Multiscale modeling approach for hierarchical aligned aggregated small area health data

Mehreteab Aregay[1*], Andrew B. Lawson[1], Christel Faes[2]
Russell S. Kirby[3], Rachel Carroll[1], Kevin Watjou[2]
[1]Department of Public Health, Medical University of South Carolina, Charleston SC USA
[2]Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University
Hasselt, Belgium
[3]Department of Community and Family Health, University of South Florida, Tampa, FL, USA
[*] Corresponding author, email: aregay@musc.edu

Abstract

When data are aggregated from a finer to a coarser geographical level, there will be loss of information known as the scaling problem in geography. To address the scaling problem, we propose to use a joint convolution model that describes the risk variation at both the finer and coarser levels simultaneously by sharing both the correlated and the uncorrelated components. We compare our model with the naive approach that ignores the scale effect in real and simulated data in a range of criteria such as deviance information criterion (DIC), Watanabe-Akaike information criterion, and mean square prediction error (MSPE). We found that our multiscale model is better than the naive model.

1 Introduction

Often, it is of interest to study the spatial distribution of diseases at different geographical levels. For example, public health workers are interested in identifying areas which have a higher risk for a certain infection. Studying the geographical variation of diseases will help policy makers to allocate public health resources in a cost effective manner, to promote educational outreach programs, and to design effective public health interventions. Researchers have studied the geographical variation of disease using standardized mortality/morbidity ratios (SMR). However, this crude approach does not accommodate the correlations between neighbors. To overcome the limitation with SMR, Besag *et al.* (1991) [3] proposed a convolution model that allows the relative risk to be statistically modeled by including spatially structured and unstructured random effects into the model. Even though the convolution model has been widely used in spatial epidemiology, it does not accommodate the spatial scaling effect associated with the aggregations of data from a finer to a coarser level.

To account for scale (aggregation) effect, Kolaczyk and Haung (2001) [5] developed a multiscale modeling approach by decomposing the coarser level likelihood into individual components of local information. Their model assumes that the hierarchical partitions correspond to successive aggregation of an initial data space. Nevertheless, their approach assumes the effects at the higher level are fixed and not random. In addition, it is not flexible enough to adjust for neighbor effects. To overcome such issues, Aregay *et al.*. (2015a) [1] proposed joint multiscale models via a shared spatially structured component. However, the shared correlated component may not be flexible enough to fully address the scale effect. To allow for additional flexibility, Aregay *et al.* (2015b) [2] considered sharing both the correlated and uncorrelated components in the framework of mixture multiscale models. In this paper, we also share both the correlated and the uncorrelated components between the finer and coarser levels in the multiscale modeling approach. In addition, we compare our shared multiscale model with the naive approach that ignores the scale effect in real and simulated data sets.

2 Georgia Oral Cancer Data

We are interested in examining the incidence of oral cancer from the state of Georgia across the county and public health (PH) districts simultaneously. In particular, we aim to investigate whether we obtain a consistent oral cancer incidence at both the county and PH district levels. We chose the state of Georgia as it provides a reasonably large set of spatial units. In Georgia, there are 159 counties which are grouped into 18 PH districts that are used for adminstration of health care resources. The outcome of interest is the number of persons discharged from non-federal acute-care inpatient facilities for oral cancer in 2008. Since a public health district contains at least one county, there may be a grouping (contextual) effect, i.e., counties (children) located within the same PH district (parent) may behave similarly (Figure 3). Our analysis of these data is deferred to Section 4.

3 Multiscale Models

Researchers have developed multiscale models to address a scaling problem due to the aggregation of data from a finer to a coarser level. It is known that the information conveyed by the maps varies with scale. Hence, to include this scale effect, Louie and Kolaczyk (2006) [6] proposed to factorize the likelihood which contains the information of the scaling effect in a multiscale fashion under the assumed Poisson model. They assumed a multinomial distribution for the data at the finer level conditioning on the coarser level. This approach is limited to the assumption of having fixed coarser level effect. Moreover, their approach does not include the neighborhood effect into the model. To address those limitations, Aregay *et al.* (2015a) [1] proposed a multiscale convolution model that jointly describes the risk variations at multiple scale levels via a shared spatially structured component. The shared spatially structured component, however, may not be flexible enough to fully address the scaling effect. In this paper, we propose to share both the unstructured and the structured components to adjust for the aggregation (scale) effect. In the next section, we present our shared multiscale modeling approach as well as the independent multiscale model that ignores the scale effect.

3.1 Model 1: Adjusting for Scaling

This model assumes that by sharing the parameters that describe the characteristics of the parent (PH district) among the children (counties), we can include the aggregation (parent) effect into the model. Our multiscale modeling approach is based on a convolution model that contains both correlated heterogeneity (CH) and uncorrelated heterogeneity (UH). The CH terms explain the similarity between neighbored regions, i.e., it handles the neighbor effects, whereas the UH terms describe the random noise in the counties. To address the PH district effect, we share both the CH (u_j^{ph}) and the UH (v_j^{ph}) random effects of the PH districts among the counties within the PH district as follows:

$$\begin{aligned}
&y_i^c \sim \text{Poisson}(e_i^c \theta_i^c), \\
&\log(\theta_i^c) = \alpha_0^c + v_i^c + u_i^c + u_j^{ph} + v_j^{ph}, \\
&y_j^{ph} \sim \text{Poisson}(e_j^{ph} \theta_j^{ph}), \\
&\log(\theta_j^{ph}) = \alpha_0^{ph} + v_j^{ph} + u_j^{ph},
\end{aligned} \qquad (1)$$

where $y_i^c, i = 1, \ldots, 159$, is the county level count of disease and $y_j^{ph} = \sum_{i \epsilon j} y_i^c, j = 1, \ldots, 18$, is the j^{th} public health (PH) district level count of disease aggregated at the county level. In this model, u_i^c and v_i^c are the CH and the UH random effects at the county level, whereas u_j^{ph} and v_j^{ph} are the CH and the UH random effects at the PH district level, respectively. In addition, α_0^c and α_0^{ph} are the intercept at the county and PH levels, respectively. The linkage between these two levels is incorporated in the model by inheriting a shared CH u_j^{ph} and UH v_j^{ph} from the PH district into the county level model. Here, e_i^c and e_j^{ph} are the expected number of cases at the county and PH level, while θ_i^c and θ_j^{ph} are the relative risk at the county and PH district, respectively. For this model and for the other model below (Model 2), we have assumed a flat prior for the intercept parameters, α_0^c and α_0^{ph}. Further, the uncorrelated heterogeneity random effects, v_j^{ph} and v_i^c, were assumed to be normally distributed, i.e., $v_j^{ph} \sim N(0, \sigma_{vph}^2)$ and $v_i^c \sim N(0, \sigma_{vc}^2)$, whereas we assumed an intrinsic conditional autoregressive (ICAR) distribution for the correlated heterogeneity random effects, i.e., $u_j^{ph} \sim \text{ICAR}(\sigma_{uph}^2)$ and $u_i^c \sim \text{ICAR}(\sigma_{uc}^2)$. For the hyperparameters, $\sigma_{vc}, \sigma_{vph}, \sigma_{uph}$, and σ_{uc}, we considered a uniform prior distribution, $U(0, 100)$ [4].

3.2 Model 2: Ignoring the scaling effect

As we have described previously, Model 1 accounts for the scaling effect due to data aggregation from a lower to a higher geographical level. To investigate the impact of ignoring the scale effect, in this section, we present the simplified version of Model 1 without a shared component. Model 2 assumes separable convolution models at both the county and PH levels. There is no linkage to accommodate for the aggregation effect. Hence, Model 2 ignores the scale effect and it is of the form

$$\begin{aligned}
&y_i^c \sim \text{Poisson}(e_i^c \theta_i^c), \\
&\log(\theta_i^c) = \alpha_0^c + v_i^c + u_i^c \\
&y_j^{ph} \sim \text{Poisson}(e_j^{ph} \theta_j^{ph}), \\
&\log(\theta_j^{ph}) = \alpha_0^{ph} + v_j^{ph} + u_j^{ph}.
\end{aligned} \tag{2}$$

We assumed the same prior distributions for the model parameters as in Model 1. Here, the CH and the UH at the county level, u_i^c and u_i^c, describe the risk variation at the county level, while v_j^{ph} and u_j^{ph} explain the risk variation at the PH district level. Note that Model 2 does not include a random effect that can serve as a bridge to jointly link the two levels as in Model 1.

3.3 Model Assessment and Goodness of Fit

To compare the models, we use the deviance information criterion (DIC [7]) as well as Watanabe-Akaike information criterion (WAIC [8]). For a predictive accuracy assessment, mean absolute prediction error (MAPE) and mean square prediction error (MSPE) were used.

3.4 Simulation Study

The goal of this simulation study is to investigate the impact of ignoring the scale effect due to data aggregation, especially during the presence of a very strong contextual effect. To achieve this goal, we generated data within the state of the Georgia at the county level by imposing a very strong PH effect as follows:

$$\begin{aligned}
&y_i^c \sim \text{Poisson}(e_i^c \theta_i^c), \\
&\log(\theta_i^c) = \alpha_0^c + v_i^c + u_i^c + u_j^{ph} + v_j^{ph}.
\end{aligned} \tag{3}$$

To obtain the data at the PH level, we summed up the simulated data at the county level within the PH district, i.e., $y_j^{ph} = \sum_{i \epsilon j} y_i^c$. We assumed the variances of the random effects at the county level (σ^2_{vc} and σ^2_{uc}) to be small relative to the variances of the random effects at the PH level, σ^2_{vph} and σ^2_{uph}. Hence, we assumed the following values for the parameters: σ_{vc}=0.01, σ_{uc}=0.01, σ_{vph}=0.3, σ_{uph}=0.3, and α^c_0=0.1. This simulation mechanism allows for a very strong PH effect as shown in Figure 2.

The models discussed above (Models 1 and 2) were fitted to 200 simulated data sets using the Markov Chain Monte Carlo (MCMC) method with 15000 samples after the first 15000 samples were discarded from the analysis. To compare the models, the bias and MSE of the parameters were calculated.

To evaluate the predictive ability of the models, MSPE and MAPE were computed at each level and averaged over the 200 data sets. Additionally, the DIC and WAIC were calculated at each level to compare model goodness of fit. Finally, the computation time (CT) was extracted to compare the execution time for the models.

4 Results

4.1 Simulation Results

The results obtained from the models fitted to the data generated within the state of Georgia are shown in Tables 1 and 2. Model 1 is better than Model 2 as measured by DIC, WAIC, and PD (the number of effective parameters) at both the county and PH levels. In addition, the prediction ability of Model 1 is better than Model 2 as measured by MAPE and MSPE, especially at the PH level. Thus, the shared multiscale model, Model 1, describes the risk variation better than the independent multiscale model, Model 2. From Table 2, we can see that Model 1 produces more unbiased and precise estimates of the standard deviations of the CH and the UH at the county level as compared to Model 2. On the other hand, Model 2 provides more unbiased and precise estimates of the intercept and the standard deviations of the CH and the UH at the PH level as compared to Model 1.

To compare the models in terms of recovering the simulated relative risk for each county (see Figure 1), we computed the average relative risk over the 200 simulated data sets for each county (see Figure 2). We can see that the naive independent multiscale model, Model 2, does not recover the pattern of the simulated risk appropriately, whereas the shared multiscale model, Model 1, recovers the pattern of the simulated risk well. Furthermore, in some of the areas, Model 2 provides inconsistent risk estimates at both the county and PH levels, while Model 1 produces consistent risk estimates at both levels.

Table 1: *Simulation Study: Model fit and predictive accuracy results averaged over the 200 simulated data sets.*

Models	PD_{dic}		DIC		PD_{waic}		WAIC		MAPE		MSPE		
	county	PH district	county	PH district	county	PH district	county	PH district	county	PH district	county	PH district	CT
Model 1	**14.66**	**7.32**	**350.09**	**79.57**	**11.51**	**2.71**	**348.28**	**75.64**	0.86	2.65	1.88	**14.36**	180.97
Model 2	18.42	9.58	362.54	88.50	15.68	6.36	362.03	87.31	0.89	2.97	1.97	17.95	183.63

4.2 Application to Data

To investigate the benefit of including shared correlated and uncorrelated random effects to handle the scale problem, we applied the models discussed above to the Georgia oral cancer data example (Figure 3). The results are shown in Table 3. Using DIC, WAIC, MAPE, and MSPE, we can see that Model 1 outperforms Model 2 at the PH level. This is an expected result because the shared components recover the lost information at the PH level due to data aggregation from the county to the PH level. In addition, Model 1 is slightly better than

Table 2: *Summary of the bias and MSE of the parameters averaged over the 200 simulated data sets.*

Models	assumed values					bias						MSE					
	α_{0c}	σ_{uc}	σ_{vc}	σ_{uph}	σ_{vph}	θ_c	α_{0c}	σ_{uc}	σ_{vc}	σ_{uph}	σ_{vph}	θ_c	α_{0c}	σ_{uc}	σ_{vc}	σ_{uph}	σ_{vph}
Model 1	0.1	0.01	0.01	0.3	0.3	0.007	-0.169	**0.202**	**0.159**	0.203	0.103	0.100	0.042	**0.048**	**0.029**	0.062	0.032
Model 2	0.1	0.01	0.01	0.3	0.3	-0.006	**-0.144**	0.367	0.228	**0.139**	**0.009**	**0.036**	**0.029**	0.159	0.064	**0.035**	**0.016**

Table 3: Model fit and predictive accuracy results for Georgia oral cancer data.

Models	PD_{dic}		DIC		PD_{waic}		WAIC		MAPE		MSPE	
	county	PH district	county	PH district	county	PH district	county	PH district	county	PH district	county	PH district
Model 1	**28.07**	**8.68**	**483.64**	**107.87**	**24.29**	**3.97**	**483.98**	**104.29**	1.39	**4.71**	5.0	**36.65**
Model 2	31.33	11.32	485.46	114.63	26.77	6.98	485.99	112.62	1.37	5.03	4.85	42.19

Model 2 at the county level. The pattern of the risk estimates obtained from both models are shown in Figure 4 indicating that Model 1 provides more consistent estimates at both the county and PH levels as compared to Model 2.

5 Conclusion

In this paper, we addressed the scaling problem due to data aggregation using a joint multiscale model by sharing both the correlated and the uncorrelated components. We also compared the shared multiscale model with the independent multiscale model that ignores the scale effect. When there is a very contextual effect, the naive approach that ignores the contextual effect results in a poor estimate of the pattern of the relative risk. On the other hand, accounting for the aggregation (scale) and contextual effects recovers the simulated risk very well. Furthermore, ignoring the scale and contextual effects produces inconsistent results at both the finer and coarser levels, whereas adjusting for the scale and contextual effects provides consistent results at both levels.

Although we managed to handle the scale effect using the shared components, our approach has the following limitations: (1) the oral cancer data set is heavily influenced by many zeros. Hence, extending our approach to account for overdispersion due to excessive zeros remains our further research, (2) our approach does not quantify the amount of scaling effect; we plan to measure the scale effect using a correlation structure, and (3) the current formulation does not allow for an evaluation of the evolution of diseases for each region, but our approach could be easily extended to accommodate spatiotemporal variation in the model.

Finally, we conclude that jointly modeling the risk variation at different geographical levels is very useful to obtain more accurate risk estimates for public health planning purposes. In addition, our method is easily implemented by public health practitioners in standard software. Hence, we recommend employing the methodology described in this paper to take into account the scale and contextual effects during spatial modeling for data collected at different geographical levels.

6 Acknowledgments

The authors would like to acknowledge support from the Nation Institutes of Health via grant R01CA172805. The third and last authors also acknowledge support from the the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

References

[1] M. Aregay, A. Lawson, C. Faes, and R. Kirby. Bayesian multiscale modeling for aggregated disease mapping data. *Statistical Methods in Medical Research*, 1:1–20, 2015.

[2] M. Aregay, A. Lawson, C. Faes, R. Kirby, R. Carroll, and K. Watjou. Spatial mixture multiscale modeling for aggregated health data. *Biometrical Journal*, 00:1–27, 2015.

[3] J. Besag, J. York, and A. Mollié. Bayesian image restoration with applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43(1):1–59, June 1990.

[4] A. Gelman. Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 3(1):515–533, 2006.

[5] E. Kolaczyk and H. Haung. Multiscale statistical models for hierarchical spatial aggregation. *Geographical Analysis*, 33(2):95–118, April 2001.

[6] M. M. Louie and E. Kolaczyk. A multiscale method for disease mapping in spatial epidemiology. *Statistics in Medicine*, 25:1287–1306, October 2006.

[7] D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Series B*, 64:583–616, 2002.

[8] S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, October 2010.

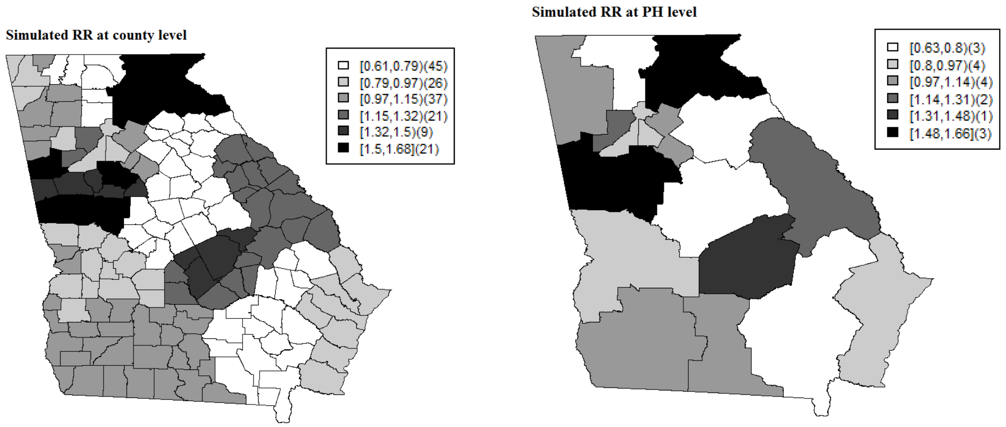

Figure 1: Simulated relative risk at county (left panel) and PH district (right panel).

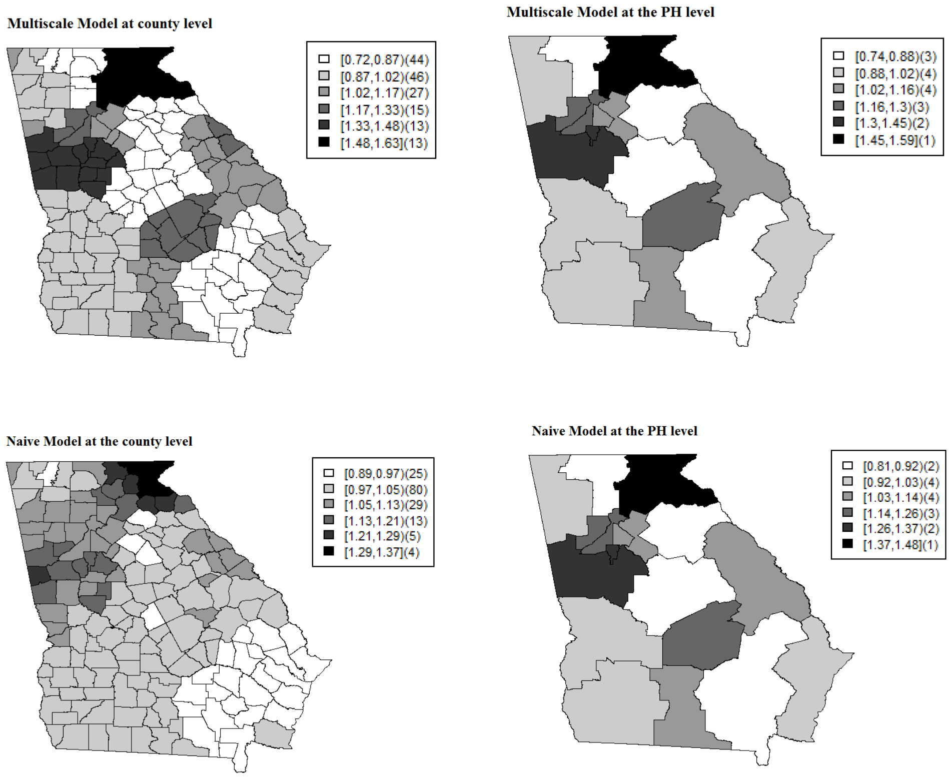

Figure 2: Fitted relative risk averaged over the 200 simulated data sets using the naive model (Model 2) and multiscale model (Model 1) at both the county and PH district levels.

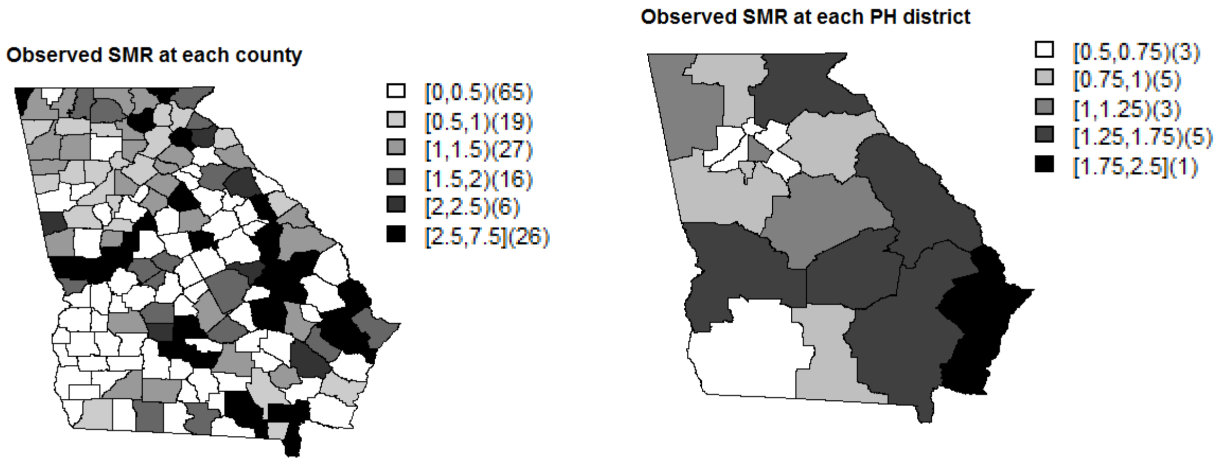

Figure 3: *Georgia oral cancer data. Observed standardized mortality ratio (SMR) at each county and public health (PH) district.*

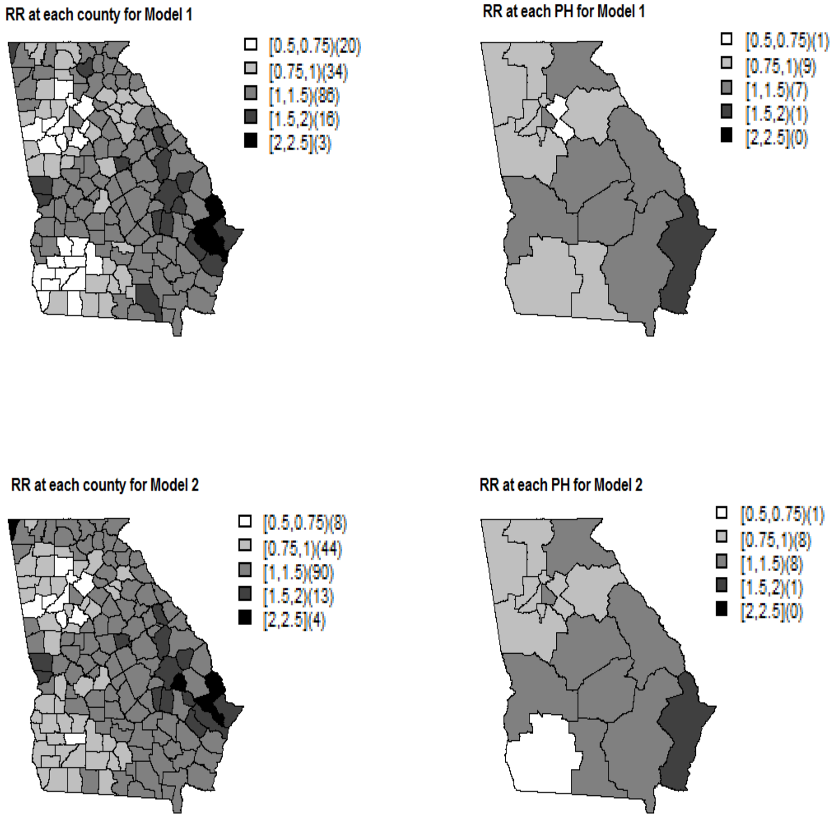

Figure 4: *Georgia oral cancer data. Relative Risk (RR) at each county and public health (PH) district.*