Small Area Estimation for the Tobacco-Use Supplement to the Current Population Survey*

Benmei Liu[1] and Aaron Gilary[2]
[1]Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive, Bethesda, MD 20892
[2]Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract

The Tobacco Use Supplement to the Current Population Survey (TUS-CPS), conducted by the Census Bureau and sponsored by the National Cancer Institute (NCI), is a key source of national and state-level data on smoking and other tobacco use in the U.S. household population. However, policy makers and cancer researchers often need county-level data to evaluate tobacco control programs, and the TUS-CPS does not have a large enough sample at the county level to support estimates with adequate precision. In such case, estimates derived through small area estimation (SAE) techniques may be preferable. Through collaboration between the Census Bureau and NCI, we propose model-based county-level estimates for several different smoking-related variables for all U.S. counties using a Bayesian framework through a Markov Chain Monte Carlo (MCMC) simulation. We applied extensive model selection and diagnosis techniques to choose the best set of auxiliary variables from a pool and the best fit models from a few candidate models. Our small area models generate a new set of estimates with improved precision over the survey-based estimates. This paper describes the methodology used and also demonstrates the accuracy of the model through data exhibits.

Key Words: credible intervals, model-based county estimates, predictor variable selection, smoking prevalence, tobacco related measures

*This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of either the National Cancer Institute or the U.S. Census Bureau.

1. Introduction

The Tobacco Use Supplement to the Current Population Survey (TUS-CPS) is a National Cancer Institute (NCI) sponsored survey of tobacco use that has been administered as part of the U.S. Census Bureau's Current Population Survey every two to four years since 1992. The Center for Disease Control co-sponsored the TUS-CPS from 2001 to 2007, and the Food and Drug Administration co-sponsors the most recent 2014-2015 cycle. The TUS-CPS is a key source of national and state-level data on smoking and other tobacco use in the U.S. household population because it uses a large, nationally representative sample that contains information on about 240,000 individuals within a given survey period. Each survey period involves the administration of TUS-CPS in three separate surveys, typically four months apart. Participating individuals must be 15 or older (this restriction becomes 18 or older from January 2007), and not in the armed forces or group quarters. For more details, we refer to the survey website: http://appliedresearch.cancer.gov/tus-cps/.

The TUS-CPS is a unique research source. It can be used to track trends in tobacco use over time, evaluate tobacco control programs, and examine tobacco health disparities and other tobacco control research. It can also be used to analyze economic aspects of tobacco use in conjunction with the CPS's occupational and economic data and other CPS supplements, e.g., Internet, American Time Use (ATUS), Cell Phone Use, Food Security, Annual Social and Economic (ASEC), etc. More importantly, TUS-CPS data can be linked to cancer and other cause-specific mortality data through the National Longitudinal Mortality Study (NLMS).

The TUS-CPS is designed to produce reliable estimates at the national and state levels. However, to better evaluate tobacco control programs, monitor progress in the control of tobacco use, and conduct tobacco-related research, policy makers, cancer control planners and researchers often need county-level data for tobacco related measures. Unfortunately, not all counties are sampled, and not all sampled counties will support estimates with adequate precision. In this case, researchers may prefer the use of Small Area Estimation (SAE) techniques to derive estimates for all counties. (Note that county identifiers are not released for TUS-CPS data.)

The key idea in SAE is to combine information from a variety of relevant sources to form indirect estimators that generally increase the effective sample size thus increasing precision for sampled counties, and make predictions in the absence of sample. These indirect estimators are based on various implicit or explicit models that provide a link to related small areas through supplementary data (e.g., census and/or administrative records) which is commonly known as borrowing strength. A comprehensive account of the range of SAE methods can be found in the definitive book on this subject by Rao (2003). Recent examples of large-scale survey data to produce small area proportions can be found in NCI's recently launched website on Small Area Estimates for Cancer Risk Factors & Screening Behaviors (http://sae.cancer.gov/), and the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program (Citro and Kalton, 2000; Bell et al., 2007), among others.

In order to support the need for county-level tobacco related data, we propose to produce model-based county-level estimates for the following five key TUS-CPS measures through SAE techniques:

- Percentage of People Who Currently Smoke (CSMOKE);
- Percentage of People Who Have Ever Smoked (ESMOKE);

• Percentage of People Who Live in a Residence Where Smoking is not Allowed (HOMEBAN);
• Percentage of People Whose Workplace Does Not Allow Smoking (WORKBAN);
• Percentage of People Who Quit Smoking for a Day of More, Among Current Smokers or Recent Former Smokers Who Quit Fewer Than 365 Days Ago (AQS).

In Section 2, we describe the direct estimates and discuss the associated issues with the direct estimates. Section 3 describes the various small area models we considered. Section 4 lays out the implementation of the model-based approach. The model evaluation and the results are described in Section 5. Section 6 provides some concluding remarks.

2. Direct Estimates and the Issues

Let N_i denote the population size in county i of the target finite population ($i = 1, \ldots, m$). Let y_{ik} be the binary response for the characteristic of interest for unit k in county i (k = 1, ..., m). The parameters to be estimated are the small area proportions $P_i = \Sigma_k y_{ik}/N_i$.

Let n_i denote the sample size in county i, and let w_{ik} denote the sampling weight for sampling unit k in county i. The standard direct survey estimator for P_i is:

$$p_{iw} = \frac{\Sigma_{k=1}^{n_i} w_{ik} y_{ik}}{\Sigma_{k=1}^{n_i} w_{ik}}, \qquad i = 1, \ldots, m. \tag{1}$$

The variance of p_{iw} can be expressed as:

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} DEFF_i, \tag{2}$$

where $DEFF_i$ is the design effect reflecting the sample efficiency of the complex sample design (Kish, 1965).

The problem is that p_{iw} is very imprecise when the sample size is small and cannot even be computed if the sample size is zero. Small area estimation procedures can address this problem.

3. Small Area Models

A commonly used area level model is the *Fay-Herriot* model (Fay and Herriot 1979), which is a two-level mixed-effects model with the following form:

Sampling model:

$$p_{iw}|P_i \sim N(P_i, D_i); \tag{3}$$

Linking model:

$$P_i = x_i^{'}\beta + v_i; \qquad v_i \sim N(0, A); \tag{4}$$

where D_i is the sampling variance and is assumed known. For small areas, the sampling variance D_i is not stable. Generalized variance function modeling techniques or scale transformations have been popularly used to smooth or stabilize the sampling variance in practice.

When measuring proportion data we commonly transform it onto the arcsine scale. For example, Carter and Rolph (1974) applied the arcsine square root transformation function $\theta_i = arcsin\sqrt{p_{iw}}$ in their false alarm probability estimation example. Efron and Morris (1975) applied a similar transformation, $\theta_i = \sqrt{n_i}\ arcsin(2p_{iw} - 1)$, to the sample proportions in order to stabilize the sampling variance in their well-known baseball data example. One advantage of the first transformation is that it is not dependent on sample size, so it can be used for small area estimation problems when prediction is needed.

Building on an extensive simulation study comparing the Fay-Herriot model to several other models (Liu and Diallo, 2013), we chose to use a Fay-Herriot model with arcsine transformation to the direct estimates. Let $z_i = arcsin\sqrt{p_{iw}}$. We applied the following *arcsine-scale* small area model:

Sampling model:

$$z_i|\theta_i \sim N(\theta_i, \frac{DEFF_i}{4n_i}); \tag{5}$$

Linking model:

$$\theta_i = x_i^{'}\beta + v_i; \qquad v_i \sim N(0, A); \tag{6}$$

The sampling model takes account the sampling error for the direct estimate of z_i. The linking model assumes the model parameter θ_i is related to a set of auxiliary variables x_i as defined in Section 4.3. Our goal is to estimate $P_i = sin^2(\theta_i)$.

In cases where the arcsine model (5)-(6) fit poorly, we considered the *probability-scale* model, a Fay-Herriot model (3)-(4) on the original (probability) scale. With $z_i = p_{iw}$, we applied the following model:

Sampling model:

$$z_i|\theta_i \sim N(\theta_i, \tau_i); \tag{7}$$

Linking model:

$$\theta_i = x_i^{'}\beta + v_i; \qquad v_i \sim N(0, A). \tag{8}$$

where the sampling variance τ_i is estimated synthetically using a design effect as in 4.1. Our goal in this case is to estimate $P_i = \theta_i$.

We also tried the *normal-logit* model studied in Liu et al. (2014), which has a similar assumption to the probability-scale model (7)-(8), except that the linking model assumes a logit-normal, instead of normal, distribution on θ_i such that $logit(\theta_i) = x_i^{'}\beta + v_i;\quad v_i \sim N(0, A)$. For this project, the predicted values generated by the normal-logit model performed worse in our diagnostic tests, so we dropped the normal-logit model from our final models.

4. Implementation of the Small Area Models

4.1 Computation of the Design Effect

Each sampling model requires an estimate of the design effect $DEFF_i$ under each outcome. The design effect is defined as the ratio of the variance under the complex design to the variance under simple random sampling. The standard survey software (SAS PROC SURVEYMEANS, SUDAAN, etc.) can be used to compute design effect when the sample size is large enough. To compute the county-level $DEFF_i$, we applied a different approach due to the small sample sizes for many counties. Because the TUS-CPS uses a clustering design, we adapted the Kish design effect formula described by Gabler, Haeder, and Lahiri (1999). Basically, this involved solving the equation:

$$DEFF_{Kish} = m\frac{\Sigma w_i^2 m_i}{(\Sigma w_i^2 m_i)^2}[1 + (\bar{b} - 1)\rho], \tag{9}$$

where m_i and w_i denote the number of observations and the weight attached to the i-th weighting class; $m = \Sigma m_i$, the total sample size; $\bar{b}$ is the average cluster size; and ρ is the intraclass correlation coefficient. All the components on the right hand side of 9 are known by design except ρ. To estimate ρ, we estimated the national level DEFF using SAS PROC SURVEYMEANS for each outcome first. We then plugged the national DEFF into (9) to solve for ρ. Once we obtained an estimate of ρ, we plugged that back into (9) and substituted state-level weights to estimate the state-level DEFF. We finally applied the state-level DEFF to the counties within the state for smoothing purposes.

4.2 Computation of the Sampling Variance

Both the arcsine-scale model (3)-(4) and the probability-scale model (7)-(8) required estimates of the sampling variance D_i. To estimate D_i, we tried two approaches following Liu et al. (2014): an MCMC approach and a synthetic approach using (2). The MCMC approach treats D_i as a function of the unknown P_i so it is estimated simultaneously with P_i. For the synthetic approach, we first fit the following logistic regression model on the m_0 counties with sample sizes $n_i > 150$ and obtained the estimate of the regression coefficient vector $(\beta_0, \beta_1, \ldots, \beta_p)'$:

$$logit(p_{iw}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, m_0;$$

where p_{iw} is the direct estimate of P_i, $x_{i1}, \ldots, x_{ip}$ are the p auxiliary variables selected for the MCMC model, and $\epsilon_i \sim N(0, \sigma_e^2)$. We then computed a synthetic estimator of p_i for all the m counties with samples, as follows:

$$p_i^{syn} = \frac{exp(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip})}, \quad i = 1, \ldots, m.$$

Finally we computed the following smoothed synthetic sampling variance of p_{iw}:

$$D_i^{syn} = \frac{p_i^{syn}(1 - p_i^{syn})}{n_i} DEFF_i,$$

where $DEFF_i$ was estimated using the procedure described in section 4.1.

4.3 Auxiliary Variables

Finding a good set of auxiliary variables is crucial for model-based SAE approaches. For this study, less than half of U.S. counties contain any TUS-CPS sample. The remaining counties rely entirely on auxiliary data from sources other than the TUS-CPS survey. Therefore we wanted to include many related predictors from other sources. As a result, the pool of the candidate auxiliary variables included thirty county-level demographic and socioeconomic variables obtained from the American Community Survey, the decennial Census, and other administrative sources. Additionally it included five state-level smoking policy variables including state smoking bans, cigarette taxes, Medicaid coverage for tobacco related treatment, overall tobacco control funding, and years since a quitting hotline was established. The list of covariates is given in Appendix A.

Including too many auxiliary variables would make the MCMC models converge slowly and potentially overfit the model. For each outcome, we applied classical backward model selection procedure to select a reduced set of auxiliary variables. These auxiliary variables were log-transformed (for the probability-scale models we included auxiliary variables under the original scale). We used these variables to develop a regression model for the θ_i term in our linking model. We also compared models where we forced in several covariates known to be related to the outcomes (based on the tobacco research literature) to models that were developed naturally through the stepwise process. The model forcing in covariates produced better results, according to the diagnostics, for the CSMOKE and AQS outcomes only.

4.4 Implementation Using Hierarchical Bayes Approach

Our models were implemented using the Hierarchical Bayes (HB) method based on the following commonly used prior assumptions for the hyper-parameters β and A:

$$\beta \propto 1; \qquad A \sim unif(0, 100)$$

The HB estimates of P_i are produced using the Markov Chain Monte Carlo (MCMC) technique (Robert & Casella, 1999; Rao, 2003, Sec. 10.2) implemented in R using the `rjags` package. We used the Deviance Information Criterion (DIC) (Lunn et al., 2002) as the main criterion to select the best model and best set of auxiliary variables from several candidate models described in Section 2, and we used Gelman and Rubin's potential scale reduction factor $\hat{R}$ as the main convergence criterion to ensure the convergence of the MCMC models.

5. Evaluation of the Different SAE Models

After obtaining the HB estimates, we did extensive model diagnosis to assess the goodness of fit for each model. We checked the overall fit of each model using the method of posterior predictive p-values (Gelman et al., 1996). Following Rao (2003, Sec. 10.2), we also assessed model fit at the individual county-level by computing two measures: one is the county-level measure providing information on the degree of consistent overestimation or underestimation of the observed value; the other is the county-level measure which is similar to a cross-validation standardized residual but uses the full predictive density. For more details, we refer to Rao (2003).

A working model must pass all of these diagnosis and convergence criteria introduced above. For both survey years, the arcsine-scale model (5)-(6) was finally chosen for estimating the percentage of people who currently smoke (CSMOKE), percentage of people who live in a house where smoking is not allowed (HOMEBAN), and the percentage of people whose workplace does not allow smoking (WORKBAN), while the probability-scale Fay-Herriot model (3)-(4) was chosen for estimating the smoking cessation rate AQS. To estimate the percentage of people who have ever smoked (ESMOKE), the probability-scale Fay-Herriot model (3)-(4) worked best for survey year 2006-2007 while the arcsine-scale model worked best for survey year 2010-2011.

As a further evaluation, we plotted the ratio of the direct estimates over the model-based estimates against the sample size. The ratio should converge to one, as large counties do not need to borrow strength. Figure 1 shows the ratio of design-based to model-based estimates (on the Y axis) against sample size scaled by log (on the X axis) for current smoking. The funnel shape reflects the convergence of direct

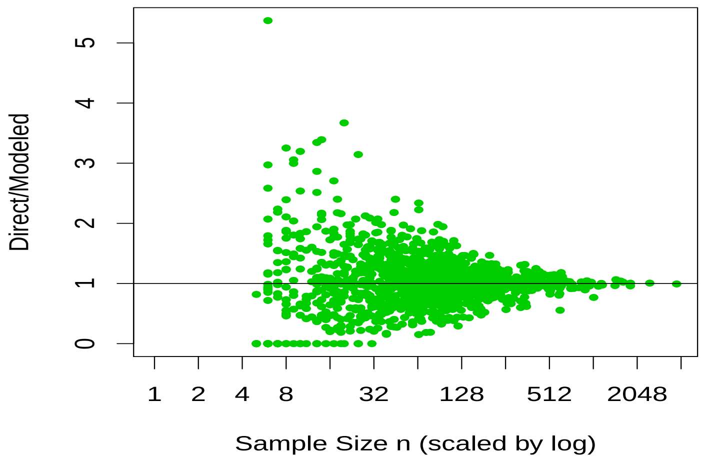

Figure 1: Ratio of Design-Based to Modeled Estimates for the Proportion of People Who Currently Smoke, for 2010-2011.

and model-based estimates, and the gradual decrease in borrowing strength, as the county sample size gets progressively larger. The pattern holds across different variables and different years.

Figure 2 shows estimates and error bars for current smoking for each county in Maryland, ordered by increasing sample size, with estimates given on the Y-axis and county index given on the X-axis. The solid red lines show direct estimates, where the dotted red lines represent (and reflect the uncertainty of estimates for) counties without TUS-CPS sample, for which direct estimates are unavailable. The blue lines show model-based estimates for the corresponding counties. Because the current smoking model was arcsine transformed, the design-based confidence intervals are obtained by finding the confidence bounds on the arcsine scale, and transforming those back to probability scale. The plot shows the following benefits of the model: it preserves the design-based means in the larger counties, and overall; it reduces variances for small counties; and it provides estimates for counties out of sample.

Table 1 compares empirical quantiles for county point estimates for direct and modeled values. We take the median and the two percentiles comprising 95% intervals. The table serves as a face validity check, as the medians stay largely consistent with new counties introduced. It also shows that the most extreme point estimates are shrunk to more reasonable values.

Table 2 shows the model-based variance reduction, as a percent reduction in the interval width for all counties in sample. CSMOKE, ESMOKE, and AQS show heavy interval reduction across the distribution, by about 50 percent. HOMEBAN and WORKBAN show reductions in general, but also a small number of interval increases (shown as negative numbers in the table) typically in cases where a direct estimate near one was adjusted downward by the model, and took on a higher variance (because the variance is based on the survey estimate). These results reinforce the overall benefit of the model, although they suggest that these intervals should be examined in context.

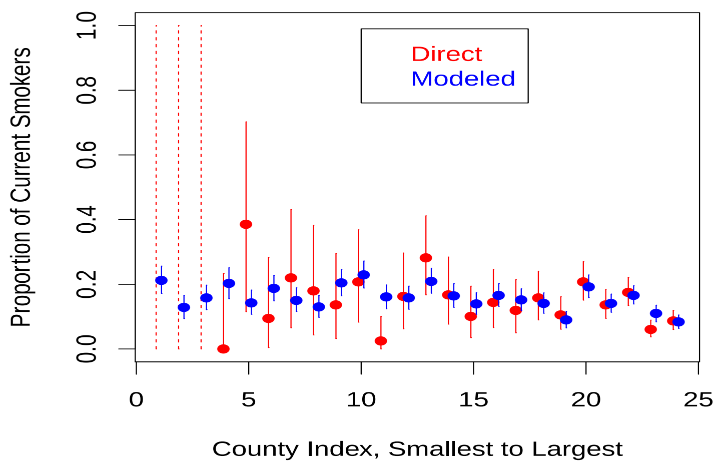

Figure 2: Modeled vs. Design-Based Estimates for the Proportion of People Who Currently Smoke, for the State of Maryland, for 2010-2011.

Table 1: Empirical Quantiles for County Point Estimates, for 2010-2011

	Direct (n=1,064)			Modeled (n=3,137)		
	2.5%	50%	97.5%	2.5%	50%	97.5%
CSMOKE	.07	.19	.39	.10	.20	.26
ESMOKE	.19	.40	.62	.25	.41	.50
HOMEBAN	.54	.82	.96	.68	.79	.91
WORKBAN	.54	.84	1	.71	.82	.92
AQS	.09	.45	.75	.34	.43	.55

Table 2: Distribution in Percent Reduction on the Length of the Confidence Intervals for the Modeled Estimate in Comparison to the Direct Estimate, for 2010-2011

	County Quantiles (n=1,064)				
	2.5%	25%	50%	75%	97.5%
CSMOKE	27	56	66	73	84
ESMOKE	12	35	45	55	69
HOMEBAN	-5	35	52	63	78
WORKBAN	-81	17	39	54	73
AQS	29	58	65	71	79

6. Summary and Future Research

Across all of the different outcomes, the modeled estimates show consistency with direct estimates in the aggregate, and reduce variance for each county in a general sense. Our small area models passed all the diagnostic tests we performed, and we believe that these estimates will be useful to researchers. Moreover, this estimation framework can be applied elsewhere, and should be useful for modeling different outcome variables, or future years of tobacco data.

Our main priority for future research is to reconcile the county-level modeled estimates, so that the aggregated lower-level estimates match direct higher-level estimates. To that end, we will benchmark our state-level modeled estimates to direct region-level estimates (using the four Census regions Northeast, Southeast, Midwest, and West), and then benchmark our county-level estimates to those benchmarked state estimates.

We will also perform further diagnostic checks on our models, to see how the modeled estimates reconcile across different years, and to edit any values that may be irreconcilable. We will release the final estimates on NCI's State Cancer Profiles website, at http://statecancerprofiles.cancer.gov. The results may be added to http://sae.cancer.gov as well.

Acknowledgments

The authors would like to thank Dr. Partha Lahiri, of the University of Maryland, for his valuable advice and assistance regarding this project.

References

Bell, W., Basel, W., Cruse, C., Dalzell, L., Maples, J., O'Hara, B., and Powers, D. (2007), *Evaluation of School District Poverty Estimates: Predictive Models Using IRS Income Tax Data.* U.S. Census Bureau, last accessed May 2014, http://www.census.gov/did/www/saipe/publications/files/report.pdf.

Carter, G. M., and Rolph, J. E. (1974), *Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities.* Journal of the American Statistical Association, 69, 880-885.

Citro, C.F., and Kalton, G. (Eds.). (2000). *Small-area income and poverty estimates: Priorities for 2000 and beyond.* National Academies Press.

Efron, B. and Morris, C. (1975), *Data Analysis Using Stein's Estimator and its Generalizations.* Journal of the American Statistical Association, 70, 311-319.

Fay, R. and Herriot, R. (1979), *Estimation of Income from Small Places: an Application of James-Stein Procedures to Census Data.* Journal of the American Statistical Association, 74, 269-277.

Gabler, S., Haeder, S. and Lahiri, P. (1999) *A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering.* Survey Methodology, 25, 105-106.

Gelman, A., Meng, X., and Stern, H. (1996). *Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies.* Statistica Sinica, 6, 733-807.

Kish, L. (1965), **Survey Sampling.** Wiley and Sons.

Liu, B. and Diallo, M. (2013). *Parametric bootstrap confidence intervals for survey-weighted small area proportions.* Proceedings of the Survey Research Methods Section of the American Statistical Association, ASA Section on Survey Research Methods, 109-121.

Liu, B., Lahiri, P., and Kalton, G. (2014). *Hierarchical Bayes modeling of survey-weighted small area proportions.* Survey Methodology, 40 (1), 1-13.

Lunn, D.J., Best, N., Thomas, A., Wakefield, J., and Spiegelhalter, D. (2002), *Bayesian Analysis of Population PK/PD Models: General Concepts and Software.* Journal of Pharmacokinetics and Pharmacodynamics, 29, 271-307.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rao, J.N.K. (2003), **Small Area Estimation**. Wiley.

Robert, C.P., and Casella, G. (1999), **Monte Carlo Statistical Methods.** Springer-Verlag, New York.

Appendices

A. The Pool of Covariates for Modeling, 2010-2011

Covariate Levels and Definitions		
#	LEVEL	DEFINITION
1	county	Average Household Size
2	county	Per Capita Income in the Past 12 Months
3	county	Median Value of Specified Owner-Occupied Housing Units
4	county	Percent Bachelor's Degree or Higher Among Age 25+
5	county	Percent Divorced Among Age 15+
6	county	Percent of Households with Female Householder
7	county	Percent Foreign Born
8	county	Percent of Households with One or More People Under 18
9	county	Percent High School Graduates with Less than a College Degree Among Age 25+
10	county	Percent Speaking Language Other Than English at Home Age 5+
11	county	Percent Male Among Age 15+
12	county	Percent Married But With Spouse Absent or Separated
13	county	Percent Never Married Among Age 15+
14	county	Percent of Single-Person Households
15	county	Percent Below Poverty Line Among Age 18+
16	county	Percent With Travel Time To Work At Least 30 Minutes
17	county	Civilian Labor Force Unemployment Rate Average
18	county	Percent of Management, Professional, and Related Occupations
19	county	Percent Widowed Among Age 15+
20	county	Percent Black or African-American (Census 2010)
21	county	Percent Hispanic or Latino (Census 2010)
22	county	Percent Living in Rural Areas (Census 2000)
23	county	Average Number of Violent and Property Crimes Known to Police and FBI
24	county	Federal Expense Per Capita
25	county	Local Government General Revenue Per Capita
26	county	Indicator of Metropolitan Statistical Area
27	county	Percent of People 65 years and Older Among the 1+ Population
28	county	Local Government General Revenue, Property Taxes Per Capita FY 2002
29	county	Retail, Eating and Drinking Expense Per Household, 2006-2007
30	county	Social Security Beneficiaries from SSA
31	state	Indicator of Ban in ALL of (Workplaces, Restaurants, Bars) by 2010
32	state	Cigarette Excise Tax in 2010 ($ Per Pack)
33	state	Medicaid Coverage for Tobacco-Related Treatment
34	state	Overall Tobacco Control Funding in 2010 ($ in Millions)
35	state	Years Since State Quitline Service was Established