Small Area Estimation Technique for the Oklahoma Health Care Insurance and Access Survey

November 2009

Foreword

This technical report describes the Small Area Estimation technique employed by SHADAC for the 2008 Oklahoma Health Care Insurance and Access Survey (OHIS). The analysis was used to develop county-level estimates of uninsurance based on data from the 2008 OHIS. This report is adapted from the technical appendix provided as a deliverable for the project. Note that a similar analysis was conducted by SHADAC for the 2004 OHIS (reported in 2005) using a different methodology; a comparative assessment is included here.

Introduction

In generating these 2008 county-level Small Area Estimation (SAE) estimates of the rate of uninsurance for the 77 counties in Oklahoma, we have employed a methodological approach that is both new and significantly different from the one used in our 2005 report. This report describes this new approach in detail and discusses the reasons why we adopted it.

Constraints

Pertinent to the application of SAE, an important feature of the OHIS in both 2004 and 2008 has been its relatively small survey sample size (3804 non-elderly in 2008) relative to its numerous counties (77). For example, in 2008 half of the counties have a sample size of 25 or fewer, 30% have 11 or fewer and 70% have sample sizes of 47 or fewer non-elderly respondents. Small sample sizes (or no sample at all) are, of course, a natural characteristic of SAE applications. These 2008 county-level data are, however, *sparse*, and that has important implications for the precision of any SAE estimates as well as for choosing the methodological approach that's appropriate for deriving a set of policy-useful SAE estimates.

Evaluation of SAE Approaches

As part of this 2009 project to generate SAE estimates for the 77 Oklahoma counties we undertook a rigorous search of the literature on SAE methodologies. We also conducted an extensive evaluation of the new candidate for a SAE methodological approach that we identified from the literature. We conducted this evaluation by assessing the SAE estimates of this new methodology alone and also relative to the set of SAE estimates that we obtained employing the old methodology used in 2005.

New SAE Approach

The new approach identified in the literature and through an Internet search is a Bayesian SAE modeling approach that has been developed and made available through a project called BIAS, short for **B**ayesian methods for combining multiple **I**ndividual and **A**ggregate data **S**ources in observational studies. This project is based at the Department of Epidemiology and Public Health, Imperial College, London. The models we have used were presented last summer as "Bayesian Small Area Estimation for policy making and policy assessment", by V. Gomez-Rubio, N. Best, S. Richardson. and P. Clarke, all of the Department

of Epidemiology and Public Health, Imperial College London. This presentation was made at the Research Methods Festival, Oxford University, England, 3 July 2008.

The Bayesian statistical methodology has been around for a long time, since 1763 in fact. But it has only been since the advent of very powerful computers that were inexpensive enough to be widely available that the Bayesian methodological approach has taken off. Indeed, its burgeoning applications have brought about a major change in how people make decisions in many areas and disciplines, such as informatics, medicine, genetics and the Internet. There are now software packages devoted exclusively to Bayesian modeling, the best known of which is WinBUGS, which stands for Windows-based **B**ayesian-Inference **U**nder **G**ibbs **S**ampling. It was developed by a team of biostatisticians at Cambridge University, England. We used WinBUGS to estimate all our Oklahoma SAE models.

The underlying guiding principle of all SAE modeling, and Bayesian SAE in particular, is the idea of 'borrowing of power, or borrowing of information" from many sources to achieve more reliable estimates for small areas—with small data representation—than would be possible from the use of just these often very small survey data samples alone. In our Bayesian models this "borrowing of power, or borrowing of information" takes three distinct forms:

- First, we have substantially more information about the true value of the uninsurance rate for the entire state than we have for individual counties within the state, especially so for the smallest counties. Bayesian models optimally balance the reliability of these individual direct estimates of counties—with their sometimes very small sample sizes—with the much larger sample available for the state as a whole. Thus Bayesian estimates by themselves are an optimal blend of the often not very reliable county estimates and the much more reliable state-wide estimate. In this way the Bayesian county-specific SAE estimates of uninsurance involve borrowing of information from the state-wide rate. In particular, when the county survey sample is large, the Bayesian SAE estimate for that county will rely heavily on the data for that large-sample county. But when the county survey sample is quite small, the Bayesian SAE estimate for that county will 'shrink' that small-sample county estimate toward the statewide mean. The amount by which the small-sample county estimate is shrunk toward the statewide mean is optimally determined by the size of the sample.

- A second source of "borrowed information" comes about from the estimation of regression models that include variables that predict the likelihood of being uninsured based on the entire state survey data set. Thus the strong, significant relationship between the likelihood of being uninsured and poverty status, or income level or other variables can be used to 'borrow information' from these relationships and bring them to bear on individual counties to achieve more reliable estimates of SAE estimates.

- Finally—and restricted to Bayesian SAE models alone—we have models that not only borrow information from the overall state-wide rate of uninsurance and from the relationship between predictors of uninsurance in our regression models, but we can also assess whether the patterns of uninsurance rates in counties surrounding the county of interest are strong enough to allow us to "borrow information" from these geographically close areas. This type of 'spatially correlated' adjustments to rates has been very effectively applied to the estimation of prevalence rates for various diseases for small areas. We use this method of 'spatially

correlated' adjustments to rates of uninsurance in our Bayesian models as well. It is, again, another type of "borrowed information".

Evaluation of Alternative SAE Models

For the second form of "borrowed information" —through the use of a regression model estimated from the entire survey data set—there are two basic types of models that one can use. One can use data on individual respondents (N =3804) and estimate the relationship at that individual-level between the probability of being uninsured and various characteristics measured in the survey such as employment status, education level attained, and employer size. [These are called *Unit Level* models.] Or one can use survey data aggregated to the county level (N =77) for the uninsurance rates and use county-level data on predictors of county uninsurance rates like county average income or county proportion of the population below poverty. [These are called *Area Level* models.] In general, each type of model has its advantages and disadvantages.

However, in the particular case of survey data sets that have many very small samples at the county level (Area), *Area Level* models provide in general more reliable estimates. This is true since it becomes in general very difficult to reliably estimate the means of the explanatory variables with very small samples. For example, the proportion of a county's residents employed in large or small employers if estimated with very small numbers of observations can and often will differ dramatically from what the true county mean might be if a larger sample were available. We also observed empirically in our Oklahoma data that this disadvantage was very important for the actual cases of many small county estimates. *Area Level* models that rely on data from outside the survey—for example census data on poverty rates or average income—do not have this problem. However, a disadvantage with these *Area Level* models that use external data is that in general you can only find one or two variables that are significant and, importantly, the external data needs to come from a year recent enough to when the uninsurance data were collected to be effective.

Our extensive assessment of these two types of models easily suggested, however, that the *Area Level* models provided better SAE estimates, and consequently we used an *Area Level* model.

As part of our full evaluation of SAE models, we also assessed the relative advantages of a Bayesian model compared to the SAE methodology used in our 2005 Oklahoma report. In that 2005 SAE analysis we employed a *Unit Level* model with a random effect for the county and we estimated this model with the statistical package MLwiN. Using a number of criteria, we judged that the Bayesian *Area Level* model outperformed the previously-employed *Unit Level* model with a random effect (MLwiN).

In addition to this advantage in terms of the greater *a priori* plausibility of the county SAE estimates—generated by the Bayesian *Area Level* model compared to the previously-employed *Unit Level* model with a random effect (MLwiN)—there are other advantages in using a Bayesian model. Specifically, with a Bayesian approach one has *direct* measures of the uncertainty of each county's SAE estimates. These Bayesian measures of uncertainty are referred to as Credible Intervals.[1] If we use $\boldsymbol{Unin}\%$ to refer to the unknown, true uninsurance rate in county A, then we can directly determine the values $\boldsymbol{Unin}\%_{\boldsymbol{L}}$ ***and*** $\boldsymbol{Unin}\%_{\boldsymbol{H}}$ for which we can say that the true, unknown value of the uninsured rate in

[1] Bayesian Credible Intervals are not the same thing as conventional Confidence Intervals, but they function in approximately the same way. They allow for more meaningful statements to be made about uncertainty than conventional Confidence Intervals allow.

county A, $\boldsymbol{Unin}\%$, has a 95% probability of falling within the bounds set by the values $\boldsymbol{Unin}\%_{\boldsymbol{L}}$ $\boldsymbol{and}$ $\boldsymbol{Unin}\%_{\boldsymbol{H}}$. More formally, we can identify from the Bayesian results the values $\boldsymbol{Unin}\%_{\boldsymbol{L}}$ $\boldsymbol{and}$ $\boldsymbol{Unin}\%_{\boldsymbol{H}}$ for which the $\boldsymbol{Prob}(\boldsymbol{Unin}\%_{\boldsymbol{L}} < Unin\% < \boldsymbol{Unin}\%_{\boldsymbol{H}}) = \mathbf{0.95}$.

As measures of uncertainty, these Bayesian Credible Intervals integrate, effectively, *all* sources of our uncertainty about our SAE estimates. They reflect the uncertainty about the underlying values of the data on the number uninsured in the county. They also reflect the uncertainty that arises concerning the magnitude of the regression model's coefficients, which can have important impacts on the SAE estimates obtained. Finally and in our special Bayesian models, they also reflect the uncertainty about the values of the 'spatially correlated' adjustments discussed above.

As such, these Credible Intervals provide important, useful guides for policy-makers. It's critical that policy-making based on these SAE estimates explicitly acknowledge the reliability of these SAE estimates as expressed in the values of these Bayesian Credible Intervals. Put another way, the size of these Bayesian Credible Intervals reflect the limits to which conclusions can and should be drawn from these SAE estimates.

For several reasons it was not possible to derive useful measures of uncertainty for the 2005 SAE estimates in our earlier report. Clearly, having the capability to derive these measures for the present study—from our Bayesian models—represents an important improvement in the policy utility of our SAE approach.

In addition to these Bayesian Credible Intervals, there's an additional benefit from using our Bayesian modeling strategy. Specifically, as described in more detail below, we have developed models for both the 2004 data and the 2008 data. Given our Bayesian orientation, we have constructed an overall model in which the 2004 data and the 2008 data are estimated in parallel, which yields SAE estimates for both years. Given the provision of these 'new' estimates from the 2004 data and the 2008 estimates, a natural question for policy is whether specific counties have experienced increases or decreases in their SAE estimates of uninsurance over this time period. The advantage of an overall model—which includes both the 2004 and 2008 data—is that the difference in county-specific estimates can be directly modeled. Consequently, we provide estimates of these estimated differences from 2004 to 2008 in county-specific SAE estimates of uninsurance. In addition—and again to guide policy-makers in their use of these estimated differences—we provide Bayesian Credible Intervals providing our uncertainty about the values of these estimated differences from 2004 to 2008 in SAE estimates of uninsurance. Of course, since both years of SAE estimates are combined through the creation of this estimated difference of uninsurance rates, the degree of uncertainty in these differences effectively combines the uncertainty from both. Consequently, and although most of these Bayesian Credible Intervals for the estimated differences are large, they provide important statements of the limitations of what can and cannot be concluded from these comparisons.

Model Description

In 2005 we used the entire Oklahoma survey to generate our SAE county estimates of uninsured, including the data from elderly respondents. Since—at the request of the State—we are estimating our SAE county estimates of uninsured in 2008 using only the non-elderly and also since we have changed SAE modeling strategies, we have generated sets of county SAE estimates of uninsurance from both the 2008 and the 2004 surveys.

Our two SAE models, for 2004 and 2008 data, have the following features:

2005 Model

Although we assessed a number of variables from census data for inclusion in our model, only the average income in the county and the % of the county's population below poverty were significantly related to the mean uninsurance rate. Since these two variables are so highly correlated, however, only one of them could be included in our model. After evaluating the sets of SAE estimates generated with both, we chose the model with county mean income in thousands (IncomeK). The table below gives the coefficient values and significance for this model.

	Coefficient	SD	Ratio Coeff/SD
Intercept	0.12	0.35	0.339
IncomeK	-0.0423	0.01	-3.91

2009 Model

Again we assessed a number of variables from census data for inclusion in our model, but as expected none were significant for the 2008 county uninsurance rates. We found, however, that the county uninsurance rate in 2004 was an important and significant predictor of the county uninsurance rate in 2008. Consequently we used this variable to 'borrow information' in a temporal sense over this 4 year period. The table below gives the coefficient values and significance for this model.

	Coefficient	SD	Ratio Coeff/SD
Intercept	-2.28	0.29	-7.76
Unins_2004	3.47	1.28	2.71

Summary Statistics of Results

While we discuss the results of both sets of 2004 and 2008 SAE estimates in detail in the report, here we present several summary measures that provide important, relevant information with which to evaluate these Bayesian SAE estimates.

We begin with descriptive statistics for the Bayesian SAE estimates of county uninsurance rates in 2008 and compare these with the estimates generated from using the 2008 survey data to *directly* estimate these rates of uninsurance (referred to as direct estimates). [The units over which these summary statistics are computed are the counties (N= 77).]

Summary Statistics of 2008 Direct Estimates and Bayesian SAE Estimates of Uninsurance Rates

	Direct Estimate	Bayesian SAE Estimates
Min	0.0%	12.3%
Max	80.0%	29.3%
Simple Mean	19.1%	19.2%
Med	17.6%	18.8%
sd	15.0%	3.2%

As can be readily seen, there is a very large range—from a min of 0% to a max of 80%—in the Direct Estimates of uninsurance rates across the 77 counties in 2008 (often with very small survey samples) . This is, of course, to be expected and is the reason why SAE modeling techniques are undertaken. This large variability is also reflected in the large standard deviation for these direct estimates of 15%-points.

As is also readily seen, the Bayesian SAE estimates 'shrink' this range down to a min of 12.3% and a max of 29.3% in accord with the optimal properties of Bayesian estimates as discussed previously. Notice that the two *simple* means[2] of these 77 estimates do not differ between the direct and the Bayesian SAE estimates, as should be the case for any SAE estimator.

One way of evaluating a set of SAE estimates involves computing the following test statistic:
Take each county's SAE estimate of the uninsurance rate and multiply this by the county's population and sum all 77 of these SAE-estimated county *number of uninsured*. This sum should come close to the aggregate number of uninsured derived for the state as a whole by multiplying the survey-weighted estimate (18.9%) by the statewide population. For our Bayesian SAE estimates, these two estimates—the county-population weighted rate of SAE estimated uninsurance and the aggregate number of uninsured derived for the state as a whole—were only *0.7%-points* different. That is, a difference of less than one percentage point. In contrast, our previously-employed *Unit Level* model with a random effect (MLwiN)—when applied to the 2008 data—yielded a county-population weighted rate of uninsurance

[2] The simple mean of these 77 county Direct Estimate rates (19.1%) differs somewhat from the survey estimate (18.9%) by virtue of the fact that the survey estimate uses the survey-weights on all 3804 observations, and consequently it is the correct estimate to use for public reporting. That is, this 19.1% rate is a simple mean of the 77 county rates, which we provide because we want to show the other descriptive statistics for these 77 different sets of estimates. This also applies to the slightly higher rate of 19.2% for the simple mean of the Bayesian SAE estimates.

that was as much as *13%-points* lower than the survey-weighted estimate times the statewide population. This was one criterion we used to judge between the Bayesian SAE model and our previously-employed *Unit Level* model with a random effect (MLwiN).

We also present summary statistics for the *widths* of the Confidence Intervals for the Direct Estimates and the Bayesian Credibility Intervals. An important feature of Bayesian estimates in general—that applies equally to SAE estimates—is that they not only provide more reliable point-estimates, but their level of uncertainty is usually lower as well. As can be readily seen, the widths of the CI's for the direct estimates range from a max of 88%-points to a min of 6%-points. The Bayesian Credible Intervals have a max width of 24%-points and also a min of 6%-points. Of importance, on average the CI's for the direct estimate are 2.5 times wider (37% vs. 15%) than the Bayesian Credible Intervals, a substantial reduction in uncertainty for the Bayesian Credible Intervals.

Summary Statistics of Widths of Confidence Intervals/Credible Intervals of Direct Estimates and Bayesian SAE Estimates of Uninsurance Rates

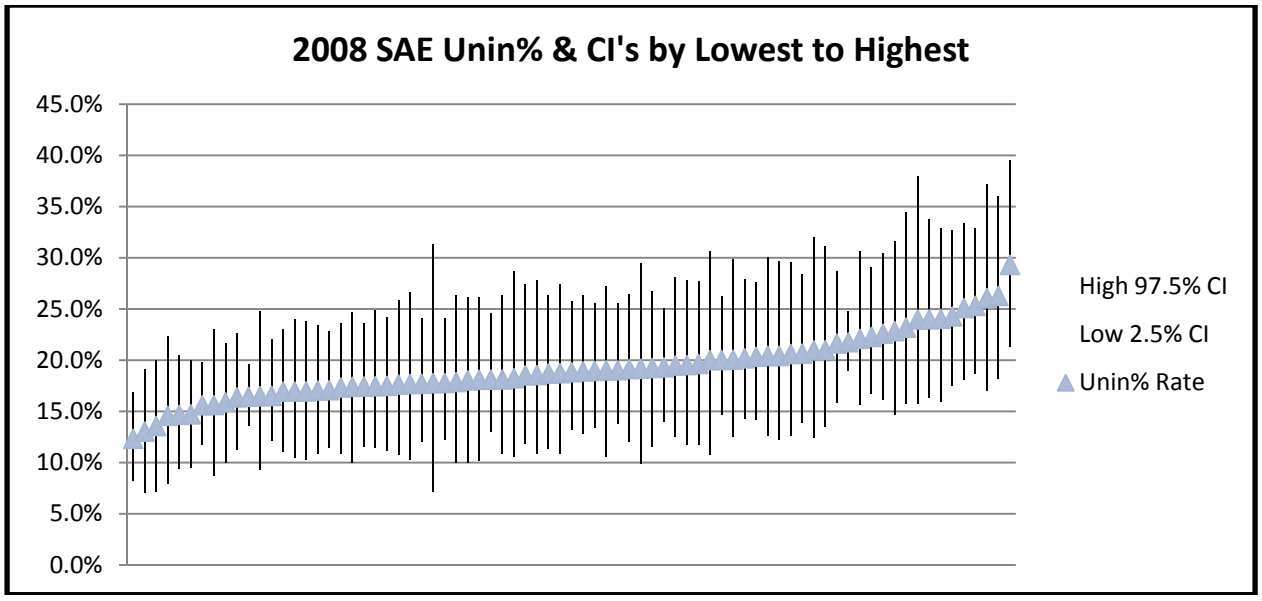
	Direct Estimate CI's	Bayesian SAE Estimate CI's
Min	6%	6%
Max	88%	24%
Mean	37%	15%
Med	32%	15%
sd	19%	3%

A visual sense of what one can say about significant differences between counties in their 2008 SAE estimates is provided by the so-called 'caterpillar graph' below. It gives in graphic format—going from the county with the lowest SAE estimate of uninsurance to the county with the highest SAE estimate of uninsurance—each county's SAE estimate together with the Credible Interval 'bar' for that county's SAE estimate. In this case the vertical length of the 'bar' indicates the level of uncertainty for this estimate.

As can be seen from this graphic, one can make statements concerning higher and lower levels of SAE uninsurance rates for relatively few counties. That is, there are relatively few pairs of counties for which the Credible Interval 'bars' do not overlap. This is a direct consequence of the sparse data described—and the level of uncertainty that necessarily inheres in these estimates of SAE uninsurance rates when they are derived from small county-level survey samples.

Using these Credible Intervals we identify which counties are significantly higher or lower than others.

As tabulated below, Adair county's SAE Unins% rate, at 29.3%, is significantly higher than the 7 counties listed in the 1st table. In addition to Canadian county having one of those 7 SAE rates below Adair, Canadian, at 12.3%, is also significantly lower than 6 additional counties, as listed in the 2nd table.



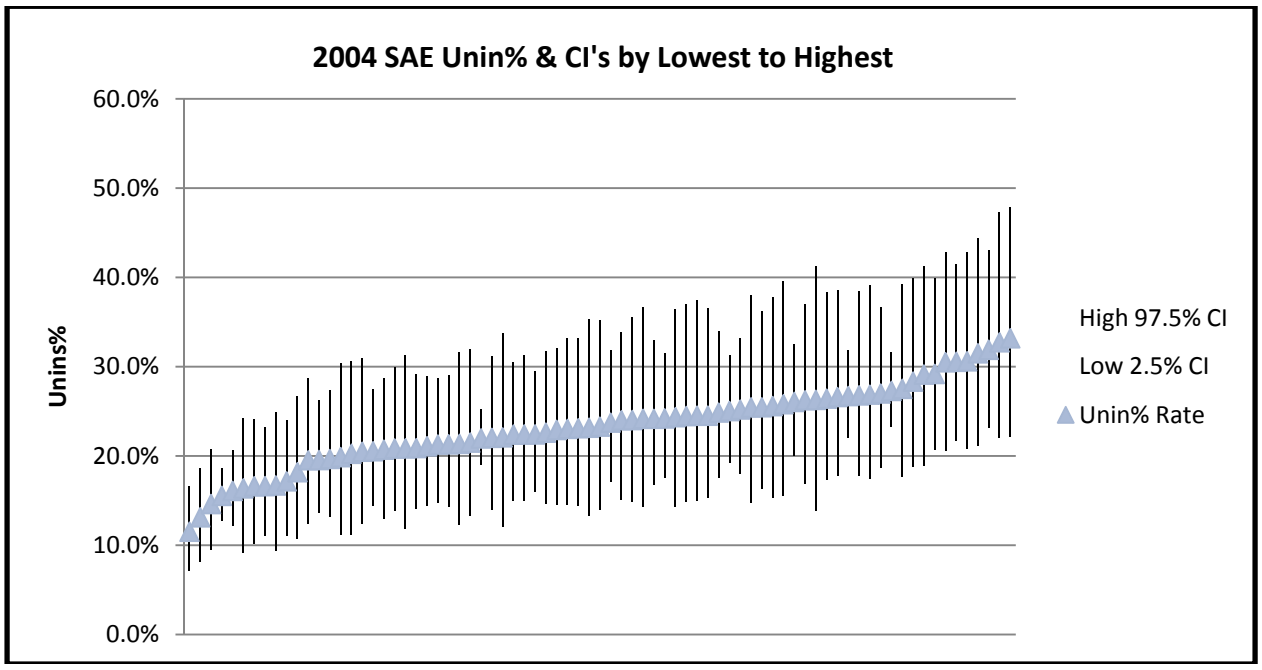
Adair County's SAE Unins%
Exceeds the following Counties'

County	node	Unins% 2008	2.50%	97.50%
Adair	p[1]	29.3%	21.3%	39.5%
	Adair has a significantly HIGHER rate of Unins% than:			
Canadian	p[9]	12.3%	8.2%	16.8%
Woodward	p[77]	13.0%	7.0%	19.1%
Jackson	p[33]	13.6%	7.2%	20.0%
Cleveland	p[14]	15.6%	11.8%	19.8%
Tulsa	p[72]	16.4%	13.6%	19.6%
Grady	p[26]	14.7%	9.5%	20.1%
McClain	p[44]	14.7%	9.4%	20.5%

Canadian County's SAE Unins%
Is Below the following additional Counties'

County	Node	Unins% 2008	2.50%	97.50%
Canadian	p[9]	12.3%	8.2%	16.8%
	Canadian has a significantly LOWER rate of Unins% than:			
Cherokee	p[11]	25.3%	18.6%	32.8%
Okfuskee	p[54]	26.3%	18.2%	36.0%
Oklahoma	p[55]	21.8%	19.0%	24.8%
Pittsburg	p[61]	25.1%	18.0%	33.3%
Coal	p[15]	26.1%	17.0%	37.2%
Delaware	p[21]	24.3%	17.5%	32.7%

For the 2004 SAE estimates, the 'caterpillar graph' indicates somewhat more significant differentiation among the counties. In part this is due to the somewhat larger sample of non-elderly (N = 4596) in the 2004 survey, approximately 21% larger than the 2008 non-elderly.



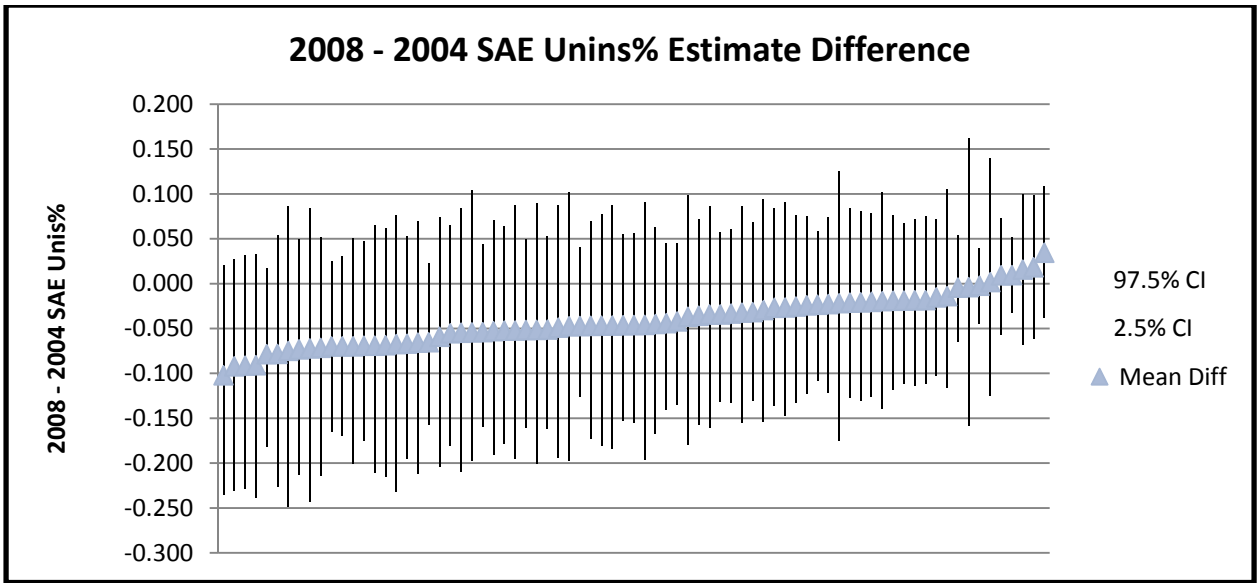
We summarize this greater differentiation in the table below by giving the names of the counties with the lowest SAE Unin% rates, and for each of these named counties we simply provide the total number of counties that had significantly higher SAE Unin% rates.

Number of Counties with 2004 SAE Unins% Rates Significantly Higher than the following Counties

County	Number of Counties With Significantly HIGHER Unis% Rates
Canadian	27
Tulsa	16
Comanche	16
Cleveland	9
Rogers	8
Wagoner	1

Finally, we present the 'caterpillar graph' for the *difference* between the 2008 and 2004 county SAE Unins% rates. As noted, our Bayesian model allows us to directly model this difference, and consequently we provide each county's *difference* in SAE estimates (2008 estimate minus the 2004 estimate) together with the Credible Interval 'bar' for that county's *difference* in SAE estimates. Again, the vertical length of the 'bar' indicates the level of uncertainty for this *difference* in SAE estimates. Also as we discussed, since *both* years of SAE estimates are combined through the creation of this estimated *difference* of uninsurance rates, the degree of uncertainty in these *differences* effectively combines the uncertainty from both. Consequently, most of these Bayesian Credible Intervals for the estimated *differences* are large. Again, however, they provide important information for policy-makers since they indicate the limits of what can and cannot be concluded from these comparisons.

In this case, either a county's *difference* in SAE estimates has a quite wide Credible Interval or when narrower, the difference itself is quite small. Thus, all 77 counties' estimates include the zero value and thus for no county can we say that it experienced either a significant decrease or increase in it SAE Unins% estimate between these two surveys. The summary statistics on these *differences* in SAE estimates are also given.



Descriptive Statistics for County Differences in SAE Unis% Estimates, 2008 - 2004

	Difference
Maximum reduction	-10.2%
Maximum increase	3.4%

We note that if we weight each county's *difference* in SAE estimates by the county population, we obtain a weighted-mean difference of a *1.85%-point reduction*. That is, the 2008 rates are on average 1.85%-points lower than the 2004 estimates. If we take the difference in the overall survey-weighted means of the estimates, (18.9% − 20.7%), we obtain a difference of a *1.87%-point reduction*. Once again, our Bayesian model's predictions of these county *differences* in SAE estimates—when properly weighted by the county population—yield for all practical purposes the same results as the differences in the two survey-weighted overall means. We provide the cautionary note, however, that a *simple, unweighted* mean of these 77 estimates of *difference* in SAE estimates would yield an estimate of average change that is larger, but this is because it does not use the proper weights and for that reason this simple mean should not be taken. The individual county estimates of this difference are useful, but their simple mean is not.

Suggested Citation

State Health Access Data Assistance Center. 2009. "Small Area Estimation Technique for the Oklahoma Health Care Insurance and Access Survey." Minneapolis, MN: University of Minnesota.

References

"Bayesian Small Area Estimation for policy making and policy assessment", by V. Gomez-Rubio, N. Best, S. Richardson. and P. Clarke. Unpublished paper available for download from http://www.bias-project.org.uk/software/

WinBUGS code for implementing SAE models is also available for download at http://www.bias-project.org.uk/software/

PowerPoint slides of the presentation given by these researchers at the Research Methods Festival, Oxford University, England 3 July 2008, is available at http://www.ncrm.ac.uk/RMF2008/festival/programme/spas/pres2/RMF08.pdf

Acknowledgement

This project was funded by the Oklahoma Health Care Authority (OHCA) and was conducted under the supervision of Kathleen Thiede Call, PhD, of SHADAC. This report was authored by Gestur Davidson, Ph.D., Research Associate at SHADAC. Dr. Davidson would like to thank Virgilio Gomez-Rubio for helpful advice on implementing their CAR model using WinBUGS.

Addendum

The WinBUGS code for the model used for generating the 2008 SAE estimates is presented here:

```
model
                {
                        for( i in 1 : N ) {

                                r_dec[i] ~ dbin(p[i],n[i])


                   logit(p[i]) <- beta1 +beta2 * un2004bayes[i]  +u[i] + v[i]

                                u[i]~dnorm(0, precu)

                        rankp[i]<-rank(p[1:N], i)

                                        }

                        v[1:N] ~ car.normal(adj[], weights[], num[], precv)

                                 precu~dgamma(0.01,0.01)
                                 precv~dgamma(0.01,0.01)


                                beta1~dnorm(0,0.001)
                                beta2~dnorm(0,0.001)


                             }
```