Published in final edited form as:

Ann Epidemiol. 2017 January ; 27(1): 42–51. doi:10.1016/j.annepidem.2016.08.014.

Space-time variation of respiratory cancers in South Carolina: A flexible multivariate mixture modeling approach to risk estimation

Rachel Carroll[A,*], **Andrew B Lawson**[A], **Russell S Kirby**[B], **Christel Faes**[C], **Mehreteab Aregay**[A], and **Kevin Watjou**[C]

[A]Department of Public Health, Medical University of South Carolina, 135 Cannon St, Charleston, SC 29425, USA

[B]Department of Community and Family Health, University of South Florida, 13201 Bruce B Downs Blvd MDC 56, Tampa, FL 33612, USA

[C]Interuniversity Institute for Statistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, 3590 Diepenbeek, Belgium

Abstract

Purpose—Many types of cancer have an underlying spatio-temporal distribution. Spatio-temporal mixture modeling can offer a flexible approach to risk estimation via the inclusion of latent variables.

Methods—In this paper, we examine the application and benefits of using four different spatio-temporal mixture modeling methods in the modeling of cancer of the lung and bronchus as well as "other" respiratory cancer incidences in the state of South Carolina, USA.

Results—Of the methods tested, no single method outperforms the other methods; which method is best depends on the cancer under consideration. The lung and bronchus cancer incidence outcome is best described by the univariate modeling formulation while the "other" respiratory cancer incidence outcome is best described by the multivariate modeling formulation.

Conclusion—Spatio-temporal multivariate mixture methods can aid in the modeling of cancers with small and sparse incidences when including information from a related, more common type of cancer.

Keywords

Spatio-temporal; Bayesian; multivariate; respiratory cancers; mixture modeling; MCMC; lung and bronchus cancer

[*]Corresponding author, mosra@musc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Respiratory cancers as a whole (ICD-9-CM codes: 160–163, and 165) are among the most common types of cancer worldwide, and lung and bronchus cancers (ICD-9-CM codes: 162.2–162.5, 162.8, and 162.9) are among the highest annual incidence cancers in the United States. While incidence of lung and bronchus cancers as well as all respiratory cancers has been decreasing nationally [1], within the state of South Carolina incidence has increased in recent years. Cancers of the respiratory system display spatio-temporal (ST) patterns [2–8] and affect people of all ages, races, and genders, though not equally [1, 9]. Thus, it is of interest to explore this occurrence via the use of complex ST statistical modeling.

Lung and bronchus cancers (LBC) differ from “other” respiratory cancers (ORC) in their risk factors as well as their rarity. The vast majority of LBCs are attributed to first and second-hand tobacco smoke exposure and, to a lesser extent, occupational exposures [10, 11]. Conversely, ORC, which includes laryngeal, nasopharyngeal, mediastinal, pleural, and related cancers are quite rare. ORC also have associations with alcohol consumption [12], ethnicity, and co-occurrence of other diseases, viruses, or neoplasms [13–15].

Since these types of cancers affect related areas of the body, it is of interest to explore the usefulness of multivariate modeling for LBC with ORC. Further, it is hypothesized that the multivariate modeling of these two diseases will improve the fit and prediction of the less common disease (ORC) by way of the additional information provided by the more common disease (LBC). This multivariate extension of a mixture model can easily allow predetermined information to be shared between the diseases of interest and ultimately lead to stronger, more robust results with the potential for selection to occur. In this exploration, information is shared via the use of common random effects between diseases, which represent unobserved effects, correlation, or both between the diseases. We believe this multivariate modeling could aid in explaining the etiology of the outcomes of interest. Most cancer epidemiologic studies focus on specific cancers, yet when studied ecologically cancer incidence often is elevated across several types of cancer. By developing multivariate mixture models, we are able to examine the joint effects of covariates on several distinct types of cancer simultaneously, within a spatial framework. Insights from these models may provide clues to environmental and social factors hitherto unexplored in more traditional approaches to cancer epidemiology.

This paper is developed as follows. First, we describe the available data. Second, we examine the different models to be applied. Next, we display the results of employing these methods to the outcomes of interest. Finally, we discuss the results and draw conclusions.

Materials and Methods

Data for this study included measures of incidence of respiratory cancers for each of the 46 counties in the state of South Carolina, USA annually for the years (1996–2009), and predictors from a variety of data sets. Since our outcome of interest was the incidence of either LBC or ORC, a conditionally independent Poisson distribution was a reasonable model for these data. This is a commonly assumed model for small area counts in disease mapping [16] and is appropriate because the Poisson distribution is a discrete frequency distribution that provides the probability of events occurring in a given area. It is described as follows:

$$y_{ij}|\mu_{ij} \sim Pois\left(\mu_{ij}\right)$$
$$\mu_{ij} = e_{ij}\theta_{ij}$$

with y_{ij} the incidence of cancers in county i at time j and μ_{ij} as the mean of the Poisson distribution such that it is the product of the expected rate of disease, e_{ij}, and the relative risk, θ_{ij}, for county i at time j. The ICD-9-CM codes are listed and described in supplementary Table A.1, and a labeled county map is included in the supplementary materials as Figure A.1.

Case Study Data

The outcomes of interest were acquired from the South Carolina Community Assessment Network data sets [17]. The respiratory cancer data was collected in two steps. First, a table of all county level respiratory cancer incidences in the state of South Carolina annually for years 1996 to 2009 was collected. Similar data for lung and bronchus cancers were also collected. Then, the ORC incidences outcome was calculated by subtracting the counts of LBC from that of the all respiratory cancer data set. Note that all county boundaries remained constant over the study time period.

Following these calculations, the resulting statistics were available for each disease classification to be utilized for analysis. Summary statistical figures are available as Figure A.2 and A.3 in the supplementary materials; these figures confirm that both incidence and rate of disease were increasing over the study time period for both LBC and ORC. For the 44668 diagnosed LBC across South Carolina during these study years, there was approximately a mean incidence of 69 cases per county per year ranging from 5–321 cases per county leading to the following rate of disease: 0.0008. For the 4077 diagnosed ORC across South Carolina during these study years, there was a mean incidence of 6 cases per county per year with a range of 0–30 cases per county leading to the following rate of disease: 0.00007. While these data had no missing values at the county level, there was some censoring where counts of 1–4 were given the value 5 and counts of 5–10 were given the value of 10. This occurs in about 5.3% (n=34) of the county level measures across the study time for the all respiratory cancer incidence table and 7.5% (n=48) of the county level measures across the study time for the lung and bronchus cancer incidence table. In these cases, the censored values were assumed to be the true value, and because this occurs in such a small portion of the data, we believe that the impact of this assumption was negligible. However, when considering stratification by age, the amount of censored data does become more of a hindrance. Further, the distribution of incidences by gender were nearly identical; thus, we performed this analysis using the county level population without stratification.

The indirect standardized incidence rate (SIR) per county over time for each disease classification are displayed in Figure 1 (LBC) and Figure 2 (ORC). The SIR was calculated as the ratio of the observed cancer incidences to the expected rates of disease for each of the 46 counties and can be useful as a first step in data analysis [18]. Therefore, an SIR of one indicates that the observed incidence is equal to that of the expected count for a particular county at a particular time. This expected rate of disease is the product of the above overall rate of disease (calculated as the total statewide incidences across the study time divided by total statewide population over the study time, this rate is also the horizontal line present in supplemental Figure A.3) and the county level population per each study year, and it is the same as that which was used in the Poisson modeling. The expected rate was not age group specific because, at the population level, age specific variation was very small across counties as well as over time. Further, if there were age effects present, their impact would be very marginal on the results. These plots are discussed further in the results section. The SIR is an estimator of the relative risk, thus estimated values for θ_{ij} under our models are comparable, albeit smoothed, to the SIR estimates.

The predictors for this analysis were obtained from a variety of resources as they were both demographic and environmental measures as well as spatial, temporal, or both in structure. The demographic predictors came from the Area Health Resource Files [19] data set and consisted of: proportion of persons with health insurance (pHI), proportion of African American population (pAA), unemployment rate of those 16 years or older (UER), and proportion of persons in poverty (pppov). The ‘proportion’ forms of the predictors were calculated by taking the ratio of a ‘number of persons’ measure to the county level populations acquired from the South Carolina Community Assessment Network data sets. The three environmental variables were average daily sunlight (sun) acquired from the North America Land Data Assimilation System [20], average in home radon concentrations (pCi/L) based on in home test kit results analyzed by the South Carolina Department of Health and Environmental Control (SCDEHC) laboratory [21], and statewide average annual rainfall from the National Oceanic and Atmospheric Administration [22]. The seven selected predictors were chosen based on availability and reasonable amount of collinearity. Additionally, variables selected as predictors have been included in previous research examining associations of environmental exposures and socio-economic status with cancer incidence both overall and more specifically with cancer of the lung and bronchus [1, 23–25]. However, these predictors were selected primarily for their ability to perform as representations of the behavior of spatial, temporal, and ST variables. These predictors were standardized (zero mean and variance of one) for the model fitting.

Among the predictor variables, radon, pHI, and pAA were set to be spatially varying predictors. They were chosen as such because of either availability or lack of temporal variation in the distribution of the predictor determined via qualitative procedures. Their distributions are displayed in supplemental Figure A.4 and this display illustrates that spatial variation. pHI and pAA were recorded in the year 2000 while radon is an average of test results that is considered current through October 31, 2014 according to South Carolina Department of Health and Environmental Control.

The predictor rainfall was chosen as an example of a temporal varying predictor because its measures of average annual rainfall were collected per county for each of the study years, and this is the way that the data source made it available. Supplemental Figure A.5 displays this predictor. The midline on this graph is the average over all of the study years; this aids in displaying the cyclical characteristic of the predictor of interest.

The remaining predictors (sun, UER, and pppov) vary spatio-temporally. The ST displays of these variables are included in supplemental Figures A.6–A.8 respectively. From these displays, one can see that these predictors offered a reasonable amount of ST variation. UER and pppov were the most collinear of the chosen predictors and the minimum and maximum values of the produced correlation matrix were 0.31 and 0.82 respectively.

Statistical Methods

Here we describe the methodology associated with the univariate and multivariate mixture models which were implemented using WinBUGS via the R package R2WinBUGS [26, 27]. Example WinBUGS code is included in the supplementary files.

We were interested in investigating whether the LBC or ORC incidence is best described by a model with spatial or ST variation. A type of mixture model first introduced by Carroll et al. [28] as a Bayesian model selection technique offered the ability to make this distinction. While that method was able to accommodate ST models, it simply selected one best ST model to be assumed across all spatial and temporal units. Hence, in this paper, we utilized a mixture parameter that can vary either spatially (p_i) or spatio-temporally (p_{ij}) for selection between two mixture components, one spatial (M_i^S) and one ST (M_{ij}^{ST}). A formula involving the ST mixture parameter is defined as follows:

$$log\left(\theta_{ij}\right) = p_{ij} M_i^S + \left(1 - p_{ij}\right) M_{ij}^{ST}$$

Alterations to these model ingredients led to the different fitted models used for analysis. Two variations of the model components were implemented in this analysis, and they involved 1) a random effects model without predictor variables (RE), which is defined as:

$$M_i^S = u_i + v_i \qquad M_{ij}^{ST} = \gamma_j + \phi_{ij}$$

and 2) a random effects model including predictors (PRED), defined as:

$$M_i^S = X_i' \boldsymbol{\beta}^S + u_i + v_i \qquad M_{ij}^{ST} = X_{ij}' \boldsymbol{\beta}_j^{ST} + \beta^T X_j + \gamma_j + \phi_{ij}$$

In both equations, the bold parameters indicate that they are vectors so that each predictor has its own parameter estimate. The parameter $\boldsymbol{\beta}^T$ is a single parameter associated with the temporally varying predictor. In these definitions, X_i' and X_{ij}' represent the i^{th} and ij^{th} values of the spatial or ST covariates. Further, the random effects were defined such that the spatial component, M_i^S, included a convolution structure by incorporating an uncorrelated heterogeneous term, $u_i \sim N\left(0, \tau_u^{-1}\right)$, and a correlated heterogeneous (CH) term, $v_i \sim N\left(\frac{1}{n_i}\sum_{i \sim l} v_l, \frac{1}{n_i \tau_v}\right)$. This CH term formulation followed that of an intrinsic conditional autoregressive (CAR) [29, 30] model where $i \neq l$, n_l is the number of neighbors for county i, and $i \sim l$ indicates that the two counties i and l are neighbors. For the ST term, M_{ij}^{ST}, the random effects were defined as a temporal random walk, $\gamma_j \sim N\left(\gamma_{j-1}, \tau_\gamma^{-1}\right)$, and an uncorrelated ST interaction term, $\phi_{ij} \sim N\left(0, \tau_{\phi j}^{-1}\right)$. Prior distributions for the regression parameters were such that $\beta \sim N\left(0, \tau_\beta^{-1}\right)$, and the precision parameters (τ) were such that $\tau^{-1/2} \sim Unif(0,4)$. This was found to be non-informative through sensitivity analysis that included comparisons to uniform distributions with larger ranges. All produced estimates related to the above parameters are means of the posterior distributions. Additionally, the PRED approach was only applied in the multivariate setting as the purpose for fitting in the univariate setting was simply to compare the separate and shared random effects. The details involved with implementation of the multivariate setting will be explained in more detail below.

Four different fitted models, F1 through F4, were produced by varying the way that the mixture parameter is structured. A summary of these variations is included in supplemental Table A.2. Models F1 and F2 were such that the mixture parameter varies spatially, while models F3 and F4 were such that the mixture parameter varies ST. F1 was a naive approach which assumes an uncorrelated linkage between the spatial and ST components of the model. F2 offered an extension of F1 via the implementation of a spatial structure on the mixture component which follows a CAR distribution. F3 allowed the mixture parameter to vary across space and time, but the correlation remained only spatially correlated. Note that the mixture parameter precisions were allowed to vary over time. Finally, F4 allowed the mixture parameter to vary across space and time.

Rather than only spatial correlation, a temporal correlation was also included in the mixture parameter structure that takes into account the values of the preceding and subsequent time points. The formulation of this temporal correlation is included in the supplemental materials as Formula A.1. Next, as indicated in Table A.2, these two random variables, with spatial and temporal correlation respectively, were averaged to create a normally distributed random variable that led to mixture parameters with correlation across both space and time.

When considering the multivariate approach to these mixture models, a separate likelihood was calculated for each disease of interest. This offered the ability for shared components to be imposed in the model formulation. In this study, we have introduced sharing between diseases via the CH effect, v_i, and the temporal random walk term, γ_j. An example of this multivariate formulation for LBC using F4 and the PRED approach to the mixture component structure is displayed below while a more complete formulation is included in the supplemental materials as Formula A.2.

$$
\begin{gathered}
log\left(\theta_{LBCij}\right)=\alpha_{LBC0}+p_{LBCij}M^{S}_{LBCi}+\left(1-p_{LBCij}\right)M^{ST}_{LBCij}\\
M^{S}_{LBCi}=X_i'\boldsymbol{\beta}^{S}_{LBC}+u_{LBCi}+v_i\\
M^{ST}_{LBCij}=X_{ij}'\boldsymbol{\beta}^{ST}_{LBCj}+\beta^{T}_{LBC}X_j+\gamma_j+\phi_{LBCij}
\end{gathered}
$$

All of these model formulation options were evaluated based on model goodness of fit (GoF) for comparison of the alternative formulation of the mixture components, F1 through F4, and the univariate versus multivariate structuring in the models. These alternative measures of GoF involved the deviance information criterion (DIC) [16] calculated using the deviance measures from the log likelihood of the Poisson distribution [31, 32] and the Watanabe-Akaike information criterion (WAIC) which makes use of the posterior predictive distribution as described by Gelman et al. [33, 34], and the mean squared error. However, the mean squared error results have been restricted to the supplementary materials. The GoF measures employed here are displayed in supplemental Table A.3 where mean() and Var() refer to the mean and variance across the MCMC samples. Note that all of the estimates produced for these measures are posterior means.

Results

Table 1 shows the goodness of fit results of the univariate RE models for each type of cancer while Table 2 shows the multivariate models broken down by model component structure (RE versus PRED) and disease. The shading in these tables represents the models that are comparable based on the disease outcome, and the bold faced numbers in these shades represent the best model in that table for that particular disease outcome according to the DIC and WAIC estimates. The italicized numbers represent the best fitting model when comparing the RE models to the PRED models within each multivariate modeling option (F1–F4 separately). Note that the predictor models were only fit for the multivariate modeling, so this only applies to Table 2. For the DIC and WAIC estimates with the univariate models, F4 was the best for all cases with the LBC data. Alternatively, the ORC results did not show such a clear best fitting model; in terms of DIC, F3 was the best, but in terms of WAIC, F2 was slightly better. When considering the multivariate models, F2 RE model appeared to be the best in terms of both DIC and WAIC. However, when looking at the individual diseases in this multivariate setting, F2 PRED model was the best for ORC in terms of both WAIC and DIC; in fact, ORC appeared to always perform best when the predictors were included in the multivariate setting.

Next, Table 3 displays the mixture parameter estimates across the South Carolina county map for F1 and F2. Supplemental Figures A.12–A.23 display the mixture parameter estimates for F3 and F4 over space as well as time for the univariate and multivariate models by disease. In these maps, a higher mixture parameter value indicates that the spatial mixture component explains the variation in cancer incidence better for that particular county, and at that time point for the multivariate setting, as that component is essentially being weighted more heavily by the mixture parameter. Alternatively, a lower mixture parameter value indicates that the ST mixture component explains the variation in cancer incidence more appropriately for that particular county.

For F1 and F2, the mixture parameter estimates were alike when comparing the univariate and multivariate models as well as when comparing the RE and PRED models in the multivariate setting. The largest difference occurred when comparing the F1 ORC RE and PRED models. Here, the range of the parameter across space became much smaller when the predictors are included in the model. Further, the estimates for F1 compared to F2 were alike with an obvious spatial smoothing present in the F2 estimates. For example, the interpretation using the mixture parameter associated with the multivariate fit of the F2 ORC RE model is as follows: The spatial mixture component appeared to explain ORC incidence better in the northwestern and tri-county (Charleston, Berkley, Dorchester counties) regions of the state while the ST mixture component appeared to explain ORC incidence better in the remainder of the state. Incidentally, those associated with the spatial component were the more populous and urban areas of the state. There were also similarities among F3 and F4 when comparing the mixture parameter from the univariate to multivariate models within each disease. Further, there was an obvious temporal smoothing effect for F4 when comparing appropriate F3 and F4 models. These space-time mixture parameters illustrate variation in the mixture parameter across time. Overall, the F4 models' results for both cancer classifications with univariate as well as multivariate modeling suggested that the relationship was best evaluated with the ST mixture component in the early study years and the spatial mixture component in the later study years. Thus, overall, the mixture parameter for these models appeared to be more temporally structured. The F3 results showed more variation for some years across the county map while other years displayed a very flat map of mixture parameter estimates.

Shared random effect estimates for the univariate and multivariate fits of F2 RE are displayed in supplemental Table A.4 and Figure A.24. The distribution of the random effect estimates were alike within each disease for all modeling combinations. CH effect in these displays indicated that there was an increase in incidence for both sets of cancers in the northern counties and that LBC had an increase in some of the southern counties as well. The temporal random walk estimates suggest that there was an overall increase in incidence over time for both diseases and that increase was more intense in LBC. These spatial and temporal random effect estimates also show that through multivariate modeling, the LBC outcome appeared to be dominating the results. Both the γ_j and v_i estimates associated with the multivariate fit of F2 RE resembled the estimates associated with the univariate fit of the F2 LBC RE model.

For the predictor models, there are a few fixed parameter estimates that proved to be well estimated based on a 95% credible interval, and their means and standard deviations are displayed in Table 4. Each of these predictors were ST in structure and they included: pppov, sun, and UER. All of the estimates associated suggest that there was a positive relationship between these three ST predictors and incidence of LBC and ORC. The positive relationship associated with the socio-economic predictors was expected as it has been noted in previous research [23, 25]. However, the positive association for sun was not. Note that these parameter estimates were also weighted by the mixture parameter, thus their exact relationship with incidence of cancer is dependent on space as well as time.

Finally, Figures 3 and 4 display $\hat{\theta}_{ij}$ for the LBC and ORC respectively. These figures are for comparison with the SIR values displayed in Figures 1 and 2 respectively. Note that the cut-points are the same for all four of these figures; however, only the first 5 were used for LBC. Figure 1 indicates that LBC incidence had a fairly strong spatial signature with lower values for the central counties. Further, it appears that LBC incidence was generally increasing with time. Alternatively, ORC had a much larger range in the SIR values indicating that its incidence rate was more variable across space and time. The LBC results are from the univariate fit of F4 while the ORC estimates are from the multivariate fit of F2 PRED as these were the best models for each disease. These estimates were calculated in nearly the same way as the SIR by substituting the estimate for $\hat{\mu}_{ij}$ $\left(\hat{\mu}_{ij}\right)$ in place of the observed incidences, thus $\hat{\mu}_{ij}/e_{ij}=\hat{\theta}_{ij}$. For LBC, $\hat{\theta}_{ij}$ was nearly identical to the SIR displayed in Figure 1. The $\hat{\theta}_{ij}$ associated with the best fit for ORC was not quite as good as those for LBC, but they were still quite alike the SIR displayed in Figure 2. However, it appeared as though the fitted model was not quite able to recover the large amount of variation present in the data as the range was much smaller and the values appear to be spatially smoothed for the $\hat{\theta}_{ij}$.

Discussion

These methods have several applications in epidemiology and public health. They offer a very flexible modeling approach which can accommodate and potentially select between a wide range of spatial or ST linear predictors that may be of interest in the study of multivariate cancer outcomes. Additionally, they offer the ability to make inference that varies spatio-temporally so that the model as a whole has a different meaning in relation to the outcome for each spatial and temporal unit. Further, they lead to improved modeling of rarer outcomes of interest.

Through the use of multivariate modeling to aid in the model fit for the rarer disease (ORC), these results also showed that the methods possess the ability to produce well estimated parameter estimates for those rarer diseases. However, they can be quite complex to interpret and the optimal transformation calculation for these parameters must be completed in WinBUGS. For example, in the ORC results presented in Table 4 for the multivariate fit of F2, the parameter estimate of $\beta^{ST}_{ORC,14,1}$ was 0.41 for year 2009 and was associated with the first ST predictor of interest, pppov. Next, the mixture parameter in the ORC results for the multivariate fit of F2 associated the first county (Abbeville County), notated as $p_{ORC,1}$, was 0.58, but because pppov was included in the ST mixture component, M^{ST}_{ij}, the weight associated was actually 1−0.58=0.42. Thus, every unit increase in the standardized version of pppov indicated that 1.19 times as many cases of ORC occur in Abbeville County during the year 2009 since 1.19= exp((1−0.58)*0.41). Yet, a transformation calculation performed in WinBUGS using the same parameters produced a posterior mean estimate of 1.77; this is the more appropriate value to interpret as it is the actual posterior mean of the transformed function rather than a transformation calculation performed with the posterior mean.

Similar calculations can be performed using each of the parameter estimates. Additionally, because this analysis was quite complex and the above calculation was somewhat intricate, these results should be interpreted with caution, for example the observed association with average daily sunlight. This positive association could have arisen because of confounding related to the fact that, for the most part, smoking must be done outdoors or that rural counties typically have more smokers and more outdoor workers. Further, these outdoor workers may also have more of the occupational exposures associated with respiratory cancers.

Based on the results above, there is a useful application of these methods for this case study both in the univariate and multivariate settings. In the univariate setting, the LBC data appeared to perform well and this is likely due to the fact that it is a more common disease with a reasonable number of incidences per county. Moreover, the multivariate results showed that modeling of LBC was not improved with the addition of predictors. This indicates that either incidence of LBC was solely explained through spatial, temporal, and ST random variation, the chosen predictors were not the best for modeling LBC, or a combination of the two. The multivariate setting led to improved model fits related to the rarer ORC outcome. Here, the additional information offered by the LBC model aided in improving the fit for the less common ORC via the implementation of shared spatial and temporal random effects.

As with any statistical method, the best model depends both on the outcome of interest and the actual model components. This was evident in the results above as the best choice in the univariate modeling of ORC was different from that of the univariate modeling of LBC as well as different from that of the multivariate modeling of the two diseases. Together, a simpler RE model where the mixture parameter varied spatially with a correlated spatial structure (F2) was important. However, when looking at the separate diseases within the multivariate setting, it appeared as though ORC itself was actually best fitting in the multivariate setting when predictors were also included in the formulation of F2.

The limitations present in this method include issues raised for mixture models in the past [28]. The first of these involves model misspecification and while this type of model formulation is better than some in this respect, it is still not without fault. Another of these issues involves the CH effect. Other studies mention that the inclusion of a correlated random effect can alter the fixed parameter estimates [35, 36], but this is still the conventional method in the disease mapping setting. Finally, this methodology is inflexible in some respects as it is limited to the a priori determination of the contents of the mixture components, M_i^S and M_{ij}^{ST}. This is an issue to be explored in future work.

Conclusion

This ST multivariate mixture model offered improved, very flexible, and informative inference for a rare disease of interest via the introduction of a more common and somewhat related disease to the modeling process. The best models in this exploration suggested that an increase in daily sunlight or percent of persons in poverty was associated with increased incidence of ORC while LBC was best explained via the inclusion of random effects only.

Further, the results indicated that ORC had a more spatial structure while LBC was more temporal. This difference in appropriate models could mean there are significant differences in the etiology between the two diseases. The approach developed here can be applied more generally to questions concerning incidence or prevalence of chronic disease outcomes across geographic areas over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by funding under grant NIH R01CA172805.

Abbreviations

LBC	Lung and Bronchus cancers
ORC	“other” respiratory cancers
pHI	proportion of persons with health insurance
pAA	proportion AA
UER	Unemployment rate
pppov	proportion of persons in poverty
sun	average daily sunlight
SIR	standardized incidence ratio
ST	spatio-temporal
RE	random effects model without predictor variables
PRED	random effects and predictors model
CH	correlated heterogeneous
CAR	conditional autoregressive
DIC	deviance information criterion
WAIC	Watanabe-Akaike information criterion

References

1. National Insitutes of Health. SEER stat fact sheets: Lung and bronchus cancer. Rockville, MD: National Institutes of Health; [cited 2016 14 Jan]
2. Hare TS. Space–Time Patterns of Respiratory Cancer Incidence and Mortality: Kentucky, 1969–2011. Pap App Geo. 2015; 1(4):333–41. DOI: 10.1080/23754931.2015.1012423
3. Kiberstis PA. Space, time, and the lung cancer genome. Science. 2014; 346(6206):204.doi: 10.1126/science.346.6206.204-d

4. Batista NE, O AA. Spatiotemporal analysis of lung cancer incidence and case fatality in Villa Clara Province, Cuba. MEDICC Rev. 2013; 15(3):16–21.

5. Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Statistics in medicine. 2000; 19(17–18):2555–67. DOI: 10.1002/1097-0258(20000915/30)19:17/18<2555::aid-sim587>3.0.co;2-# [PubMed: 10960871]

6. Knorr-Held L, Besag J. Modelling risk from a disease in time and space. Statistics in medicine. 1998; 17(18):2045–60. DOI: 10.1002/(sici)1097-0258(19980930)17:18<2045::aid-sim943>3.0.co;2-p [PubMed: 9789913]

7. Waller LA, Carlin BP, Xia H, Gelfand AE. Hierarchical Spatio-Temporal Mapping of Disease Rates. J Am Stat Assoc. 1997; 92(438):607–17. DOI: 10.1080/01621459.1997.10474012

8. Xia H, Carlin BP, Waller LA. Hierarchical Models for Mapping Ohio Lung Cancer Rates. Environmetrics. 1997; 8(2):107–20. DOI: 10.1002/(sici)1099-095x(199703)8:2<107::aid-env241>3.0.co;2-e

9. Mayo Clinic. About lung cancer: long term. Rochester, MN: Mayo Foundation for Medical Education and Research; [cited 2016 11 February]. Available from: http://www.mayoclinic.org/diseases-conditions/lung-cancer/multimedia/vid-20078621

10. Mayo Clinic. About lung cancer: causes. Rochester, MN: Mayo Foundation for Medical Education and Research; [cited 2016 11 February]. Available from: http://www.mayoclinic.org/diseases-conditions/lung-cancer/multimedia/vid-20078621

11. American Cancer Society. Lung cancer prevention and early detection. Atlanta, GA: American Cancer Society; [cited 2016 11 February]. Available from: http://www.cancer.org/cancer/lungcancer-non-smallcell/moreinformation/lungcancerpreventionandearlydetection/index

12. National Cancer Institute. General information about laryngeal cancer: Key points. Rockville, MD: U.S. Department of Health and Human Services, National Institutes of Health; [updated 19 May 2015; cited 2016 11 February]. Available from: http://www.cancer.gov/types/head-and-neck/patient/laryngeal-treatment-pdq

13. Cleveland Clinic. Mediastinal Tumor. Cleveland, OH: Cleveland Clinic; [updated 18 February, 2011; cited 2016 11 February]. Available from: http://my.clevelandclinic.org/services/heart/disorders/hic_mediastinal_tumors

14. National Cancer Institute. General information about nasopharyneal cancer: Key points. Rockville, MD: U.S. Department of Health and Human Services, National Institutes of Health; [updated 12 August 2015; cited 2016 11 February]. Available from: http://www.cancer.gov/types/head-and-neck/patient/nasopharyngeal-treatment-pdq

15. National Cancer Institute. Malignant mesothelioma. Rockville, MD: U.S. Department of Health and Human Services, National Institutes of Health; [cited 2016 11 February]. Available from: http://www.cancer.gov/types/mesothelioma

16. Lawson, AB. Bayesian disease mapping: Hierarchical modeling in spatial epidemiology. 2. Boca Raton, FL: CRC Press; 2013.

17. Cancer Incidence. Columbia, SC: South Carolina Community Assessment Network, South Carolina Department of Health and Environental Control;

18. Breslow, NE., Day, NE. The Design and Analysis of Cohort Studies. New York: Oxford University Press; 1987.

19. Area Health Resource Files (AHRF). Rockville, MD: US Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Workforce; 2003.

20. North America Land Data Assimilation System (NLDAS). Daily Sunlight (insolation) for years 1979–2011 on CDC WONDER Online Database. Centers for Disease Control and Prevention; 2013.

21. South Carolina Department of Health and Environmental Control. Average in home radon concentrations (pCi/L). Columbia, SC: South Carolina Department of Health and Environmental Control (SCDHEC); 2014.

22. National Oceanic and Atmospheric Administration. Climate at a Glance. Ashville, NC: National Centers for Environmental Information; [updated 6 March 2015; cited 2016 20 January]. Available from: http://www.ncdc.noaa.gov/cag/

23. American Cancer Society. Cancer facts & figures 2015. Atlanta, GA: American Cancer Society; 2015. [cited 2016 13 Jan]. Available from: http://seer.cancer.gov/statfacts/html/lungb.html

24. Giovannucci E. The epidemiology of vitamin D and cancer incidence and martality: A revew (United States). Cancer Causes Control. 2005; 16(2):83–95. [PubMed: 15868450]

25. National Cancer Institute. Cancer health disparities. Rockville, MD: National Institutes of Health; 2008. [cited 2016 14 Jan]

26. Thomas A, O'hara B, Ligges U, Sturtz S. Making BUGS Open. R News. 2006; 6(1):12–7.

27. Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D. The BUGS book: A practical introduction to Bayesian analysis. 1. Boca Raton, FL: CRC Press; 2013.

28. Carroll R, Lawson AB, Faes C, Kirby RS, Aregay M, Watjou K. Spatio-temporal Bayesian model selection for disease mapping. Environmetrics. 2016 Submitted.

29. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991; 43(1):1–20. DOI: 10.1007/bf00116466

30. Besag J, Green PJ. Spatial Statistics and Bayesian Computation. J Roy Stat Soc B. 1993; 55(1):25–37.

31. Gelman, A., Carlin, JB., Stern, HS., Rubin, DB. Bayesian Data Analysis. 2nd. CRC Press; 2004.

32. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J Roy Statist Soc B. 2002; 64(4):583–639. DOI: 10.1111/1467-9868.00353

33. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res. 2010 Dec.11:3571–94. 10.1.1.407.7976.

34. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Stat Comp. 2014; 24(6):997–1016.

35. Reich BJ, Hodges JS, Zadnik V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics. 2006; 62(4):1197–206. DOI: 10.1111/j.1541-0420.2006.00617.x [PubMed: 17156295]

36. Hodges JS, Reich BJ. Adding spatially-correlated errors can mess up the fixed effect you love. Am Stat. 2010; 64(4):325–34. DOI: 10.1198/tast.2010.10052

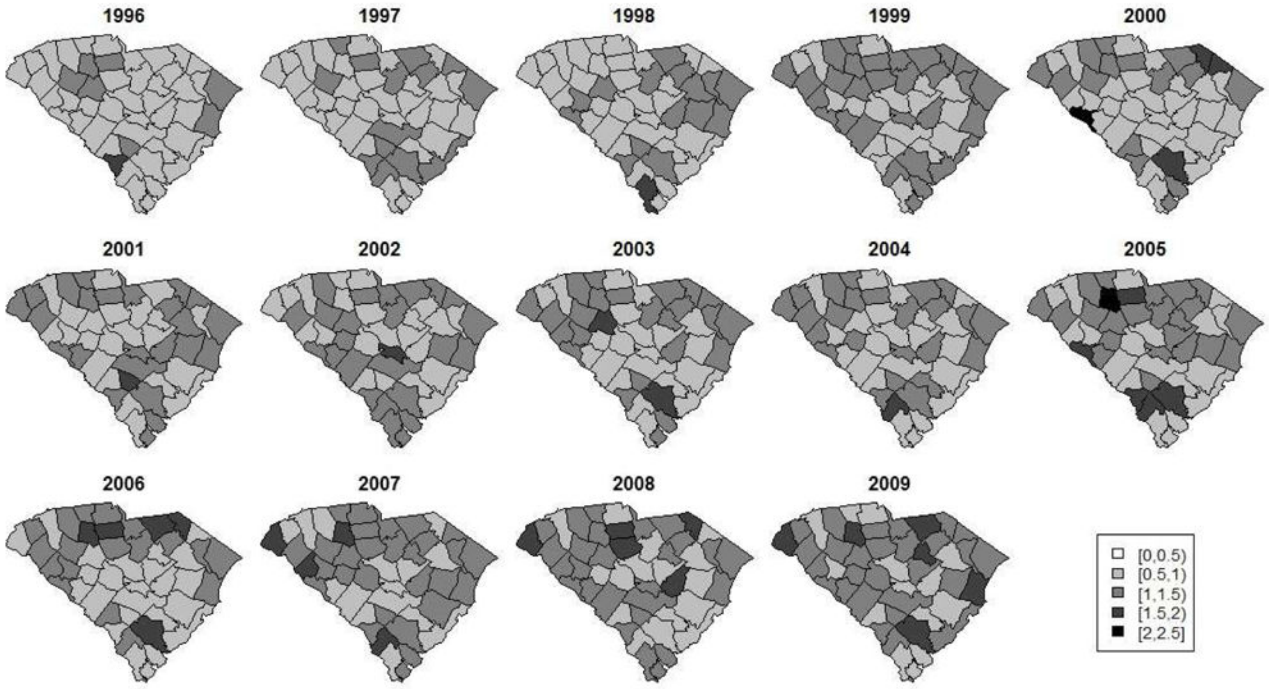


Figure 1.

SIR for LBC over time where an SIR of one indicates that the observed incidence is equal to that of the expected count for a particular county at a particular time.

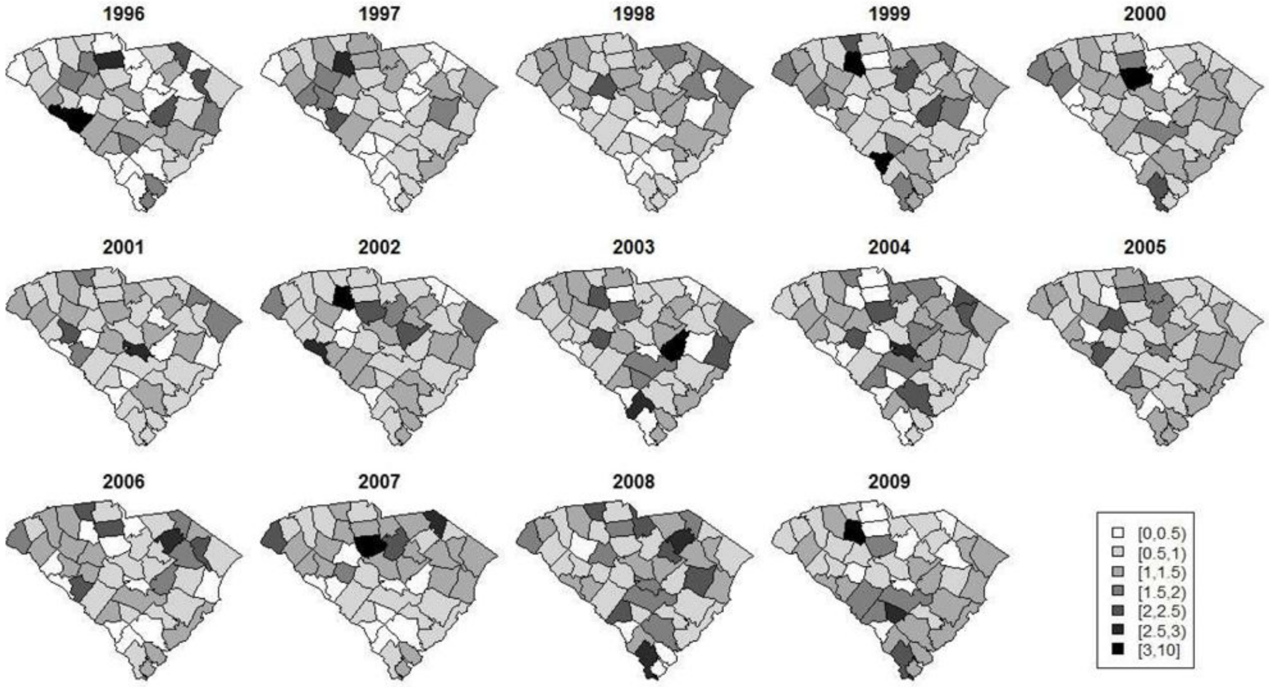


Figure 2.

SIR for ORC over time where an SIR of one indicates that the observed incidence is equal to that of the expected count for a particular county at a particular time.

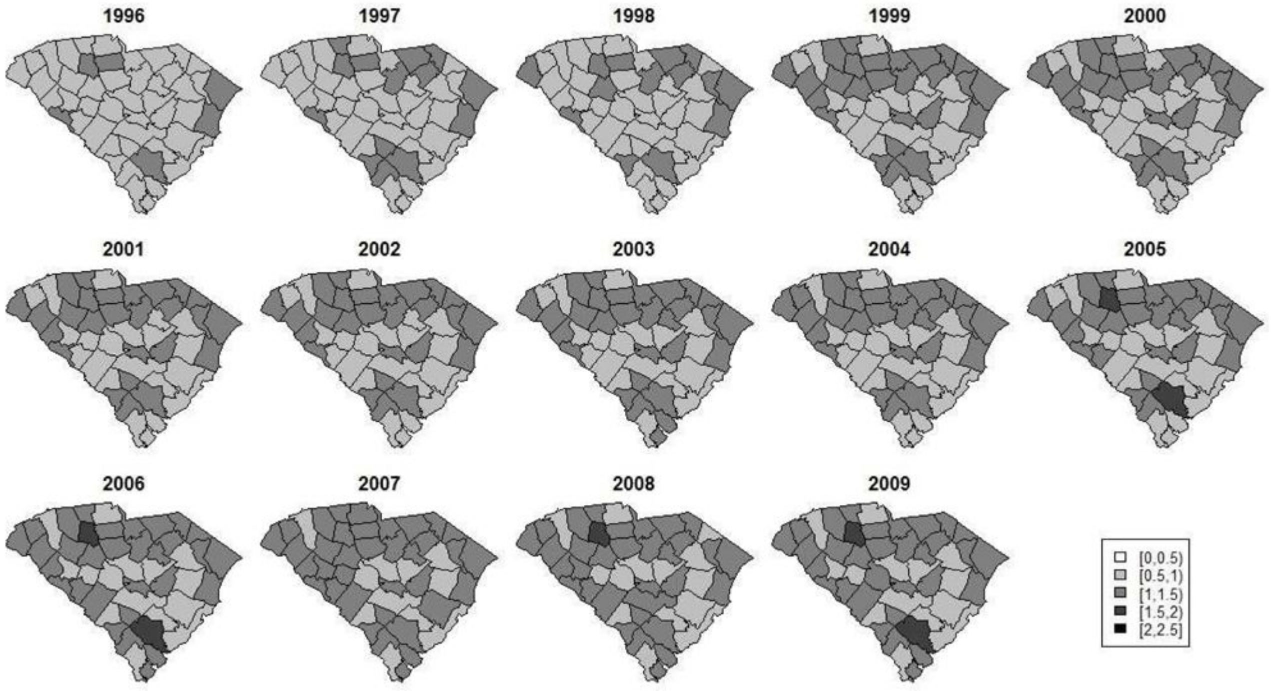


Figure 3.

$\hat{\theta}_{ij}$ for LBC over time with the univariate fit of F4 where $\hat{\theta}_{ij}=1$ indicates that the estimated value of μ is equal to the expected count for a particular county at a particular time.

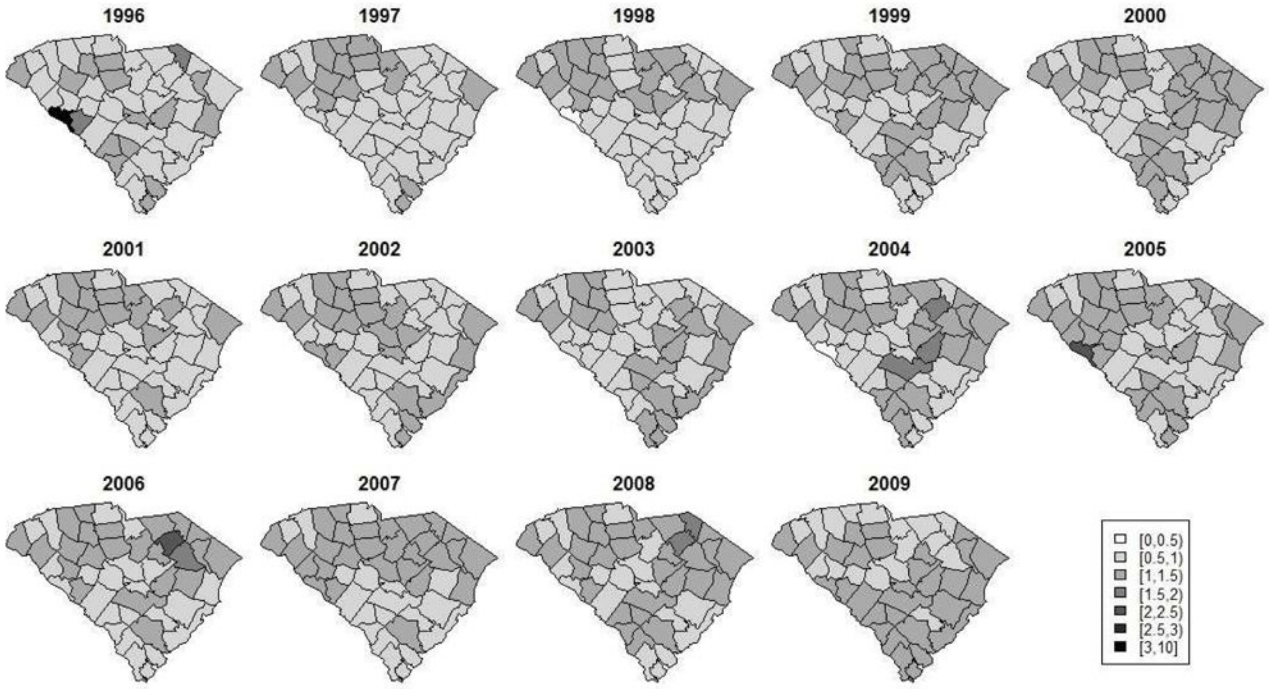


Figure 4.

$\hat{\theta}_{ij}$ for ORC with the multivariate fit of F2 PRED where $\hat{\theta}_{ij}=1$ indicates that the estimated value of μ is equal to the expected count for a particular county at a particular time.

Table 1

GoF measures for the univariate models.

Cancer	**Model**	***DIC***	$\frac{Var\left(\overline{D}\right)}{2}$	$\overline{D}$	***WAIC***	$\boldsymbol{pD_{WAIC}}$
ORC	F1	2955.54	309.05	2646.49	2768.39	105.87
	F2	2923.53	280.55	2642.98	**2759.20**	101.72
	F3	**2905.06**	256.20	2648.86	**2762.80**	99.81
	F4	2936.18	275.48	2660.70	2771.08	96.48
LBC	F1	4798.46	539.70	4258.76	4426.19	139.07
	F2	4555.94	301.72	4254.22	4398.97	123.98
	F3	4572.67	335.94	4236.74	4410.48	143.15
	F4	**4507.77**	271.39	4236.39	**4391.05**	129.11

The bold numbers indicate the best model when comparing models in the same shade (disease classification).

Table 2

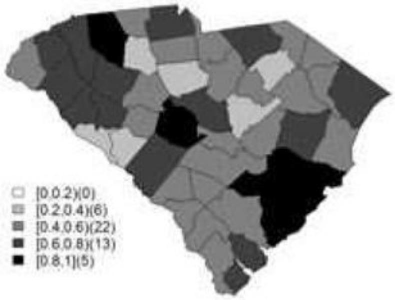
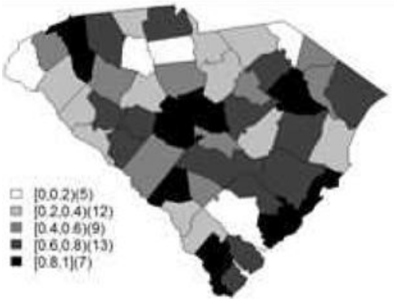
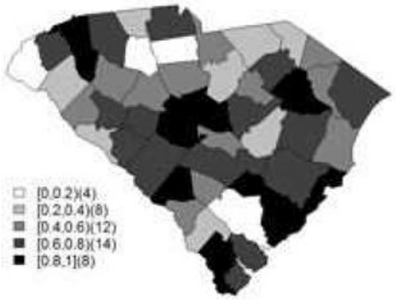
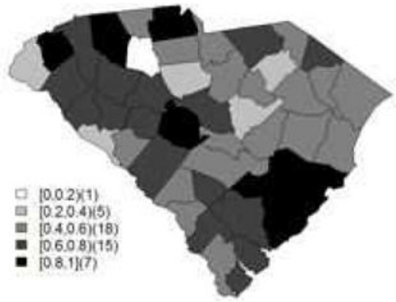
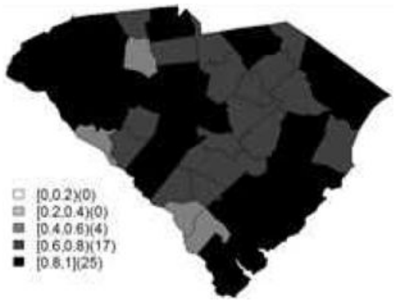
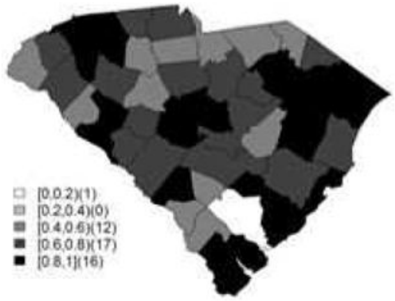
Goodness of fit measures for the multivariate models.

Model			***DIC***	$\frac{Var\left(\overline{D}\right)}{2}$	$\overline{D}$	***WAIC***	***pD***$_{WAIC}$
F1	RE	All	*7719.45*	808.71	6910.73	*7197.36*	244.42
		ORC	2922.65	284.64	2638.00	2755.33	101.86
		LBC	4871.34	598.61	4272.72	*4442.03*	142.57
	PRED	All	7738.87	885.62	6853.24	7210.14	292.98
		ORC	*2890.79*	298.40	2592.39	***2732.71***	119.23
		LBC	*4865.95*	605.08	4260.87	4477.43	173.75
F2	RE	All	***7382.22***	480.85	6901.37	***7148.08***	213.57
		ORC	2877.23	230.63	2646.60	2752.68	93.80
		LBC	***4539.08***	274.31	4254.77	***4395.40***	119.77
	PRED	All	7454.53	588.44	6866.10	7157.58	245.71
		ORC	***2844.60***	228.06	2616.54	***2742.48***	107.56
		LBC	4575.61	326.03	4249.57	4415.10	138.15
F3	RE	All	7468.32	568.97	6899.35	*7172.57*	232.08
		ORC	2896.68	246.67	2650.01	2754.20	91.97
		LBC	4615.48	366.14	4249.34	4418.37	140.11
	PRED	All	*7410.88*	550.41	6860.46	7410.88	550.41
		ORC	***2840.55***	211.68	2628.87	*2739.78*	95.65
		LBC	4578.17	346.57	4231.59	4420.98	153.62
F4	RE	All	7478.20	570.98	6907.22	***7152.13***	211.39
		ORC	2905.85	243.11	2662.74	*2758.13*	85.61
		LBC	*4552.83*	308.35	4244.81	***4394.00***	125.78
	PRED	All	*7413.01*	537.74	6875.27	7173.53	248.79
		ORC	*2877.20*	242.86	2634.34	*2754.23*	102.41
		LBC	4565.61	324.69	4240.92	4419.30	146.38

The bold numbers indicate the superior model when comparing models in the same shade (disease classification). The italicized numbers indicate the superior model when comparing RE and PRED for the same multivariate fitted model within the same shade.

Table 3

Posterior mean mixture parameter estimates for F1 and F2 models.

Model		Param	Univariate	Multivariate
F1	RE	p_{ORC}	 □ [0,0.2)(0) ▨ [0.2,0.4)(6) ▨ [0.4,0.6)(22) ■ [0.6,0.8)(13) ■ [0.8,1](5)	 □ [0,0.2)(5) ▨ [0.2,0.4)(12) ▨ [0.4,0.6)(9) ■ [0.6,0.8)(13) ■ [0.8,1](7)
		p_{LBC}	 □ [0,0.2)(4) ▨ [0.2,0.4)(8) ▨ [0.4,0.6)(12) ■ [0.6,0.8)(14) ■ [0.8,1](8)	 □ [0,0.2)(1) ▨ [0.2,0.4)(5) ▨ [0.4,0.6)(18) ■ [0.6,0.8)(15) ■ [0.8,1](7)
	PRED	p_{ORC}		 □ [0,0.2)(0) ▨ [0.2,0.4)(0) ▨ [0.4,0.6)(4) ■ [0.6,0.8)(17) ■ [0.8,1](25)
		p_{LBC}		 □ [0,0.2)(1) ▨ [0.2,0.4)(0) ▨ [0.4,0.6)(12) ■ [0.6,0.8)(17) ■ [0.8,1](16)

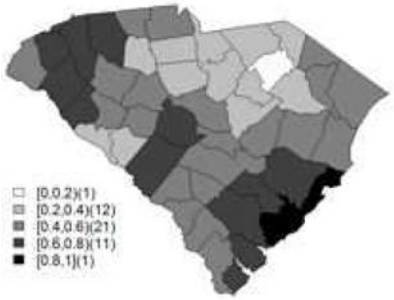
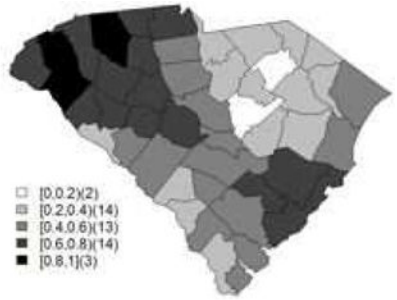
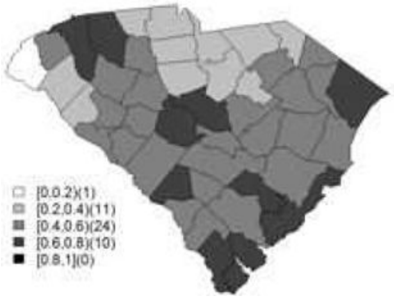
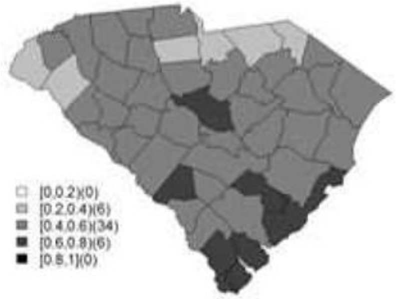
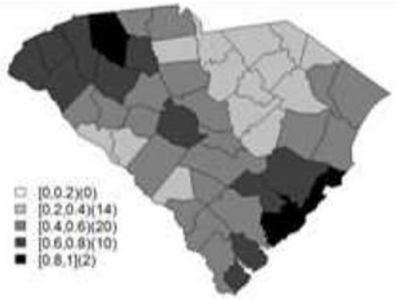
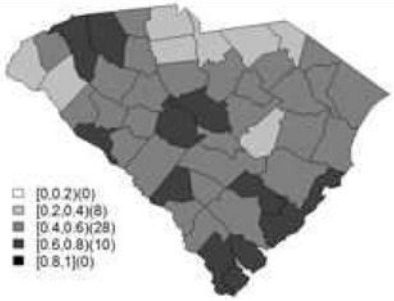
Model		Param	Univariate	Multivariate
F2	RE	p_{ORC}	 [0,0.2)(1) [0.2,0.4)(12) [0.4,0.6)(21) [0.6,0.8)(11) [0.8,1](1)	 [0,0.2)(2) [0.2,0.4)(14) [0.4,0.6)(13) [0.6,0.8)(14) [0.8,1](3)
		p_{LBC}	 [0,0.2)(1) [0.2,0.4)(11) [0.4,0.6)(24) [0.6,0.8)(10) [0.8,1](0)	 [0,0.2)(0) [0.2,0.4)(6) [0.4,0.6)(34) [0.6,0.8)(6) [0.8,1](0)
	PRED	p_{ORC}		 [0,0.2)(0) [0.2,0.4)(14) [0.4,0.6)(20) [0.6,0.8)(10) [0.8,1](2)
		p_{LBC}		 [0,0.2)(0) [0.2,0.4)(8) [0.4,0.6)(28) [0.6,0.8)(10) [0.8,1](0)

Table 4

Well estimated fixed parameter estimates for the multivariate PRED models written as posterior mean (standard deviation).

Parameter	Associated Predictor	F1	F2	F3	F4
$\beta^{ST}_{ORC,14,1}$	pppov	—	0.41 (0.19)	—	0.28 (0.23)
$\beta^{ST}_{ORC,9,2}$	sun	—	0.51 (0.22)	0.69 (0.23)	0.62 (0.27)
$\beta^{ST}_{ORC,13,2}$	sun	—	0.47 (0.22)	—	0.63 (0.31)
$\beta^{ST}_{ORC,1,3}$	UER	—	—	0.56 (0.25)	0.52 (0.25)
$\beta^{ST}_{LBC,2,2}$	sun	0.30 (0.15)	0.19 (0.10)	0.19 (0.09)	0.13 (0.07)
$\beta^{ST}_{LBC,3,2}$	sun	—	0.18 (0.09)	0.18 (0.08)	0.13 (0.07)
$\beta^{ST}_{LBC,14,2}$	sun	—	—	0.14 (0.07)	—
$\beta^{ST}_{LBC,9,3}$	UER	—	—	0.18 (0.08)	—