DataFirst Technical Papers

Re-weighting South African National Household Survey Data to create a consistent series over time: A cross entropy estimation approach

by

Nicola Branson and Martin Wittenberg

Technical Paper Series
Number 15

About the Author(s) and Acknowledgments

Nicola Branson is a PhD student in the School of Economics at the University of Cape Town.
Martin Wittenberg is an Associate Professor in the School of Economics at the University of Cape Town, a Research Associate within SALDRU and Acting Director of DataFirst.

This paper is based on Nicola Branson's Master's thesis of the same title. We thank Matthew Welch and Lynn Woolfrey of DataFirst for access to and help with the Statistics South Africa data sets. We thank Cally Ardington and an anonymous referee for helpful comments. We acknowledge support from the Mellon Foundation through their grant to DataFirst.

This is a joint DataFirst and SALDRU paper as part of the Mellon Data Quality Project

Recommended citation

Branson, N., Wittenberg, M. (2011). Re-weighting South African National Household Survey Data to create a consistent series over time: A cross entropy estimation approach. A DataFirst Technical Paper Number 15. Cape Town: DataFirst, University of Cape Town

---

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Re-weighting South African National Household Survey Data to create a consistent series over time: A cross entropy estimation approach[1]

Nicola Branson and Martin Wittenberg*

DataFirst Technical Paper 15
University of Cape Town
February 2011

Abstract

In the absence of established longitudinal panel surveys in South African, national cross-sectional household survey data are frequently used to analyse change. When these data are stacked side-by-side, however, inconsistencies both in time trends and between household and person level data are found. This study uses a new set of weights calibrated to the ASSA 2003 model totals using a cross entropy estimation approach. This approach is favoured because the calculated weights are similar to the initial sample weights (and hence retain the survey design benefits) but match to a series of age-sex-race and province marginal totals that are consistent over time. The weights are publicly available for a fourteen year period between 1994 and 2007.

JEL classification: J11, C81, C83
Keywords: Post-stratification, Cross entropy, South Africa, October Household Survey, General Household Survey, Labour Force Survey

*School of Economics and SALDRU University of Cape Town.

[1] This paper is based on Nicola Branson's Master's thesis of the same title. We thank Matthew Welch and Lynn Woolfrey of DataFirst for access to and help with the Statistics South Africa data sets. We thank Cally Ardington and an anonymous referee for helpful comments. We acknowledge support from the Mellon Foundation through their grant to DataFirst.

1. Introduction

One main focus of post apartheid research in South Africa is change. Questions include the progress of South Africa in the economic, social and political arena. National datasets such as the October Household Surveys (OHS), Labour Force Surveys (LFS) and the General Household Surveys (GHS) provide a rich source of information on both economic and social variables in a cross sectional framework. These datasets are repeated annually or biannually and therefore have the potential to highlight changes over time. Yet to treat the cross sectional national data as a time series requires that, when stacked side by side, the data produce realistic trends. Since these data were not designed to be used as a time series, it is possible for changes in sample design, the interview process and shifts in the sampling frame to cause unrealistic shifts in aggregate numbers over a short period of time. This raises concerns about the validity of using these datasets as a time series to examine change.

One particular source of spurious shifts is a change in the way in which the survey weights are calibrated (see for instance Casale, Muller and Posel 2004, p.984). The purpose of survey weights is to inflate the sample to represent the population and therefore the weights play an important role in creating consistent aggregates over time. Statistics South Africa (StatsSA) household and person weights are not simple design weights i.e. inverse inclusion probability weights. StatsSA post-stratifies the design weight 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.

This paper highlights two concerns with the weights released by StatsSA. First, the auxiliary data used to benchmark the surveys are inconsistent as a series over time. This results in temporal inconsistencies even at the aggregate level. Second, until 2003 the post-stratification adjustment was made at the person level, resulting in person weights differing within the same household. Thus household weights were either left unadjusted or the weight of a representative person (for example the household head) in the household assigned as the household weight. Given that the household is the unit that is sampled it makes more sense for person weights to be common within a household. This would also lead to hierarchical consistency between the person and household weighted series.
We therefore advocate the use of a new set of weights created using entropy estimation for two reasons. First, these weights are calibrated to consistent demographic and geographic trends. While this has immediate importance for aggregates over time we also illustrate how shifts in the relative population shares across age-sex-race and province can impact on analyses. We demonstrate this by examining the trend in children's completed education over time. Second, the entropy approach makes it straightforward to incorporate constraints that do not include marginal totals. Thus the entropy person weights are constrained to be identical within household and can therefore be applied to both person and household level analyses.

The re-weighting does not however ameliorate all oddities in the data. Specifically, the re-weighting procedure does not deal with specific measurement changes in the data series

that are unrelated to the weights. This highlights an important characteristic of the entropy weights; they deal with representation errors only.

The remainder of this paper is organised as follows. Section 2 briefly introduces the theoretical basis for weights and post-stratification and highlights the South African interest in data quality issues. Section 3 describes StatsSA post-stratification procedure and how this has changed over time. Section 4 describes how the entropy weights were constructed. Section 5 assesses the impact of the new weights on aggregate trends, the age-sex-race and province distributions and trend in educational attainment between 1994 and 2007. Section 6 concludes.

2. Data Quality in South Africa

Awareness of South African data quality issues among researchers is fairly common. Researchers often present a caveat to their findings: results are subject to data quality[2]. Bhorat and Kanbur (2006, p. 2) cite "data quality and comparability" as one of three key aspects to research and debate in South Africa. They give the example of the 'jobless growth' debate[3], to highlight how much controversy statistics from incomplete and flawed datasets can generate.

Sample design problems and changes in the South African datasets are relatively well documented in the literature. Posel and Casale (2003) compare changes in the definition of a household and who is classified as resident, with particular attention to migrant members. Muller (2003) and Casale et al. (2004) look at the change in the framing of hurdle questions and their impact on sample selection bias. Wilson et al. (2004) note the improved ability of the Labour Force Surveys (LFS's) to capture employment and labour force participation compared to that of the October Household Surveys (OHS's). Wittenberg and Collinson (2007) find the national household surveys have a far higher proportion of single person households than the Agincourt demographic surveillance data would suggest is plausible. Keswell and Poswell (2004) and Ardington et al. (2006) discuss the effect incomes incorrectly captured as zero can have on an analysis.

It is, however, common for researchers to use multiple cross sectional data sets to elicit trends in variables over time with minimal concern given to whether these data are comparable even on basic demographic and geographic variables at the aggregate level. To date, the South African literature that assesses the sensitivity of economic trends to comparability of cross sectional surveys is limited. We only found two papers which correct for comparability by adjusting the sample weights. Simkins (2003) generates a set of weights for the 1995 and 2000 Income Expenditure Survey (IES) data resulting in comparable inequality estimates. A raking procedure is used to adjust the 1995 and 2000 province and population totals to the accepted 1996 census proportions. Ozler (2007) uses a procedure similar to Simkins (2003) to adjust the 2000 IES to the 2001 Census. These adjusted weights

[2] Bhorat & Kanbur (2006), Branson & Wittenberg (2007), Burger &Yu (2006) Casale, Muller & Posel (2004), Cronje & Budlender (2004), Wittenberg & Collinson (2007), Kingdon & Knight (2007) and others.

[3] The Standardised Employment and Earnings (SEE) dataset was used to show declining employment since the 1990s. This dataset does not however capture all economic activity and a reverse in the trend was found in the LFS.

are found to have a significant effect on mean expenditure, but have a limited effect on measured poverty changes. They conclude that while the direction of their findings is not significantly affected by which sample weights are used, the magnitudes of the results do change. These weights apply to the IES data and are not available in the public domain.

A new set of weights has been created using entropy estimation that is available in the public domain (Branson, 2010). These weights apply to 14 years of consecutive data from the OHS, LFS and GHS datasets and result in an aggregate series between 1994 and 2007 which is consistent with demographic and geographic totals from the ASSA 2003 estimates. We will illustrate that it is important for researchers examining trends over time to assess whether their findings are robust to the use of these consistent weights to rule out the possibility that their results are impacted by fluctuations in the StatsSA survey weights.

3. Weighting in the National household surveys

The OHS, the LFS and the GHS data provide a unique opportunity to examine a range of development and poverty indicators in the first fourteen years post apartheid. The surveys collected information on a variety of topics including unemployment, work details, education and access to resources and infrastructure. If accurately weighted, the surveys can reflect the national population and hence have the potential to produce aggregate data to assess progress and formulate projections. Accurately weighted data are important as a policy tool and to complement the national accounts.

Censuses are extremely costly and a well designed survey is equally useful and less expensive. The sample weights play a key role. The principle behind sample weights is to inflate the sample to reflect the population. The individual/household design weight is the inverse probability that the person/ household is included in the sample and is therefore defined by the sample design.

A common survey design is two-stage sampling. The sampling frame provides a complete list of households in the population grouped into areas or clusters. A two-stage design initially randomly selects clusters from the sampling frame and selects households within these clusters as a second step. Frequently, to aid the representation of certain subgroups within the population, a stratification step is included. The sampling frame is stratified by the defining characteristic of the subgroup (e.g. geographical region, population group) and clusters are drawn from these sub-samples. This guarantees that enough observations for each subgroup are selected in the total sample. The probability of inclusion in each stage is calculated and the household weight constructed by multiplying these inclusion probabilities together.

Divergences between the sample and the population come from differences in selection probabilities due to both planned (as per the survey design) and unplanned factors. Unplanned differences arise due to measurement errors and sampling errors, for example an out-of-date sampling frame or non-response. To obtain accurate population estimates the sample needs to be weighted with weights that reflect actual inclusion probabilities, in other words, accounts for both planned and unplanned differences. The design weights

only account for the survey design and do not account for unplanned differences in inclusion probability. Adjustment of the survey design weights to account for these unplanned differences can be done using post-stratification.

Post-stratification incorporates any data adjustment procedure which organises data into homogenous groups post-data collection and is usually undertaken to benchmark the data to external totals (Smith, 1991). The main function of post-stratification is therefore to adjust the design weights to account for sampling errors (out-of-date sampling frame and non-response) with the aim of improving the representation of the sample.

Post-stratified estimation is, however, only as good as the auxiliary data used. This exposes adjustments to two potential sources of error. First, population totals at the post-strata level may be unavailable or unreliable. Post-stratification adjustments are based on adjusting the sample estimates to what is assumed to be the 'true population'. If the 'population' data available are unreliable or out of date, the adjustment is made to incorrect frequencies and can introduce bias. Thus if auxiliary data are of poor quality or form an inconsistent series over time, the value of post-stratification may offset the gains from increased precision (Smith, 1991).

Second, auxiliary information is generally only available at the person level. Yet in most household surveys the household is the unit that is sampled and the individuals are enumerated within it. Consequently the probability of including an individual conditional on the household being selected is one. This suggests that the weight attached to every individual within a household should be equal. This constraint should therefore be included in the post-stratification adjustment. Not all post-stratification methods can include constraints which are not related to marginal totals. In particular, not all methods can constrain person weights to be common within a household. Since auxiliary data at the household level are hardly ever available, household weights are often derived from the person weights inappropriately or left uncalibrated (Neethling & Galpin, 2006). This can result in different inference when analyses are done using household versus person data.

The survey weights supplied by Statistics South Africa (StatsSA) in the national household surveys are *adjusted* design weights. StatsSA benchmark their data to external population estimates/projections in an attempt to address unplanned differences in inclusion probabilities due to non-response and other sampling problems. Since the OHS's, LFS's and GHS's are cross sectional datasets, the purpose of their benchmarking is to produce representative data for the particular year in question. The focus is not on producing a consistent series over time.

Table 1 provides details on which variables were used as benchmarks, the source of the benchmark and the Census on which these benchmarks were based as well as the calibration method used in each year by StatsSA. Four areas can be identified as possible sources of inconsistencies over time. First the base population from which the population is constructed in each survey year differs. Until 2002 the 1996 Census was used[4], while post 2002 this was updated to the 2001 Census. Second, the method used to extrapolate/project

[4] The weights for surveys pre 1996 calibrated to the 1996 Census were released after the completion of the Census.

the census population to the month of the census differs. The OHS data were benchmarked to the "1996 Census, adjusted for growth[5]" to the year of the OHS. The LFS's and GHS's (with the exception of 2002) use the mid-year population estimates adjusted to the month of the survey. The mid-year population estimates are produced by StatsSA's demography division. They are projected from the base population under assumptions about fertility and mortality. Dorrington & Kramer (2007) call into question the 'correctness' of the assumptions used in the mid-year population projections. They find that the mid-year estimates are internally inconsistent across years and not in line with other model projections[6]. In addition, they note inconsistencies between the 1996 and 2001 Censuses.

The third potential source of inconsistency comes from the choice of marginal totals over time. The LFS's and GHS's (with the exception of 2002) use demographic variables in the calibration process while the OHS's use both geographic and demographic variables. Lastly, the post-stratification method used changed over time. From 2003 CALMAR 2 has been used. This SAS macro also allows person weights to be common within households and therefore has the benefits detailed below. However, prior to 2003, CALMAR (not CALMAR 2) and, before that, relative scaling were used for post-stratification. These approaches made adjustments at the person level without consideration of household factors.

<Table 1 >

Thus while the current post stratification method used by StatsSA, CALMAR 2, has many advantages which will be carried forward in the calibration of future datasets, the methods used prior to 2003 result in inconsistencies which should be addressed when constructing a time series of data. In addition, inconsistency in the mid-year benchmarks both in isolated years and as a time series, affect the aggregates in all years.

4. The entropy weights

The StatsSA sample weights are internally inconsistent. First, the external data used as benchmarks do not present realistic trends over time. Second, prior to 2003, consistency between the person and household level data is not found since adjustments were made at the person level without constraining person weights to be constant within households. On the other hand, the StatsSA weights contain valuable information about the sample design. Thus when constructing a new set of weights, we wish to retain this information. The aim is therefore to create a new set of weights which inflate the sample to a consistent series of aggregate external data, which are hierarchically consistent between person and household level files in all years but which are otherwise as similar to the original StatsSA weight as possible. The cross-entropy estimation technique (Golan, Judge and Miller 1996, p.29) is consistent with the re-weighting estimation problem described above.

[5] No further information is given.

[6] Dorrington and Kramer (unpublished) replicate the StatsSA projection model and compare the mid-year estimates they would have got for 2001 with the Census 2001. They find, among other things, an over-representation of men and women in the mid-year estimates between age 15-35, with a 10% over-representation of males between the ages of 20 and 29. This is accompanied by a deficit of people over 60.

Wittenberg's (2010) maxentropy command in Stata was used to construct a set of internally consistent weights for the OHS, LFS and GHS cross sectional datasets. This section briefly outlines how these weights were constructed. Refer to Wittenberg (2010) for further details on the method.

Define the cross-entropy measure as

$$\sum_{i=1}^{n} p_i \, ln \frac{p_i}{q_i}$$

where p_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). The aim is to minimize the cross-entropy measure through the choice of a set of $p_i^{'}s$. StatsSA person weights contain a large amount of information about the sample design and demography of the population. These were used as the starting point for the estimation, as the set of ex-ante weights. While it would have been ideal to use the design weights, these are not publicly available. The minimisation is done subject to the set of constraints imposed on the problem, i.e.

$$\sum_{i=1}^{n} p_i = 1$$

$$y_i = \sum_{i=1}^{n} x_{ij} p_i$$

In this case y_i is a particular population proportion (e.g. the proportion of people in the Western Cape) and x_{ij} is a dummy variable indicating whether the i-th individual in the data set is in state j. The constraints information came from the ASSA 2003 model. Altogether there were 146 constraints[7]: nine 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. The set of weights p_i obtained through the cross-entropy estimation were converted to "raising weights" by multiplying them by the population total in each year as given by the ASSA 2003 model population estimates. The weights (Branson, 2010) and the program used to calculate the weights (Wittenberg 2010) are available.

Wittenberg (2009) shows that the cross-entropy solution is equivalent to the solution that would be obtained by rescaling the proportions iteratively until convergence is achieved. The CE weights therefore present a new set of weights which are in a sense as close to the original person StatsSA weights, but which at the same time satisfy the moment constraints from the ASSA 2003 aggregate data and are common within the household.

[7] 145 in the case of the OHS 1994, 1995 and 1997 which have no missing age-sex-race cells. Note, no weights were constructed for the OHS 1996 data due to data errors in the original file. There is no unique household identifier and there are individuals without households.

5. Assessing the Entropy weights

5.1 Consistency over time

Figure 1 present estimates of the population for each available OHS, LFS and GHS survey between 1994 and 2007. Estimates using both the original StatsSA person weight and the new cross-entropy weights are presented. When placed side by side and weighted by the original person weights the surveys do not present a consistent series. The series can be divided into three parts each section with a differing slope. 1995-2000, 2001 to 2003 and 2004 to 2007. The cross entropy weights produce a smooth trend in the population over time.

< Figure 1 >

The distinction between the cross entropy weights and the original person weights is even clearer when the population is assessed at provincial level. Figure 2 presents population estimates for two large provinces, the Eastern Cape and Gauteng. While the cross entropy weights form a smooth series, the original survey totals are not consistent when placed back to back. For example, the Gauteng population increases by over one million people between the LFS 2002_2 and LFS 2003_1 data.

< Figure 2 >

Figure 3 presents the trend in the number of households weighted using the household weights. It is clear that the household data was not benchmarked to an external series prior to 2003. The number of households follows a distinctively step-wise function until 2003 with increases in 1999 and 2003. 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. Post 2003, CALMAR 2 was used and hence the household weights were calibrated and the trend is more realistic. The cross entropy weights present a relatively smooth increase in the number of households over time, a function of restricting person weights to be common within a household.

< Figure 3 >

Although the ASSA 2003 totals do not present a 'gold standard', the importance of benchmarking to a consistent series should not be understated, especially given the frequency with which comparisons are made between cross sectional surveys in South Africa. While the new weights cannot be said to produce better population estimates within a specific year, analyses investigating changes over time will benefit from using these weights. The new weights will provide researchers with the confidence that the surveys are representing the same population over time and therefore that shifts observed are not a result of spurious shifts in the population.

5.2 A relative shift in the distribution

Figures 1-3 illustrate the effect of the new weights on aggregate numbers. However, many research questions are concerned with relative changes. For example, has there been a change in the proportion of the population living in poverty? Is the average educational attainment of children increasing, decreasing or staying the same? Such analyses will be affected by the new weights if the new weights change the relative importance of a sub group of people within the population of interest. For example, if the new weights increase the representation of Gauteng, a largely urban province, and decrease the representation of the Eastern Cape, a province with a higher rural contingent, in the relative distribution of the population across the provinces, this will have an impact on estimates of the proportion of people living in poverty. On the other hand, if the new weights only increase/decrease the population but have no effect on the distribution across age-sex-race and province cells, then analyses of this type will not be affected.

Table 2a-2c illustrate the shift in the distribution across the age-sex-race and province cells when the new versus the old weights are used. The numbers presented are percentages and reflect the relative under representation of that cell relative to the ASSA 2003 model when the original weights are used (negative number represent a relative over representation). For example, 2.55 for the OHS 1994 0-9 age category indicates that the share of children age 0-9 in the population is 2.55 percentage points higher in the ASSA model than in the population produced by the original StatsSA person weights.

<Table 2a-2c >

Observing the last three columns of Table 2a, it is clear that the surveys over represent older people when compared to the ASSA model in all years. In most years this is offset by an under representation of 20-59 year olds, with the difference increasing over time. In addition, the surveys tend to have an under representation of 0-9 year old children and an over representation of 10-19 year old children when compared with the ASSA model distribution.

Table 2b presents the population group distributions. While the early OHS surveys under represent Africans and over represent Whites relative to the ASSA model, most of the other surveys have a larger share of Africans than other population groups when compared to the ASSA model.

Finally Table 2c presents the province distributions. Before 2003, there does not appear to be any systematic difference in the relative representation of the survey provinces relative to the ASSA model. For example, the OHS 1994, 1995 and LFS 2001 over represent Gauteng, but the OHS 1999 and LFS 2002 (Feb) under represent Gauteng. However, the direction of the representation difference for the surveys from LFS (March) 2003 to LFS (March) 2004 is the same. EC, NC, FS are systematically under represented and KZN, NW and GT systematically over represented relative to the ASSA model. Similarly, the later surveys (from GHS 2004 onwards) can be grouped; the WC, GT and MP are underrepresented and the EC FS NW systematically over represented relative to the ASSA model.

Age, sex, population group and province are correlated with socioeconomic factors. Thus any analysis where there is a shift in the representation of a certain group could be affected by the new weights. For example, we saw that in some surveys the share of children age 10-19 decreases when the new entropy weights are used, i.e. they are over represented in the surveys relative to the ASSA model, while the share of children aged 0-9 increases. Hence an analysis of a chosen childhood characteristic, for example years of education, will be less strongly weighted to the characteristics of the older age group and more strongly weighted to the characteristics of the younger group when the new weights are used. Since educational attainment and age are correlated, this would imply a decrease in educational attainment when the new weights are used. This is investigated in section 5.3.

5.3 Assessing the effect of the new weights

This section illustrates two points. First, the weights can affect the substantive findings of analyses. We therefore advocate that researchers who use the cross sectional surveys to elicit changes over time, assess the sensitivity of their finding to these new consistent weights. The second point is that the weights are not a panacea for all errors in the cross sectional surveys. They deal with one important source of survey error, representation, but do not address measurement errors in the surveys unrelated to representation.

Table 2a illustrated that the new weights decrease the share of 10-19 year olds in most of the survey years, with the share of 0-9 year olds increased fairly substantially in the early surveys and to a lesser extent in the later surveys. We assess the effect of the new cross entropy weights on the trend in educational attainment in Figure 4. Point estimates including confidence bands are presented for the original person weights (left hand panel) and the cross entropy weights (right hand panel).

<Figure 4 >

While both graphs illustrate that educational attainment has been increasing over the past 14 years, the cross entropy weights present a much more conclusive picture. This is particularly clear if we break the series into estimates from the OHS's, GHS's, LFS's (Feb/March) and LFS's (Sept).

<Figure 5 >

Figure 5 presents the year-on-year change in mean educational attainment within survey type. The dropline represents estimates using the original person weights and the dashed line represents estimates using the new entropy weights. Points above the zero line show an increase in educational attainment while points below the line show a decrease in educational attainment. If educational attainment has been increasing consistently over the period we would expect the points to be consistently above the zero line. It is clear from Figure 5, that this is more closely achieved with the entropy weights than with the original person weights. The series weighted by the original person weights shows increases followed by decreases year-on-year, while when the series is weighted by the entropy weights, it either tracks the zero line fairly closely or shows an increase. Observing the

confidence bands of the estimates in Figure 4, within survey type, this consistency is reiterated by the overlap in most years of these bands when the new weights are used. This is not the case when the original person weights are used.

<Table 3>

Table 3 illustrates the second point. The percentage of children age 6-18 that have no education, grade 1, 2 or 3 in addition to the percentage with grades 1 through 3 is presented. As noted by Ardington (2008), the percentage with no education is much higher in the OHS 1997 and 1998 surveys than all other surveys, with a deficit in grades 1 and 2. It is possible that this is a function of a difference in the relative representation of young versus older children in these surveys. The right hand panel of Table 3 illustrates that this, however, is not the case. The high frequency of children with no education and the deficit in grade one and two is evident when the cross entropy weights are used. This error is not related to how the weights are constructed.

Ardington (2008) surmises that this inconsistency is a result of the different structure of the highest educational attainment question, in other words a result of measurement error. In all the surveys besides the OHS 1997 and 1998, educational categories where included on the questionnaire and interviewers had to tick off the corresponding box. In the 1997 and 1998 OHSs respondents were asked the open ended question “What is the highest school class/standard that (the person) completed?” with the additional interviewer instruction “If no schooling, or currently in Sub A/Grd 1 write none.” It is possible that respondents and/or fieldworkers did not consider Grades one and two as a completed standard or grade since these grades were previously referred to as Sub A and B or Class one and two (Ardington, 2008). The point we illustrate from the table is that the cross entropy weights does not ameliorate this type of error.

5.4 Discussion

We conclude that the cross entropy weights present an appropriate alternative to the original StatsSA person and household weights with noticeable advantages over the originals. First, the weights are calibrated in a consistent manner to a consistent benchmark series in each year. They therefore produce time trends in demographic, geographic and other variables which are more realistic. At the same time, the cross entropy weights are similar to the original person weights and therefore preserve the information about the original sample design contained in them. Second, the household and person entropy weights are internally consistent and therefore enable analyses at the household level using weights calibrated to external totals. Finally, if the analysis looks at the proportion of the population in a particular state over time and this subpopulation is differentially represented by the demographic or geographic variables used as restrictions in the re-weighting procedure when compared to their representation under the original weights, then the new weights will affect these analyses as well.

6. 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 internal 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 with the household weight either left uncalibrated or the weight of a representative household member assigned to the household. Thus trends in household variables are inconsistent over time. In this paper a comparison is made between trends calculated using the original StatsSA weights and a new set of publicly available consistent entropy weight that are benchmarked to aggregate numbers from the ASSA 2003 model.

The cross entropy weights are found to be appropriate as an alternative to the StatsSA person and household weights and have added advantages. The main advantage of the cross entropy weights is that they create consistent aggregates over time. For many analyses, and to limit confusion, it is important that the demographic and geographic variables in the national household surveys produce realistic aggregate trends and are in line with other aggregates such as those found in the ASSA model. When comparing different years of the OHS, LFS and GHS as a time series, results will be more realistic if the benchmarks are consistent over time and if the post-stratification method is consistent in each year. In other words, working with data calibrated in a similar manner to a smooth series of benchmarks reduces biases in trends due to inconsistencies in calibration totals and post-stratification methodologies. The entropy weights therefore take care of one potential source of error, faulty weights. Thus the researcher can be assured that shifts observed over time are not a result of post-stratification inconsistencies.

In addition, the entropy person weights are common within a household. While this reduces complexity i.e. the researcher does not have to use different weights for household and person level analyses, it also makes intuitive sense. Mismatches between the sample and the population are a result of non response at the household level, not the individual level. Hence it makes sense for all weights within a household to be the same. Up until 2003 the StatsSA household weights were not calibrated to external totals and as a result trends in household level variables are erratic over time.

We showed that the new weights can have an effect on the substantive findings of an analysis. We showed that the trend in educational attainment is far more systematic and realistic when the cross entropy weights are used.

Finally, we show that the entropy weights do not deal with specific measurement errors. The OHS 1997 and 1998 question on completed education resulted in a higher percentage of respondents and\or interviewers classifying the respondent as having no completed education. This cannot be corrected via weighting.

References

ARDINGTON, C. (2008). *Parental death and schooling outcomes in South Africa.* (Doctoral dissertation). University of Cape Town.

----------, LEIBBRANDT, M., LAM, D. and WELCH M. (2006). "The sensitivity of estimates of post-apartheid changes in South African poverty and inequality to key data imputations", *Economic Modelling*, 23: 822-835.

ASSA. (2003). AIDS Demographic Model 2003. Actuarial Society of South Africa. Main demographic model and ProvOutput, Version 051129.

BHORAT, H. and KANBUR, R. (2006). Introduction: Poverty and well-being in post-apartheid South Africa. In H. Bhorat, & R. Kanbur, *Poverty and Policy in Post-Apartheid South Africa* (p. 512). South Africa: HSRC Press.

BRANSON, N. (2010). Cross entropy weights OHS 1994-LFS 2007 September. http://www.datafirst.uct.ac.za/catalogue3/index.php/catalog

BRANSON, N. and WITTENBERG, M. (2007). "The Measurement of Employment Status in South Africa using Cohort Analysis, 1994-2004", *South African Journal of Economics*, 75(2): 313-326.

BURGER, R. and YU, D. (2006). "Wage trends in post-apartheid South Africa: Constructing an earnings series from household survey data", *Stellenbosch Economics Working Papers: 10/06*.

CASALE, D., MULLER, C. and POSEL, D. (2004). "'Two million net new jobs': A reconsideration of the rise in employment in South Africa, 1995-2003", *South African Journal of Economics*, 72(5): 978-1002.

CRONJE, M. and BUDLENDER, D. (2004). "Comparing Census 1996 and Census 2001", *South African Journal of Demography*, 9(1): 67-89.

DORRINGTON, R. and KRAMER, S. (2007). The 2004 mid-year estimates: Method, Reliability and Implication. Unpublished .

GOLAN, A., JUDGE, G. and MILLER, D. (1996). *Maximum Entropy Economics, Robust Estimation with Limited Data*. West Sussex, England: John Wiley and Sons Ltd.

KESWELL, M. and POSWELL, L. (2004). "Returns to education in South Africa: A retrospective sensitivity analysis of the available evidence", *The South African Journal of Economics*, 72(4): 834-860.

KINGDON, G. and KNIGHT, J. (2007). "Unemployment in South Africa, 1995-2003: Causes, Problems and Policies", *Journal of African Economies*, 16(5): 813-848.

OZLER, B. (2007). “Not Separate, Not Equal: Poverty and Inequality in Post-Apartheid South Africa”, *Economic Development and Cultural Change*, 55(3): 487-529.

MULLER, C. (2003). “Measuring South Africa's Informal Sector: An Analysis of National Household Surveys”, Development Policy Research Unit, Working Paper 03/71.

NEETHLING, A. and GALPIN, J. (2006). “Weighting of Household Survey Data: A Comparison of Various Calibration, Integrated and Cosmetic Estimators”, *South African Statistics Journal*, 40(2): 123-150.

POSEL, D. and CASALE, D. (2003). “What has been happening to internal labour migration in South Africa, 1993-1999”, *The South African Journal of Economics*, 71:3, 455-479.

SIMKINS, C. (2003). “A Critical Assessment of the 1995 and 2000 Income and Expenditure Surveys as a Source of Information on Incomes”, University of the Witwatersrand , Unpublished.

SMITH, T. (1991). “Post-stratification”, *The Statistician*, 40(3): 315-323.

STATISTICS SOUTH AFRICA metadata for various surveys

WILSON, R., WOOLARD, I., & LEE, D. (2004). “Developing a national skills forecasting tool for South Africa”. South Africa: Human Sciences Research Council.

WITTENBERG, M., & COLLINSON, M. (2007). “Household transitions in rural South Africa, 1996-2003”, *Scandinavian Journal of Public Health*, 35(3): 130-137.

WITTENBERG, M. (2009). “Sample Survey Calibration: An Information-theoretic perspective”. *School of Economics and SALDRU, University of Cape Town*.

---------- (2010). “An introduction to maximum entropy and minimum cross-entropy estimation using Stata”, *The Stata Journal*, 10(3): 315-330.

Tables and Figures:

Table 1: StatsSA Sampling and Post-stratification details –Sample frame, marginal totals, auxiliary data and calibration method

Survey	Census used as base for sample frame	Marginal totals	Auxiliary data source	Base population from which the mid year estimates were constructed	Calibration method
OHS 1994	No information	No information	No information	No information	No information
OHS 1995	1991 Census	Province, gender, age groups, race.	1996 Census adjusted for growth	1996 Census	Relative scaling
OHS 1996	1996 Census post-enumeration survey	Province, gender, age groups, race.	1996 Census	1996 Census	Generalised raking with a linear distance function
OHS 1997	1996 Census	Province, urban/rural, gender, age group, race	1996 Census adjusted for growth	1996 Census	Relative scaling
OHS 1998	1996 Census	Province, urban/rural, gender, age group, race	1996 Census adjusted for growth	1996 Census	Relative scaling
OHS 1999	1996 Census	Province, gender, age groups, race	1996 Census adjusted for growth		Relative scaling
LFS 2000_1	1996 Census	Province, gender, age groups, race	2000 midyear estimates	1996 Census	Relative scaling
LFS 2000_2-LFS 2002_2	1996 Census	Gender, race, age group	Midyear estimates	1996 Census	CALMAR
LFS 2003_1-LFS 2007_2	2001 Census	Gender, race, age group	Midyear estimates	2001 Census	CALMAR2
GHS 2002	2001 Census	Province, gender, age groups, race	Exponential extrapolation from the 1996 and 2001 censuses	1996 Census	CALMAR
GHS 2003-2007	2001 Census	"Population estimates"	Midyear estimates	2001 Census	CALMAR2

Notes to Table 1: There are multiple typographical errors in the metadata files. This is particularly true for the LFS's metadata files where it appears that the documentation has been updated each year from the previous year, often without all the necessary details correctly changed. While most years explicitly state that gender, race and age group were used as the post stratification cells, 2003_1 and 2004_1 through 2007_2 just say 'population estimates.' 2003_2 says 'age, gender and age group', which we assume is a typographical error and should be race, gender and age group. Midyear population estimates are published by StatsSA demography department annually. These estimates were adjusted to the month of the survey.

Table 2a: Comparing the population distribution using the original person versus entropy weights – the age distribution

		Age group											
		0-9	*10-19*	*20-29*	*30-39*	*40-49*	*50-59*	*60-69*	*70-79*	*80+*	*0-19*	*20-59*	*60+*
Survey	*Year*												
OHS	1994	2.55	-0.71	0.02	0.14	-0.84	-0.60	-0.36	0.06	-0.25	1.84	-1.28	-0.56
OHS	1995	3.23	-0.57	0.08	-0.18	-0.93	-0.83	-0.45	-0.08	-0.26	2.66	-1.86	-0.79
OHS	1997	0.91	-1.56	0.92	1.59	0.33	-0.33	-0.96	-0.57	-0.34	-0.64	2.51	-1.87
OHS	1998	1.16	-1.40	0.94	1.60	0.37	-0.82	-0.96	-0.42	-0.47	-0.23	2.09	-1.86
OHS	1999	1.74	-1.44	0.75	0.61	0.00	-0.47	-0.53	-0.28	-0.38	0.30	0.90	-1.19
LFS (Sept)	2000	-0.17	0.07	-0.99	1.20	0.50	-0.27	-0.31	0.17	-0.21	-0.10	0.44	-0.34
LFS (Feb)	2001	2.91	-1.55	0.35	0.03	-0.05	-0.41	-0.57	-0.23	-0.50	1.37	-0.07	-1.29
LFS (Sept)	2001	2.23	-1.70	0.70	0.16	-0.01	-0.29	-0.45	-0.20	-0.43	0.52	0.56	-1.08
LFS (Feb)	2002	1.74	-1.41	0.60	0.16	0.09	-0.20	-0.31	-0.24	-0.44	0.34	0.65	-0.99
GHS	2002	1.71	-1.05	0.12	0.03	-0.03	-0.01	-0.11	-0.17	-0.48	0.65	0.11	-0.76
LFS (Sept)	2002	2.63	-2.15	0.86	1.03	-0.18	-0.61	-0.68	-0.37	-0.53	0.49	1.10	-1.58
LFS (Mar)	2003	1.89	-1.27	0.25	0.06	-0.11	-0.02	-0.14	-0.19	-0.48	0.63	0.19	-0.81
GHS	2003	1.94	-1.36	0.32	0.01	-0.14	0.03	-0.12	-0.25	-0.42	0.58	0.22	-0.80
LFS (Sept)	2003	2.23	-1.84	0.61	0.31	-0.16	-0.13	-0.26	-0.30	-0.45	0.40	0.62	-1.02
LFS (Mar)	2004	2.09	-1.55	0.41	0.02	-0.23	0.08	-0.10	-0.28	-0.45	0.54	0.29	-0.83
GHS	2004	0.05	-0.36	-0.33	0.53	0.82	0.06	-0.55	0.07	-0.31	-0.30	1.09	-0.78
LFS (Sept)	2004	0.08	-0.40	-0.27	0.51	0.83	0.07	-0.57	0.07	-0.32	-0.32	1.14	-0.82
LFS (Mar)	2005	0.15	-0.54	-0.12	0.40	0.87	0.11	-0.61	0.01	-0.29	-0.38	1.27	-0.89
GHS	2005	0.20	-0.62	-0.02	0.32	0.90	0.15	-0.63	0.00	-0.30	-0.42	1.34	-0.93
LFS (Sept)	2005	0.24	-0.62	-0.02	0.29	0.91	0.15	-0.65	-0.03	-0.27	-0.38	1.34	-0.95
LFS (Mar)	2006	0.09	-0.63	0.05	0.22	1.02	0.17	-0.62	0.03	-0.33	-0.54	1.46	-0.92
GHS	2006	0.12	-0.65	0.08	0.15	1.05	0.19	-0.63	0.00	-0.32	-0.53	1.47	-0.94
LFS (Sept)	2006	0.08	-0.64	0.09	0.10	1.10	0.19	-0.65	0.04	-0.32	-0.55	1.48	-0.92
LFS (Mar)	2007	0.09	-0.62	0.17	0.01	1.13	0.23	-0.66	0.00	-0.34	-0.53	1.54	-1.01
GHS	2007	0.08	-0.61	0.21	-0.05	1.15	0.25	-0.67	-0.03	-0.32	-0.53	1.55	-1.02
LFS (Sept)	2007	0.09	-0.60	0.23	-0.09	1.17	0.25	-0.68	0.02	-0.38	-0.51	1.56	-1.05

Table 2b: Comparing the population distribution using the original person versus entropy weights – the population group distribution

		Population Group			
		African	*Coloured*	*Indian*	*White*
Survey	*Year*				
OHS	1994	2.08	0.12	0.00	-2.20
OHS	1995	2.89	-0.17	-0.13	-2.60
OHS	1997	-0.79	0.57	-0.02	0.24
OHS	1998	0.06	0.34	-0.28	-0.12
OHS	1999	0.11	-0.18	0.00	0.07
LFS (Sept)	2000	-0.39	0.09	0.08	0.22
LFS (Feb)	2001	-1.38	-0.15	0.06	1.47
LFS (Sept)	2001	-0.94	-0.26	-0.05	1.25
LFS (Feb)	2002	1.02	-0.01	-0.58	-0.43
GHS	2002	-0.58	0.04	0.04	0.50
LFS (Sept)	2002	0.69	-0.96	0.03	0.24
LFS (Mar)	2003	-0.61	-0.01	-0.01	0.62
GHS	2003	-0.69	-0.03	0.06	0.66
LFS (Sept)	2003	-0.52	-0.26	-0.15	0.93
LFS (Mar)	2004	-0.79	-0.01	0.06	0.73
GHS	2004	-0.27	0.05	0.05	0.16
LFS (Sept)	2004	-0.15	0.03	0.03	0.08
LFS (Mar)	2005	-0.26	0.06	0.05	0.16
GHS	2005	-0.27	0.07	0.04	0.17
LFS (Sept)	2005	-0.28	0.06	0.03	0.19
LFS (Mar)	2006	-0.30	0.08	0.03	0.19
GHS	2006	-0.30	0.07	0.06	0.17
LFS (Sept)	2006	-0.32	0.08	0.04	0.20
LFS (Mar)	2007	-0.30	0.08	0.03	0.19
GHS	2007	-0.33	0.10	0.04	0.19
LFS (Sept)	2007	-0.39	0.09	0.04	0.26

Table 2c: Comparing the population distribution using the original person versus entropy weights – the province distribution

		Province								
		WC	*EC*	*NC*	*FS*	*KZN*	*NW*	*GT*	*MP*	*LP*
Survey	*Year*									
OHS	1994	-0.04	0.64	-0.05	0.58	0.82	-0.40	-2.23	0.28	0.40
OHS	1995	0.06	0.13	0.11	0.20	0.56	0.04	-2.11	-0.05	1.05
OHS	1997	0.42	-0.24	0.08	0.31	-0.70	0.22	-0.20	-0.04	0.14
OHS	1998	0.20	-1.42	0.01	0.39	0.25	0.04	0.81	-0.05	-0.22
OHS	1999	0.38	-0.77	-0.13	-0.17	0.21	-0.22	1.05	0.08	-0.43
LFS (Sept)	2000	0.67	-0.87	-0.04	-0.26	0.18	-0.22	0.92	0.06	-0.43
LFS (Feb)	2001	0.02	0.17	0.25	0.19	0.09	0.65	-2.91	0.59	0.96
LFS (Sept)	2001	-0.05	-0.29	0.29	0.49	0.41	0.94	-3.88	0.91	1.18
LFS (Feb)	2002	0.67	-1.35	-0.08	-0.24	0.32	-0.13	1.82	0.01	-1.03
GHS	2002	0.08	0.06	0.08	0.09	-0.17	-0.20	-0.06	0.01	0.12
LFS (Sept)	2002	-1.46	-0.94	0.03	-0.04	1.87	0.58	-1.25	0.39	0.81
LFS (Mar)	2003	0.07	0.19	0.10	0.12	-0.26	-0.20	-0.26	0.04	0.18
GHS	2003	0.06	0.22	0.12	0.12	-0.28	-0.21	-0.23	0.00	0.20
LFS (Sept)	2003	0.06	0.23	0.11	0.13	-0.30	-0.20	-0.27	0.00	0.22
LFS (Mar)	2004	0.06	0.30	0.14	0.14	-0.29	-0.21	-0.39	-0.01	0.26
GHS	2004	0.57	-1.00	-0.06	-0.39	0.04	-0.24	1.15	0.11	-0.17
LFS (Sept)	2004	0.52	-0.96	-0.06	-0.38	0.04	-0.18	1.04	0.12	-0.13
LFS (Mar)	2005	0.52	-0.96	-0.06	-0.39	0.02	-0.23	1.06	0.12	-0.09
GHS	2005	0.50	-0.94	-0.06	-0.39	0.01	-0.23	1.02	0.13	-0.04
LFS (Sept)	2005	0.49	-0.92	-0.06	-0.39	0.01	-0.23	0.99	0.13	-0.02
LFS (Mar)	2006	0.47	-0.87	-0.06	-0.38	0.00	-0.23	0.91	0.12	0.04
GHS	2006	0.46	-0.85	-0.06	-0.38	-0.01	-0.23	0.86	0.13	0.08
LFS (Sept)	2006	0.45	-0.50	-0.45	-0.38	-0.42	0.79	0.17	-0.41	0.75
LFS (Mar)	2007	0.42	-0.79	-0.06	-0.38	-0.04	-0.22	0.76	0.13	0.18
GHS	2007	0.40	-0.43	-0.44	-0.38	-0.46	0.80	0.06	-0.40	0.85
LFS (Sept)	2007	0.39	-0.41	-0.44	-0.38	-0.47	0.80	0.03	-0.40	0.87

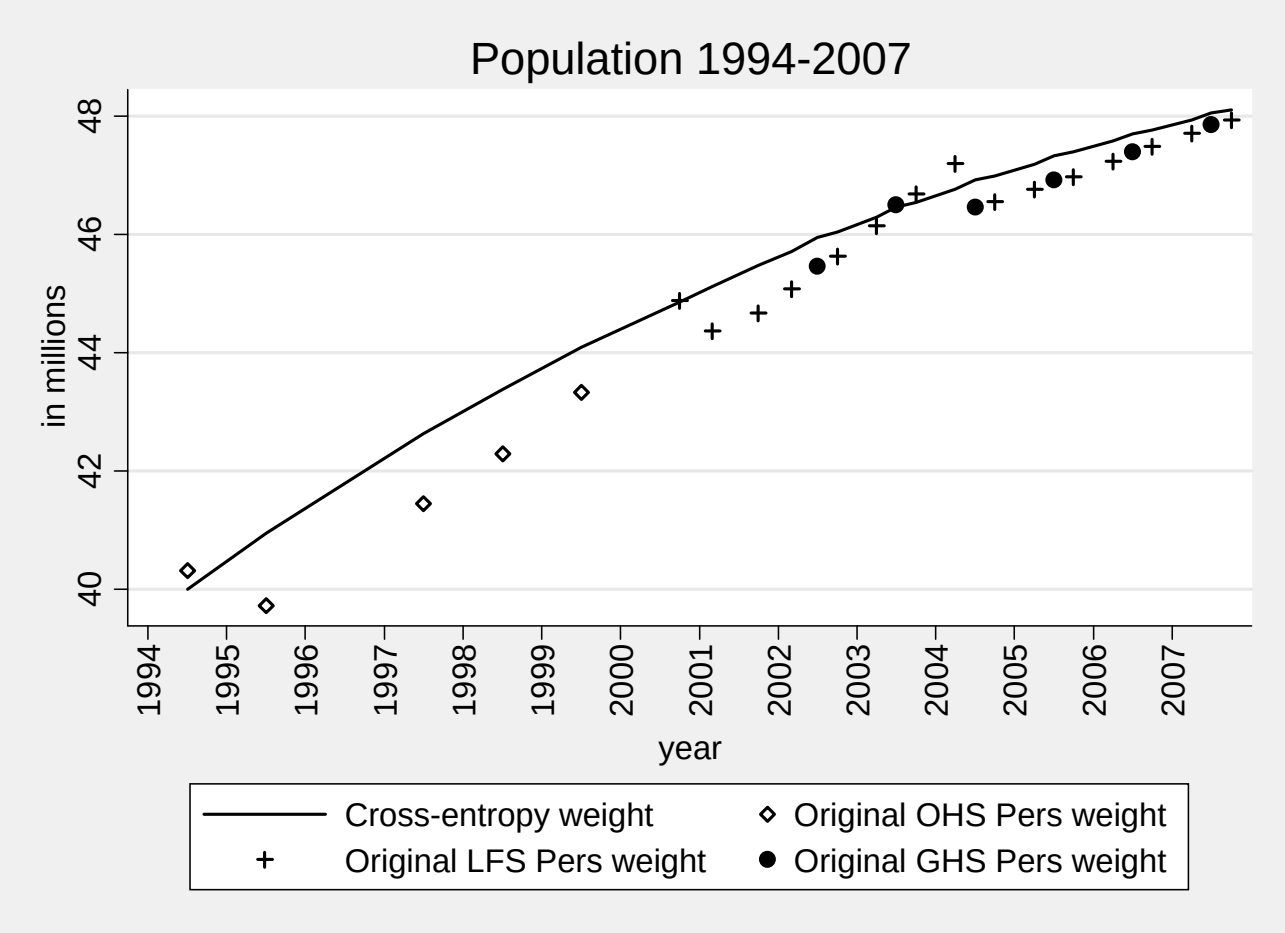
Notes to Table 2a-2c: Table 2a-2c illustrate the shift in the distribution across the age-sex-race and province cells when the new versus the old weights are used. The numbers presented are percentages and reflect the relative under representation of that cell relative to the ASSA 2003 model when the original weights are used (negative number represent a relative over representation). For example, 2.55 for the OHS 1994 0-9 age category indicates that the share of children age 0-9 in the population is 2.55 percentage points higher in the ASSA model than in the population produced by the original StatsSA person weights.

Table 3: Percentage of children 6-18 with no education or who have completed grades 1, 2 or 3

		Original Person Weight					Cross Entropy Weight				
Survey	*Year*	No education	Grade 1	Grade 2	Grade 3	Grade 1-3	No education	Grade 1	Grade 2	Grade 3	Grade 1-3
OHS	1994	10.07	31.41			31.41	10.23	31.58			31.58
OHS	1995	8.58	30.30			30.30	8.54	30.03			30.03
OHS	1997	20.64	4.06	5.91	10.80	20.77	20.77	4.08	5.89	10.81	20.78
OHS	1998	20.66	4.58	7.49	10.86	22.94	19.80	4.37	7.35	10.79	22.52
OHS	1999	15.05	10.33	9.62	10.19	30.14	14.42	9.81	9.21	9.98	29.00
LFS (Sept)	2000	13.99	9.61	9.60	9.74	28.95	13.43	9.24	9.37	9.59	28.20
LFS (Feb)	2001	12.73	8.46	8.29	9.38	26.14	13.13	8.70	8.49	9.40	26.59
LFS (Sept)	2001	14.94	8.17	8.61	9.39	26.17	15.55	8.42	8.66	9.26	26.34
LFS (Feb)	2002	13.79	8.42	7.89	8.67	24.98	14.45	8.66	7.98	8.62	25.26
GHS	2002	14.71	8.65	7.75	8.80	25.19	15.28	8.93	7.80	8.81	25.54
LFS (Sept)	2002	15.69	8.44	8.07	8.89	25.40	16.60	8.79	8.23	8.94	25.95

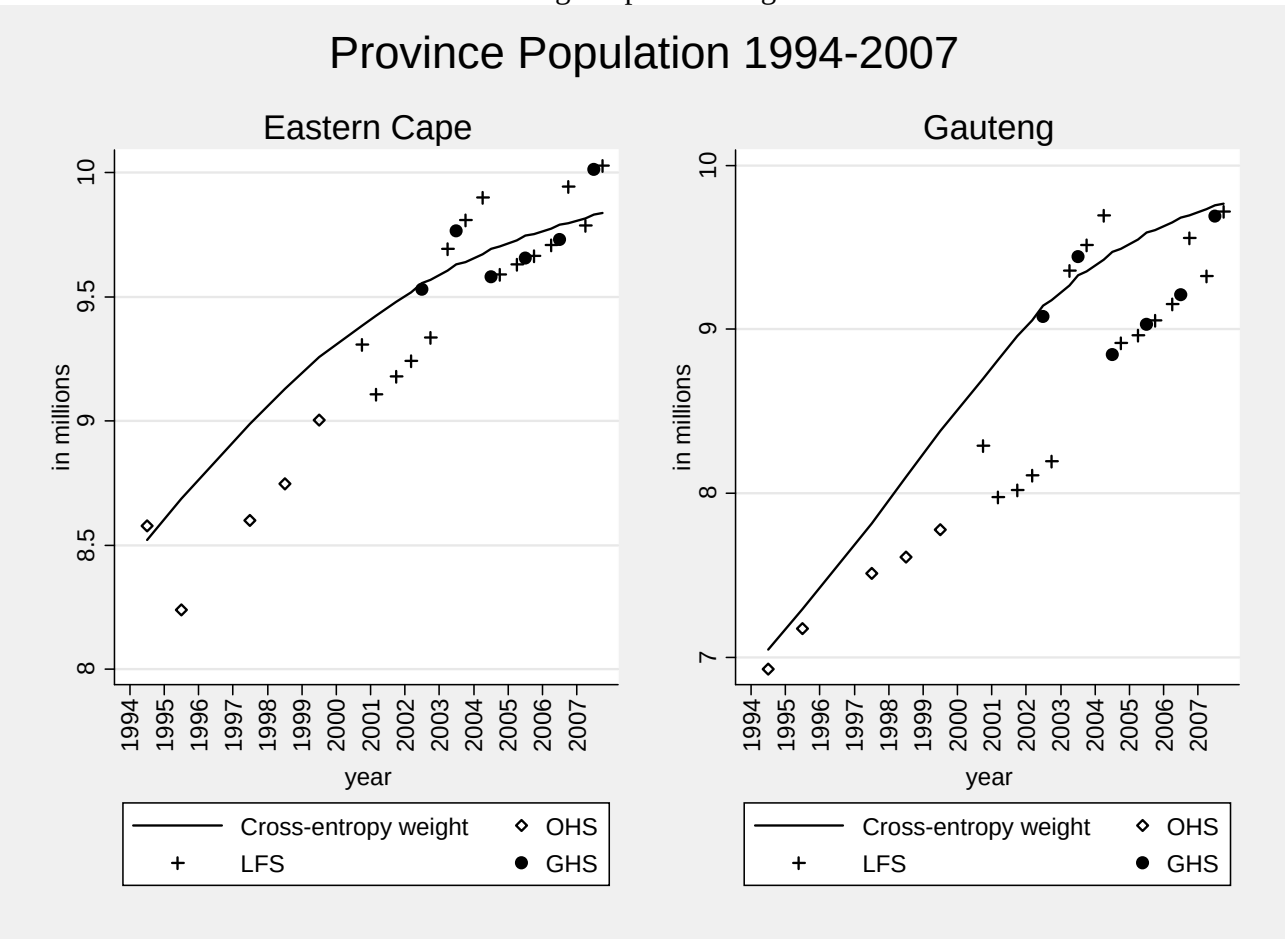
Notes to Table 3: Table 3 presents the percentage of children age 6-18 that have no education, grade 1, 2 or 3 in addition to the percentage with grades 1 through 3. The left panel of the table presents estimates weighted using the original person weights and the right panel of the table presents estimates weighted using the cross entropy weights.

Figure 1: Population counts using the entropy versus the original person weights



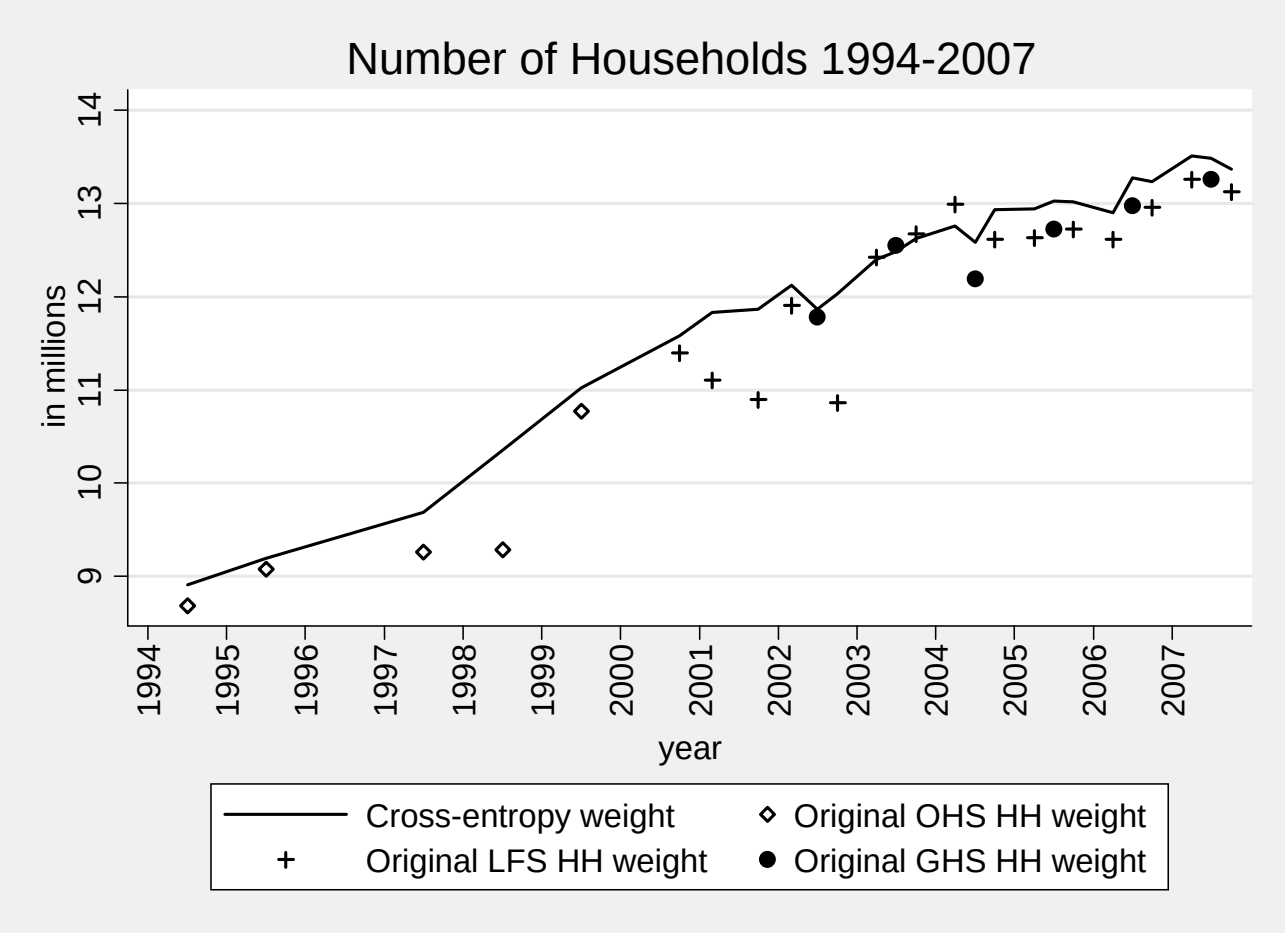
Notes to Figure 1: Figure 1 present estimates of the population using both the original StatsSA person weight and the new cross-entropy weights for each available OHS, LFS and GHS survey between 1994 and 2007. When placed side by side and weighted by the original person weights the surveys do not present a consistent series. The series can be divided into three parts each section with a differing slope. 1995-2000, 2001 to 2003 and 2004 to 2007. The cross entropy weights produce a smooth trend in the population over time.

Figure 2: Population counts for the Eastern Cape and Gauteng, a comparison using the entropy versus the original person weights.



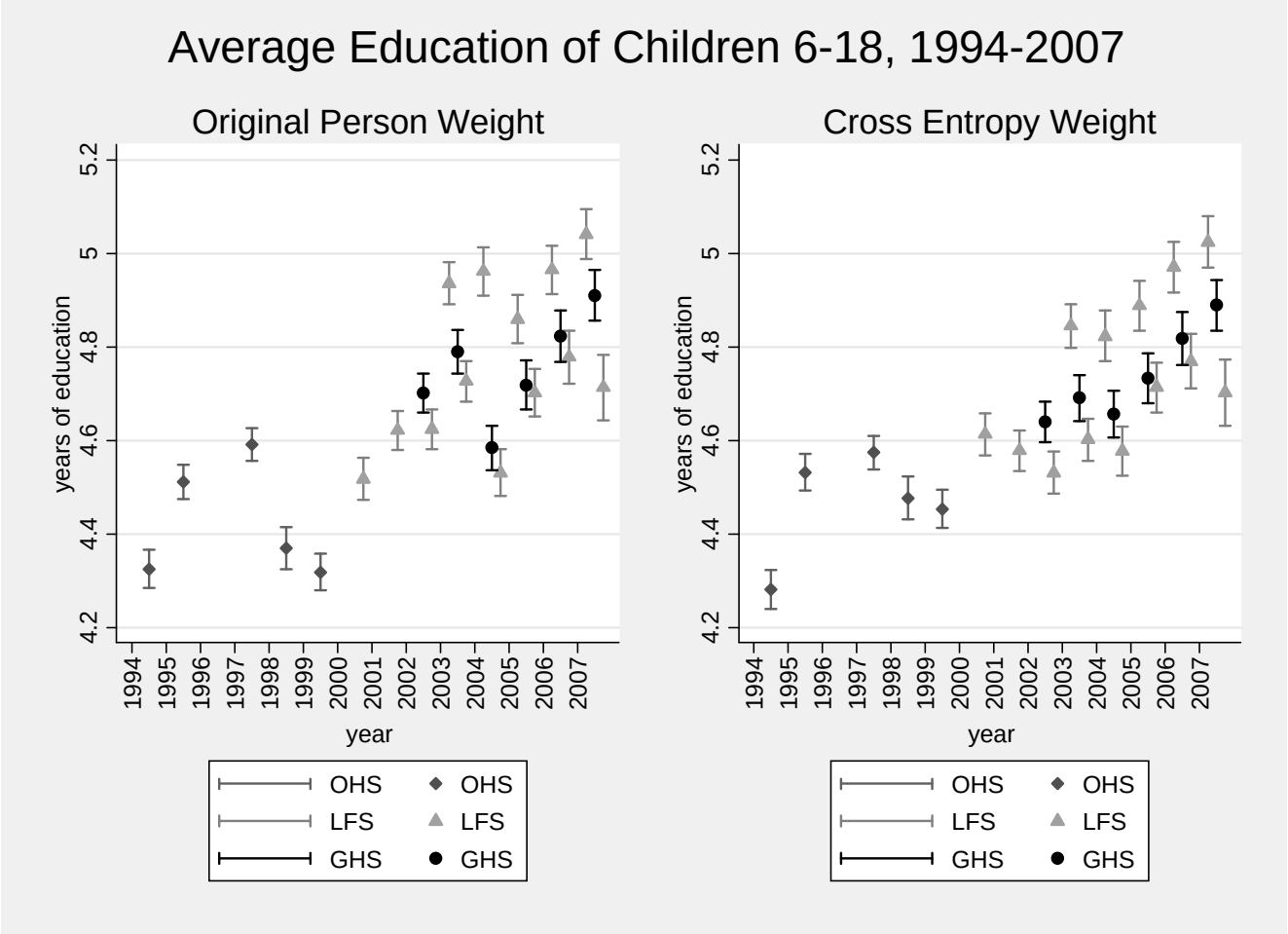
Notes to Figure 2: Figure 2 presents population estimates for the Eastern Cape and Gauteng using the original StatsSA person weight and the new cross-entropy weights. While the cross entropy weights form a smooth series, the original survey totals are not consistent when placed back to back.

Figure 3: Number of households, comparison using the entropy versus the original household weights



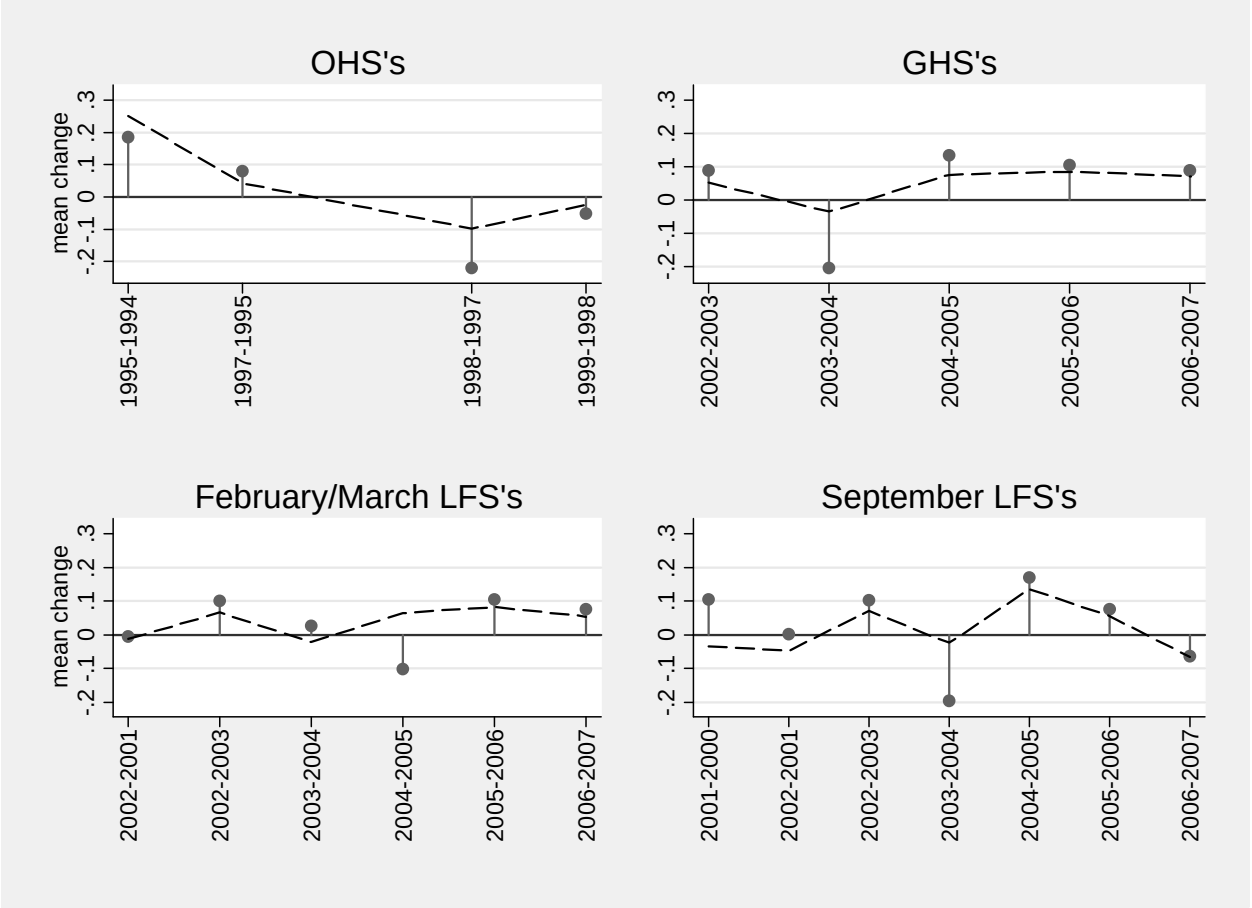
Notes to Figure 3: Figure 3 presents the trend in the number of households weighted using both the original StatsSA household (HH) weights and the cross entropy weights. It is clear that the household data was not benchmarked to an external series prior to 2003. The number of households follows a distinctively step-wise function until 2003 with increases in 1999 and 2003. 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.

Figure 4



Notes to Figure 4: Figure 4 presents *p*oint estimates and confidence bands for average educational attainment among children age 6-18. Estimates in the left hand panel are weighted using the original person weights and estimates in the right hand panel are weighted using the cross entropy weights.

Figure 5: Change in mean educational attainment between survey years – an assessment using the original versus the entropy person weights



Notes to Figure 5: Figure 5 presents the year-on-year change in mean educational attainment within survey type. The dropline represents estimates using the original person weights and the dashed line represents estimates using the new cross entropy weights. Points above the zero line show an increase in educational attainment while points below the line show a decrease in educational attainment.

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708