Hierarchical models for state-level AK estimators in the Current Population Survey (CPS)

Yuan Li[1], Michael D. Larsen[1]
[1]The George Washington University, Washington, D.C. 20052

Abstract

The Current Population Survey (CPS) is a multistage household probability sample that produces monthly labor force estimates in the U.S. Adults in a household are interviewed for four months in a row, left out for eight months, and then included for four more months. This 4-8-4 rotation design produces overlap in the sample. Several weighting steps are used to adjust the ultimate sample to be representative of the population. In order to produce efficient estimates of labor force levels, an estimator, called the AK composite estimator, combines current estimates from 8 rotation panels and the previous month's estimate, is applied. Finding the optimal values for the parameters, A and K, of AK composite estimator can be very helpful for estimating the employment and unemployment counts. Our method is to build a hierarchical model for each state. We contrast a univariate model with a bivariate model that includes correlation between A and K. The Gibbs sampler with multiple independent sequences is used for computations. Under the model the 51 state-level values of A and K experience shrinkage toward the overall mean values. Final unemployment estimates can use the modeled A and K values.

Key Words: composite estimation, labor force, panel survey, unemployment, Gibbs sampler

1. Introduction

The document is organized in five sections. The first section introduces the Current Population Survey and its sampling design. The second section gives formulas for several estimators. The third section presents the way of finding optimal coefficients of AK estimate for each state. Section 4 presents hierarchical models for use with (A, K) values estimated at the state level. Initial univariate and bivariate models are presented. Section 5 presents results of some computations using public monthly CPS data from January 2007 to April 2014, the latest period with constant values of (A, K) in use. Estimates, variance estimates, state-level optimal (A,K) values, and a clustering of states are presented. Section 6 discusses planned research work.

1.1 Current Population Survey

The Current Population Survey (CPS), a monthly household survey conducted by the U.S. Census Bureau and the U.S. Bureau of Labor statistics (BLS), is the primary source of labor force statistics (LFS) for the population of the United States. It produces monthly labor force and related estimates for the total U.S. civilian population and provides details by age, sex, race and so on. Important CPS estimates include estimates of the number of persons in three major force categories: employment, unemployment, and people "not in the labor force". In addition, estimates for the number of other population subdomains are produced on either a monthly or quarterly or yearly basis.

1.2 CPS Sampling design

The sample design for the CPS is a two-stage sample of more than 72,000 housing units to measure the number of employment, unemployment and person not in labor force in the United States. The first stage involves dividing the U.S. into about 2,000 Primary Sampling Units (PSUs), which consist of a large county or a group of smaller counties. In the second stage of sampling, a sample of housing units within the sampled PSUs is drawn. Clusters of 4 housing units with similar demographic composition and geographic proximity are grouped together to form the Ultimate Sampling Units (USUs).

In order to balance reliability requirements for estimators of monthly level and month-to-month change, the CPS employs a "4-8-4" sample rotation scheme: a panel of households called a rotation group is interviewed for four months in a row, left out for eight months, then interviewed for another four months. This 4-8-4 rotation design ensures that there is continuity in the sample from month to month, quarter to quarter and year to year, which includes personal interview (CAPI) and telephone interview (CATI) based on sampling in the different rotation groups.

Table1: Months in sample by year and month for CPS samples and rotations

Year and Month		Sample 88						Sample 89								Sample 90			
		C	D	E	F	G	H	A	B	C	D	E	F	G	H	A	B	C	D
13	Jan	**8**	**7**	**6**	**5**									**4**	**3**	**2**	**1**		
	Feb		8	7	6	5									4	3	2	1	
	Mar			8	7	6	5									4	3	2	1
	Apr				8	7	6	5									4	3	2
	May					8	7	6	5									4	3
	Jun						8	7	6	5									4
	Jul							8	7	6	5								
	Aug								8	7	6	5							
	Sep									8	7	6	5						
	Oct										8	7	6	5					
	Nov											8	7	6	5				
	Dec												8	7	6	5			
14	Jan													**8**	**7**	**6**	**5**		
	Feb														8	7	6	5	

Table1 above shows how this 4-8-4 rotation works. Eight rotation groups, which approximately equal in size, make up each monthly CPS sample. The eight rotation groups in sample for a given month can also be considered "Month-In-Sample" groups. Each month, interviewers collect data from the sampled housing units. And each housing units entering the CPS remains in sample for 4 consecutive months, leave the sample for the following 8 months, and then reenter for another 4 months. Therefore, a sampled housing unit is interviewed eight times. The rotation scheme ensures that in any single month, one-eighth of the housing units are interviewed for the first time, another one-eighth is interviewed for the second time, and so forth.

2. Estimators and Variance Estimation:

Based on the 4-8-4 rotation design, for each month, there are six out of eight rotation groups in the survey for the previous month. That is to say, there is a 75 percent overlap from month to month, and also there is a 50 percent overlap from year to year. The sample overlap can improve the estimates of change over time. Through composite estimation, the positive correlation among CPS estimators for different months is increased. This increase in correlation improves the accuracy of monthly labor force estimates.

2.1 Estimators

For each person in the monthly CPS sample, the Census Bureau calculates a weight, which is a rough estimate of the number of actual persons the sample person represents. This is accomplished by means of ratio adjustment, which followed by four categories: the first-stage ratio adjustment, the national coverage adjustment, the state coverage ratio adjustment, and the second-stage ration adjustment. (The data we are using is the public data from DataFerrett of Census Bureau, in which the weight is an available variable. So we do not need to be concerned about the calculation procedure of the weighting adjustment.)

Since the main interest of the research is the estimation of monthly level unemployment and their month to month change, define Y_t = the unknown total number of people who are unemployed in month t and $\hat{Y}_{t,i}$= weighted estimate of the number of unemployed in the i^{th} rotation group at month t, where rotation group i=1,2,3,4,5,6,7,8.

(1) Ratio Estimator

Based on the weights after the second-stage ratio adjustment, for each month t, we have a form of the estimation for the CPS, which is called the ***Ratio Estimator***. And the Ratio estimate of unemployment in month t is the summation of weights of the eight rotation groups (month-in-sample) in month t, and is denoted by $\hat{Y}_t^{(1)}$, which has the format:

$$\hat{Y}_t^{(1)} = \sum_{i=1}^{8} \hat{Y}_{t,i}$$

This type of estimate has been variously referred to as a two-stage Ratio Estimates or a simple weighted estimate. However, the most official CPS labor force estimates that are based on the information collected each month from the sample is a composite estimate.

(2) Difference Estimator

Besides the estimation of monthly level unemployment, the month to month change is also an important quantity that needs to be considered. We decide another estimate called ***Difference Estimator***, which is the most efficient method to measure the change between months. Let $S_2 = \{2,3,4,6,7,8\}$, the set of indicators of the month-in-sample groups in the CPS sample for a given month t that were also in sample in month $t-1$. And denote $\hat{Y}_t$ be the estimate of Y_t.As there is a 6/8 (75%) overlap of rotation group from month to month, so by using the overlap groups, we are able to estimate the difference, $\hat{\Delta}_t$, which is called the estimate of change:

$$\hat{\Delta}_t = \frac{4}{3} \sum_{i \in S_2} \left(\hat{Y}_{t,i} - \hat{Y}_{t-1,i-1} \right) .$$

So the Difference Estimator is

$$\hat{Y}_t^{(2)} = \hat{Y}_{t-1} + \hat{\Delta}_t = \hat{Y}_{t-1} + \frac{4}{3} \sum_{i \in S_2} \left(\hat{Y}_{t,i} - \hat{Y}_{t-1,i-1} \right) .$$

(3) Composite Estimator

In fact, the 4-8-4 Rotation Design compromises between the longitudinal and independent sample design. Ratio Estimate is an estimation of the total monthly level unemployment, and Difference Estimate is usually used in a longitudinal survey design. Since Ratio Estimate $\hat{Y}_t^{(1)}$ does not contain the information of the previous month and Difference Estimate $\hat{Y}_t^{(2)}$ does not use all of the data, so there is an estimate called Composite Estimate, $\hat{Y}_t^{C}$, compromises both $\hat{Y}_t^{(1)}$ and $\hat{Y}_t^{(2)}$:

$$\begin{aligned}\hat{Y}_t^{C} &= (1-K)\hat{Y}_t^{(1)} + K\hat{Y}_t^{(2)} \\ &= (1-K)\hat{Y}_t^{(1)} + K(\hat{Y}_{t-1}^{C} + \hat{\Delta}_t)\end{aligned}$$

where K is between 0 and 1, inclusive of the endpoints. The Composite Estimator is a weighted average of Ratio Estimator and Difference Estimator. The two terms in the composite estimator were given equal weight before 1985. The use of this Composite Estimator can reduce the variance for estimates of month to month change; however, it might not help reduce the month in sample bias. So, in next section, we will emphatically introduce the main estimate that is used in our research.

2.2 AK Composite Estimator

In addition to the weighted average of the two estimates, ratio estimate and difference estimate, the CPS composite estimate incorporates an adjustment which reduces the variance while partially correcting for the bias associated with month in sample at the same time. It applies the method from Gurney and Daly (1965) by adding more weight in month in sample 1 and 5, which leads to the form:

$$\left\{\sum_{i\in S_1}(1-K+A)\hat{Y}_{t,i} + \sum_{i\in S_2}(1-K-\frac{1}{3}A)\hat{Y}_{t,i}\right\} + K\hat{Y}_t^{(2)}$$

Then, we get the estimate:

$$\begin{aligned}\hat{Y}_t^{AK} &= (1-K)\hat{Y}_t^{(1)} + K\hat{Y}_t^{(2)} + A\hat{\beta}_t \\ &= (1-K)\hat{Y}_t^{(1)} + K(\hat{Y}_{t-1}^{AK} + \hat{\Delta}_t) + A\hat{\beta}_t\end{aligned}$$

where $\hat{\beta}_t = \sum_{i\in S_1}\hat{Y}_{t,i} - \frac{1}{3}\sum_{i\in S_2}\hat{Y}_{t,i}$ is called the bias adjustment term, $S_1 = \{1,5\}$ and A, K are both constant parameters that should satisfy the restriction (according to the formula by Gurney and Daly (1965)):

$$\begin{cases} 1-K+A>0 \\ 1-K-\frac{1}{3}A>0 \\ 0<K<1 \end{cases} \quad\Rightarrow\quad K-1<A<3(1-K)$$

So the range of A and K can be from -1 to 3 and 0 to 1, respectively. As we can see from the format of the AK Composite estimate, the coefficient *K* determines the weight of the ratio estimate and difference estimate, while coefficient *A* determines the weight of $\hat{\beta}_t$, an adjustment term, that reduce the both the variance of this composite estimate and the month in sample bias. This estimate is often called “AK composite estimate”.

2.3 Variance Estimation

Estimate the variance of the survey estimates for use in various statistical analyses is of the major statistics of interest for the CPS. The replication methods are able to provide satisfactory estimates of variance for a wide variety of designs using probability sampling. A practical alternative is to draw a set of random subsamples from the full sample of each month, using the principles of selection as those used for the full sample, and to apply the regular CPS estimation procedures to those subsamples, which are called replicates.

The current approach to estimate the variances is called Successive Difference Replication (SDR). The theoretical basis for the successive difference method was discussed by Wolter (1984). Fay and Train (1995) proposed a successive difference replication (SDR) variance estimate:

For a series of ordered estimates $\hat{y}_i$'s, $i=1,2,...,n$, they define the estimate of each replicate as $\hat{Y}(r)=\sum_{i=1}^{n} f_{i,r}\hat{y}_i$, where $f_{i,r}=1+(2)^{-3/2}a_{i+1,r}-(2)^{-3/2}a_{i+2,r}$ is the replicate factor for each $\hat{y}_i$, and $a_{i,r}$ equals a number in the Hadamard orthogonal matrix $(+1$ or $-1)$.

Then, the variance estimate for the character of interest is a sum of squared differences between each replicate estimate ($\hat{Y}(r)$) and the full sample estimate ($\hat{Y}$):

$$\hat{V}_{SDR}(\hat{Y})=4(4k)^{-1}(1-f)\sum_{r=1}^{4k}(\hat{Y}(r)-\hat{Y})^2 ,$$

To increase the precision of the variance estimation, the CPS is currently using 160 replicates for the Fay method of variance estimation and the replication factors are calculated using a 160×160 Hadamard orthogonal matrix. So consider using the AK Composite Estimate, the variance estimation for the AK estimate of unemployment in month t is: $\hat{V}_{SDR}(\hat{Y}_t^{AK})=\frac{4}{160}\sum_{r=1}^{160}(\hat{Y}_t^{AK}(r)-\hat{Y}_t^{AK})^2$. Although this variance estimation is imperfect, the procedure is accurate enough for all practical uses of the data, and captures the effect of sample selection on the total variance.

3. Optimal Coefficients for the AK Composite Estimate

The current value of coefficients A and K for the AK Composite Estimate of the unemployment, used by CPS currently, was introduced in Lent, Miller, Cantwell (1994). In their study, they select the A, K pairs based on comparisons of the variances (by considering three measurements: monthly level, month to month change and annual average). As no one pair yielded the smallest variances for all three measurements, they adopt a compromise and use $A=0.3$ and $K=0.4$. These choices are optimal for measuring the monthly level unemployment and close to optimal when measuring the month to month change and annual average.

In our study, the data we have is all the survey information from January 2007 to March 2014, and Section 2 provides us with the approach of simulating the AK Composite Estimate and Variance Estimation of each month: for each month t (from Jan 2007 to March 2014), we will have $\hat{Y}_t^{AK}$ and $\hat{V}_{SDR}(\hat{Y}_t^{AK})$. So despite of the current value of coefficients A and K, our ***goal*** is to find out the optimal coefficients A and K based on this AK Composite Estimate and Fay method of variance estimation, as well as the data.

Since *K* is from 0 to 1 and *A* is correlated with *K* (in section 2.2), so according to the current value of *A* and *K*, we would like to choose both *A, K* from the set M, where M is the set of values from 0.00 to 1.00 by increments of 0.05.

3.1. Optimal coefficients for the national level

There are 87 months from Jan 2007 to Mar 2014. And for each month t, the AK estimates ($\hat{Y}_t^{AK}$) and Variance Estimations ($\hat{V}_{SDR}(\hat{Y}_t^{AK})$) are function of A, K and would vary by different combinations of *A* and *K*. Our way of selecting the optimal pair (*A*, *K*) for the national level is the following: Choose the pair (*A*, *K*) that minimizes the total variance estimation, which is the summation of the Variance Estimation $\hat{V}_{SDR}(\hat{Y}_t^{AK})$ for the 87 months (i.e. minimizing $\sum_t \hat{V}_{SDR}(\hat{Y}_t^{AK})$, where t is from Jan 2007 to Mar 2014). Thus, $(A,K)_{opt} = \arg\min_{(A,K)} \sum_t \hat{V}_{SDR}(\hat{Y}_t^{AK})$. We will compare the current values of A and K with the values that we generate, and then find out where is difference is.

3.2.Optimal coefficients for the state level

For each combination of A and K, we are interested in finding the optimal pair (*A*, *K*) for state *j* in the following three steps:

Step 1: Calculate $\hat{V}_{SDR}(\hat{Y}_{tj}^{AK})$ by plugging in *A* and *K* both from the set M;

Step 2: Obtaining the optimal value of the parameters:

$$(A_{tj}, K_{tj}) = \arg\min_{(A_{tj}, K_{tj})} \hat{V}_{SDR}(\hat{Y}_{tj}^{AK})$$

Step 3: The optimal (A_j, K_j) for each state can be calculated by taking the average of (A_{tj}, K_{tj}) over months from Jan 2007 to March 2014: $(A_j, K_j)_{opt} = (\bar{A}_{tj}, \bar{K}_{tj})$

Although taking into account the compromise of all the three measurements (unemployment monthly level, month to month change and annual average) seems to be the best and perfect way of selecting the optimal *A* and *K*, minimizing the variances of monthly level also has been proved to be a good compromise choice.

4. State-level Hierarchical Models

Currently for the CPS, the AK Composite estimate for the national level is applied with $A = 0.3$ and $K = 0.4$ when estimates the number of unemployment, as $A = 0.4$ and $K = 0.7$ for estimating the employment. For any labor force characteristic (employment or unemployment), we usually consider three measurements when choosing the parameter *A* and *K:* monthly level, month to month change and annual average, so each of these pairs represents a compromise across these three important measurements.

Our approach to generate the optimal values for the parameter A and K mainly contains two parts: for each of the 51 states, build a univariate hierarchical model and a bivariate hierarchical model, and find out the appropriate method of calculating A and K by contrasting these two models.

4.1 Univariate Hierarchical Model

For each state j, we have 6 variables: Y_{tj}^{AK}, A_j, K_j, σ_j^2, σ_{aj}^2, σ_{kj}^2,which represents the observed values of AK estimators, A value, K value, variance of Y_{tj}^{AK}, variance of A and variance of K, respectively. Then for each of these 6 parameters, we produce an appropriate distribution. And under the assumption that all these variables are independent, we then have the Univariate Hierarchical Model:

$$\begin{cases} \hat{Y}_{tj}^{AK} \mid A_{j,}K_j,\sigma_j^2 \sim N(\hat{Y}_{tj}^{(1)} + K_j(\hat{Y}_{tj}^{(2)} - \hat{Y}_{tj}^{(1)}) + A_j\hat{\beta}_{tj},\sigma_j^2) \\ A_j \mid \sigma_{aj}^2 \sim N(a_j,\sigma_{aj}^2) \\ K_j \mid \sigma_{aj}^2 \sim N(k_j,\sigma_{kj}^2) \\ \sigma_j^2 \sim Inv-Gamma(\alpha_{0j},\beta_{0j}) \\ \sigma_{aj}^2 \sim Inv-Gamma(\alpha_{1j},\beta_{1j}) \\ \sigma_{kj}^2 \sim Inv-Gamma(\alpha_{2j},\beta_{2j}) \end{cases},$$

where $\hat{Y}_{tj}^{AK}$, A_j, K_j are normal distributed, while σ_j^2, σ_{aj}^2 and σ_{kj}^2 have the Inverse Gamma distribution. The hyper-parameters: a_j, k_j, α_{0j}, β_{0j}, α_{1j}, β_{1j}, α_{2j}, β_{2j} can be estimated by the data we observed. And the estimators, as mentioned in section 2.1, $\hat{Y}_{tj}^{(1)}$ (ratio estimator), $\hat{Y}_{tj}^{(2)}$ (difference estimator), $\hat{\beta}_{tj}$ (bias adjustment estimator) will also be estimated.

The Gibbs Sampler with multiple independent sequences (with random starting points) of the iterative simulation is used for the computation, and finally will give us the result of the values for *A* and *K* of the State *j*.

4.2 Bivariate Hierarchical Model

The Univariate Hierarchical Model is under the assumption of independence of the variables. However, as of section 2.2, the parameters *A* and *K* are correlated. To solve this problem, we will take *A* and *K* to be multivariate normal distributed instead of having normal distribution separately. So the Bivariate Hierarchical Model is:

$$\begin{cases} \hat{Y}_{tj}^{AK} \mid A_{j,}K_j,\sigma_j^2 \sim N(\hat{Y}_{tj}^{(1)} + K_j(\hat{Y}_{tj}^{(2)} - \hat{Y}_{tj}^{(1)}) + A_j\hat{\beta}_{tj},\sigma_j^2) \\ \sigma_j^2 \sim Inv-Gamma(\alpha_j,\beta_j) \\ \begin{pmatrix} A_j \\ K_j \end{pmatrix} \mid \Sigma_j \sim N(\begin{pmatrix} a_j \\ k_j \end{pmatrix},\Sigma_j) \\ \Sigma_j \sim Inv-Wishart(\Sigma_j^0) \end{cases},$$

with variables: Y_{tj}^{AK}, A_j, K_j, σ_j^2, Σ_j, and hyper-parameters: a_j, k_j, α_j, β_j, Σ_j^0. With the same process as of the Univariate Model, we will have the values of the parameters *A* and *K* of the Bivariate Model for each state.

After the simulation is done, it is important to check the convergence of the iteration of the two models. Once the convergence of the models is monitored, we can find out the better model, whether it is reasonable to assume independence of the variables (*Univariate Model*) or the correlation between *A* and *K* should be considered (*Bivariate Model*), by comparing the Goodness of fit of the two models.

Then to find out the difference in unemployment estimators, we will do the shrinkage for the small areas (51 states) for both of the univariate model and bivariate model. Compare the after shrinkage values of *A* and *K* with the original *A*, *K*.

5. Computation

5.1 Computation of the estimates for national level:

The CPS values of A and K are updated every 10 year and the current values of these two parameters are valid since 2006. So by using the current A=0.3 and K=0.4, we regularly produce the number of unemployment for different estimates, as mentioned in Section 2.1 and 2.2.

Figure1 shows the number of unemployment of the United States from January 2006 to March 2014 when we applied Ratio Estimates, Difference Estimates, Composite Estimates and AK composited Estimates. The Ratio Estimates, the orange line is stays higher than other estimates for most of the time, while the Difference Estimates, the green line, appears lower. And the other two composite estimates are very close and sit between the orange and green lines, which accord with the fact that the composite estimates are weighted average for Ratio and Difference Estimates.

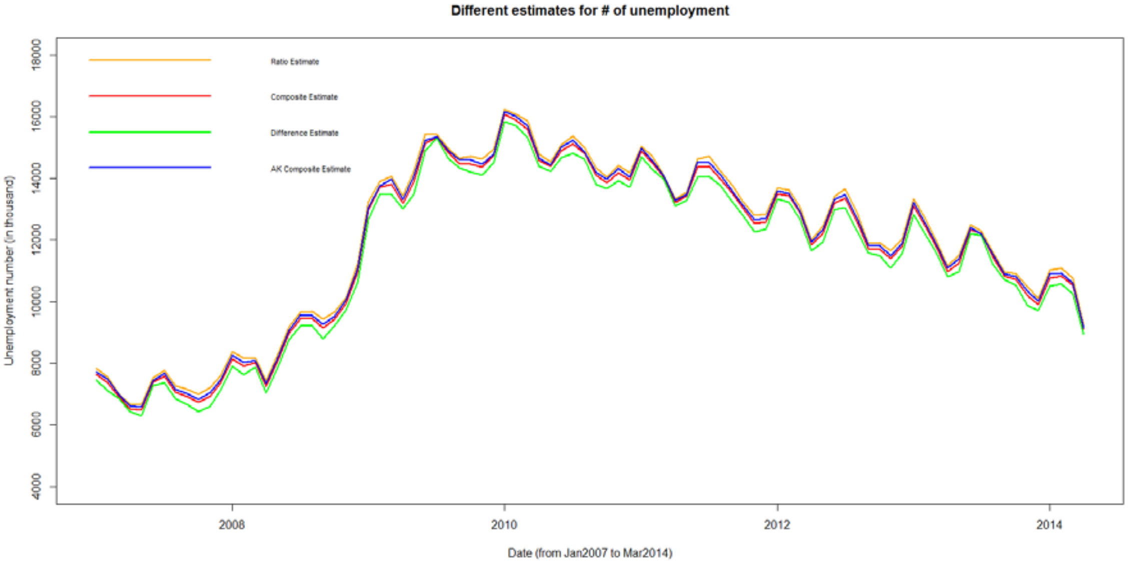

Figure1: Comparison among different estimates for the national unemployment

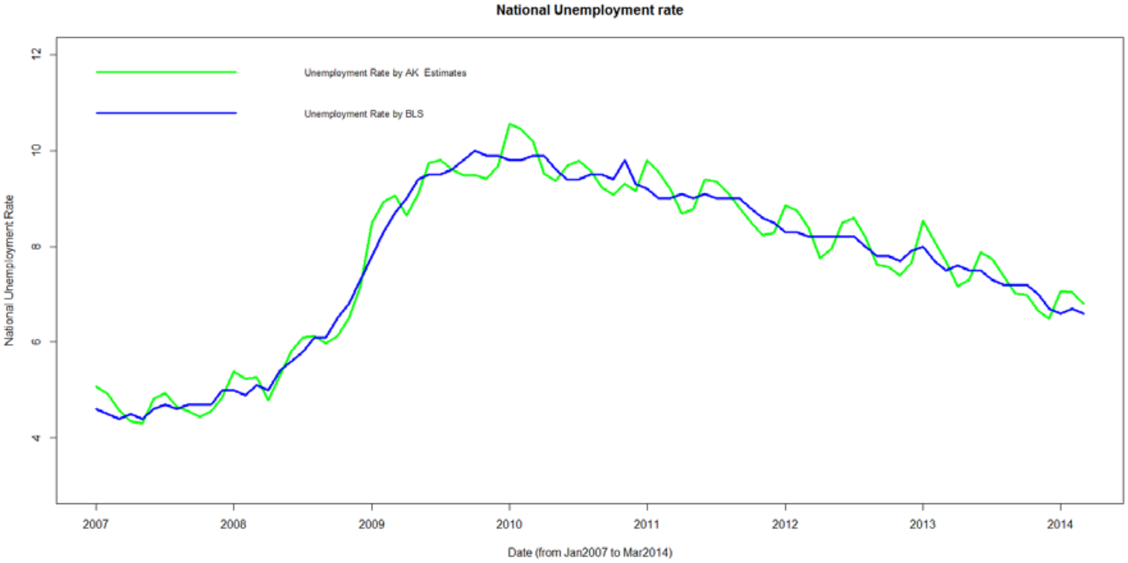

Figure2. National Unemployment rate

Figure2 compares the national unemployment rate by using AK Composite Estimates with the national unemployment rate after the seasonal adjustment produced by BLS. Our result of the unemployment rate by AK Composite Estimate matches well with the official data. And obviously, the curve with the seasonal adjustment is smoother than that of the after second stage.

Then, we will take a look at the variance estimation for the AK Composite Estimates. Figure 3 (a) gives us the result of the value of the Variance Estimation (simulation by the SDR method in Section 2.3), which is around the range from 10^{10} to 5×10^{10}. And Figure 3 (b) shows the coefficient of variation (CV). The values are all stay much lower than 5% which provide us with the evidence that our AK Composite estimates and Variance Estimation works well under the circumstances as well as the data we observe.

5.2 Optimal AK values for the national level:

Despite of the current A, K values of CPS, we want to find out the optimal A and K by using the criteria we mentioned in section 3.1 base on the data from Jan 2006 to Mar 2014. The simulation of the total Variance Estimation, $\sum_t \hat{V}_{SDR}(\hat{Y}_t^{AK})$, by different combination of A and K is given in Table 2. Table 2 shows only part of the result for the 441 combinations. However, it gives us the minimum value of the total Variance Estimation, which can be obtained by choosing $(A, K) = (0.3, 0.25)$.

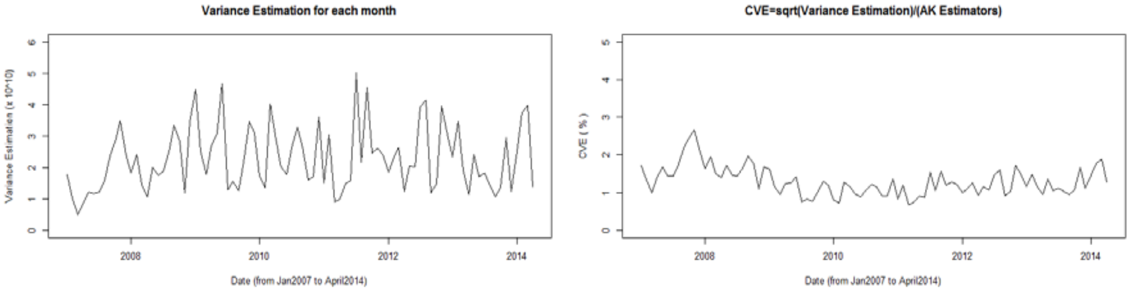

Figure 3: Variance estimation for the AK composite estimates. Left panel (a) Variance Estimation. Right panel (b) Coefficient of Variation

Table 2. Total Variance Estimation for different combination of A and K ($\times 10^{11}$)

		K				
		0.15	**0.20**	**0.25**	**0.30**	**0.35**
A	**0.20**	*0.8316*	*0.7723*	*0.8077*	*0.9169*	*1.1034*
	0.25	*0.7982*	*0.7132*	*0.7121*	*0.8017*	*0.9197*
	0.30	*0.7869*	*0.6950*	***0.6721***	*0.7392*	*0.8270*
	0.35	*0.8555*	*0.7266*	*0.6789*	*0.7064*	*0.7783*
	0.40	*0.9310*	*0.7914*	*0.7243*	*0.7001*	*0.8076*

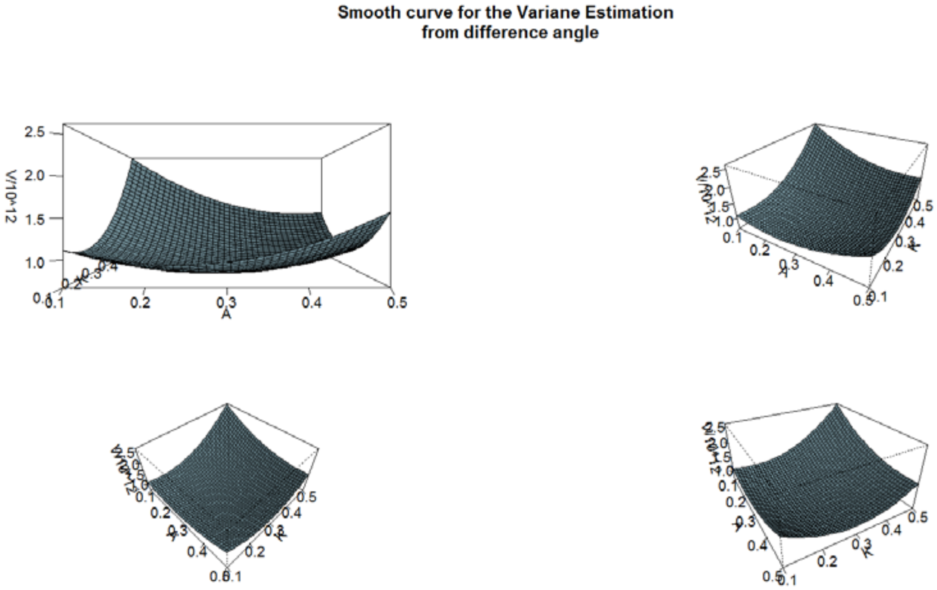

Figure 4: Smooth curve for the Variance Estimation from different angle

Figure 4 is the smooth curve for the total Variance Estimation in terms of A and K. And the curve gives us the rough result of where the minimum value of total Variance Estimation locate, which is around $(A, K) = (0.315, 0.245)$. Note that although $A = 0.3$ and $K = 0.4$ are the current value for the AK Composite Estimate, the total variance estimation would be larger.

To conclude: firstly, the measure here is the variance of monthly level unemployment instead of month to month change or annual average or a compromise between all of the three; secondly, with $A = 0.315$ almost the same as it of the current value, $K = 0.245$ is smaller than the current value $K = 0.4$, which means there is less emphasis on Difference Estimate and more on the Ratio Estimate.

5.3 Computation of the Estimates for the State levels

Similarly, we generate the optimal A, K for each state by applying the criteria we mentioned in Section 3.2. Figure 5 shows the scatter of $(A_j, K_j)_{opt}$ for the 51 states.

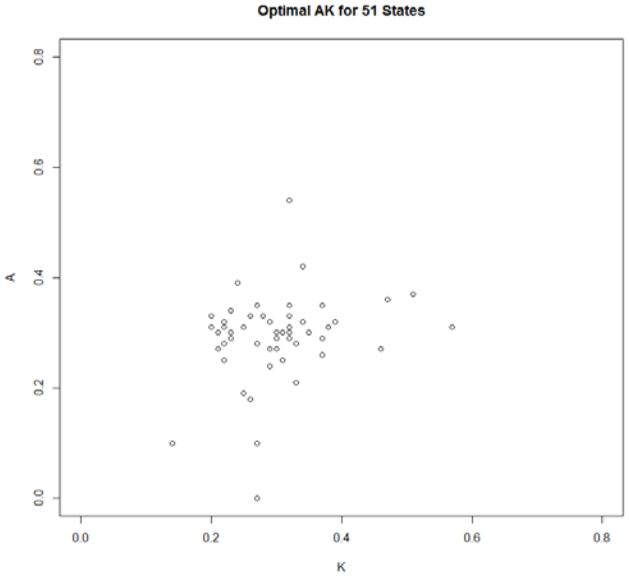

Figure 5. Optimal (A, K) for the Fifty-one States

Though the estimates and variance estimation show great results on the national level, we mainly focus on the small areas. Figure 6 shows the Ratio Estimates (upper left), Ratio & AK Estimates (upper right), Relative difference of Ratio and AK Estimates (lower left) and Unemployment rate (lower right) for Texas. At this point, the parameter $(A, K) = (0.3, 0.4)$ for estimating the unemployment will not be applied due to the way it is calculated. So we are using the optimal A, K values for Texas.

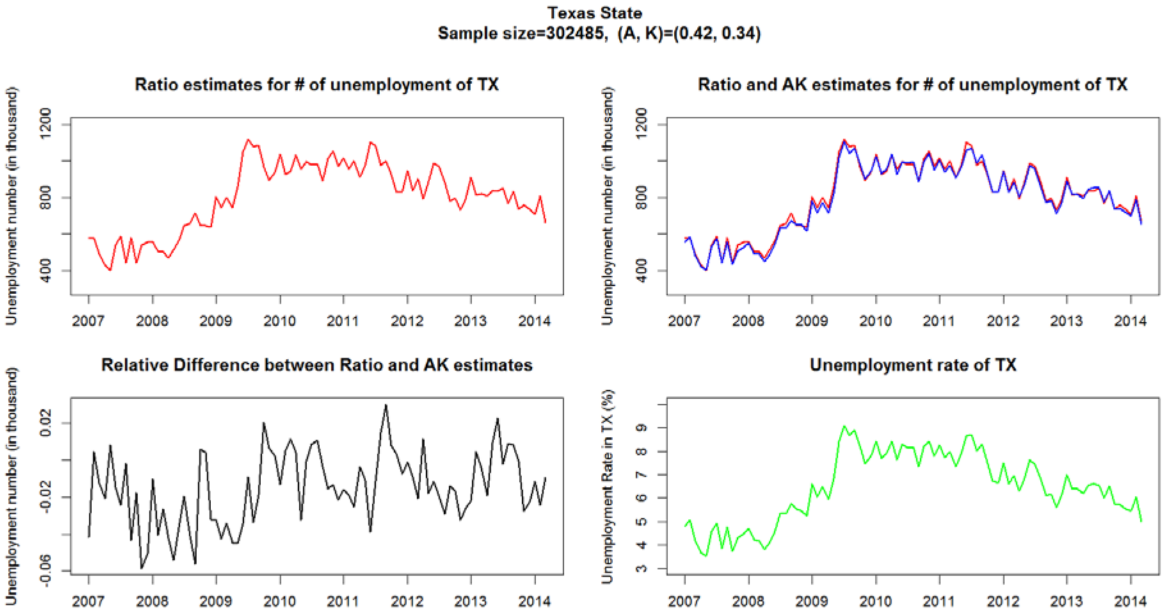

Figure 6. State of Texas: plots over time using optimal (A,K) for Texas.

Note that the Relative differences for these states are around the range from -0.1 to 0.05 and in the negative percentage mostly, which means the Ratio Estimates are usually a little larger than the AK Composite Estimates. Also note the rate of the unemployment has almost the same shape with the unemployment AK estimates.

Figure 7 below shows the unemployment rate for the United States with four states: California, Texas, Michigan and New York. The unemployment rate of California and Michigan are the highest and above the average (national unemployment curve), and Texas has the lowest employment rate among these four states.

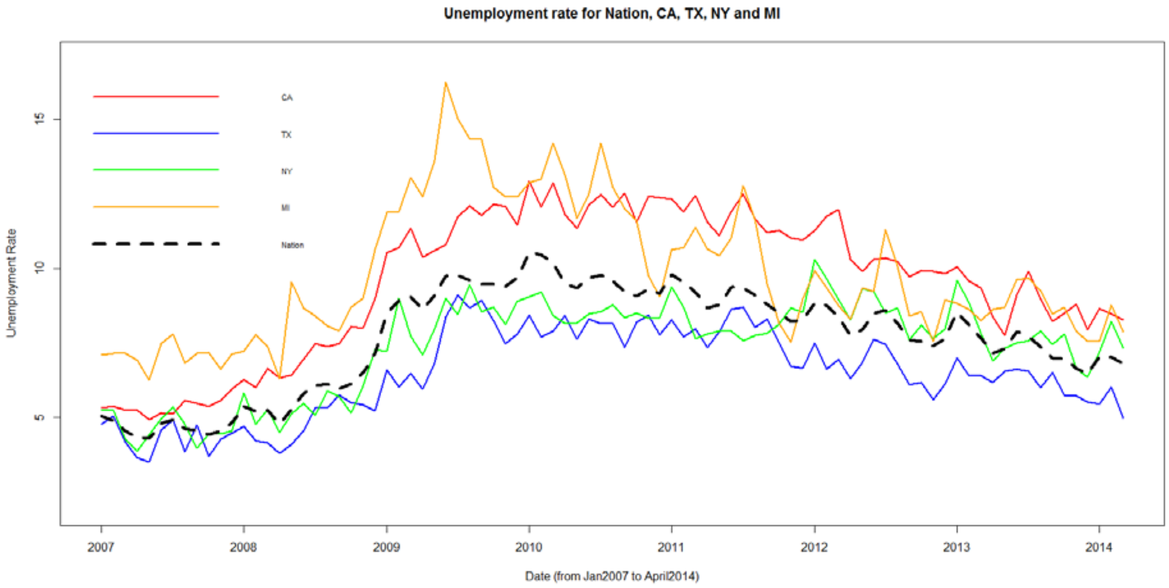

Figure 7. Unemployment rate for California, Texas Michigan and New York

In this work, we have noticed potential clusters of the states. Since the unemployment in 2007 fluctuate around 5 percent and around 10 percent in 2010, so we will cluster the states base on these two factors.

Table 3. Unemployment rate for States (%)

State	Unemployment Rate in 2007	Unemployment Rate in 2010	above 5% in 2007	above 10% in 2010
CA	5.378346	12.21924	T	T
FL	4.115864	11.09876	F	T
KY	5.415782	10.32309	T	T
MA	4.660328	8.542498	F	F
MD	3.615926	7.754744	F	F
NV	4.671826	14.48302	F	T
NY	4.652118	8.535516	F	F
OK	4.485344	7.148713	F	F
TN	4.658436	9.453955	F	F
TX	4.322707	8.04798	F	F
UT	2.622853	8.203455	F	F
WA	4.64582	10.21469	F	T

Table 3 above shows the unemployment rate of both 2007 and 2010 for 12 States. And Table 4 below shows the cluster result. Florida, Nevada and Washington may hugely affected by the bad economy, while Massachusetts, Maryland, New York, Oklahoma, Tennessee, Texas and Utah may not affected as much. Most importantly, no state has unemployment rate higher than 7% in 2007 could control the rate to lower than 10% in 2010.

Table 4. Cluster of the 13 States

	Lower than 10% in 2010	***Higher than 10% in 2010***
Lower than 5% in 2007	MA, MD, NY, OK, TN, TX, UT	FL, NV, WA
Higher than 5% in 2007		CA, KY

The Table 4 above shows that if the unemployment rate in 2007 is lower than 5%, then in 2010 it could be either lower or higher than 10% due to the economy effect. But if the unemployment rate in 2007 is higher than 5%, it will only be higher than 10% in 2010 mostly. In another words, the states with higher unemployment rates usually would stay higher.

6. Planned Research Work

This section describes in brief planned research work. The first major extension of current work is to fit a hierarchical model to data from multiple states simultaneously. Such as multi-state model should have better shrinkage properties than single-state models and reduce dependence on prior distributions. Work involved in this task includes specifying, fitting, and summarizing both the simple and multilevel models.

A second major extension of current work is to carefully evaluate the estimates under different models and then to simulate performance of procedures. One way to compare estimates is in terms of mean squared error. As no true value is known for the unemployment rate one must compare to direct survey estimates or use an approach to estimate mean squared error. A leave-one-out Jackknife approach has been used in other small area estimation contexts to estimate mean squared error under models similar to the ones planned for the (A, K) values.

Simulation can be approach many ways. One approach that could be valuable in this situation is to fit a model to unemployment rates by state over time and then simulate a population level unemployment rate for a state. One then can sample from the population assuming a given unemployment rate over time. Such a simulation has complexities in implementation. Once the sampling process is implemented it will be possible to apply estimation approaches and compare performance of estimators. The outline of the simulation study then will be as follows. Simulate data under a model for a state. Fit a model to a state's estimates using chosen values of (A, K). Then use the model results to generate unemployment values for the state. Compare performance of estimators in terms of bias, variance, and MSE. Also compute coverage of confidence intervals to the values used to setup the simulation. The simulation will need to be expanded to multiple states to study larger hierarchical models.

References

[1] Bailar, B. (1975), ‘‘The Effects of Rotation Group Bias on Estimates from Panel Surveys,’’ *Journal of the American Statistical Association*, Vol. 70, pp. 23−30.

[2] Breau, P. and L. Ernst (1983), ‘‘Alternative Estimators to the Current Composite Estimator,’’ *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 397−402.

[3] Breidt, FJ; Fuller, WA. (1999). “Design of supplemented panel surveys with applications to the National Resources Inventory,” *Journal of Agricultural Biological and Environmental Statistics*, 4(4): 391-403.

[4] Cheng, Yang. (2012). “Overview of Current Population Survey Methodology,” *Proceedings of the Survey Research Methods Section*, *American Statistical Association,* pp. 3965-3979.

[5] Fay, R., Train, G., (1995). "Aspect of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," *Proceedings of the Section on Government Statistics, American Statistical Association, 154-159.*

[6] Lent, J., S. Miller, and P. Cantwell (1994), ‘‘Composite Weights for the Current Population Survey,’’ *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 867−872.

[7] Lent, J., S. Miller, P. Cantwell, and M. Duff (1999), ‘‘Effect of Composite Weights on Some Estimates From the Current Population Survey,’’ *Journal of Official Statistics*, Vol.15, No. 3, pp. 431−438.

[8] Mansur, K.A. and H. Shoemaker (1999), ‘‘The Impact of Changes in the Current Population Survey on Time-in-Sample Bias and Correlations Between Rotation Groups,’’ *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 180−183.

[9] Shoemaker, H. (2004). ‘‘Redesign of the Sample for the Current Population Survey,’’ *Employment and Earnings*, 51(12): 4-8. Washington, DC: Government Printing Office, December 2004.

[10] Tegels, Robert and Lawrence Cahoon, (1982), ‘‘The Redesign of the Current Population Survey: The Investigation Into Alternate Rotation Plans,’’ *Proceedings 1982 Joint Statistical Meetings, American Statistical Association*.

[11] U.S. Census Bureau, (2006). “Design and Methodology: Current Population Survey, Technical Paper 66”, October 2006.