County-Level Small Area Estimation using the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS)

Van L. Parsons, Nathaniel Schenker
Office of Research and Methodology, National Center for Health Statistics,
Centers for Disease Control and Prevention, 3311 Belcrest Rd, Hyattsville, MD, 20782

Abstract

This paper looks at two modeling approaches to a small area estimation problem, a hierarchical Bayesian procedure and a generalized linear mixed model. The problem is to produce U.S. county-level estimates for mammography prevalence using combined data from the NHIS and BRFSS surveys. Several comparisons of the two approaches are provided.

Key Words: Bayesian, random effects, mixed models

1. Introduction[1]

A recent collaborative project among statisticians at the National Center for Health Statistics (NCHS), the National Cancer Institute (NCI), the University of Michigan and the University of Pennsylvania focused on developing Bayesian methodologies to produce small area estimates for the prevalence of cancer risk and cancer screening factors for counties in the United States. Data sources for this project were the Behavioral Risk Factor Surveillance System (BRFSS), an ongoing telephone survey of the health behaviors of adults in the United States, and the National Health Interview Survey (NHIS), a face-to-face household interview survey. An outcome of this project was the development of a hierarchical Bayesian model for combining data from the BRFSS and NHIS surveys to produce small area county-level estimates of prevalence. The details of this model appear in Raghunathan et al. (2007).

As there are numerous modeling methodologies for producing small area estimates, see Rao (2003), it was decided to compare the proposed Bayesian approach to that of using a Generalized Linear Mixed Model (GLMM). The GLMM is a modern, multi-level modeling technique that has gained popularity due to computational advancements; see Gelman and Hill (2007). The recent development of the R (2008) package *lme4* provided a flexible means for fitting a wide range of models and was chosen for this study. The results of this Bayesian and GLMM comparison were presented as a poster session, and this paper provides a condensed summary of the tables, graphs and maps of the poster.

2. A Comparison of a Bayesian and GLMM Small Area Models for County-Level Mammography Screening Prevalence

2.1 Comparison of BRFSS and NHIS Surveys

Table 1 provides summary information, with an emphasis on the contrasts, of these two surveys. Detailed information about BRFSS and NHIS is available at http://www.cdc.gov/brfss/ and http://www.cdc.gov/nchs/nhis.htm, respectively.

[1] The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention

For these two surveys we were interested in county-level estimates of the percentage of women age 40 and over who have had a mammogram in the past 2 years. The mammography screening questions were asked during different years for the two surveys, so the data years 1997-1999 and 2000-2003 were pooled to define two collapsed time periods. Using the SUDAAN software, the direct county-level estimates for each survey were computed along with the standard errors. For the NHIS estimates the weights were not poststratified to any control totals. For the proposed Bayesian and GLMM modeling, the standard errors were used to estimate effective sample sizes for the county-level estimates (Table2). For each county, socio-economic-demographic (SED) covariates were available (Table3).

The BRFSS direct county-level estimates of mammography prevalence tend to be greater than the corresponding NHIS direct estimates over both time periods. At the National level this is also true, with the difference about 7 percent between BRFSS and NHIS estimates. We attributed this significant discrepancy to a bias within the BRFSS system. Our modeling procedures discussed in the next section attempt remove to this bias.

2.2 A Bayesian and GLMM small area approach to county-level estimation

The BRFSS survey system provides sample for about 99% of all counties, but the BRFSS direct estimates have possible deficiencies: bias for landline-only sample, system biases (51 operationally different surveys, large non-response, telephone sample, and telephone-mode effect) and small samples for many counties. While the NHIS samples only about 800 counties, for those that are sampled we can make the assumption that the county direct NHIS estimates listed in Table 2 are unbiased. This assumption that the direct NHIS estimates are unbiased along with associated county-level SED covariates allows us to create models that separate a hypothesized BRFSS bias and borrow strength from the totality of data to produce model-based county-level estimates.

A starting point was to make the following conditional distributional statement about the entries of Table 2: Given some model-specified stochastic parameters, say $\boldsymbol{B} = \boldsymbol{b}$, as fixed, the county-level response p times the effective sample of size n has a binomial distribution. The stochastic nature of the variable $\boldsymbol{B}$ will be modeled under Bayesian and GLMM frameworks.

2.2.1 Bayesian Model

For this approach the arcsine transformation was used on the p's of Table 2. The specifics of the model chosen are described in detail in Raghunathan et al. (2007), but the following hierarchical model, (Bayesian Model Equation) provides an indication of the structures imposed on the direct estimates listed in Table 2. The posterior distributions of the model parameters, θ, ϕ, δ, were computed at the county level using an MCMC sampling algorithm. Custom-programming was done using the Gauss software package (Aptech Systems, 2003). We shall abbreviate this model in the remaining text by BAYES.

2.2.2 GLMM model

We used the GLMM framework in the R package *lme4* (Version: 0.999375-20) using the logit as the link function for the binomial. In terms of fixed and random effects the GLMM model can be specified by

Fixed Effects: intercept, **x (**covariates), survey and time,

County-level Random Effects: intercept, survey and time, and

State-level Random Effects: intercept and survey.

Reference points: survey = 0, 1 for NHIS and BRFSS
time = 0, 1 for 1997-1999 and 2000-2003.

Table 1. Comparisons of BRFSS and NHIS Surveys

	BRFSS	NHIS
Type	State-level, **Telephone only**	National, face-to-face, multistage cluster sample
Sample size/year	150-250 K Households 2000+ Households per state	30-40 K Households 50-5000 Households per State (Non-state design)
Cost/response	Low	High
Sponsor and Data Collection Organizations	CDC/States	NCHS/Census
Response rate	**Lower**	Higher
Coverage	**Landline** Telephone Residential Households, **About 99%** of counties	Households (including dormitories and group quarters) Sample contains about **800** of 3000+ counties
Available Geographical Information	State (public) County (on request)	4 Regions (public), State/County (restricted access) NCHS Research Data Center

Table 2. Direct Estimates of Proportions for County *d* in sample (1997-1999 and 2000-2003)
(time period index "*t*" omitted**)**

Domain	Direct Estimates		Effective Sample Sizes	
	NHIS	BRFSS	NHIS	BRFSS
Telephone Households	p_{Td}	p_{Bd}	n_{Td}	n_{Bd}
Non-Telephone Households	p_{Nd}		n_{Nd}	
All Households	p_d		n_d	

Table 3: County-Level SED Covariates from Multiple Sources (independent of NHIS and BFSS)

Small MSA status	Percent persons high school graduates
Non MSA status	Percent persons college graduates
Per capita property taxes	Percent persons below poverty
Per capita expenditures	Civilian labor force unemployment rate
Total Social Security benefit recipients	Per capita income
Number of serious crimes	County population density
Number of social service establishments per capita	Population size
Newspaper readership rate	Percentage of persons work commute 30+ min
Median effective buying income index	Percentage over 65
Per capita personal income	Percentage 1-person households
Percent blue collar workers	Percentage households with children
Percent black	Percent persons living in urban area
Percent Hispanic	Median home value

Bayesian Model Equation (BAYES) (details omitted)

$$\begin{pmatrix} \sin^{-1}\sqrt{p_{Tdt}} \\ \sin^{-1}\sqrt{p_{Ndt}} \\ \sin^{-1}\sqrt{p_{Bdt}} \end{pmatrix} \sim N_3\left[\begin{pmatrix} \theta_{dt} \\ \phi_{dt} \\ (1+\delta_{dt})\theta_{dt} \end{pmatrix}, \begin{pmatrix} \frac{1}{4n_{Tdt}} & \frac{\rho_t}{4\sqrt{n_{Tdt}\, n_{Ndt}}} & 0 \\ & \frac{1}{4n_{Ndt}} & 0 \\ & & \frac{1}{4n_{Bdt}} \end{pmatrix} \right] \leftarrow \text{Within-area model}$$

$$\begin{pmatrix} \theta_{dt} \\ \phi_{dt} \\ \delta_{dt} \end{pmatrix} \sim N_3(X_{dt}\,\beta, \Sigma), \quad \longleftarrow \quad \text{Between-area model}$$

where

X_{dt} = area-level covariates and Indicators for period,

$\boldsymbol{\theta} = (\beta, \Sigma)$ = unknown parameters,

vague prior distribution for $\boldsymbol{\theta}$.

The parameters θ, ϕ, δ correspond to telephone sample, non-telephone sample and bias parameters.

2.2.3 Differences in the Bayesian and GLMM models

The Bayesian methods allow for a finer level of parameterization than traditional methods. Counties with sparse data are compensated by the prior distribution. In the GLMM framework the non-telephone NHIS data was too sparse at the county level to allow for a distinct non-telephone component. Thus, only the direct full-county NHIS estimates were modeled using the GLMM approach. Furthermore, the lack of NHIS county data limits the GLMM models to main effects. The "state" random effect helps to express the nature of the BRFSS state survey system and was included for GLMM. At the time of this research the customized program we used for Bayesian analysis was too rigidly defined to easily include any additional state parameters in a timely manner and they were not included. The fixed "survey" component in the GLMM model must be considered as a somewhat overall bias parameter aggregating all types of bias discussed in section 2.2. Since BAYES and GLMM models are slightly different, we will not make any statements as to which one is preferred, but focus on the strengths and weaknesses of each approach. Furthermore, both the BAYES and GLMM models should be considered as exploratory attempts, with many alternative options still to be investigated.

3.0 Findings

The state of California is somewhat representative in showing the types of characteristics the model-based estimates will have. BRFSS has samples in most counties. By the nature of having highly populated areas, many of the California counties are in the NHIS sample with certainty; others are sampled, so it is possible to have counties with small or no samples. Figure 1 shows the BAYES and GLMM California county level estimates along with the original data. The horizontal reference lines of 0.65 and 0.70 represent the national prevalence at time periods 1 and 2 respectively. Figure 2 shows estimated levels of the coefficients of variation (CV) of the county-level estimators. Figure 3 shows the totality of the model-based estimates for the 1997-1999 time-period.

3.1 Observations

- Figure 1: Both models demonstrate a high degree of smoothing of the original data. The GLMM model had strong global "survey" and "time" fixed effects upon which the corresponding random effects had only moderate impact. For example, all differences in the county-level time prevalence are positive. The BAYES did not smooth the data as much as the GLMM model; it tracked the NHIS data somewhat more closely.
- Figure 2: The second time period had larger sample sizes which explains the general tendency of having smaller CV's in the second time period. CV patterns within BAYES or GLMM method were not as consistent for BAYES as for GLMM. For example, the GLMM second time period county-level CV was always smaller than the CV from the first time period, while the BAYES time period CV's reversed order for some counties. For those counties with NHIS sample less than 20, the CVs of the GLMM are less than those of BAYES within each time period, which is an evidence of a higher degree of smoothing in the GLMM.
- Figures 1 and 2: For the counties with large NHIS samples, the two methods appear to be somewhat consistent with each other.
- Figures 3a and 3b: The maps show a substantial smoothing over the map of the original BRFSS data (omitted for space reasons). One apparent difference between the GLMM and BAYES is the inclusion of the "state" random effect in the former model. Some states show a distinct separation by boundary, for example, the adjacent states of Mississippi and Alabama show a distinct separation for GLMM, but not for BAYES.

3.2 Conclusions

- Both models appear to produce estimates in some proximity to each other. Independent estimates are not available to assess the degree of bias.
- The selected GLMM is a stronger data "smoother" than the BAYES.
- Both models provide examples of the methods but not the final model. Refinements are needed.

Acknowledgments:

This project was a spinoff of the project mentioned in section 1.0. The authors thank the team listed in Raghunathan et al. (2007) for the Bayesian estimates.

References

Aptech Systems (2003). *Gauss: Advanced Mathematical and Statistical Systems*. Version 5, Black Diamond, CA.

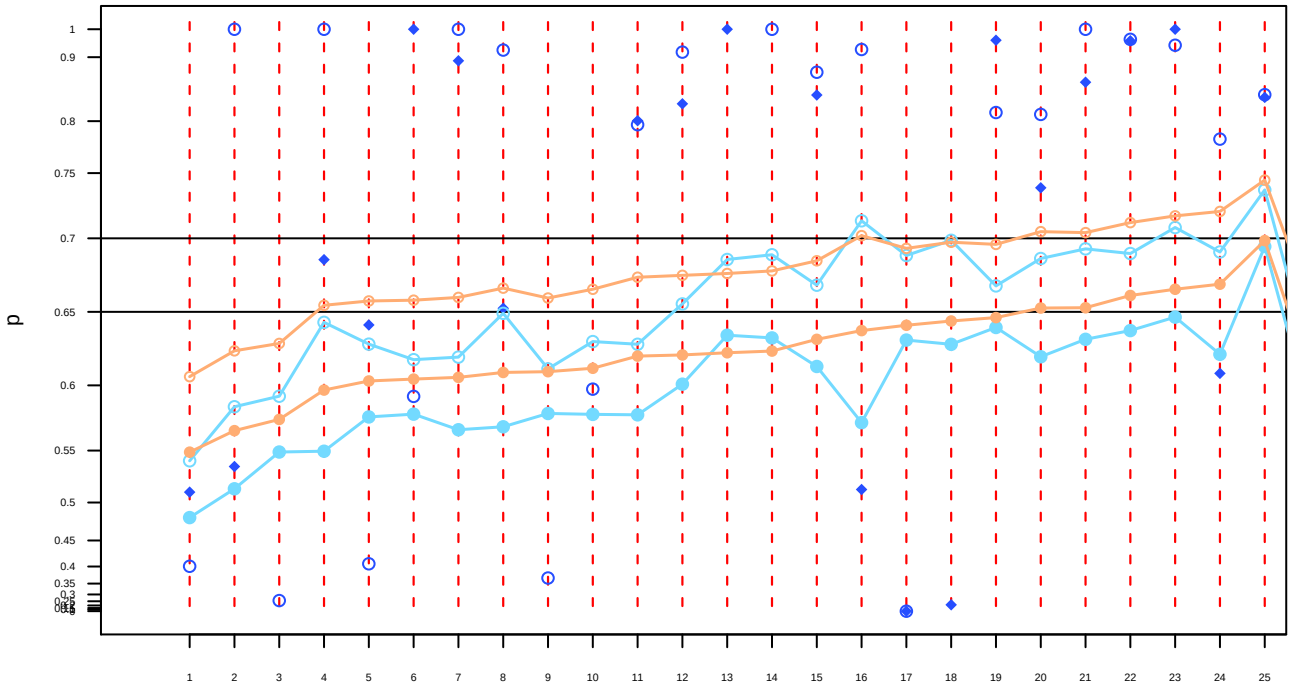
Gelman A, Hill J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raghunathan TE, Xie D, Schenker N, Parsons V, Davis WW, Dodd K, Feuer EJ. (2007). Combining information from multiple surveys for small area estimation: a Bayesian approach. *Journal of the American Statistical Association,* 102, 474-486.

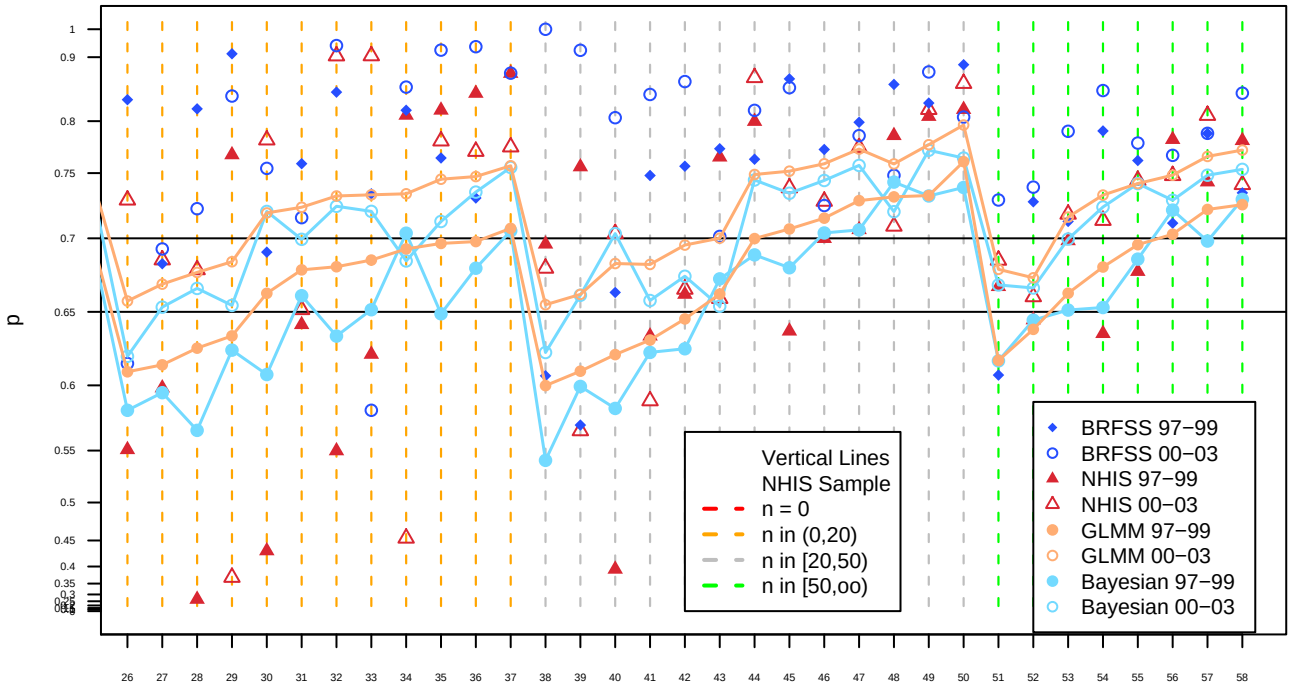
Rao, JNK. (2003) *Small Area Estimation*, Wiley

Figure 1: Bayesian and GLMM Model Fittings for CA County Estimates
Women 40+, Had Mammography Within Past 2 years
1997−1999 and 2000−2003



CA Counties − Blocked by NHIS Sample Size − Ordered by GLMM Fitted p

Figure 1: Bayesian and GLMM Model Fittings for CA County Estimates
Women 40+, Had Mammography Within Past 2 years
1997−1999 and 2000−2003



CA Counties − Blocked by NHIS Sample Size − Ordered by GLMM Fitted p

Figure 2: Bayesian and GLMM Coefficients of Variation (CV) for Fittings for CA County Estimates
Women 40+, Had Mammography Within Past 2 years
1997−1999 & 2000−2003

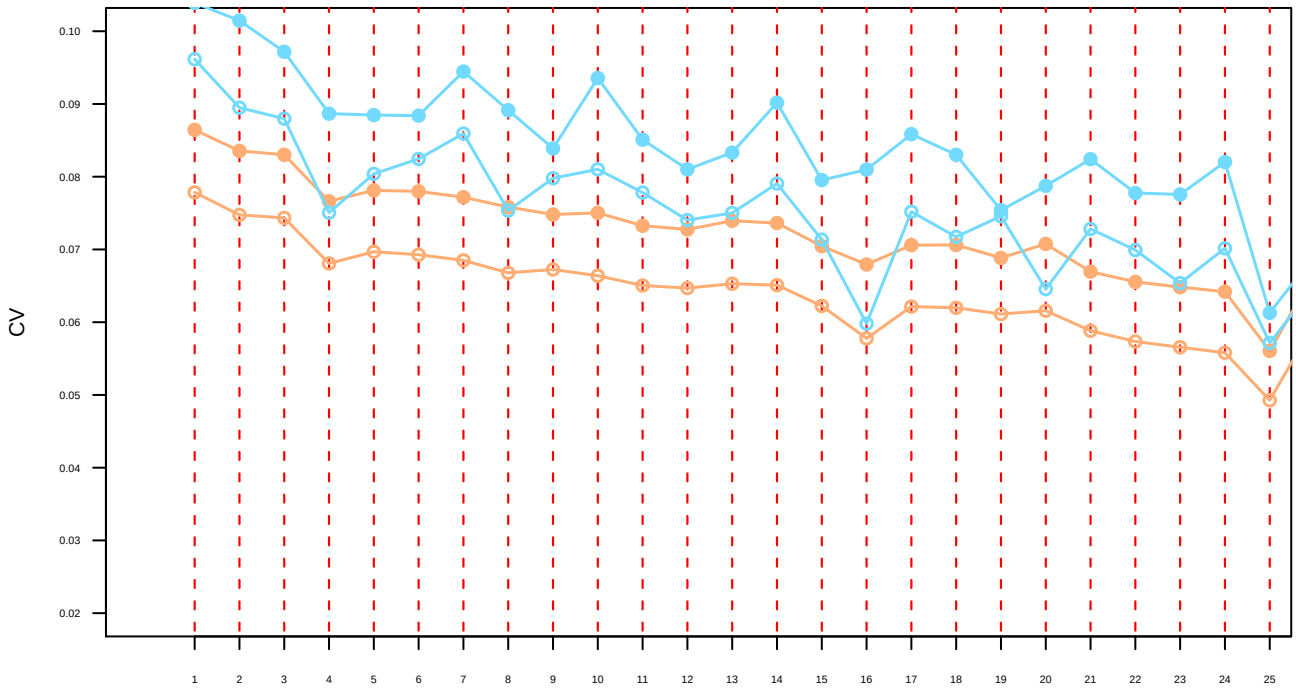


Figure 2: Bayesian and GLMM Coefficients of Variation (CV) for Fittings for CA County Estimates
Women 40+, Had Mammography Within Past 2 years
1997−1999 & 2000−2003

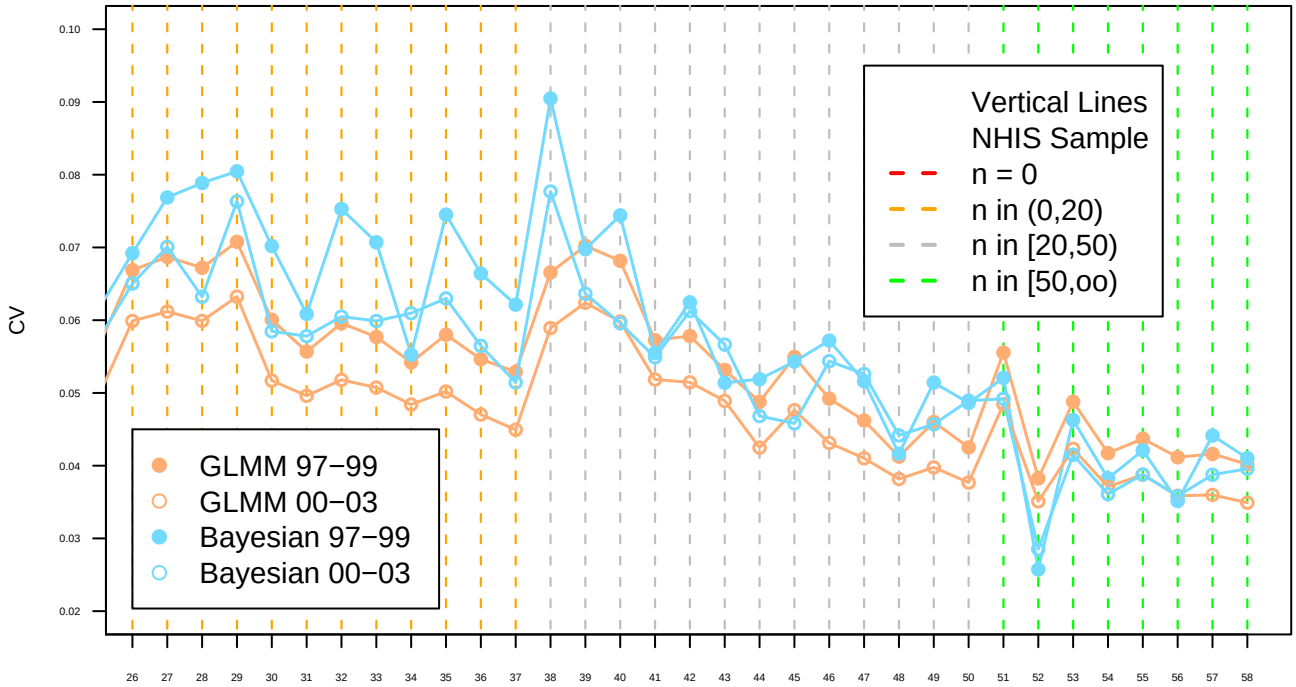
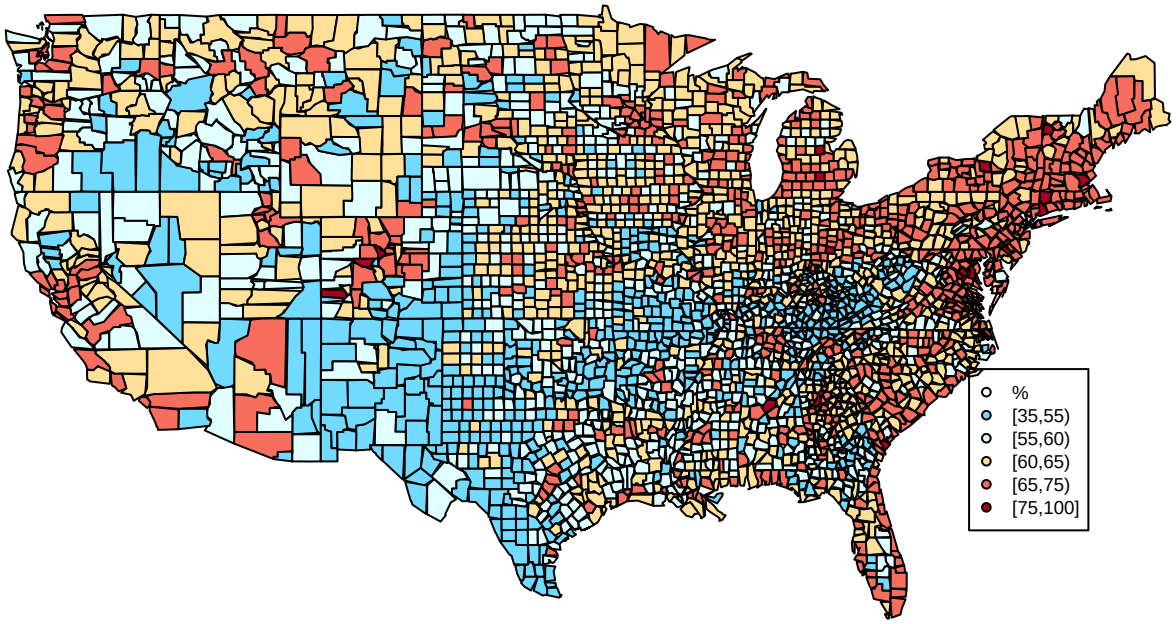
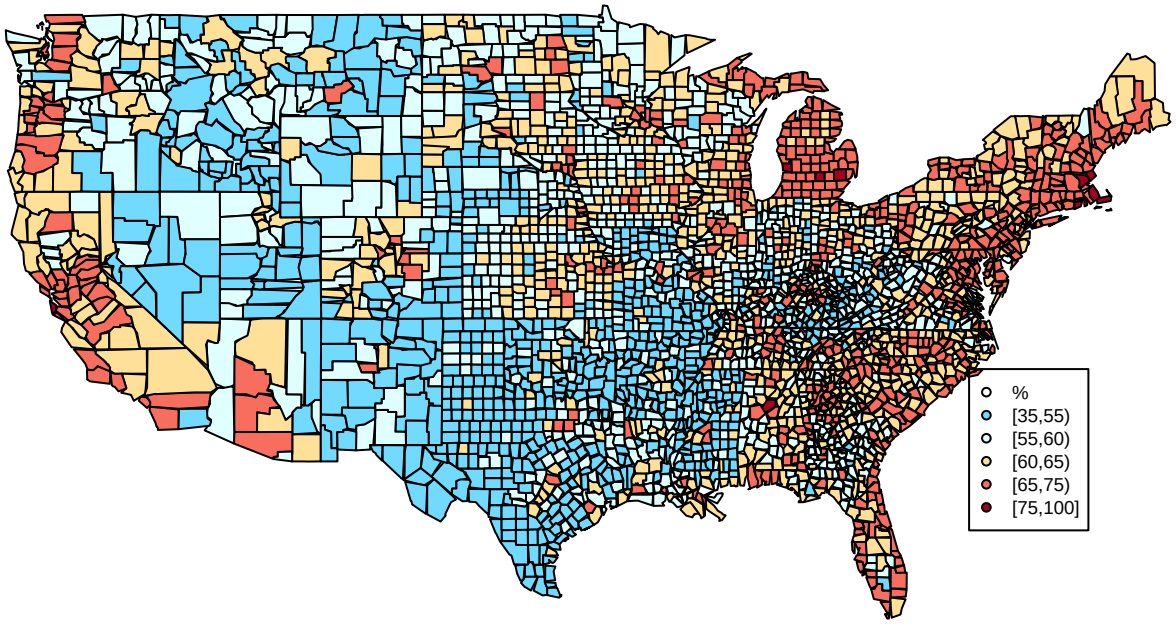


Figure 3a: Bayesian Model: Percent of Women 40+ having Mammography in past 2 years, 1997−1999



Bayesian County Quartiles : 56% ,61% , 65%, range: 35% − 82%

Figure 3b: GLMM Model: Percent of Women 40+ having Mammography in past 2 years, 1997−1999



GLMM County Quartiles : 55% ,60% , 64%, range: 36% − 78%