

# A Cost-Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach

Ali-Reza Khoshgooyanfar and Mohammad Taheri Monazzah<sup>1</sup>

## Abstract

This paper primarily aims at proposing a cost-effective strategy to estimate the intercensal unemployment rate at the provincial level in Iran. Taking advantage of the small area estimation (SAE) methods, this strategy is based on a single sampling at the national level. Three methods of synthetic, composite, and empirical Bayes estimators are used to find the indirect estimates of interest for the year 1996. Findings not only confirm the adequacy of the suggested strategy, but they also indicate that the composite and empirical Bayes estimators perform well and similarly.

Key Words: Composite estimator; Design-based estimator; Empirical Bayes estimator; Indirect estimator; Non-sampling error; Synthetic estimator; Post-strata.

## 1. Introduction

Each year, sample surveys are conducted in Iran to obtain statistical information required for decision and policy making. However, these surveys cannot fulfill all statistical requirements because of two factors. The first one is related to the governmental and non-governmental sectors' demand for comprehensive statistical information not only at national and regional but also at small area levels. Further, they need the information at shorter periods of time per year, say monthly or quarterly. The second factor is that the main source of statistical data in Iran is surveys, and there are financial limitations for conducting surveys several times per year at small area levels. These two factors challenge statistical agencies to find efficient strategies to balance both cost and statistical information quality. The work presented here is an endeavor to overcome this challenge by using small area estimation (SAE) methods.

The purpose of SAE methods is to provide acceptable estimates for some subpopulations in a sample design planned for the "whole" population regardless of the subpopulations. For example, a sample design is planned for estimating population parameters for the "country" and after data collection the parameters are estimated by the national sample data. If simultaneously "provincial estimates" of the parameters are needed, it is not possible to conduct separate provincial sample surveys. The provinces are unplanned subpopulations in the sense that the available sample design has been planned just for estimating the parameters for the country without considering the provincial level. In the nationwide sample, few or no sample units may be available for some provinces. Hence, acceptable estimates for such provinces (subpopulations) cannot be produced.

Before the availability of SAE methods, such subpopulation estimates were obtained by direct design-based estimation. If there were data from a given subpopulation in the nationwide sample, an estimate would be directly calculated according to the nationwide sample design by using "the available data". The direct estimate may differ substantially from the actual subpopulation parameter due to large sampling errors owing to small sample size.

Statisticians and demographers have developed ways of estimating for such subpopulations. Indirect estimators have been suggested and applications have been increasing over the last twenty years. However, the SAE methods are still an active topic of study. See Purcell and Kish (1979, 1980), Ghosh and Rao (1994), Schaible (1995), Marker (1999), Pfeffermann (2002) and especially Rao (2003a) for problem definition and a review of the SAE methods.

For a number of years, the Statistical Center of Iran (SCI) implemented annually a national one-stage cluster sample in order to estimate the intercensal unemployment rate at the country level. For sixteen years, separate one-stage cluster samples for all provinces have been conducted to estimate provincial unemployment rates. A weighted combination of provincial estimates then yields the unemployment rate for the total country. The increasing need for estimation of the unemployment rate at a provincial level on a monthly, or at least seasonal basis, and the lack of administrative records in Iran at both small and national levels persuaded SCI to try the SAE methods as the core of a revised strategy to meet the provincial need.

The revised strategy consists of designing a sample survey only at the national level and producing the provincial estimates by SAE methods. A province in the strategy is a small area. This strategy demands a smaller sample size than that for aggregating provincial samples. If the revised

1. Ali-Reza Khoshgooyanfar, The Center for Research, Studies and Program Assessments of IRIB. E-mail: khosh\_ar@yahoo.com; Mohammad Taheri Monazzah, The Central Bank of Iran. E-mail: Taheri53@yahoo.com.

strategy proves practicable, time and cost of the data collection can be reduced, and produce provincial estimates on a monthly basis. The smaller sample is easier to control in the field, and estimates are less affected by nonsampling errors.

This paper is intended to answer the following questions:

1. Can a nationwide sample substitute for separate provincial samples for making estimates of the provincial unemployment rates?
2. From the three SAE methods – synthetic, composite and empirical Bayes estimators – which one produces the best estimates?

To answer empirically these two questions, estimates were produced for the year 1996 when the actual values of the provincial unemployment rates are available from the 1996 Census. Consequently, the actual bias of each provincial estimate can be computed.

The process includes the following three stages. First, a sample of size 13,000 from the whole country is selected (the 1996 Census data file). The sample size is determined at the national level, and is allocated to all provinces proportionally to population. The allocation provides sample from each province enabling direct estimates of the unemployment rate for each province. Direct estimates are not necessarily acceptable for all provinces because of the large sampling errors due to small sample sizes in some provinces. Second, applying three SAE methods, indirect estimates are produced for each province. Third, the indirect estimates are evaluated by comparing them with corresponding actual values, computing MSEs, mean of absolute errors (MAE), and mean of errors (ME).

In addition to this introduction, the paper takes in three more sections. Section 2 offers a short review of the three estimators used in this paper, including the estimation methods, their corresponding MSEs, and properties of the estimators. The estimates and corresponding computational aspects are presented in section 3, where performances of the estimators are tentatively appraised. Section 4 is devoted to final remarks and recommendations about the estimators and the merit of the SAE strategy.

## 2. A Glance Over the Estimators

Indirect estimators used in the study are introduced briefly. However, an excellent discussion of the SAE methodology is in Rao (2003a). First, the synthetic estimator is considered, and then the composite estimator. The empirical Bayes (EB) estimator as a model-based estimator is also considered.

### 2.1 Synthetic Estimator

There is a family of small area estimators characterized as synthetic, see Rao (2003a, chapter 4). The traditional and simplest is discussed here. For this estimator,

1. The country is partitioned into six post-strata on the basis of six age groups, see Table (1).
2. Next, the number of unemployed persons is estimated in each province, providing the numerator in expression (1).
3. The synthetic estimate of the  $i^{\text{th}}$  province is obtained by dividing the estimated number of unemployed persons in province  $i$  by its Economically Active Population (EAP), namely

$$\hat{P}_i^S = \left( \sum_{j=1}^6 N_{ij} \hat{P}_j \right) / N_i \quad (1)$$

where  $\hat{P}_j$  is a direct design-based estimate of the unemployment rate in post-stratum  $j$ ,  $N_i$  is the EAP in province  $i$ , and  $N_{ij}$  is the EAP in the intersection of province  $i$  and post-stratum  $j$ , cell  $(i, j)$ . The synthetic estimate of the  $i^{\text{th}}$  province is according to the official definition of the unemployment rate in Iran.

The synthetic estimate shares all national sample data by using national direct estimates of the unemployment rate from the post-strata. It uses the six estimated “post-strata” unemployment rates computed over all provinces rather than specific estimates of the six “cells”. This process thus **borrow strength** because each province contributes to the national sample by pooling provincial sample units to overcome small sample sizes in each province.

This estimator has three limitations:

1. The smaller the inter-post-stratum variation is, the better synthetic estimator performs. It means that all provinces should have a rather equal unemployment rate in each age group. Using the national post-strata direct estimates equally for all provinces is allowable only under this assumption. If the homogeneity assumption is not satisfied, the synthetic estimator cannot reflect specific small area variation, and the estimates could be severely biased.
2. If there are several variables that are important in post-stratification, the synthetic estimator cannot often use all of them because post-strata (after cross-classification of the several variables) have sample sizes that are too small and yield unacceptable direct estimates of the post-strata. Generally speaking, many post-strata give rise to poor direct estimates for some of the post-strata. This can create serious problems for synthetic estimation when a poor direct estimate receives a large EAP for a cell.

- Quality of the EAPs can affect the synthetic estimates. Owing to lack of timely data sources such as administrative records, out of date EAPs from the 1986 Census data are used here in order to produce the synthetic estimates for the year 1996.

### 2.2 Composite Estimator

The composite estimator of the  $i^{\text{th}}$  province combines the synthetic and direct estimators of that province, namely

$$\hat{P}_i^C = W_i \hat{P}_i^D + (1 - W_i) \hat{P}_i^S \tag{2}$$

where  $\hat{P}_i^D$  is the direct design-based estimator for the  $i^{\text{th}}$  province, and  $0 \leq W_i \leq 1$ . Expression (2) improves upon (1) by exploiting both estimators. That is, provincial differences may take into account in the composite estimator via the provincial unbiased direct estimates and instability of the direct estimator may be reduced via the synthetic estimator.

The weight  $W_i$  can be specified so as to minimize mean square error of  $\hat{P}_i^C$ ,  $MSE(\hat{P}_i^C)$ . Assuming  $Cov(\hat{P}_i^D, \hat{P}_i^S) \cong 0$ , the weight is simplified as

$$W_i^{\text{opt}} = \frac{1}{(V(\hat{P}_i^D) / MSE(\hat{P}_i^S)) + 1} \tag{3}$$

where  $V(\hat{P}_i^D)$  and  $MSE(\hat{P}_i^S)$  are the variance of  $\hat{P}_i^D$  and the mean square error of  $\hat{P}_i^S$ , respectively. In expression (3), the weights of the direct and synthetic estimators in (2) are proportional to the MSEs of the two estimators. See Schaible (1978) and Rao (2003a, page 58) for properties of the estimator and weight.

In practice, we should estimate  $MSE(\hat{P}_i^S)$  and  $V(\hat{P}_i^D)$  to generate an estimate of the weight (3). If there are some sample data from the  $i^{\text{th}}$  province, according to the sample design, an unbiased design-based estimator of  $V(\hat{P}_i^D)$  can be computed by using only the sample data. Therefore, only an estimator for  $MSE(\hat{P}_i^S)$  is required. Under the assumption that  $Cov(\hat{P}_i^D, \hat{P}_i^S) \cong 0$ , Ghosh and Rao (1994) proposed the unbiased estimator

$$M\hat{S}E(\hat{P}_i^S) = (\hat{P}_i^S - \hat{P}_i^D)^2 - \hat{V}(\hat{P}_i^D). \tag{4}$$

Under the same assumption, one can easily show that

$$MSE(\hat{P}_i^C) = W_i^2 V(\hat{P}_i^D) + (1 - W_i)^2 MSE(\hat{P}_i^S). \tag{5}$$

The estimator (4) may result in negative estimates for some provinces, and the weight in expression (3) is no longer computable. In this case, instead of (3) and (4), we have used respectively the combined weight in (6) and  $AMSE = (1/I') \sum_{i=1}^{I'} M\hat{S}E(\hat{P}_i^S)$  where  $I'$  is the number of small areas having positive estimated MSE (see Gonzalez and Waksberg (1973) for more details):

$$W^C = \frac{1}{\left( \frac{\sum_i \hat{V}(\hat{P}_i^D)}{\sum_i M\hat{S}E(\hat{P}_i^S)} \right) + 1}. \tag{6}$$

In addition to expressions (3) and (6), Copas (1972), Ghosh and Rao (1994), and Thompsen and Holmoy (1998) suggest alternative weights.

### 2.3 Empirical Bayes (EB) Estimator

Model based SAE methods have received more attention than the synthetic and composite esitmaters. Marker (1999) regarded the SAE methods as having a common element expressed through regression models. The EB method is of the regression type. Consider the following mixed model (see Rao (2003a, page 76)):

$$\mathbf{g} = X\boldsymbol{\beta} + \mathbf{v} + \boldsymbol{\varepsilon} \tag{7}$$

where

$$\mathbf{g}' = (Ln \frac{\hat{P}_1^D}{1 - \hat{P}_1^D}, \dots, Ln \frac{\hat{P}_I^D}{1 - \hat{P}_I^D}),$$

$X$  is an  $I \times k$  design matrix of supplementary variables,  $\boldsymbol{\beta}$  is a  $k \times 1$  vector of unknown parameters, and  $\mathbf{v}$  and  $\boldsymbol{\varepsilon}$  are  $I \times 1$  random vectors ( $I$  is the number of provinces). Assume that:

- $\mathbf{v}$  and  $\boldsymbol{\varepsilon}$  are independent.
- $E(\boldsymbol{\varepsilon}) = 0$  and  $Var(\boldsymbol{\varepsilon}) = \text{Diag}(d_1^2, \dots, d_I^2)$ .
- $\mathbf{v} \sim N(0, \boldsymbol{\Sigma})$  where  $\boldsymbol{\Sigma} = \text{Diag}(t^2, \dots, t^2)$ .

Ghosh and Meeden (1997) show that the EB estimate of the  $i^{\text{th}}$  element of  $\mathbf{g}$  is:

$$\hat{g}_i^{\text{EB}} = \hat{W}_i \mathbf{x}'_i \hat{\boldsymbol{\beta}} + (1 - \hat{W}_i) g_i \tag{8}$$

where  $\mathbf{x}'_i$  and  $g_i$  are the  $i^{\text{th}}$  row and the  $i^{\text{th}}$  component of  $X$  and  $\mathbf{g}$  respectively, and  $\hat{W}_i$  is an estimate of

$$W_i = \frac{d_i^2}{d_i^2 + t^2}. \tag{9}$$

Consequently, the EB estimate of the  $i^{\text{th}}$  rate is:

$$\hat{P}_i^{\text{EB}} = \frac{\exp(\hat{W}_i \mathbf{x}'_i \hat{\boldsymbol{\beta}} + (1 - \hat{W}_i) g_i)}{1 + \exp(\hat{W}_i \mathbf{x}'_i \hat{\boldsymbol{\beta}} + (1 - \hat{W}_i) g_i)}. \tag{10}$$

It is obvious that (10) needs two estimates for  $\boldsymbol{\beta}$  and the weight in (9). On the other hand, the weight in (9) relies on the estimates of  $t^2$  and  $d_i^2$ . By applying the delta method,  $(g'_i)^2 \hat{V}(\hat{P}_i^D)$  generates an estimate of  $d_i^2$  where  $g'_i$  through the first derivative of  $g_i = Ln(\hat{P}_i^D / 1 - \hat{P}_i^D)$ . Based on Chand and Alexander (1995), estimates of  $\boldsymbol{\beta}$  and  $t^2$  are found by simultaneously solving

$$\begin{cases} t^2 = (\mathbf{g} - X\boldsymbol{\beta})'V^{-1}(\mathbf{g} - X\boldsymbol{\beta})/(I - k) \\ \boldsymbol{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{g} \end{cases} \quad (11)$$

where  $V = \text{Diag}(d_1^2 + t^2, \dots, d_j^2 + t^2)$ . Note that the equations in (11) are solved by numerical iteration with an initial value for  $t^2$ .

The EB and composite estimators have similarities although they arise from different approaches. Both estimators have two components; a direct component ( $\hat{P}_i^D$  in (2) and  $g_i$  in (8)) computed from the provincial sample data, and an indirect component ( $\hat{P}_i^S$  in (2) and  $\mathbf{x}'_i \hat{\boldsymbol{\beta}}$  in (8)) constructed from the national sample data and supplementary information. Both estimators (2) and (8) give more weight to the indirect component when it is reliable. Otherwise the direct component receives more weight. Additional details are given in Cressie (1989), Ghosh *et al.* (1998) and Rao (2003 a, b).

### 3. Estimation for Iran

Estimates were produced for the year 1996 because the 1996 actual unemployment rate of each province is known from the 1996 Census. As a result, the actual bias of each estimate can be computed.

In 1996 the country consisted of 26 provinces. However, 21 provinces are studied here because supplementary information from the 1986 Census was available for 21 geographically unchanged provinces between the years 1986 and 1996. To make the three indirect estimates, at the national level, a sample was planned and its sample size was determined for estimating the unemployment rate of the country as a whole. Each province is a small area. The national sample was allocated among the 26 provinces proportional to population in order to have sample data from each province (a top-down approach). This enabled direct design-based estimates for each province and its corresponding variance required for both the EB and composite estimators. The sample design is able to produce good estimates for the country and for some provinces.

#### 3.1 Computational Aspects

To construct synthetic estimates, six age groups formed the post-strata. The estimated unemployment rate of each group based on the national sample and its corresponding actual value based on the 1996 Census are presented in Table (1), which also contains absolute errors of the estimates.

The estimates for the first two groups have very large error. Therefore, if a province in expression (1) gives large EAPs to these age groups, its synthetic estimate may not

perform well. The 1986 Census data were used in computing the EAPs for all provinces and cells ( $N_i$  and  $N_{ij}$  in expression (1)) because, in the absence of administrative records, the nearest census to the year 1996 is the main source of data at any level.

**Table 1**  
Post-strata Characteristics

Age Group	Estimated Rate ( $\hat{P}_i$ )	Actual Value	Absolute Error
10–15	0.3240	0.2826	0.0414
16–20	0.2402	0.2629	0.0227
21–25	0.1868	0.1856	0.0012
26–30	0.0811	0.0802	0.0009
31–50	0.0363	0.0366	0.0003
More than 50	0.0653	0.0648	0.0005

To construct the composite estimates, provinces were divided into two groups. The first consists of 14 provinces that used the weight in expression (3), and the second of seven provinces used the common weight  $W^C = 0.873184$  based on (6). Because the estimator in expression (4) produces negative estimates for  $\text{MSE}(\hat{P}_i^S)$  for these seven provinces, the AMSE was used.

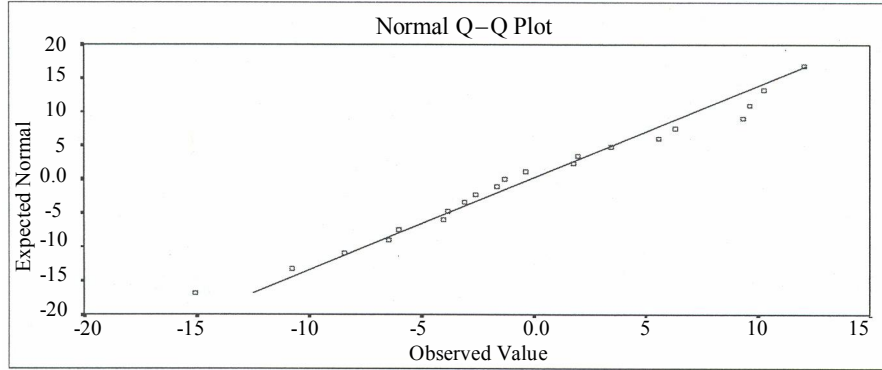
To construct the EB estimates,  $d_i^2$  was estimated by using the delta method and then  $t^2$  was estimated following Prasad and Rao (1990) by using a SAS/IML program (the program is available from the authors). An initial estimate for  $t^2$  is required in this program, and was calculated by the moment estimation method as  $t^2 = 0.3117194$ . To solve the equations in (11), the following  $21 \times 2$  design matrix was used, whose first and second columns are 1s and EAPs, respectively:

$$X = \begin{bmatrix} 1 & 133,449 \\ 1 & 141,124 \\ 1 & 883,653 \\ 1 & 795,714 \\ \vdots & \vdots \\ 1 & 522,976 \\ 1 & 162,892 \end{bmatrix}$$

The estimated  $t^2$  and  $\hat{\boldsymbol{\beta}}$  are

$$\hat{t}^2 = 0.5596389, \hat{\boldsymbol{\beta}} = \begin{pmatrix} -2.066874 \\ -1.273 \times 10^{-7} \end{pmatrix}$$

To test normality, a normal Q–Q plot and a Shapiro-Wilk’s test for standardized residuals of the fitted model were examined. The points in the Q–Q plot are close to a straight line, and the test did not reject the null hypothesis of normality ( $p$ -value = 0.851).



### 3.2 Results

The results are organized into four parts. First, bias in the forms of error and absolute error is examined using two criteria, ME and MAE. Second, MSEs are compared among methods. Third, efficiencies of the indirect estimators relative to the direct estimator are evaluated. Finally, the weights of the direct components in expressions (2) and (8) are analyzed. All the results are depicted in appropriate figures, however, details can be found in Table (2).

Suppose  $S_a$  is the allocated sample size from the national sample to a given province and  $S_r$  the required sample size if the sample size is separately determined for the province. In other words, if there is a sample of size  $S_r$  from the province, an acceptable direct estimate can be then computed for the province. Therefore,  $(S_a/S_r) \times 100$  measures how much the available sample size ( $S_a$ ) is adequate for a given province. This measure is used on horizontal axes of all plots as a basis for comparison sample size effects.

The synthetic estimator has the highest MAE, which was even larger than that of the direct estimator (see Figure 1). Conversely, MAEs of the composite and EB estimators are the lowest, and very similar to one another. Based on ME, there is a slight overestimation of the actual value for all estimators. The direct estimator has the lowest ME because it is unbiased. MEs of the composite and EB estimators are close and the synthetic estimator has the highest ME.

For the direct, composite, and EB estimators, all provinces with  $S_a/S_r \geq 10\%$  have absolute errors less than 0.02. The highest absolute errors belong to Ilam and Kohgiluyeh & Boyerahmad which have the smallest populations and very small  $S_a/S_r$ . Plots of these three estimators have relatively similar patterns. The story is different for the synthetic estimator because the “national” sample data are only used in making synthetic estimates through the post-strata direct estimates and then the “national” sample size (not  $S_a/S_r$ ) affects the synthetic estimate of a province through the cell EAPs. In other words, if a post-stratum does not have “enough” national

sample data to yield acceptable direct estimates, and a province gives large EAP to the post-stratum direct estimate, the province has a poor synthetic estimate. This is the case for Sistan & Baluchestan, Bushehr, Tehran and Lorestan because of poor direct estimates for the first two post-strata (the age groups of 10–15 and 16–20) and large young populations of these provinces.

The lowest MSE always belongs to the composite or EB estimator (see Figure 2). However, MSE of the composite estimator is often lower than that of the EB estimator. The MSE of the synthetic estimator is always higher than those of the other estimators, even the direct estimator.

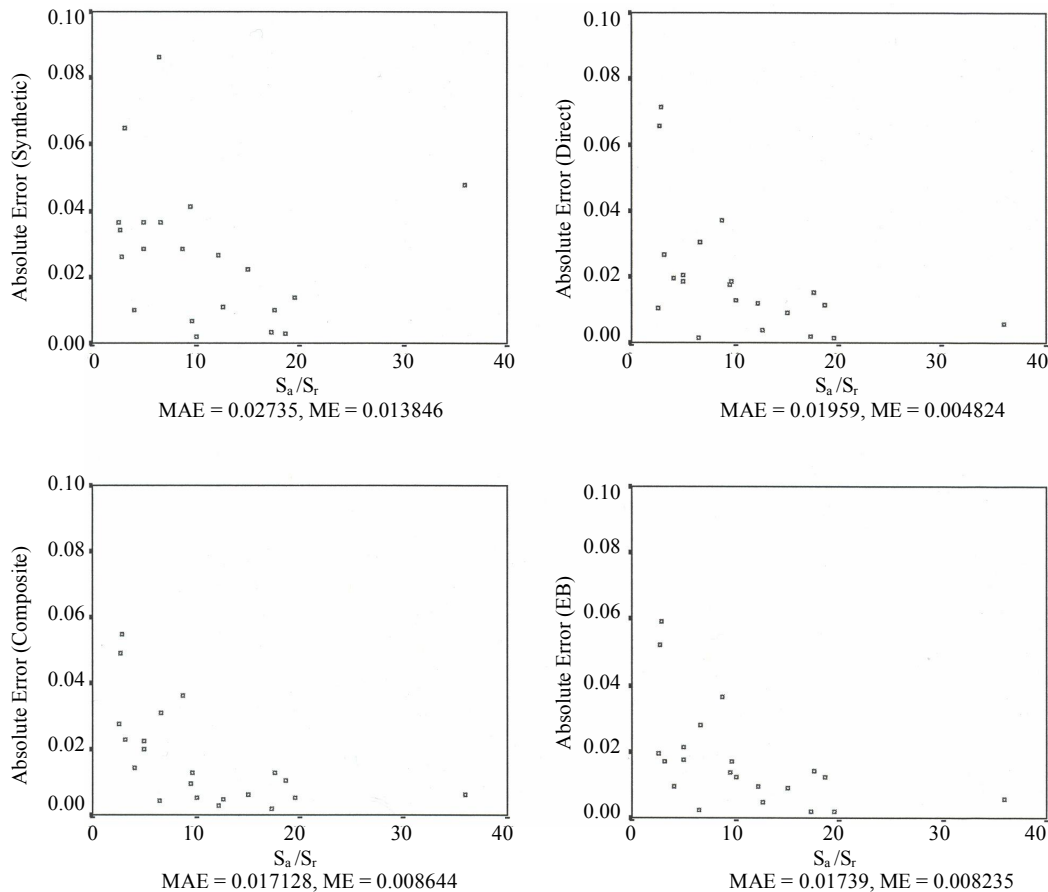
As the  $S_a/S_r$  increases, the MSE decreases for the direct, composite and EB estimators (see the descending trend in Figure 2). This effect is very drastic for Tehran ( $S_a/S_r = 36\%$ ). Again, there are two exceptions for the three estimators, Ilam and Kohgiluyeh & Boyerahmad, both having the smallest populations and very small  $S_a/S_r$ . The pattern of Figure 2 for the synthetic estimator may be misleading because seven provinces used AMSE. However, the four previous provinces (Sistan & Baluchestan, Bushehr, Tehran and Lorestan) also do not conform to the pattern. As a general rule for an estimator, the greater the dependency on the provincial direct estimates the stronger the relationship between the MSE and the ratio  $S_a/S_r$ .

The relative efficiencies (RE) of the three indirect estimators compared to the direct estimator for all provinces are often smaller than or equal to one for the composite and EB estimators and greater than one for the synthetic estimator. Some composite estimates have good REs: Semnan (0.34), West Azarbajejan (0.46), Khorasan (0.70), Kermanshah (0.75) and Hamadan (0.77). Means of REs ( $\overline{RE}^S = 13.6$ ,  $\overline{RE}^C = 0.8595$  and  $\overline{RE}^{EB} = 0.9951$ ) indicate that the composite estimator is the most efficient estimator among the three indirect estimators. Further, in Figure 3 as  $S_a/S_r$  increases  $RE^{EB}$  approaches one. Figure 3 as well as Figure 2 may be misleading for the synthetic estimator.

**Table 2**  
Provincial and Estimator Characteristics

Province	EAP	S <sub>a</sub>	S <sub>r</sub>	S <sub>a</sub> /S <sub>r</sub>	RE <sup>C</sup>	RE <sup>EB</sup>	RE <sup>S</sup>	AE <sup>C</sup>	AE <sup>EB</sup>	AE <sup>S</sup>	AE <sup>D</sup>	MSE <sup>C</sup>	MSE <sup>EB</sup>	MSE <sup>S</sup>	MSE <sup>D</sup>
Bushehr	133,449	146	4,550	3.2%	0.96	1.17	25.57	0.03300	0.01687	0.06501	0.02644	<b>0.0003030</b>	0.000368	0.0080483	0.0003148
Chaharmahal & Bakhtiari*	141,124	203	4,063	5%	0.87	0.95	6.52	0.02136	0.02135	0.03644	0.02031	<b>0.0003813</b>	0.000417	0.0028670	0.0004397
Esfahan	883,653	1032	5,850	17.6%	0.90	1.00	9.56	0.01268	0.01421	0.00990	0.01504	<b>0.0000533</b>	0.000059	0.0005631	0.0000589
Fars	795,714	925	6,175	15%	0.91	0.99	9.69	0.00610	0.00886	0.02235	0.00904	<b>0.0000836</b>	0.000091	0.0008931	0.0000922
Gilan*	734,196	683	5,364	12.7%	1.04	0.97	17.25	0.00484	0.00460	0.01107	0.00393	0.0001728	<b>0.000162</b>	0.0028670	0.0001662
Hamadan	387,517	439	4,550	9.6%	0.77	1.00	3.36	0.01294	0.01701	0.00675	0.01880	<b>0.0001155</b>	0.000150	0.0005030	0.0001498
Hormozgan*	168,268	198	4,063	4.9%	0.84	0.93	5.12	0.01984	0.01734	0.02821	0.01862	<b>0.0004731</b>	0.000519	0.0028670	0.0005600
Ilam	84,210	111	4,063	2.7%	0.83	0.87	4.94	0.04901	0.05201	0.03395	0.06579	<b>0.0013919</b>	0.001450	0.0082747	0.0016734
Kerman*	312,768	450	5,200	8.7%	0.96	0.97	12.00	0.03615	0.03672	0.02864	0.03724	<b>0.0002283</b>	0.000231	0.0028670	0.0002389
Kermanshah	357,096	436	3,575	12.2%	0.75	0.96	3.07	0.00265	0.00928	0.02641	0.01210	<b>0.0002747</b>	0.000349	0.0011190	0.0003640
Khorasan	1,410,863	1,587	8,125	19.5%	0.70	0.99	2.36	0.00515	0.00193	0.01353	0.00160	<b>0.0000298</b>	0.000042	0.0000999	0.0000424
Khuzestan*	609,044	786	4,225	18.6%	1.03	0.97	16.83	0.01034	0.01247	0.00308	0.01140	0.0001760	<b>0.000166</b>	0.0028670	0.0001704
Kohgiluyeh & Boyerahmad	90,655	105	3,575	2.9%	0.83	0.86	4.83	0.05486	0.05932	0.02630	0.07165	<b>0.0013629</b>	0.001408	0.0079493	0.0016449
Kordestan*	276,575	341	5,200	6.6%	0.91	0.95	9.22	0.03105	0.02814	0.03641	0.03027	<b>0.0002833</b>	0.000297	0.0028670	0.0003111
Lorestan	310,918	341	3,575	9.5%	0.86	0.95	6.22	0.00943	0.01383	0.04101	0.01754	<b>0.0004090</b>	0.000451	0.0029534	0.0004747
Mazandaran*	917,259	1,043	6,013	17.3%	1.30	0.98	33.57	0.00199	0.00188	0.00310	0.00183	0.0001112	<b>0.000084</b>	0.0028670	0.0000854
Semnan	110,166	121	4,713	2.6%	0.34	1.08	0.51	0.02776	0.01929	0.03661	0.01042	<b>0.0001524</b>	0.000491	0.0002317	0.0004542
Sistan & Baluchestan	272,752	318	4,875	6.5%	0.96	0.97	26.53	0.00431	0.00228	0.08606	0.00123	<b>0.0002519</b>	0.000254	0.0069347	0.0002614
Tehran	2,343,290	2,913	8,125	35.9%	0.99	1.00	83.08	0.00605	0.00573	0.04767	0.00555	<b>0.0000209</b>	0.000021	0.0017530	0.0000211
West Azarbayegan	522,976	654	6,500	10.1%	0.46	0.98	0.85	0.00505	0.01247	0.00182	0.01309	<b>0.0000552</b>	0.000118	0.0001024	0.0001199
Yazd	162,892	207	5,038	4.1%	0.82	1.36	4.52	0.01414	0.00968	0.01008	0.01950	<b>0.0001299</b>	0.000215	0.0007164	0.0001586

\*Denote provinces for which expression (3) produces negative estimates for MSEs.  
 EAP: Economically Active Population  
 S<sub>a</sub>: Allocated Sample Size  
 S<sub>r</sub>: Required Sample Size  
 RE: Relative Efficiency  
 AE: Absolute Error  
 MSE: Mean Squared Error (the lowest MSE is bold for each province)  
 C, EB, S and D stand for Composite, Empirical Bayes, Synthetic and Direct estimators, respectively.



**Figure 1.** Absolute errors of the estimates against  $S_a/S_r$ .

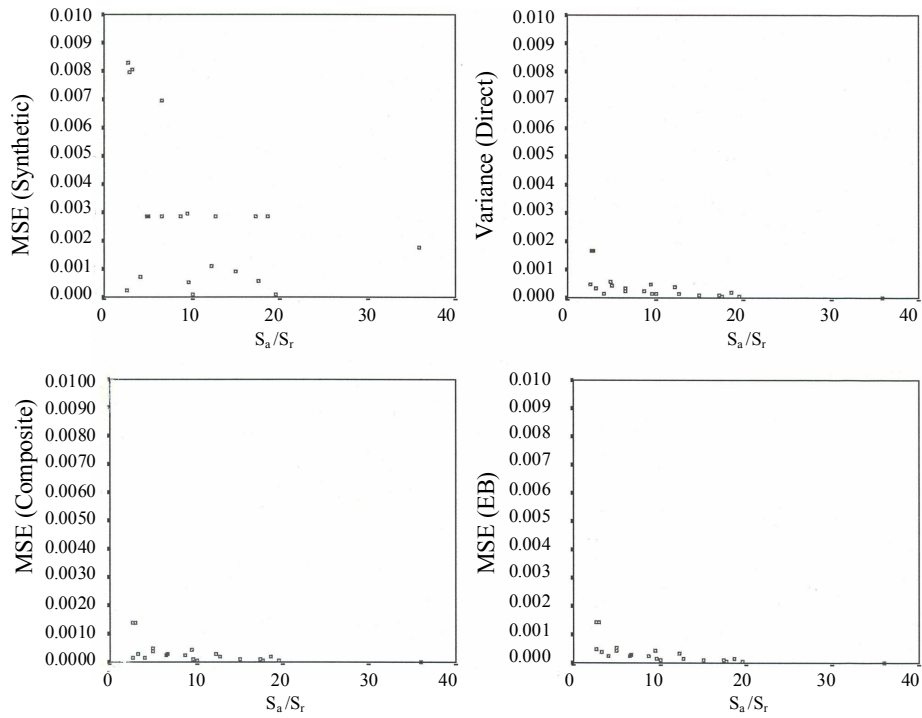


Figure 2. MSEs of the estimates against  $S_a/S_r$ .

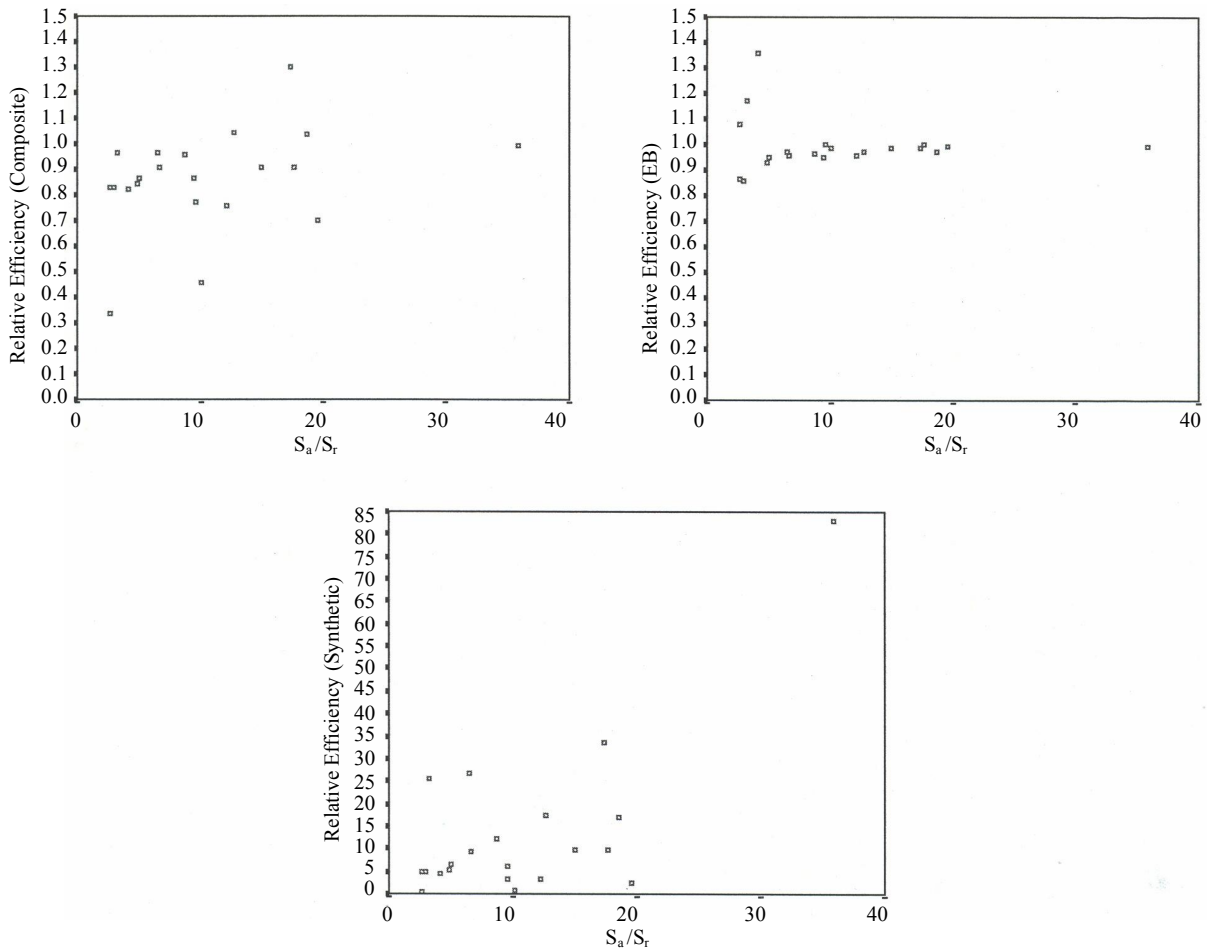


Figure 3. Relative efficiency (Estimated MSE of indirect estimator/Estimated variance of direct estimator) against  $S_a/S_r$  (the plot of synthetic estimator has a different scale on the vertical axis for legibility).

The direct component of the estimator in expression (8),  $g_i$ , always receives more weight than the indirect component. This is the case for the composite estimator, except for two provinces of Semnan and West Azarbayejan. For the composite estimator, Rao (2003a, page 58) states that “the optimal weight  $W_i^{opt}$  will be close to zero or one when one of the component estimators has a much larger MSE than the other, that is, when  $f_i = \text{MSE}(\hat{P}_i^C) / \text{MSE}(\hat{P}_i^S)$  is either large or small. In this case, the estimator with larger MSE adds little information and therefore it is better to use the component with smaller MSE.” This comment is clearly illustrated for Bushehr ( $W = 0.962355$ ,  $RE^S = 25.27$ ), Sistan & Baluchestan ( $W = 0.963670$ ,  $RE^S = 26.53$ ) and Tehran ( $W = 0.988083$ ,  $RE^S = 83.08$ ), because the direct estimates of these provinces have smaller MSEs than the synthetic estimates. Figure 4 clearly shows an ascendant relationship between the weight and  $S_a/S_r$  for the EB estimator. For the composite estimator, the lowest and highest weights pertain to the provinces with the lowest  $S_a/S_r$  and the highest  $S_a/S_r$  respectively.

In general, the synthetic estimator performs poorly based on the MAE, ME, MSE and RE criteria, even though the synthetic estimates of some provinces are individually closer to actual values than other estimates. However, the synthetic estimates have been computed under the most disadvantageous conditions. The EAPs applied to construct the synthetic estimates are based on the 1986 Census (ten years before the year when the estimates were produced). In addition, the direct estimates of the first two post-strata are quite different from the other post-strata, causing poor synthetic estimates.

To address the first problem, administrative records should be developed; for the second, post-strata estimation should be handled in the sample design in advance. If not only post-strata estimation but also provincial classifications

in planning the sample design are considered in advance, good direct estimates for the post-strata can be expected. Consequently, good synthetic estimates for the provinces can be expected. The provincial classifications can increase homogeneity by putting similar provinces in classes together and using only sample data of the classified provinces to make the direct post-strata estimates to construct the synthetic estimates of those provinces.

The composite and EB estimators usually perform well when  $S_a/S_r$  is 10% or larger for a province because the direct components of the estimators (2) and (8) are relatively stable and receive a larger weight, especially for the EB estimator. Tehran, Khorasan, Khuzestan and Esfahan are of this type, while Bushehr, Ilam, Kohgiluyeh & Boyerahmad and Semnan are not.

#### 4. Final Remarks

In developing countries like Iran, administrative records are not often available both at small and large area levels. Surveys may yield satisfactory estimates for large areas but not for small areas. Periodic censuses do not meet all demands for effective policies and planning. These limitations lead to deficiencies in official statistics. Therefore, the statistical planning activities of SCI are directed towards compensating these deficiencies by using new methods and strategies. The present study proposes a cost-effective strategy to overcome some of the limitations.

In this study, the findings support the idea that a nationwide sample design can be used instead of the separate provincial sample designs by applying suitable SAE methods. The nationwide sample design consists of nearly 13,000 persons, whereas the twenty-one separate provincial sample designs totally consist of almost 100,000 persons.

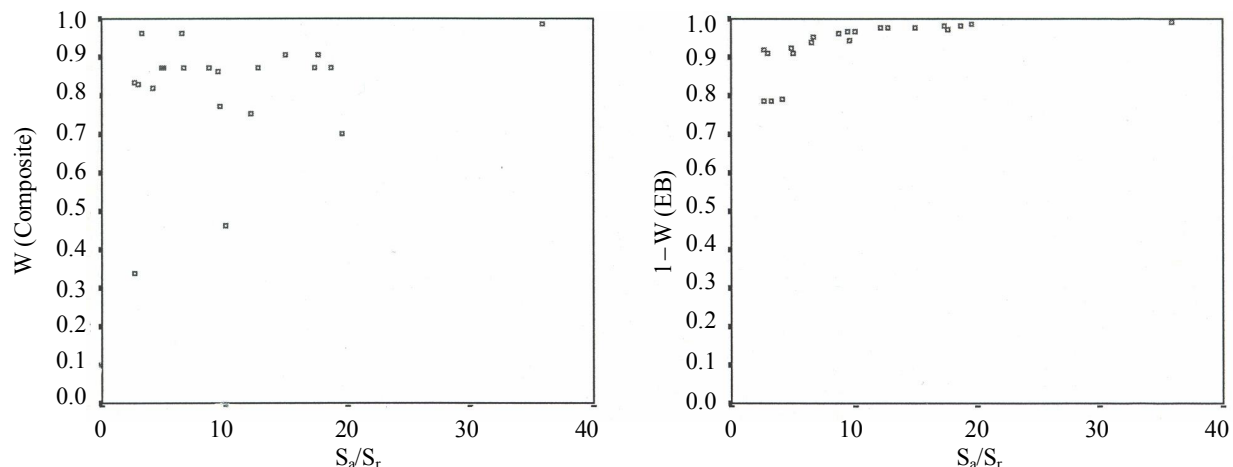


Figure 4. The weights of the direct components of composite and EB estimates against  $S_a/S_r$ .

The provincial design is the method currently used to produce provincial estimates by SCI. Using the national sample design decreases costs more than 80 percent. In addition, note that:

1. Although some SAE methods do not rely on existing sample data from all small areas, the strategy for producing provincial estimates is more appropriate when the small areas of interest are pre-specified. Therefore, the nationwide sample can be allocated to all small areas of interest to produce direct design-based estimates. It is important to adjust the sample design to accommodate the SAE methods before data collection begins. As Singh, Gambino and Mantel (1994, page 3) note

“small area needs should be recognized at the early stages of planning for large scale surveys. The sample design should include special features that enable production of reliable small area data using design or model estimators”.

Therefore, SCI must re-plan sample designs to reflect small area needs.

2. The SAE estimators usually perform well as the sample size increases. To improve provincial estimates, the nationwide sample size can be enlarged to have larger sample sizes from each province. Also, the provinces can be classified into groups with similar characteristics, such as unemployment rates, socio-demographic variables, and so on. Separate sample size would then be determined for each group.
3. Appropriate supplementary variables, which are related to the variable of interest, play a central role in improving the estimators.
  - The synthetic estimator used only one variable (age) for partitioning but it may be possible to use another variable or a combination of variables for partitioning. The post-strata in the synthetic estimator should be formed by variables that reduce variation in each post-stratum. These variables can indirectly affect the composite estimator as well.
  - The EB model can be improved with better supplementary information. Therefore it is important to try different supplementary variables to find the best model. In this work only EPA was used as the independent variable in the model, but there may be other variables that produce better estimates.
4. The composite estimator performs relatively better than the synthetic and EB estimators. However, the

results are only meant to provide a first impression of the utility of SAE methods. More research is needed to develop a generic SAE methodology in Iran. Further, the SAE methods should be applied not only in estimating unemployment rates but also in estimating other parameters, and SAE methods should be compared with the estimates coming from separate sample designs.

## Acknowledgements

Research for this paper was partially supported by the Statistical Research Center of Iran. The authors are grateful for many useful and helpful comments from the referees and the Associate Editor. My heartfelt thanks should go to Jim Lepkowski for his friendly helps. The views expressed are the authors' and do not necessarily reflect those of SCI.

## References

- Chand, N., and Alexander, C.H. (1995). Indirect estimation of rates and rates for small areas with continuous measurement. In *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 549-554.
- Copas, J.B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika*, 59, 349-360.
- Cressie, N. (1989). Empirical bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 65-93.
- Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.
- Gonzalez, M.F., and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimation. *Journal of the American Statistical Association*, 73, 7-15.
- Gonzalez, M.F., and Waksberg, J. (1973). Estimation of the errors of synthetic estimates. Paper presented at the first meeting of the International Association of Survey Statistician, Vienna, Austria, 18-25 August.
- Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. In *Proceedings of the American Statistical Association, Social Statistics Section*, 328-331.
- Marker, D.A. (1995). *Small area estimation: A Bayesian perspective*. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.
- Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.
- Pfeffermann, D. (2002). Small area estimation-New developments and directions. *International Statistical Review*, 70, 125-143.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean square error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

- Purcell, N.J., and Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.
- Rao, J.N.K. (2003a). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rao, J.N.K. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 2, 145-169.
- Schaible, W.L. (1978). Choosing weight for composite estimators for small area statistics. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.
- Schaible, W.L. (1995). Ed. *Lecture Notes in Statistics: Indirect Estimators in U.S. Federal Programs*, New York: Springer.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.
- Thompsen, I., and Holmoy A.M.K. (1998). Combining data from surveys and administrative record system: The Norwegian experience. *International Statistical Review*, 66, 201-221.