Calibrating StatsSA's National Household Survey weights to a consistent series over time

1. *Why post-stratify?*

The purpose of survey weights is to inflate the sample to represent the entire population. These weights therefore play an important role in creating consistent aggregates over time. Statistics South Africa's (StatsSA) household and person weights are not simple design weights i.e. inverse inclusion probability weights. The weights presented in the StatsSA National Household surveys are the design weight post-stratified to external population totals. Since the data are cross sectional the intention of the post-stratification adjustment is to produce best estimates of the population given the information available at the time and temporal consistency is not considered.

The StatsSA weights presented in the data are problematic for analyses over time for two main reasons. First, the auxiliary data used as a benchmark in the post-stratification adjustment are unreliable and inconsistent over time and hence result in temporal inconsistencies even at the aggregate level. Second, since the adjustments were made at the person level until 2003, there is no hierarchical consistency between the person and household weighted series until 2003. Thus estimates at the household and person level may disagree. We therefore advocate the use of a new set of weights created using entropy estimation. These weights result in consistent demographic and geographic trends and can be used at both the person and household level.

2. *Method*

2.1 The constraints

The sample weights were adjusted so that for each year, the October Household Survey (OHS), the Labour Force Survey (LFS) and General Household Survey (GHS) populations conformed to the age-sex-race distribution of the population estimates as calculated by the ASSA 2003 model[1]. A separate constraint required the distribution by provinces to correspond with the ASSA population estimates[2]. Further constraints required that the total weights add up to the estimated total population in each year and that the weights be constant within households. This latter constraint is based on the assumption that the mismatch is due to the fact that the surveys disproportionately missed certain types of households, rather than disproportionately under-enumerated particular age groups within the households. Individuals whose age, sex or race was missing were all allocated to a residual category. We imposed the condition that the proportionate weight of these individuals (below 0.5% of the sample in each year, see table 1) should not change due to the reweighting.

[1] The ASSA model estimates are mid-year estimates. These were used directly for the OHS and GHS calculations. Since the LFS is administered biannually, the ASSA model estimates were adjusted to the month of the survey. Interpolation using exponential growth was used to adjust the mid-year estimates to February for the 2000_1-2002_1 surveys, to March for the 2003_1-2007_1 and to September for the 2000_2-2007_2 surveys.

[2] ProvOutput_051129.xls

Table 1: Percentage with age, sex or race missing

Survey	OHS	OHS	OHS	OHS	OHS	OHS	OHS	LFS_2	LFS_1	LFS_2	GHS	LFS_1	LFS_2	GHS
Year	1993	1994	1995	1996	1997	1998	1999	2000	2001	2001	2002	2002	2002	2003
% missing	0.00	0.00	0.00	0.00	0.00	0.13	0.35	0.30	0.33	0.27	0.23	0.22	0.25	0.14

Survey	LFS_1	LFS_2	GHS	LFS_1	LFS_2	GHS	LFS_1	LFS_2	GHS	LFS_1	LFS_2	GHS	LFS_1	LFS_2
Year	2003	2003	2004	2004	2004	2005	2005	2005	2006	2006	2006	2007	2007	2007
% missing	0.14	0.12	0.14	0.12	0.20	0.16	0.22	0.42	0.20	0.19	0.22	0.26	0.28	0.33

2.2 The technique

We calculated the post-stratified weights by the "cross-entropy" estimation procedure (Golan, Judge and Miller 1996, p.29). The idea is to minimise the cross-entropy measure

$$\sum_i w_i \ln\left(\frac{w_i}{q_i}\right)$$

where w_i is the set of weights to be chosen (one for each individual) and q_i is the set of ex-ante weights (rescaled to sum to one). We used the StatsSA weights presented in the data (as we had no access to the design weights). The minimisation is done subject to the set of constraints imposed on the problem, i.e.

$$\sum_i w_i = 1$$

$$\sum_i w_i d_{ij} = p_j$$

In this case p_j is a particular population proportion (e.g. the proportion of people in the Western Cape) and d_{ij} is a dummy variable indicating whether the i-th individual in the dataset is in the Western Cape or not.

Altogether there are 146 constraints in total[3]: 9 provincial proportions, 136 age-sex-race proportions plus the proportion "missing". Two of these constraints are redundant, since the province proportions add up to unity, as do the age-sex-race plus "missing" proportions. It is relatively straightforward to show that the cross-entropy solution is equivalent to the solution that would be obtained by rescaling the proportions iteratively until convergence is achieved (Wittenberg 2009b). In a sense the weights w_i are those as close to the original weights q_i as possible, while obeying all the constraints. The set of weights w_i obtained through the cross-entropy estimation were converted to "raising weights" by multiplying them by the population

[3] 145 in the case of the OHS 1993-1997 which have no missing age-sex-race cells

total in each year as given by the ASSA 2003 model population estimates. The program used to calculate the weights is available (Wittenberg 2009a).

3. *Assessing the New Weights*

Figure 1 and 2 present estimates of the population and the number of households for each year 1993 to 2007. Each figure presents an estimate using the original StatsSA weight and the new cross-entropy weights. It is clear in the figures that the cross-entropy weights produce a more consistent trend in the population and the number of households over time.

Figure 1: Population counts using the old and new weights

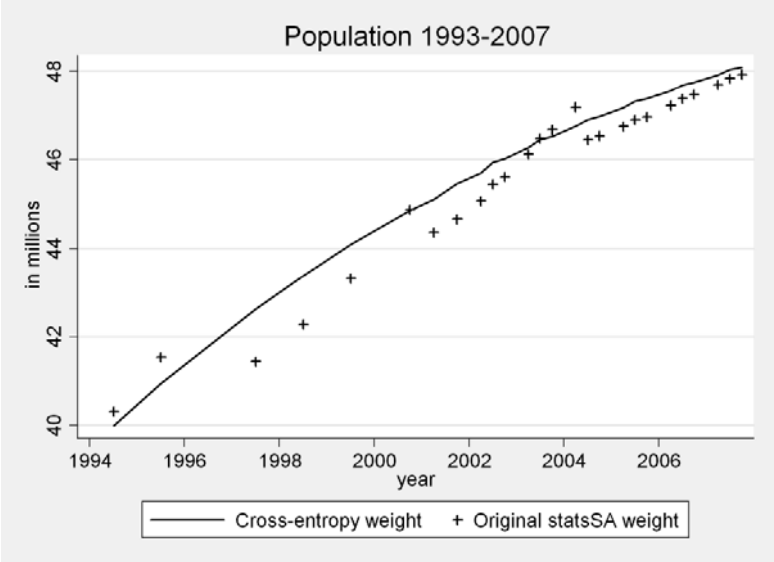
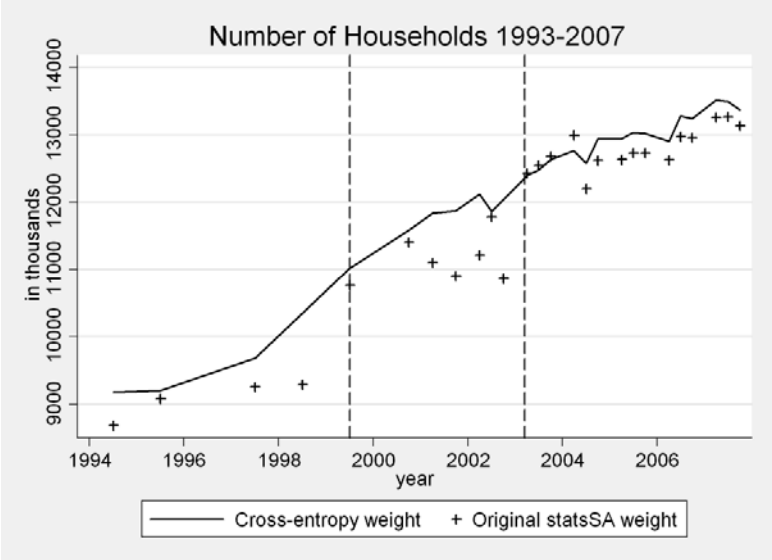
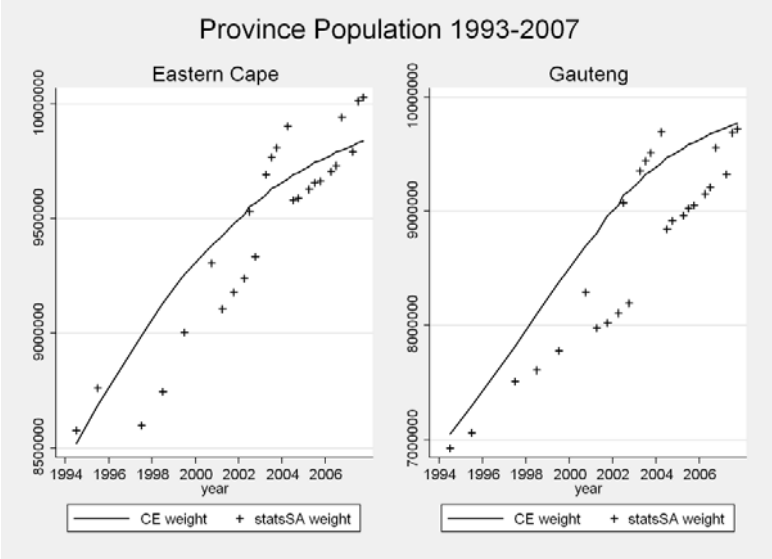


Figure 2: Household counts using the old and new weights



This is particularly the case for the trend in the number of households. The number of households follows a distinctively step-wise function with increases in 1999 and 2003 when the original household weights are used. This is not an accurate depiction of reality. The large increase in number of households in 1999 and 2003 coincide with the implementation of the 1996 and 2001 Census sampling frames which replaced the previously used 1991 and 1996 Census sampling frames respectively. The trend in the number of households is far more realistic when the cross-entropy weights are used.

Figure 3: Population counts in the Eastern Cape and Gauteng



If the attributes that a researcher is trying to measure are correlated with characteristics used in the post-stratification adjustment, the cross-entropy weights can improve the survey

representation. For instance, if income is distinct across province, using the cross-entropy weights which are post-stratified to a benchmark series where the trend in the proportion of the population in each province is realistic, (see figure 3 for the Eastern Cape and Gauteng) would produce a more consistent series of income over time.

4. Conclusion

OHS, LFS and GHS data are frequently stacked side-by-side to create time series data. These data are however, designed as cross sections with no emphasis on consistency in the series over time. As a result the series shows large fluctuations even at the aggregate level. In addition, until 2003, post-stratification was done at the person level which results in inconsistencies between the person and household files. The new set of publicly available consistent cross-entropy weights are benchmarked to aggregate numbers from the ASSA model and therefore present aggregates which are more consistent over time. In addition, these weights can be used at both the person and household level.

5. Accessing the Cross-entropy weights

The cross entropy weights are publicly available and can be accessed from the DataFirst website. Data files containing the cross-entropy weight and the unique household identifier are available for each survey used (e.g. OHS1999_cewgt.dta). These files are at the individual level, but since the weight is common across households, the weight can be used as either a household or person weight.

Cross entropy weights were not originally created for OHS 1996 because this dataset has numerous duplicates in the household file. These weights were created during work on the first version of the Post-Apartheid Labour Market Series (PALMS) and have now been included with Branson's weights.

Once these data errors have been corrected a set of weights will be created. Cross entropy weights were also not created post 2007 as the urban/rural classification was no longer available.

References:

Golan, Amos, George Judge, and Douglas Miller, Maximum Entropy Econometrics: Robust Estimation with Limited Data, Chichester: Wiley, 1996.

Horvitz, D.G. and D.J. Thompson, "A Generalization of Sampling Without Replacement From a Finite Universe," Journal of the American Statistical Association, 1952, 47 (260), 663—685.

Wittenberg, Martin, "An introduction to maximum entropy and minimum cross-entropy estimation using Stata," School of Economics and SALDRU, University of Cape Town 2009.

Wittenberg, Martin, "Sample Survey Calibration: An Information-theoretic perspective," School of Economics and SALDRU, University of Cape Town 2009.