: implications for the study of social disparities in health. Ann Assoc Am Geogr 100(1):137–158

Downing A, Forman D, Gilthorpe M, Edwards K, Manda S (2008) Joint disease mapping using six cancers in the Yorkshire region of England. Int J Health Geogr 7(1):41

Fortunato L, Abellan JJ, Beale L, LeFevre S, Richardson S (2011) Spatio-temporal patterns of bladder cancer incidence in Utah (1973–2004) and their association with the presence of toxic release inventory sites. Int J Health Geogr 10(1):16

Georgia Department of Public Health (2008) Cancer program and data summary (trans: Health GDoP). Atlanta, GA

Georgia Department of Public Health (2011) Georgia comprehensive cancer registry. http://www.health.state.ga.us/programs/gccr/. Accessed 4th Oct 2011

Goodchild MF, Lam NS (1980) Areal interpolation: a variant of the traditional spatial problem. Geo-Processing 1:297–312

Held L, Natário I, Fenton SE, Rue H, Becker N (2005) Towards joint disease mapping. Stat Methods Med Res 14(1):61–82

Kelsall J, Wakefield J (1999) Discussion of ‘ Bayesian models for spatially correlated disease and exposure data’, by Best et al. In: Bernardo J, Berger J, Dawid A, Smith A (eds) Bayesian statistics 6. Oxford University Press, Oxford, p 151

Knorr-Held L (2000) Bayesian modelling of inseparable space-time variation in disease risk. Stat Med 19(17–18):2555–2567

Knorr-Held L, Best NG (2001) A shared component model for detecting joint and selective clustering of two diseases. J R Stat Soc Ser A (Stat Soc) 164(1):73–85. doi:10.1111/1467-985x.00187

Krieger N (1992) Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. Am J Public Health 82(5):703

Krieger N, Williams DR, Moss NE (1997) Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health 18(1):341–378

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian S, Carson R (2002) Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? Am J Epidemiol 156(5):471

Lawson AB (2009) Bayesian disease mapping: hierarchical modeling in spatial epidemiology, vol 20. CRC, Boca Raton, FL

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10(4):325–337

Mao Y, Hu J, Ugnat A-M, Semenciw R, Fincham S (2001) Socioeconomic status and lung cancer risk in Canada. Int J Epidemiol 30(4):809–817

Matukala Nkosi T, Parent M-É, Siemiatycki J, Rousseau M-C (2012) Socioeconomic position and lung cancer risk: how important is the modeling of smoking? Epidemiology 23(3):377

Menvielle G, Boshuizen H, Kunst AE, Dalton SO, Vineis P, Bergmann MM, Hermann S, Ferrari P, Raaschou-Nielsen O, Tjønneland A (2009) The role of smoking and diet in explaining educational inequalities in lung cancer incidence. J Natl Cancer Inst 101(5):321–330

Mollié A (2001) Bayesian mapping of Hodgkins disease in France. Spat Epidemiol 1(9):267–286

Openshaw S, Taylor PJ (1981) The modifiable areal unit problem. In: Wrigley N, Bennett R (eds) Quantitative geography: a British view. Routledge, London, pp 60–69

Population Estimates Program (2011) County intercensal estimates (2000–2010). http://www.census.gov/popest/data/intercensal/county/county2010.html. Accessed 22nd Feb 2012

Richardson S, Thomson A, Best N, Elliott P (2004) Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect 112(9):1016

Richardson S, Abellan J, Best N (2006) Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). Stat Methods Med Res 15(4):385

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B (Stat Methodol) 64(4):583–639

Tassone EC, Waller LA, Casper ML (2009) Small-area racial disparity in stroke mortality: an application of Bayesian spatial hierarchical modeling. Epidemiology 20(2):234–241

van Loon AJM, Burg J, Goldbohm RA, van den Brandt PA (1995) Differences in cancer incidence and mortality among socio-economic groups. Scand J Public Health 23(2):110–120

Vena JE (1983) Lung cancer incidence among nonwhites in Erie County, New York. J Natl Med Assoc 75(12):1229

Wakefield J, Best N, Waller L (2001) Bayesian approaches to disease mapping. Spat Epidemiol 1(9):104–128

Waller L, Carlin B, Xia H, Gelfand A (1997) Hierarchical spatio-temporal mapping of disease rates. J Am Stat Assoc 92(438):607–617

Wheeler DC, Waller LA, Elliott JO (2008) Modeling epilepsy disparities among ethnic groups in Philadelphia, PA. Stat Med 27(20):4069–4085

Wheeler DC, Ward MH, Waller LA (2012) Spatial-temporal analysis of cancer risk in epidemiologic studies with residential histories. Ann Assoc Am Geogr 102(5):1049–1057

Xia H, Carlin B (1998) Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. Stat Med 17(18):2025–2043