DataFirst Technical Papers

Re-weighting the OHS and LFS National Household Survey Data to create a consistent series over time: A Cross Entropy Estimation Approach

by

Nicola Branson

Technical Paper Series

Number 12

About the Author(s) and Acknowledgments

I would like to thank Professor Martin Wittenberg for his contributions to this paper.

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

Recommended citation

Branson, N. (2009). Re-weighting the OHS and LFS National Household Survey Data to create a consistent series over time: A Cross Entropy Estimation Approach. A DataFirst Technical Paper Number 12 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 the OHS and LFS National Household Survey Data to create a consistent series over time: A Cross Entropy Estimation Approach

Nicola Branson

Abstract

In the absence of South African longitudinal data for the ten years post apartheid, national cross-sectional household survey data is frequently used to analyse change over time. When these data are stacked side-by-side however, they reveal inconsistencies both in trends across time and between the household and person level data. These inconsistencies can introduce biases into research which analyse change. This study calculates a new set of person and household weights for the October Household Surveys between 1995 and 1999 and the Labour Force Surveys between 2000 and 2004. A cross entropy estimation approach is used. This approach is favoured because the calculated weights are similar to the initial sample weights (and hence retain the survey design benefits) while simultaneously being consistent with aggregate auxiliary data. A consistent series of aggregates from the Actuarial Society of South Africa (ASSA) model and the 1996 and 2001 South African Census data are used as benchmarks. The new weights result in consistent demographic and geographic trends over time and greater consistency between person and household level data.

Key words: *South African national household survey data, Post-stratification, Re-weighting, Cross entropy estimation.*

- I would like to thank Professor Martin Wittenberg for his contributions to this paper

1. Introduction

One main focus of post apartheid research in South Africa is on change. Questions include the progress of South Africa in the economic, social and political arena. National datasets such as the October Household Surveys (OHS) and Labour Force Surveys (LFS) 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, there are changes in sample design, the interview process and shifts in the sampling frame which can cause unrealistic changes in aggregates over a short period of time. This raises concerns about the validity of using these datasets as a time series to examine change.

The purpose of the survey weights is to make the sample represent the population and therefore the weights play an important role in creating consistent aggregates over time. Surveys select different households with different inclusion probabilities as a result of both designed and unplanned factors. Some households are therefore overrepresented relative to other households and in order for the sample estimates to accurately reflect the population it is necessary to weight each household according to its ‘true’ inclusion probability (Deaton, 1997). Design weights reflect the sample design and therefore would inflate the sample to the population in a world without non-coverage, item and unit non-response. Post-stratification adjustment is an adjustment to the weights after data collection which attempts to account for these errors by benchmarking the survey data to external aggregate data. Yet unlike design weights, the post-stratification adjustment is not well-defined, but rather open to judgement and hence error.

Ten years of data from the OHS and LFS are stacked side-by-side and it is found that the aggregate trends calculated from the survey weights are both temporally and internally[1] inconsistent. Examining the weights given in the datasets, in addition to the public documentation, it is clear that the Statistics South Africa (StatsSA) household and person weights are not simple design weights i.e. inverse inclusion probability weights. StatsSA post-stratifies the person 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 creates problems when the data are used as a time series. This paper highlights and addresses two concerns with the original StatsSA weights. First, the auxiliary data used as a benchmark in the post-stratification adjustment is unreliable and inconsistent over time and hence results in temporal inconsistencies even at the aggregate level. Second, since the adjustment is made at the person level until 2003, there is no hierarchical consistency between the person and household weighted series. This means that analyses done at the household and person level will not necessarily agree.

A new set of person and household weights is generated using an entropy estimation technique. The new weights result in consistent demographic and geographic trends and greater consistency between person and household level analysis. The benefit of the entropy post-stratification approach is that it preserves the survey design by

[1] Household and person weights produce numbers which are inconsistent with each other

selecting the new weights to be as similar to the original weights as possible while simultaneously meeting the restrictions. To test the new weights, they are used in a simple employment analysis and the results compared with those found when using the old weights. The trends are smoother when the new weights are used. In particular, the LFS employment series shows more consistency over time and a household level series, the number of households with piped water, is far more realistic. The re-weighting does not however mediate the large increase in economic activity between 1997 and 2000, the employment spike in 2000 or the apparent over-representation of households with piped water in 1995. This highlights an important aspect of this paper. The re-weighting procedure does not deal with specific measurement changes in the data series. Any changes observed when the new weights are used indicate that the variable being analysed is influenced by the original weights distorting the distribution of one or more of the variables used as restrictions in the entropy calculation. For instance, since the large increase in economic activity between 1997 and 2000 is not mediated by the new weights, this signals that this shift is not driven by faulty weights but rather by something internal to the questionnaire, for instance, how it was administered or another uncontrollable factor.

The remainder of this paper is organised as follows. Section 2 introduces the theoretical basis for weights and post-stratification and highlights the South African interest in data quality issues. This section also includes a description of the approach StatsSA takes with respect to post-stratification and introduces entropy estimation as an alternative approach which can be implemented in Stata. Section 3 motivates the need to re-weight ten years of national household data to be consistent with demographic and geographic numbers presented by the Actuarial Association of

South Africa (ASSA) model and Census data. Section 4 explains the entropy concept and theoretical framework for the re-weighting procedure. Section 5 presents the results including both an assessment and comparison of the old and the new weights and the affect the new weights have on aggregate trends. The cross entropy weights are found to be an appropriate alternative to the original StatsSA person and household weights with some added advantages over the originals. They present consistent time trends in demographic, geographic and other variables while preserving the benefits of the original sample design. In addition, the household and person entropy weights are more internally consistent. Section 6 concludes.

2. Literature

2.1 An overview of weighting the OHSs and LFSs

The purpose of the OHS and the LFS is to collect data on the circumstances of people in South Africa with the LFSs, as their name suggests, focusing primarily on variables related to the labour force. While data collected on the sample tell a story of the living condition of the people in the sample, the survey design weights allow the researcher to make inferences about the national population. Thus the sample data has the potential, if correctly weighted, to produce aggregate data which can be used in assessment and projections. These weighted data can thus be used to inform policy and to complement the national accounts (Deaton, 1997).

The markets respond to changes in the aggregate numbers. Therefore a series of incorrectly or inconsistently weighted data can depict an inaccurate picture of aggregate changes over time. For instance, favourable changes in aggregate numbers can be used as political leverage. One such instance which was pointed out by Posel and Casel (2004), was the African National Congress's (ANC) comment in the run up to the election that the "economy (had) created two million net new jobs since 1995". If part of this increase is driven by shifts in the weights, for instance an increased representation of employment, then discussing an aggregate increase between two years without assessment of the comparability of the datasets could lead to erroneous claims. It is therefore important that National surveys which form a series over time, such as the LFS's, be carefully weighted to reflect realistic changes over time. This comment also illustrates the importance of accurate aggregate numbers over time in addition to realistic changes in proportions.

The principle behind sample weights is simply to inflate the sample to reflect the population. If one had a complete list of all available households in the population one could randomly draw a sample and each household would have the same probability of selection. If each household selected was willing to participate, then each household would represent an equal proportion of households in the population. This form of sampling is called simple random sampling. (Deaton, 1997, pp. 9-18)

In most cases however, due to cost restrictions, research demands and sampling error, the survey design is more complex than this. One common approach is a two-stage sampling design. In this case the sampling frame provides, in principle, 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. Even in this more complex design, if the clusters are randomly selected with probability proportional to the number of households within them and the same number of households is drawn from each cluster, the sample design will be self-weighting and each household would have an equal probability of inclusion in the sample. (Deaton, 1997, pp. 9-18)

Research frequently requires that the representation of subgroups within the population, often minority groups, be large enough to produce robust estimates. Stratification by the defining characteristic of the subgroup, be it geographical region, population group or other, divides the sample into sub-samples each one representing a subgroup. This guarantees that enough observations for each subgroup will be included in the total sample. In addition, stratification has the ability to reduce the variance of estimates and hence make the point estimates more reliable. Since the strata are independent the overall variance is the sum of the individual strata variances only; the covariance across groups is zero. In designing a survey there is often information known about the target population prior to data collection. If this information indicates that groups are similar within group but different across group, then stratification reduces the within group variance and hence the total variance.

While it is possible to have a clustered and stratified survey design which still results in households which have equal inclusion probabilities, it is more likely that households will differ in inclusion probability. This is a result of design features (limiting cost and to attain accurate measurement of small strata) as well as due to non-response and other sampling errors. When inclusion probabilities differ across households, each household in the sample represents a different number of households in the population. As such, when using the data to make inferences about the population it becomes important to weight the sample correctly such that each subgroup is correctly represented in the population. Straight averages calculated from the sample will be biased estimates of the population and weighted averages which account for the survey design should rather be used. Each household is weighted by the inverse of its probability of inclusion in the sample (Deaton, 1997). This makes intuitive sense since a household with a low probability of selection represent a large number of households in the population and a household with a high selection probability represents a minority-type household in the population. These weights are often referred to as “raising” or “inflation” factors since they inflate the sample to look like the total population.

Divergences in weights across households come from differences in selection probabilities due to both planned (the survey design as discussed above) and unplanned factors. Unplanned differences arise due to measurement errors as well as sampling errors, like an out-of-date sampling frame or non-response. To obtain accurate population averages the sample needs to have weights that reflect actual inclusion probabilities, in other words account 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. The adjustment of the weights to account for unplanned differences is, in character, a less controlled process and involves judgement and modelling. The survey weights which accompany the national household surveys have been adjusted to account for unplanned differences under certain assumptions chosen by the survey agency (StatsSA). These assumptions might not be correct and/or in line with an individual researcher’s view.

2.2 Data Quality in South Africa

Awareness of South African data quality issues among researchers is not new, but is growing. Researchers often present a caveat to their findings: results being subject to data quality issues[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. While documentation of these issues is still in its infancy, the University of Cape Town Datafirst centre has been awarded a grant by the Mellon Foundation, specifically dedicated to assessing and documenting South African data quality issues.

Sample design problems and changes in the South African datasets are relatively well documented in the literature. Posel and Casale (2003) compare changes in household definition and who is classified as resident, with particular attention to migrant members. Muller (2003) and Posel *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. Keswell and Poswell (2004) and Ardington *et al* (2006) discuss the effect incomes incorrectly captured as zero can have on an analysis.

[2] Bhorat, H. and Kanbur, R.(2006), Branson, N and Wittenberg, M (2007), Burger, R and Yu, D. (2006) Casale, D., Muller, C. and Posel, D. (2004), Cronje, M and Budlender, D (2004), Wittenberg, M. and Collinson, M. (2007), G. Kingdon and J. 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.

Little research goes further by attempting to address the observed data problems. Posel and Casel (2003) attempt to ensure consistency of migrant household membership by imposing the stricter migrant membership definition from the 1997 and 1999 OHS's on the 1995 OHS and 1993 PSLD data. Ardington *et al* (2006) assess the effect that different treatment of missing and outlying income data from the 2001 Census have on poverty measures. These adjustments do make a difference. For instance, while Ardington *et al* (2006) find that their use of multiple regression imputations for missing data results in similar conclusions regarding poverty to when the missing data are ignored (implicitly assuming that the missing data are missing completely at random), the adjustment for outliers results in a significant increase in mean income and thus a more optimistic picture of poverty and inequality.

South African literature that assesses the sensitivity of economic trends to weighting issues is even further limited. Simkins (2003) generates a set of weights for the 1995 and 2000 Income Expenditure Survey (IES) data sets 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 are found to have a significant effect on mean expenditure, but 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 these results do change.

2.3 Statistics South Africa Weights

The survey weights supplied by Statistics South Africa (StatsSA) in the national household surveys are adjusted inverse sample inclusion probability (design) weights.

It is worthwhile to review the approach taken by StatsSA in constructing these weights during both the sample design and post-stratification.

The sample design of the OHS and LFS data is a two stage procedure. Take for instance the LFS 2002 September data. Initially 3000[4] primary sampling units (PSU's), the clusters, are drawn from a Census[5] master sample (the sampling frame) and from these ten dwellings (households) per PSU are selected. PSU's are explicitly stratified by province and area type (urban/rural)[6] and a systematic sample of PSU's are drawn by probability proportional to size within stratum.

The household weight is created as a function of the PSU inclusion probability and the household inclusion probability[7]. Each person within a household is assigned the same person weight. Due to sampling and measurement errors, these weights do not inflate the sample to accurately reflect the population and therefore they need to be adjusted. The adjustment procedure undertaken by StatsSA is not clearly documented, but the following guideline is presented in the metadata files of these datasets[8]. In the LFS data sample person weights are assessed for outliers using a SAS procedure called Univariate. Next a SAS calibration estimation macro called CALMAR (CALibration to MARgins) is used which adjusts the data to population proportions defined by population sex, race and age group marginal totals in the mid-year estimates[9]. These weights are said to be trimmed but the method used is not detailed. Exponential projection is used to adjust these weights to the date of the LFS, for

[4] All years had 3000 PSU's, except 1996 (1600) and 1998 (2000)
[5] 1995-2002 use the 1996 Census and 2003 and 2004 use the 2001 Census
[6] Resulting in 18 stratum
[7] See LFS Metadata for details
[8] See page 28 (Table 1) for variations in the post-stratification procedure between years
[9] Mid-year estimates are produced by Stats SA

example the June (mid-year) estimates are adjusted to September in the case of the LFS September data sets.

2.4 Post-Stratification

StatsSA uses the CALMAR (2) approach (referred to as generalised raking by Deville *et al* (1993)) which is a form of re-weighting known as post-stratification. Post-stratification incorporates any data adjustment which organises data in homogenous groups post-data collection, but is usually done where external information on these groups is available (Smith, 1991). Post-stratification adjusts the survey design weight within chosen subgroups (called post-strata) such that the sample reproduces the known population proportions.

a. The purpose of post-stratification

Post-stratification has three main functions. The first and chief function is to adjust the design weights to account for sampling errors and hence enable the sample to represent the population. In other words, the main purpose of post-stratification is to reduce biases from coverage and non response error (Smith, 1991, p. 322). Take for instance non-response. When non-response is not completely random, the probability sampling scheme of the non-respondents actually depends on the variable of interest, i.e. the sampling scheme is informative about the non-respondents. The role of post-stratification is to make the non-response scheme uninformative and thus eliminate the non-response bias (Smith, 1991).

Second, post-stratification can be used as part of the sample design. When a stratified sample is constructed, knowledge prior to sampling of the stratifying variable is required. If the stratifying variable is not available at the time of selection or is too difficult or expensive to use, post-stratification is a useful alternative (Little, 1993 &

Smith, 1991). Lastly, post-stratification has the potential to increase the precision of estimates highly correlated with the auxiliary information (Zhang, 2000). As such, post-stratification combines survey data and aggregate population estimates and hence imposes a consistency between survey results and those from other sources, a highly beneficial characteristic of any data.

b. The disadvantages of post-stratification

Post-stratified estimation does not necessarily present a robust approach to improving the representation of a sample. There are some potential drawbacks. First, in any stratification there is the potential to create strata which have too few data points (or none) for robust estimation. This is called the small or empty cell problem. Second, population totals at the post-strata level may be unavailable or unreliable. Third, auxiliary information is generally available at the person level and hence adjustments are made to the person weights. Auxiliary data at the household level is more limited and hence in practise, household weights are often derived from the person weights inappropriately. This can result in different inference when analyses are done using household versus person data (Neethling & Galpin, 2006). Lastly, some re-weighting techniques do not control the range of the adjusted weight which can result in negative, zero and/or very large weights. Negative weights are clearly illogical, while zero and very large weights result in a part of the sample being significantly under or over influential.

The small or empty cell problem arises when the cross classification of post strata variables results in cells with small sample size or even completely empty cells. Weighting of these small sample cells to reflect a proportion of the population is imprecise. One remedial approach is to collapse cells which have small size with neighbouring cells. The main aim is to collapse neighbouring post-strata that are as homogenous as possible. However, when the assumption of missing completely at random (MCAR) does not hold[10] it is important not to collapse post-strata with significantly differing response rates. Since post-strata with low response rates are most prone to being collapsed, the gain in precision can be offset by an increase in bias. Calibration estimation provides a methodology for dealing with the small cell problem and is applicable even when MCAR does not hold (Deville & Sarndal, 1992).

Post-stratification adjustments are based on adjusting the sample estimates to what is assumed to be the 'true population'. This requires knowledge of the exact population distribution or marginal distributions. If the 'population' data available are unreliable or out of date, frequently the case when using census data, adjusting to the incorrect frequencies introduces bias. Thus if auxiliary data are of poor quality (or in the case of a series of data are inconsistent over time), the value of post-stratification is questionable since the potential bias introduced may offset the gains from increased precision (Smith, 1991).

Little (1993) notes that most of the post-stratification literature approaches post-stratification from this randomisation perspective[11], where benefits to the sampling distribution are assessed, taking the population estimates as fixed or true. This approach implicitly assumes that known population estimates are without error. This assumption is unlikely to hold in most cases, with the possible exception of countries

[10] See Little (1993) for other suggestions on how to deal with the small cell problem when MCAR does not hold

[11] Post-stratified estimation (Holt & Smith, 1979), regression estimation (Bethlehem & Wouter, 1987), calibration estimation (Deville & Sarndal, 1992) and generalised raking (Deville, Sarndal, & Sautory, 1993) are a few examples.

with detailed population registries. In an attempt to address this deficiency, Little (1993) takes a “predictive modelling perspective” (Little, 1993, p. 1001), a Bayesian approach where the population estimates themselves are random variables and are allowed to follow a distribution.

Neethling and Galpin (2006) investigate the extent of the bias introduced when post-stratification is done at the person level without a control for household level factors. When adjustments are made at the person level the person weights frequently differ across people in the same household. This creates two problems. First, the person weight does not account for household size or within household homogeneity, in other words for the fact that certain people are in the same household and should be treated as a cluster. Second, since person weights differ across people within the same households it is not immediately obvious which weight should be used to represent the adjusted household weight (Neethling & Galpin, 2006). Integrated linear weighting developed by Lemaitre and Dufour (1987) deals with the problem of consistency between person and household[12].

StatsSA’s approach to post-stratification has both strengths and weaknesses. StatsSA currently use the SAS macro CALMAR 2 to post-stratify the design weights. This approach has the ability to address the small sample cell and negative weight problems, the availability of population totals and consistency between household and individual data problems. It is therefore a beneficial approach to take. There are however, a few issues which need to be addressed when constructing a data series from these cross sectional data.

[12] See Neethling and Galpin (2006) for a clear example

First, the reliability of the auxiliary data (the mid-year estimates) used by StatsSA in the calibration procedure, remains questionable. Dorrington and Kramer (unpublished) highlight the problems present in the mid-year estimates produced by StatsSA. Inconsistency within the mid-year estimates across years and in relation to other model projections is found. Thus the auxiliary data used in the post-stratification adjustment is of poor quality especially when used as a time series. Thus the approach taken to re-weight the national household survey sample weights introduces bias in trends across time with consequences for statistical inference.

Second, prior to 2003 CALMAR and, before that, relative scaling were used for post-stratification. These approaches made adjustments at the person level without consideration of household factors. This results in inconsistency between person and household level datasets. Finally, the metadata from earlier years indicates that a trimming adjustment was required. This signals that the procedures used did not ensure that the calculated weights fall within a realistic range.

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

2.5 Re-weighting the Series: Entropy Estimation

Entropy estimation has many of the advantages outlined for the CALMAR approach. In addition, entropy estimation can be simply applied to effectively address some of the time and household-person hierarchical inconsistencies observed in the series from 1995 to 2004. Entropy estimation is becoming popular in economics due to its ability to deal with ill-posed (data points less than unknowns) and ill-conditioned (unstable parameter estimates, for instance due to collinearity) problems (Fraser, 2000). The underlying principle, based on the work of Jaynes (1957) and Shannon (1948), is to find a solution consistent with the data without imposing extraneous assumptions on the data (Golan, Judge, & Miller, 1996).

Consider the information available to a user when creating a series of data between 1995 and 2004: national sample data, adjusted design weights (which contain the survey design information), and a time consistent set of external aggregates (benchmarks) from an external source such as the ASSA model. The adjusted design weights are biased due to time inconsistent benchmarks, hierarchical inconsistency between household and person files and trimming error in the earlier years. The cross entropy approach re-calculates the weights to account for these errors, i.e. makes the sample represent the population, but at the same time keeps the adjusted weights as similar to the original weights as possible, hence preserving the sample design benefits.

Thus the advantages of entropy estimation, like the CALMAR approach, are three fold. First, entropy estimation adjusts to marginal totals and therefore the small/empty cell problem does not affect the estimation procedure and benchmarks from multiple sources can be used simultaneously. Second, the entropy approach does not require that the re-calculated weights be trimmed since the functional form of the entropy problem does not allow negative weights (Merz (1994) and Merz and Stolze (2006)).

Third, because the constraint set (the external benchmarks) can contain information at different hierarchical levels, the cross entropy weights can potentially be calculated to be consistent across person and household files (although this advantage is not utilised in this paper).

Entropy estimation has two further advantages over the CALMAR approach. First, the procedure can be programmed in Stata using the ml command and hence avoids Stata users the additional cost of purchasing CALMAR(2). Second, the basic cross entropy approach can be extended to allow for measurement error in the external aggregates[13]. Allowing for measurement error in the aggregate data, recognises that population level data (with the exception of a complete population registry) is not free from error. Robilliard and Robinson (2003) use a generalised cross entropy (GCE) estimation method in an attempt to reconcile Madagascan household survey data and macro data. The authors favour this approach because it allows them to estimate a new set of household weights which are consistent with the aggregate data while simultaneously allowing the aggregate data to be measured with error (Robilliard & Robinson, 2003, p. 2).

This paper calibrates the StatsSA survey weights to external benchmarks from the ASSA model and census data. A cross entropy estimation approach is used. The most efficient way to make this adjustment would be to use the original design weights (pre StatsSA's post-stratification adjustment) in the estimation. These weights are however, not publicly available and hence the adjusted design weights are used. These weights generate an aggregate series which is time consistent, i.e. enables the OHS

[13] Not utilised in this paper

and LFS data to be used as a consistent series at the aggregate level. In addition, consistency between the household and person level data is increased.

3. Motivation for paper

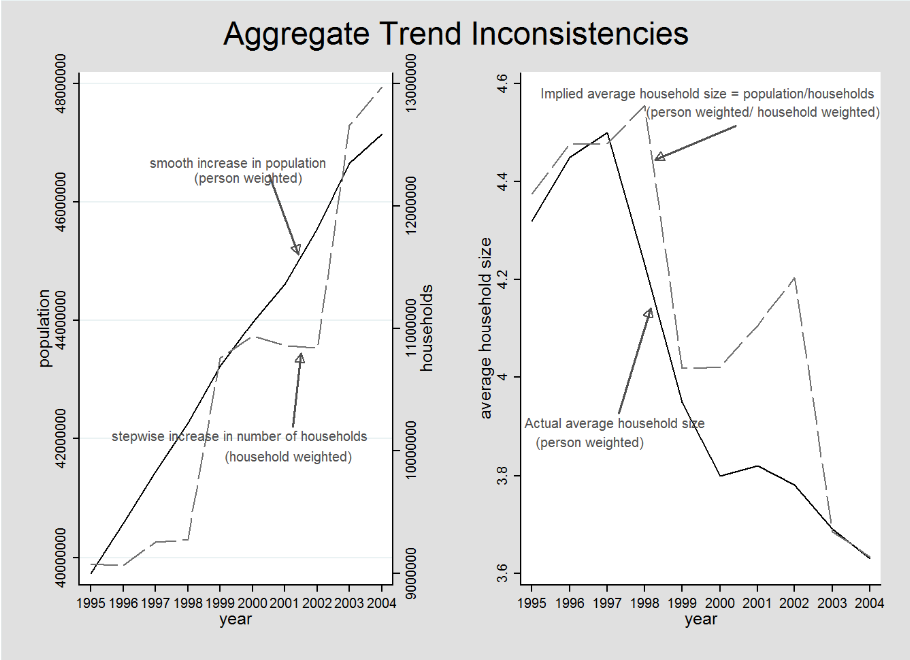
The OHS and LFS national household surveys are cross sectional surveys which have common features over time (similar questionnaires and sample designs). They are not however, designed to be used as a time series and hence unconditional stacking of these surveys year-on-year to create a time series (a practise commonly undertaken by researchers) can result in problems. This section illustrates that when the October Household Surveys (OHS) from 1995 to 1999 followed by the Labour Force Surveys (LFS) from 2000 to 2004 are stacked side-by-side the resulting trends show non-trivial inconsistencies over time. Even at the aggregate level, for instance population by province, the data display large fluctuations over time. The volatility in the series could be a result of various differences in survey design, the way the survey weights are calculated or other measurement changes over time. This paper focuses on the effect of the survey weights. The external benchmarks used by StatsSA in their post-stratification procedure produce inconsistent aggregates over time. In addition, there are inconsistencies between the person and household level files.

The following section takes a closer look at the data in an attempt to illustrate inconsistencies that can potentially be addressed through re-weighting the national household datasets to a demographically and geographically consistent external data series. Inconsistencies both at the aggregate household and strata level and between the household and person level files are shown. While survey data are used at the aggregate level to inform and assess policy and progress at the macro level, most research focuses on analyses that look at the changes in the proportion of the population in a certain state. Section 3.3 discusses inconsistencies observed in the proportion of the population classified as living in a single person household. In section 3.4 an attempt is made to find reasons why these inconsistencies were not addressed through the post-stratification procedure undertaken by StatsSA. Throughout the discussion the example of the large increase in the economically active female population between 1995 and 2004 is used to illustrate the potential effect incorrect weighting can have on a proportionate analysis.

3.1 Aggregate trend inconsistencies

Figure 1 displays trends in the population, the number of households and the average household size (implied and actual). The figure illustrates how inconsistent the time series of household survey data is even at the aggregate level. While the population trend appears realistic, increasing steadily over time, the number of households follows a distinctively step-wise function with increases in 1999 and 2003. These are not an accurate depiction of reality. One explanation of the large increase in number of households in 2003 is the implementation of the 2001 Census as the sampling frame, replacing the 1996 Census sampling frame. Similarly, the increase in 1999 could be a result of the introduction of the 1996 Census.

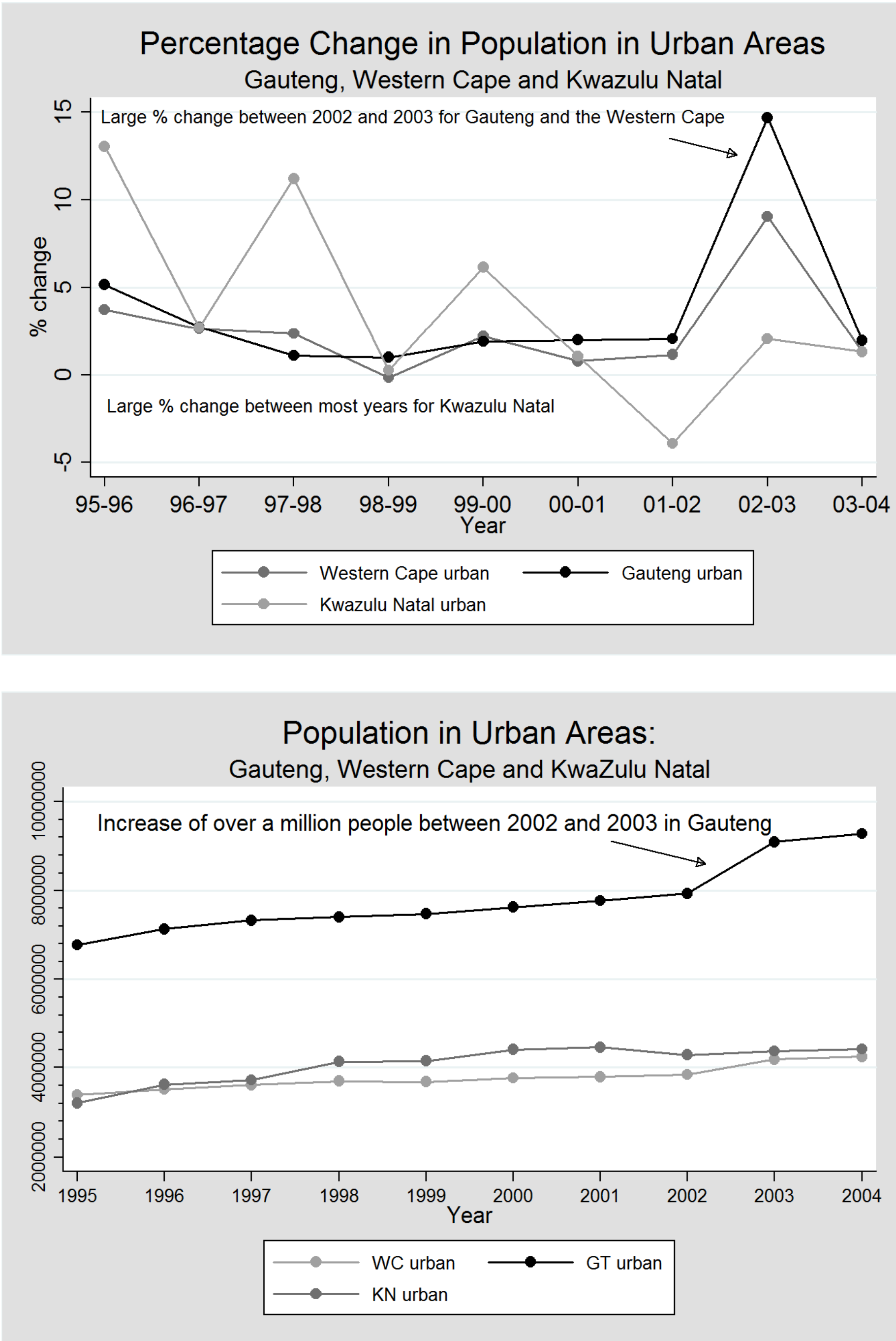
Figure 1: Household, population and household size trends



Notes: *The household data has not been post-stratified (adjusted to meet external population aggregates). This results in a stepwise increase in the number of households over time. In addition, the household and person level data give different analytic conclusions. The right hand panel illustrates that when the population is divided by the number of households (implied average household size) from the household level data the series is very different from the actual trend in average household size observed in the person-level data*

At the strata (province by urban/rural) level, the inconsistency of the household surveys as a time series is further illustrated. Figure 2 shows that the largest urban only province, Gauteng, increases by over a million people between 2002 and 2003, this represents a 15% increase in the population of Gauteng in one year. Similarly, the Western Cape shows a large increase between these years and Kwazulu-Natal shows large changes between most years.

Figure 2: Population Trend in Three Large Urban Areas



Notes: *The OHS and LFS data are not designed as a time series. Stacking the data can result in large year-on-year shifts. The figure illustrates that even at the province level there are large temporal shifts*

Of the rural provinces (displayed in Appendix A figure A.1), two of the largest mainly rural provinces show decreases between 2002 and 2003; both the Eastern Cape and Limpopo show around a five hundred thousand person decrease, a 8-10 percent decrease between 2002 and 2003, where in previous year the trend was upwards. The Kwazulu-Natal rural trend is U-shape over the ten year period. Between 2000 and 2004 an increase of a million people is observed.

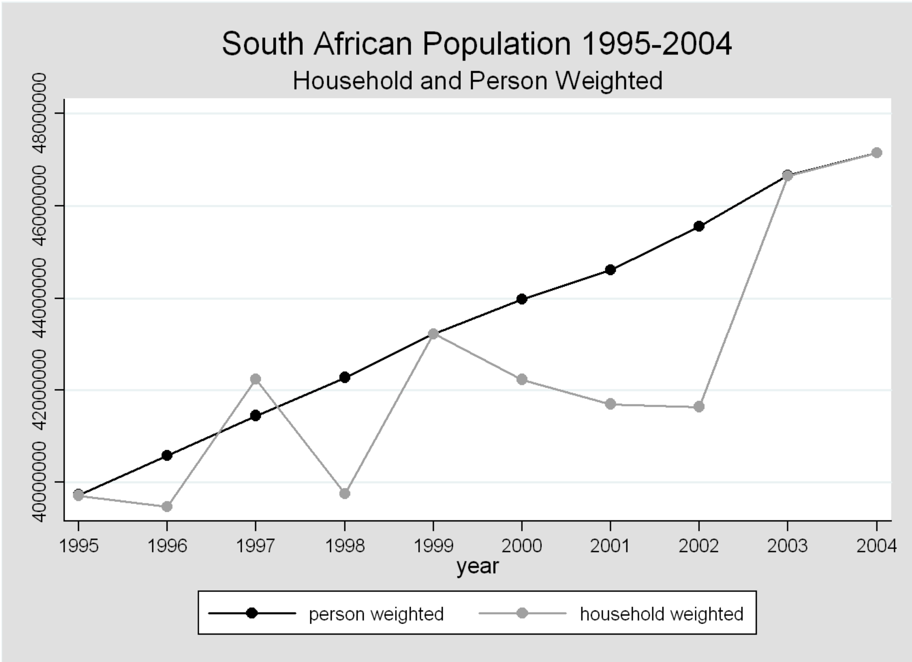
While the increase in the population over time looks believable, the strata trends show that the placement of people within the country looks unrealistic between years in some provinces. These shifts have important implications for analyses given the varying levels of poverty and inequality, resource access and other social factors between provinces. In the 2003 case, by inflating Gauteng's representation and deflating rural Eastern Cape and Limpopo representation the likelihood of over-representating resource access in the aggregates is high.

Take for instance the proportion of the population economically active. Large increases in economic activity are found between 1998 and 2001 (Branson & Wittenberg, 2007). Is this a result of an actual shift, i.e. people increasing their propensity to seek work, or is it that provinces with high economic activity are initially under represented and/or later over represented? It is not completely obvious which provinces would represent areas of high or increasing economic activity. While it might be assumed that over or under estimation of large urban areas would have the main or only impact on changes in the proportion of the population classified as economically active, areas with a large proportion of informal employment should also be considered. If, as suggested by Casale *et* al (2004), later surveys became increasingly astute at finding marginal forms of employment, and hence economic activity, it is likely that an over representation of these areas would confound this effect. This example illustrates the importance of a consistent series in the basic geographic (and demographic) variables to rule out distortions on analyses done at the proportionate level. Benchmarking the weights to meet a series inherently consistent over time, can rule out the effect of a once-off shift in the weights.

3.2 Between household and person file inconsistencies

In addition to consistency over time, it is also important that inferences made at the household level tell the same story as inferences made at the person level. The person design weights would have been common within household (due to stratification at the household level) and thus the household design weight and the person design weight are the same. Post-stratification at the person level however, results in differing person weights within household which can result in inconsistent inference between person and household level data. StatsSA post stratified at the person level until 2003, thereafter the introduction of CALMAR 2 allowed household level calibration. Figure 4 presents the population trend for the OHS and LFS data. A comparison is made between the population count generated from the original StatsSA person weights and the population count generated by assigning the StatsSA household weight to each member in the household. There is no consistency between analyses using the household and person files up until 2003 when CALMAR 2 was introduced. It is clear that the person weights have been benchmarked to an external series such that the population trend is uniform over time, while the household weights do not appear to have been adjusted and hence display an erratic trend.

Figure 4: Inconsistent population trends



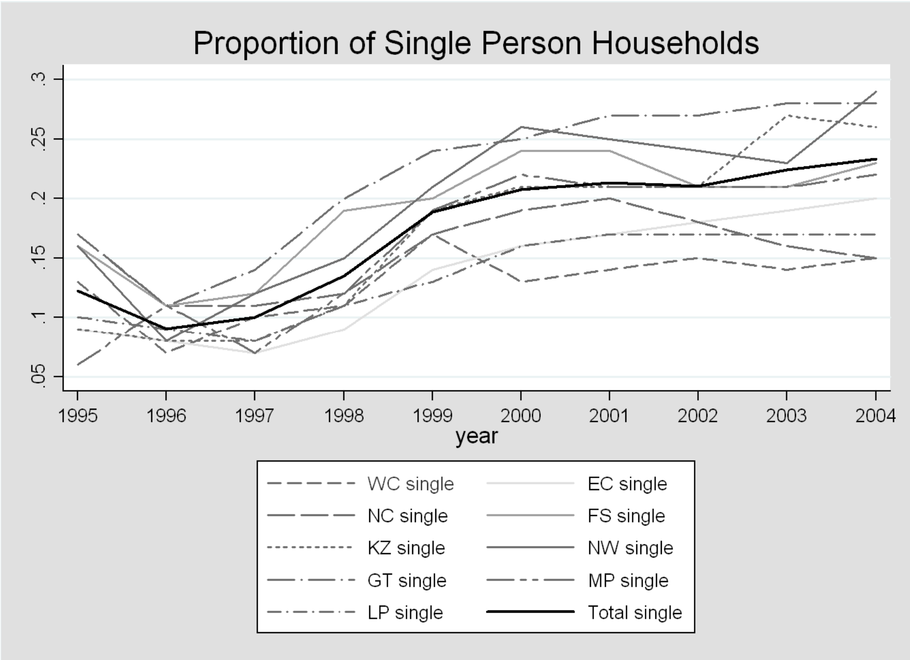
Notes: *The household weight is assigned to each person within the household and the population calculated. This trend is compared to the population when the person weight is used. It is clear from the figure that StatsSA undertook post-stratification at the person level until 2003. Thus until this point, the household and person weighted trends diverae.*

The OHS and LFS household level data have important variables reflecting economic and social well-being at the household level, for instance access to water, electricity, materials used to construct the dwelling, and number of rooms in the dwelling. Aggregates of these variables are used to assess of economic and social well-being and progress at the country level and hence correctly weighted household data are essential.

3.3 Proportionate Analyses

Most research examining change over time focuses on changes in the proportion of the population in a certain state. Part of the motivation behind using proportions instead of numbers is to avoid the inconsistencies discussed above. Yet, even at the proportionate level the data series display inconsistencies.

Figure 3: Proportion of single person households



Notes: *The figure illustrates a rapid increase in the proportion of single person households, much of the increase takes place between 1997 and 2000. This is a common feature in most of the national household surveys but is not found in other data sets.*

Take for example the trend in the proportion of single person households. Wittenberg and Collinson (2007) find the proportion of single person households increases more rapidly in the national household surveys than in other data. The implication of increasing the prevalence of single person households on an analysis could be significant. Take for instance the proportion of economically active females. People who live alone are more likely to be economically active (with the exception of the elderly) simply because they have no immediate financial support network. Figure 3 shows an increase in single person households from 12% to 23% over the ten year period, with most of the increase taking place between 1997 and 2000. This rapid increase coincides with the large increase in economic activity.

3.4 The Benchmarks

StatsSA benchmark their data to external population projections in an attempt to address unplanned differences in inclusion probabilities due to non-response and other sampling problems. Since the OHS's and LFS's are cross sectional datasets, the purpose of the benchmarking is to produce representative data for the particular year in question. The focus is therefore not on producing a consistent series over time. The problem however, is that the data are frequently stacked year-on-year by users to create a time series without questioning the consistency of the data as a series. This is only reasonable if the surveys are annually representative.

Table 1 details which variables were used as benchmarks and the source of the benchmark in each year. There is a clear distinction between the OHS's and the LFS's. The OHS data were benchmarked to the "1996 Census, adjusted for growth[14]" to the year of the OHS. The LFS's use the mid-year population estimates[15] adjusted to the month of the LFS. The LFS's use demographic variables in the calibration process while the OHS's use geographic variables as well. Thus there is a break in the benchmark series.

[14] No further information is given
[15] produced by StatsSA's demography division

Table 1: StatsSA Post-stratification information

Survey	Calibration method	Auxiliary data source	Post-stratification variable
OHS 1995 re-weighted	Relative scaling	1996 Census adjusted for growth	Province, gender, age groups, race.
OHS 1996	Generalised raking with a linear distance function	1996 Census	Province, gender, age groups, race.
OHS 1997	Relative scaling	1996 Census adjusted for growth	province, gender, urban/rural, age group, race
OHS 1998	Relative scaling	1996 Census adjusted for growth	province, gender, urban/rural, age group, race
OHS 1999	Relative scaling	1996 Census adjusted for growth	Province, gender, age groups, race.
LFS 2000	CALMAR	2000 mid year estimates	Gender, race, age group
LFS 2001	CALMAR	2001 mid year estimates	Gender, race, age group
LFS 2002	CALMAR	2002 mid year estimates	Gender, race, age group
LFS 2003	CALMAR2	2003 mid year estimates	Gender, race, age group
LFS 2004	CALMAR2	2004 mid year estimates	Gender, race, age group

*source: OHS and LFS metadata

The mid-year estimates are projected from a base population under certain assumptions about fertility and mortality. The 2000, 2001, and 2002 mid-year estimates used the 1996 Census as the base population, while the 2003 and 2004 estimates were projected from an adjusted version of the 2001 Census. Under 'correct' assumptions of mortality and fertility these benchmarks would be consistent over time. Unfortunately the 2001 Census calls into question the 'correctness' of these assumptions; inconsistencies between both the 1996 and 2001 Census and the mid-year estimates and the 2001 Census are found.

The consistency of the national household survey data over time will in part depend on the reliability of the benchmarks as a series itself. For example, if the early OHS's are benchmarked to aggregates underestimating those most likely to be economically active, namely men of prime working age, and/or the later OHS's and LFS's over represent this sub-group, then a correction to the series of benchmarks used could help correct the apparent rapid increase in economic activity. In addition, peculiarities in particular years can be mediated.

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.

We therefore conclude that the series of benchmarks used over the ten-year period from 1995 to 2004 does not result in a consistent trend with respect to demographic variables. As a result the StatsSA data cannot produce a series which is consistent over time. The ASSA model estimates are proposed as an alternative benchmark series. The ASSA model projects a consistent time series which will therefore control the level of demographic and geographic variables in the national household surveys over time. Province, urban/rural, age group, sex and the proportion of single person households are used as benchmarks. The urban/rural and single person proportions are calculated from the Census 1996 and 2001.

3.5 Summary of motivation

The motivation for this paper is three-fold. First, to estimate a set of person weights which meet a consistent set of aggregate demographic and geographic trends. The ASSA model has been chosen for these benchmarks. In creating a series that is consistent at the aggregate level over time one potential source of error is removed, shifts in the survey weights. Second, the inconsistencies observed between the person and household level datasets will be reduced. Finally, the introduction of consistent demographic and geographic trends has the potential to affect analyses done at the proportionate level. If the proportion of population economically active is affected by the adjustment of the variables used in the post-stratification procedure, for instance an over representation of a province with a high proportion of economically active people, then the trend in the proportion of the population economically active could change.

The ASSA model accompanied by the 1996 and 2001 Census points is used to generate a smooth series over time. In addressing these factors through external benchmarking a set of weights which are demographically and geographically consistent between 1995 and 2004 are generated. A cross entropy (CE) estimation approach is used to estimate a new set of person weights that are consistent with the ASSA model estimates and Census data. The CE approach results in a set of weights which is consistent with the auxiliary data provided by the ASSA model and Census data while being as similar to the original StatsSA person weights as possible. Stata's maximum likelihood estimation procedure is used to programme the unconstrained dual CE problem presented in Golan *et al* (1996) (Wittenberg, unpublished).

4. Methodology

4.1 A Brief Introduction to the Entropy Concept

The purpose of weighting is to recover population estimates from a sample dataset. While sampling methodologies deal with many representation issues, errors arise due to sampling errors, non-response and coding errors. These need to be adjusted for in the weights. This is the objective of post-stratification: calibrate the sample weights to some known external population. When formulated as a classical estimation problem this estimation procedure results in an ill-posed problem since the number of unknown parameters to be estimated, the individual weights, exceeds the number of data points presented by the auxiliary data. There is no unique solution and no clear rule to choose the most appropriate solution. While the classical approach of dealing with an ill-posed problem is to reduce the number of possible solutions by introducing assumptions, these assumptions are often arbitrary and inconsistent with the data. Entropy estimation is an approach that is not subject to the ill-posed problem[16].

In information theory, entropy is a measure of uncertainty. Intuitively, "information contained in an observation is inversely proportional to its probability" of occurring (Fraser, 2000). The occurrence of an event with a high probability of occurring is unsurprising, while observing an event with a low probability elicits far more information about the underlying process (Fraser, 2000). Shannon (1948) defined a function, the entropy measure, to measure the uncertainty of the occurrence of a group of events. Following the notation of Golan, Judge and Miller (1996), let **x** be a random variable with possible outcomes x_k, $k = 1, 2, ..., K$ and probabilities, $\mathbf{p} = (p_1, p_2, ..., p_K)'$ then the entropy measure is:

$$H(\mathbf{p}) \equiv -\sum_k p_k \ln p_k$$

[16] Maximum entropy estimation is also not subject to the ill-conditioned problem

where $0 \cdot \ln(0)$ is defined to be 0. $H(\mathbf{p}) = 0$ presents the degenerate solution, one possible outcome with certainty. Note that $H(\mathbf{p})$ reaches a maximum when the probability distribution is uniform, in other words since the uniform distribution is least informative it maximises the uncertainty measure. Jaynes (1957) uses Shannon's entropy measure to recover the unknown probabilities, $\mathbf{p}$. Jaynes's maximum entropy principle chooses the distribution which is least informative but just sufficient to meet the probability constraints (Golan, Judge, & Miller, 1996). The solution to this problem thus uses all and only the available information without the need for extraneous assumptions.

The maximum entropy principle can be generalised to include prior information about the probability distribution with the aim to improve the accuracy of the estimates. This approach is called the principle of Cross Entropy (CE)[17]. Let **q** be the prior distribution, then the CE estimate of **p** is that estimate which minimises the difference from **q**, given the constraints of the problem. As such, the estimate which is as close to our prior knowledge as possible while being consistent with the data is chosen. The CE principle is defined as follows (Golan, Judge, & Miller, 1996):

$$\underset{p_k}{Min}\, I(p,q) = \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k} \right) \right)$$

$$= \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln p_k - \sum_{k=1}^{K} p_k \ln q_k \right)$$

4.2 Parameter Estimation

In this paper, an entropy estimation approach to reconciling household surveys with aggregate data from the Census and the ASSA model is presented. The aggregates are

[17] According to Kullback (1959)

taken as the population totals and the person data is reconciled to these totals using the cross entropy principle. The problem therefore is to re-estimate the person weights such that the survey data are consistent with the aggregates presented in the ASSA model and censuses while simultaneously being as similar to the original person weights as possible. Maximum entropy weights are also generated for illustrative purposes.

Information used in our approach comes from three sources. First, the StatsSA person weights detail a large amount of information about the sample design and demography of the population. These will be used as the starting point for the estimation; as the prior distribution of the weights, $\mathbf{q}$. The second source of information is the survey data itself. Lastly, the ASSA model and Census aggregates represent known moments of the population distribution. The ASSA model estimates by province, age-group and sex are used. Smoothed series of urban/rural distribution and the proportion of single person households are generated from the 1996 and 2001 Census points. See Appendix B Table b.1 for a detailed description of the restrictions.

The estimation problem is therefore to estimate a new set of sampling probabilities (person weights) which are as close as possible to a prior set of sampling probabilities given by the StatsSA person weights, while satisfying the moment constraints from the aggregate data.

Consider a survey sample of K individuals with prior to adjustment probabilities q_k, i.e. the initial person weights converted into proportions. Each individual has a vector of x_k observed characteristics, age group, province by urban/rural, sex and whether they live in a single person household. The ASSA model has aggregate population information about the province, age group and sex distributions. The Census 1996 and 2001 provide information about the urban/rural distribution in each province and the proportion of single person households. Pre, intervening and post-census year information was calculated using exponential interpolation and extrapolation.

We minimize the CE measure[18] of the distance between the new sampling probabilities p_k and the prior distribution q_k (Golan, Judge, & Miller, 1996)

$$\underset{p_k}{Min}\, I(p,q) = \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k} \right) \right)$$

$$= \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln p_k - \sum_{k=1}^{K} p_k \ln q_k \right)$$

subject to the moment consistency constraints

$$\sum_{k=1}^{K} p_k x_t = y_t \qquad t \in [1,\ldots,T]$$

and adding-up normalization constraint

$$\sum_{k=1}^{K} p_k = 1$$

Each x_t is a person level indicator, indicating which strata and age group the individual is in, the individual's sex and whether they live in a single person household. T represents the total number of restrictions. In our case T=36, (18-1) strata, (18-1) age groups (2-1) sexes and (2-1) household types (single or other). As the restrictions cover the complete dataset, i.e. each person is in an age group, of a particular sex, in a certain strata and either from a single person household or not, one category from each restriction class had to be left off to avoid linear dependencies.

[18] See Golan *et al* (1996)for the formulation of the ME problem

Cases with missing information on one or more of the restriction variables were not included in the calibration. Appendix C presents a table with the number of cases not included in each year due to missing data on the restriction variables. *K,* the number of people in the sample, is very large[19], while *T,* the number of constraints, is small. Thus there are not enough degrees of freedom to support a unique solution using a classical estimation procedure such as Ordinary Least Squares. A CE approach is therefore used.

The new probability weights are estimated as follows (Golan, Judge, & Miller, 1996):

$$\underset{p_k}{Min}\, L = \underset{p_k}{Min}\left(\sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k}\right) + \sum_{t=1}^{T} \lambda_t \left(y_t - \sum_{k=1}^{K} p_k x_k \right) + \mu\left(1 - \sum_{k=1}^{K} p_k\right)\right)$$

The first-order conditions are:

$$\frac{dL}{dp_k} = \ln p_k - \ln q_k + 1 - \sum_{t=1}^{T} \lambda_t x_t - \mu = 0 \qquad k \in [1,\ldots,K]$$

$$\frac{dL}{d\lambda_t} = y_t - \sum_{k=1}^{K} p_k x_k = 0 \qquad t \in [1,\ldots,T]$$

$$\frac{dL}{d\mu} = 1 - \sum_{k=1}^{K} p_k = 0$$

The solution to which can be written as (Golan, Judge, & Miller, 1996):

$$\tilde{p}_k = \frac{q_k}{\Omega\left(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_T\right)} \exp\left[\sum_{t=1}^{T} \tilde{\lambda}_t x_k\right] \qquad k \in [1,\ldots,K]$$

where

$$\Omega\left(\tilde{\lambda}\right) = \sum_{k=1}^{K} q_k \exp\left[\sum_{t=1}^{T} \tilde{\lambda}_t x_k\right]$$

1

[19] See appendix C

The estimation problem as specified above has no closed form solution. However, the unconstrained dual approach initially formulated by Agmon *et al* (1979) and later generalised by Miller (1994) and Golan *et al* (1996) presents a simple solution.

The dual objective as a function of the Lagrange multipliers λ_t is:

$$\begin{aligned} L(\lambda) &= \sum_{k=1}^{K} p_k \ln\left(\frac{p_k}{q_k}\right) + \sum_{t=1}^{T} \lambda_t \left(y_t - \sum_{k=1}^{K} p_k x_k\right) \\ &= \sum_{k=1}^{K} p_k(\lambda)\left[\sum_{t=1}^{T} \lambda_t x_{tk} - \ln\left(\Omega(\lambda)\right)\right] + \sum_{t=1}^{T} \lambda_t \left(y_t - \sum_{k=1}^{K} p_k x_k\right) \\ &= \sum_{t=1}^{T} \lambda_t y_t - \ln\left(\Omega(\lambda)\right) \equiv M(\lambda) \end{aligned}$$

The adding up constraint is satisfied in the optimal $\mathbf{p}(\lambda)$.

$M(\lambda)$ is just equation 1 with $\tilde{\lambda}$ substituted for λ, thus by maximising $M(\lambda)$ with respect to λ, we get $\tilde{\lambda}$ and hence the solution to our problem, $\tilde{p}_k$. The new person weights are calculated by means of the above formulation using Wittenberg's (unpublished) dual CE Stata programme. The dual CE is programmed using the Stata Maximum Likelihood (ml) macro.

The optimal approach to generate a new set of weights would be to calculate household entropy weights and assign a common weight to each person within the household. This could theoretically be achieved by including a restriction under the moment consistency constraints which restricts the person weight to be common within household during the estimation process. One formulation of this restriction, illustrated for the two household case would be:

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} p_{11} \\ p_{21} \\ p_{31} \\ p_{12} \\ p_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Where p_{ij} represents person *i's* weight in household *j* and the 3x5 matrix is part of the restriction matrix previously called x_k. Since $p_{1j} - p_{ij} = 0$ for all with $i = 2,...,n_j$ and *j*, where n_j is the size of household *j*, the person weights are restricted to be equal within household. This computation requires an additional (K-H) restrictions where K is the number of people in the sample and H is the number of households in the sample and is therefore very computationally intensive. Having experimented with this procedure and found it not to be feasible with the current Stata ml formulation, the household entropy weight is set to be equal to the mean entropy person weight within household.

This post-stratification approach therefore deals with some of the concerns posed about the survey series. Most importantly, it allows us to adjust the sample to meet aggregate trends which appear realistic over time while simultaneously diverging as little as possible from the StatsSA weights which contain important information with regards to the sample design. In addition, since the entropy approach adjusts to marginal totals, different data sources (here the Census and the ASSA model) can be used as external benchmarks in unison. This gives greater flexibility to the post-stratification adjustment procedure. The use of marginal totals has the added benefit of avoiding the small or empty cell problem. Finally, the functional form of the entropy problem guarantees positive weights.

5. Results

In evaluating the validity of the new entropy weights three areas of interest are assessed. First, does the entropy estimation procedure generate weights that meet our expectations, i.e. do the weights meet the external restrictions and are the CE-weights similar to the prior weight distribution? Second, are the entropy household weighted data consistent with the person weighted data? Finally, how realistic is the trend in other key aggregates (aggregates not used as restrictions) over time? Noting the validity of the entropy weights in all areas concerned, the impact of the new weights on a simple employment status analysis is investigated.

5.1 A look at the New Weights

Table 2a shows the distribution of the three sets of weights; the original StatsSA weights, the maximum entropy (ME) weights and the cross entropy (CE) weights. It is clear that both the entropy weight distributions are very similar to the original StatsSA person weight distribution. This is especially true for the CE-weights. Table 2b presents a regression of the new weights on the original weights. The CE-weight model has a strong fit and shows that the original StatsSA weight and the new weight are very similar. This is in line with expectation given that the StatsSA weights are included as prior information (to be met as closely as possible given the restrictions) in the estimation of the cross entropy weights. The mean entropy weight increases up until 2003, after which it decreases. This is because the ASSA population totals are greater than the population totals calculated using the original StatsSA weights between 1995 and 2002 and lower thereafter.

The ME-weight model has a far worse fit. This is expected since the ME-weight distribution is only affected by the restrictions in the calculation and not the prior weight distribution. Appendix D presents the regression including all the restrictions used in the entropy estimation as controls. The model fit is very good once the restriction variables are added and it is clear that the ME-weight distinguishes people primarily on the basis of strata, sex, age group and single versus non-single person households. While the coefficient on the StatsSA weight is highly significant, it is small.

Table 3 displays some aggregate results calculated using the different weights. The estimation procedure results in totals which match the aggregates from the ASSA model and Census data (See Appendix B table b.1 for restriction values). Table 3 includes the number of households and the proportion of single person households[20], both variables not used as restrictions in the estimation procedure, to illustrate that two independent trends are not distorted by the use of the entropy weights but rather show more realistic trends.

Tables 2 and 3 illustrate that the entropy estimation procedure appears accurate. The restrictions are met in both the CE and ME cases. The weight distribution for the CE-weights approximates the prior distribution given by the StatsSA person weight, preserving the benefits of the sample design, while simultaneously meeting the external aggregates. The entropy weights result in demographically and geographically consistent trends between 1995 and 2004.

[20] The proportion of people in single person households is used as a constraint, not the proportion of households

Table 2a: Comparing different weight distributions

		Original StatsSA Person Weight (prior)				Maximum Entropy Weight				Cross Entropy Weight			
	Sample size	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
1995	130787	303.74	149.83	0.07	1759.65	313.36	127.37	42.68	925.15	313.36	160.04	0.07	1826.34
1996	72889	556.77	290.47	62.00	6053.00	574.60	146.48	110.77	1761.45	574.60	316.28	52.73	8219.42
1997	140015	295.99	125.99	42.00	1834.00	304.92	92.70	53.47	1141.96	304.93	139.12	23.77	2558.38
1998	82263	513.94	251.36	47.66	2629.73	528.33	176.40	75.72	1469.75	528.35	269.26	46.38	2459.60
1999	106424	406.14	214.59	12.71	2387.87	415.15	126.79	77.99	830.51	415.15	223.43	8.66	2640.16
2000	105242	417.86	229.93	47.15	1667.20	426.16	132.75	73.07	790.23	426.16	236.85	26.91	1828.61
2001	106300	419.63	235.54	53.88	1367.22	427.77	136.58	68.97	863.00	427.77	239.16	29.78	1741.44
2002	102334	445.19	246.99	53.58	1396.32	449.86	159.59	59.70	1060.19	449.86	247.20	22.47	1752.79
2003	98695	472.74	241.98	5.37	3434.45	471.48	171.71	53.85	996.26	471.48	243.06	3.36	2576.55
2004	98174	480.29	387.66	6.65	21533.18	478.40	172.64	56.31	1051.92	478.40	386.03	3.99	23477.04

Table 2b: Comparing different weight distributions

Maximum entropy weight

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
StatsSA weight	0.620***	0.175***	0.415***	0.411***	0.303***	0.309***	0.316***	0.387***	0.449***	0.181***
	[0.0016]	[0.0018]	[0.0016]	[0.0020]	[0.0016]	[0.0015]	[0.0015]	[0.0016]	[0.0017]	[0.0013]
Constant	125.2***	477.2***	182.0***	316.9***	292.2***	296.9***	295.1***	277.8***	259.1***	391.5***
	[0.55]	[1.10]	[0.52]	[1.13]	[0.71]	[0.72]	[0.72]	[0.82]	[0.93]	[0.80]
Observations	130787	72889	140013	82261	106424	105242	106300	102334	98695	98174
R-squared	0.53	0.12	0.32	0.34	0.26	0.29	0.30	0.36	0.40	0.16

Cross entropy weight

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
StatsSA weight	1.023***	1.054***	1.059***	1.040***	1.013***	1.007***	0.982***	0.949***	0.960***	0.973***
	[0.00085]	[0.0010]	[0.00084]	[0.00090]	[0.00074]	[0.00066]	[0.00079]	[0.00099]	[0.00094]	[0.00068]
Constant	2.555***	-12.38***	-8.535***	-6.064***	3.908***	5.242***	15.80***	27.39***	17.69***	11.08***
	[0.29]	[0.63]	[0.27]	[0.51]	[0.34]	[0.32]	[0.38]	[0.51]	[0.50]	[0.42]
Observations	130787	72889	140013	82261	106424	105242	106300	102334	98695	98174
R-squared	0.92	0.94	0.92	0.94	0.95	0.96	0.93	0.90	0.91	0.95

Standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1

Table 3: Selected aggregate results

Some Person File Results

		StatsSA person weight (Prior)	Maximum Entropy weight	Cross entropy weight
Population**	*1995*	39725179	40982879	40982879
	1996	40582538	41882359	41882359
	1997	41443101	42693725	42693725
	1998	42276946	43462350	43462350
	1999	43223018	44182313	44182313
	2000	43976533	44850114	44850114
	2001	44606761	45472430	45472430
	2002	45606383	46035622	46035622
	2003	46657408	46532612	46532612
	2004	47152334	46966648	46966648
Number of households	*1995*	8927190	9451628	9398360
	1996	9065041	9688253	9781228
	1997	9151934	9728271	9878893
	1998	9881735	10315931	10424111
	1999	10740554	10770615	10903595
	2000	11264763	11113775	11322202
	2001	11326814	11451587	11547375
	2002	11672293	11774650	11840755
	2003	12660109	12294181	12284284
	2004	12974226	12079599	12401814
Share of single person households*	*1995*	11.94	14.35	14.43
	1996	9.54	15.17	15.03
	1997	10.56	16.37	16.12
	1998	13.80	16.68	16.51
	1999	18.32	17.27	17.06
	2000	19.83	18.04	17.71
	2001	19.70	18.86	18.71
	2002	20.33	19.67	19.56
	2003	22.37	19.11	19.13
	2004	23.26	19.63	19.12
Share Urban**	*1995*	51.11	53.04	53.04
	1996	53.66	53.73	53.73
	1997	54.18	54.51	54.51
	1998	54.08	55.28	55.28
	1999	53.87	56.04	56.05
	2000	55.20	56.80	56.80
	2001	54.39	57.54	57.54
	2002	53.17	58.21	58.21
	2003	54.75	58.29	58.29
	2004	54.99	58.34	58.34
Share Male **	*1995*	48.01	48.56	48.56
	1996	48.06	48.53	48.53
	1997	48.19	48.50	48.50
	1998	48.27	48.48	48.48
	1999	48.37	48.46	48.46
	2000	48.05	48.44	48.44
	2001	48.06	48.43	48.43
	2002	48.14	48.42	48.42
	2003	47.60	48.41	48.41
	2004	47.65	48.40	48.40

**used as a restriction

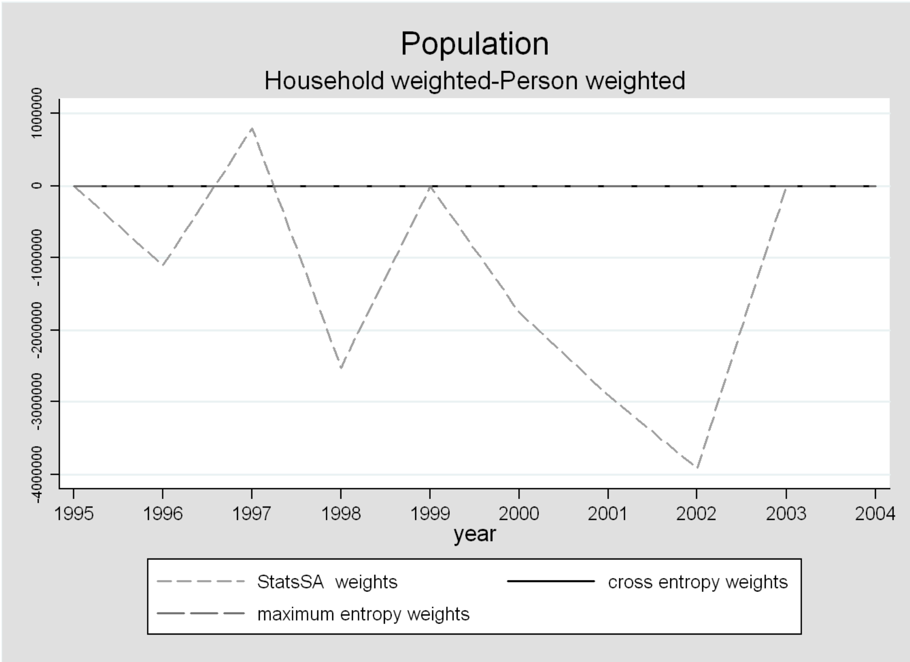
*the proportion of people living in single person households was used as a restriction i.e. at the person level, not the share of single person households

5.2 Internal consistency

One of our concerns was that aggregates calculated at the household and person level were inconsistent with each other when the original StatsSA person and household weights were used. While the most ideal approach would be to restrict the person weight to be common within a household during the entropy estimation, this requires many additional restrictions and hence computational time and was not feasible with the present Stata entropy estimation procedure. As an alternative, the mean person weight within a household was assigned to each person in the household. Thus no explicit restrictions were included in the estimation procedure.

The method used to calculate the entropy household weights ensures a consistency between the household and person level files. Figures 7 and 8 illustrate the increased internal consistency when the entropy weights are used compared to the original StatsSA weights. The figures plot the difference in the population and the number of households when the person weight versus the household weight was used. Each graph presents this difference for the original StatsSA weights, the ME-weights and the CE-weights. In each case, when the value is calculated at the person level, household weighted implies that the household weight is assigned to each person in the household, while person weighted uses the individual person weights. When measurement is at the household level, person weighted signifies that the mean person weight within household is assigned to the household.

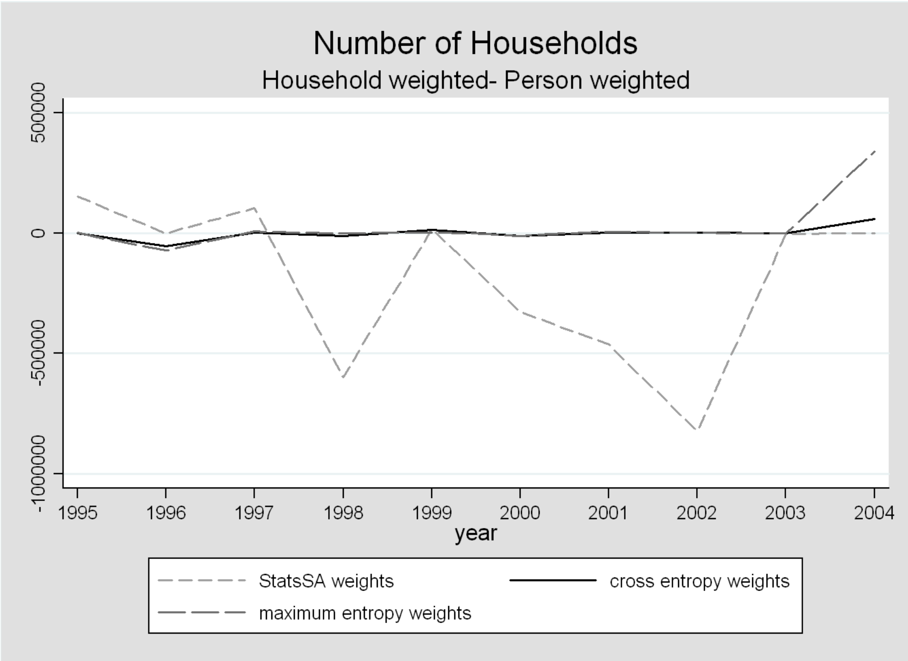
Figure 7: Population: difference between the person and household weights



Notes: *The entropy household and person weights result the same population numbers by design. The figure illustrates the improved consistency over the original StatsSA weights.*

Figure 7 presents the difference in the population as measured by the person and household weights within the person file. The difference between the population count using the original StatsSA person weights versus the original StatsSA household weights is large in most years. The entropy household and person weights result in a consistent series at the population level by design which is beneficial to the original series when household and person level data are being used simultaneously in an analysis. A similar picture is observed in figure 8 for differences in the number of households. There are very small differences between the household and person entropy weights with the exception of 1996 and 2004 while the StatsSA weights show large divergences between household and person weights in most years.

Figure 8: Households: differences between person and household weights



Notes: *The figure illustrates the improved consistency between the household and person level data when the entropy weights are used.*

The entropy weights show far greater consistency between person and household weighted analyses[21] than the original StatsSA weights, with the exception of 2003 and 2004. 2003 marked StatsSA's introduction of CALMAR2 as the post-stratification procedure used in calibrating the design weights. The main advantage of CALMAR2 over CALMAR is that it ensures consistency between household and person level data. This is evident in figures 7 and 8 for 2003 and 2004.

5.3 Consistency over time

We have established that the entropy weights meet the external restrictions and since the ASSA model produces consistent estimates over time and the Census data points were exponentially smoothed, the new weights generate consistent estimates with respect to the restrictions by default. It is however, important to assess whether other variables not used as restrictions are consistent over time.

[21] Similar results were found for the trend in the proportion of single person households and the share of the population living in urban areas

Figure 9: Household numbers over time

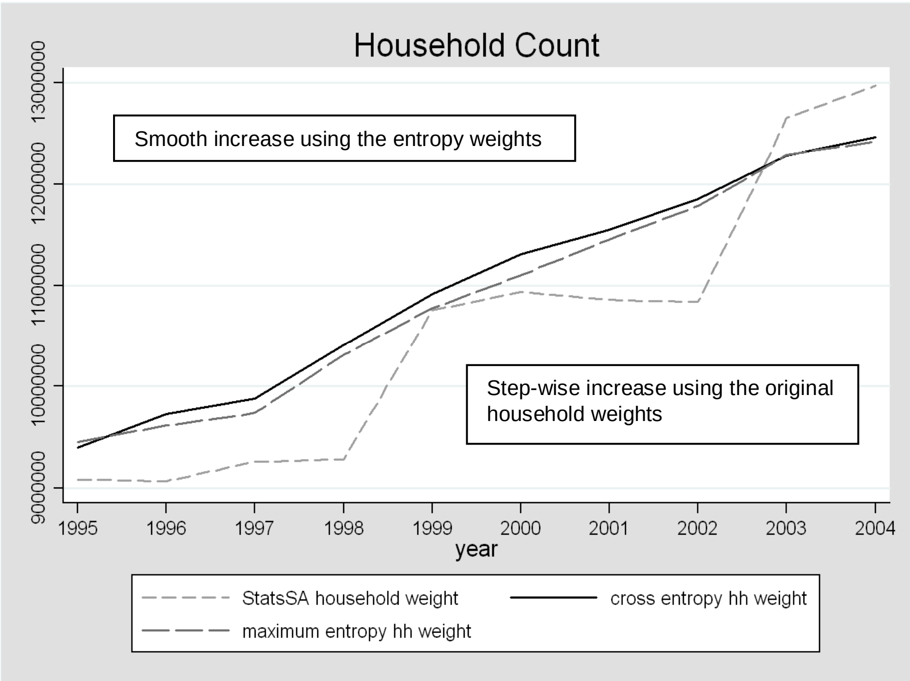
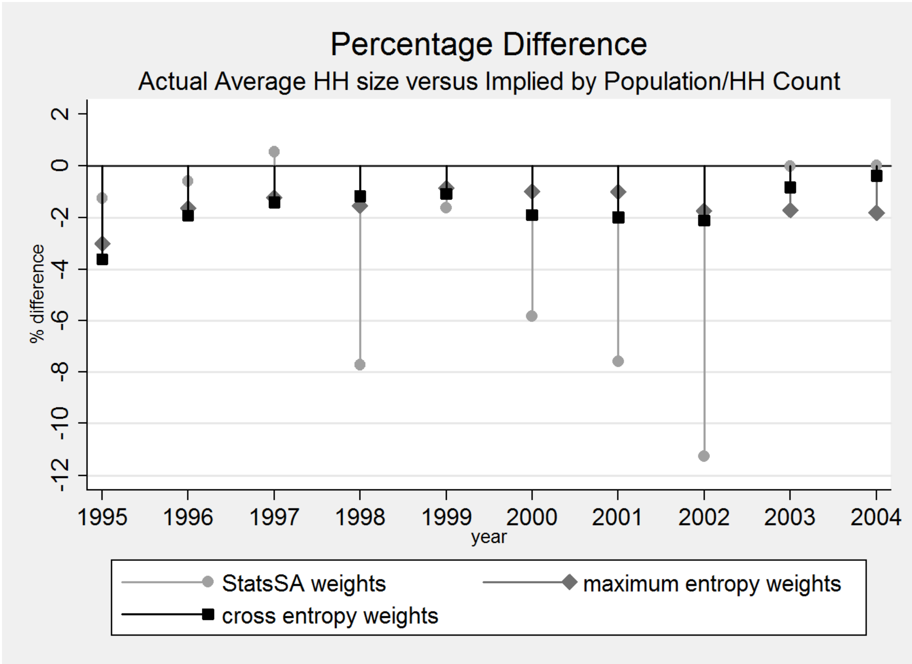


Figure 9 plots the number of households in each year between 1995 and 2004 calculated using the original StatsSA household weight, the maximum entropy household weight and the cross entropy household weight. The trend shows a fairly constant increase when the entropy weights are used. This is not only more realistic than the stepwise function evident when the original StatsSA household weights are used, but also creates consistency between the person and household files. In figure 10 the average household size calculated in the person file is compared with the implied average household size when the population is divided by the number of households calculated using the household weights. While the inconsistency between these two measures is marginally increased in 1995-1997 and 2003 and 2004 when the entropy weights are used, the overall effect of the entropy weights is increased consistency trend over the ten year period.

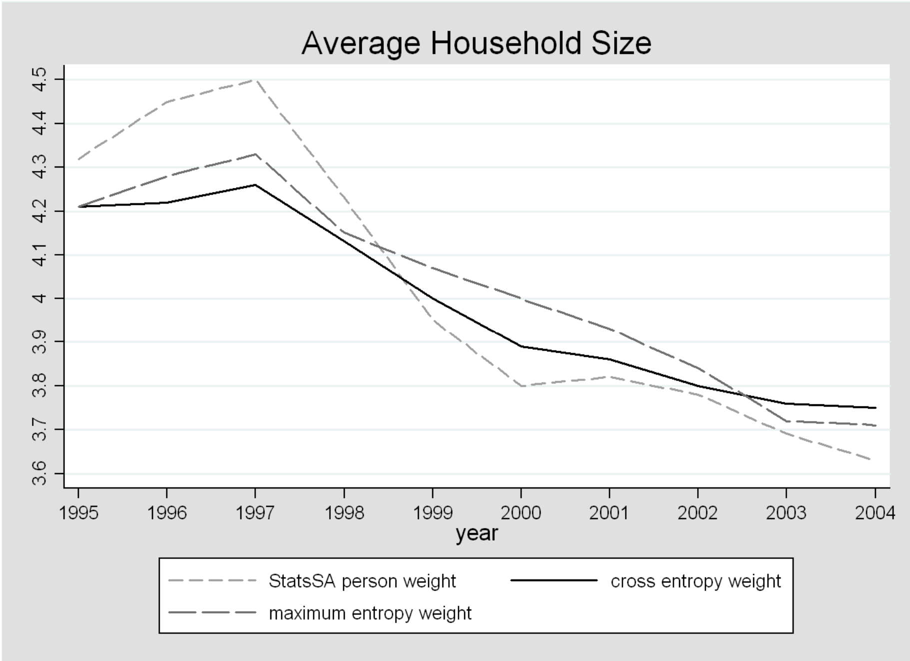
Figure 10: Aggregate trend consistency



Notes: *The figure illustrates the increased internal consistency between the person and household level file.*

Figure 11 shows the trend in average household size. The entropy weights, especially the CE-weights, result in a more realistic trend. The CE-weights show a relatively constant decline in average household size over the ten-year period. The impact of increasing the proportion of single person households in the earlier years and reducing them in later years mediates the large decrease in average household size between 1997 and 2001 observed when using the original StatsSA weights.

Figure 11: Consistency average household size trends



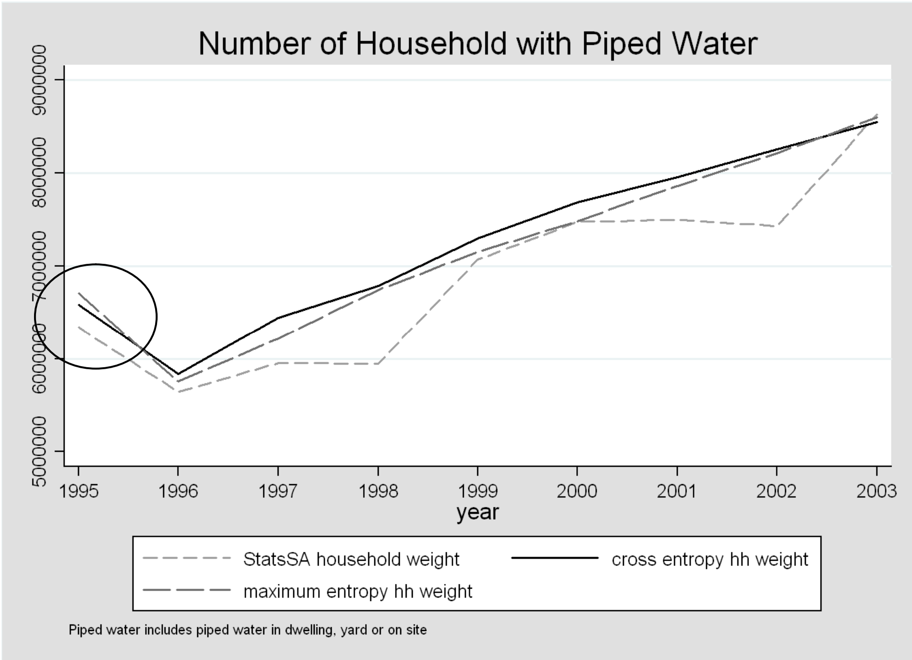
Notes: *The entropy weights result in a more realistic decrease in average household size, mediating the large decrease between 1997 and 2000.*

Figure 12 presents the trend in the number of people with piped water[22]. The original StatsSA household weight creates a trend which is unlikely, while the entropy weights produce trends which are smoother over time. 1995 appears to be an outlier finding far too many households with piped water. This points to 1995 being different from the other years as has been discussed in the literature.

The entropy weights show strong consistency of demographic and geographic variables, both internally between the household and person files, and over time. In addition, the trend in the number of households with piped water, an indicator of service delivery, is more realistic.

[22] Piped water includes piped water in dwelling, yard or on site

Figure 12: Number of households with piped water



Notes: *The original household weights result in a step-wise increase in the number of households with piped water. The entropy weighted trend is smooth. 1995 represents an outlier, finding far more households with piped water than in subsequent years*

We conclude that the entropy weights, especially the CE-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 in each year and therefore produce time trends in demographic, geographic and other variables which are more realistic. At the same time, the CE-weights are similar to the original weights and therefore preserve the benefits of the original sample design. Second, the household and person entropy weights are more internally consistent and therefore enable analyses that combine household and person level data. Finally, if the variable of interest in a proportionate analysis is affected by the over or under representation of the demographic or geographic variables used as a restriction in the re-weighting procedure, then the new weights will affect this analysis as well.

Our final area of interest is to assess the sensitivity of the labour market variables to the new weights. In particular, do the new demographically and geographically consistent weights reduce the large year-on-year shifts observed in labour force variables? For instance, are the large increases in economic activity between 1997 and 2000 reduced and is the level of employment in 1995 for males and 2000 for females more in line with the overall trend? No significant mediation of these effects is found and therefore the conclusion that the observed shifts are a result of shifts in measurement of the labour force variables, in other words measurement changes, and not a result of once-off shifts in the survey weights is drawn.

Figure 13: The trend in the economically active population

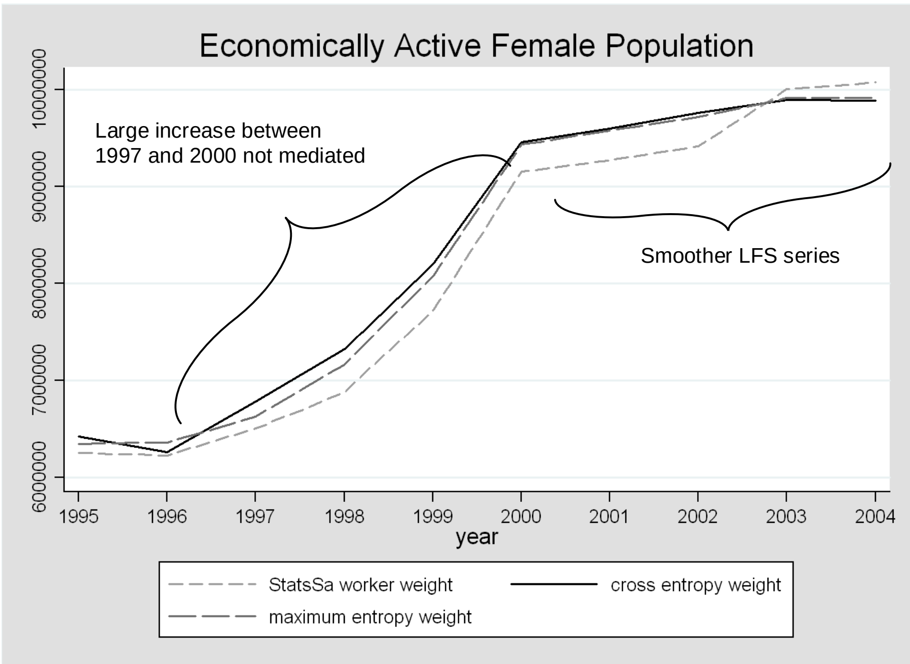
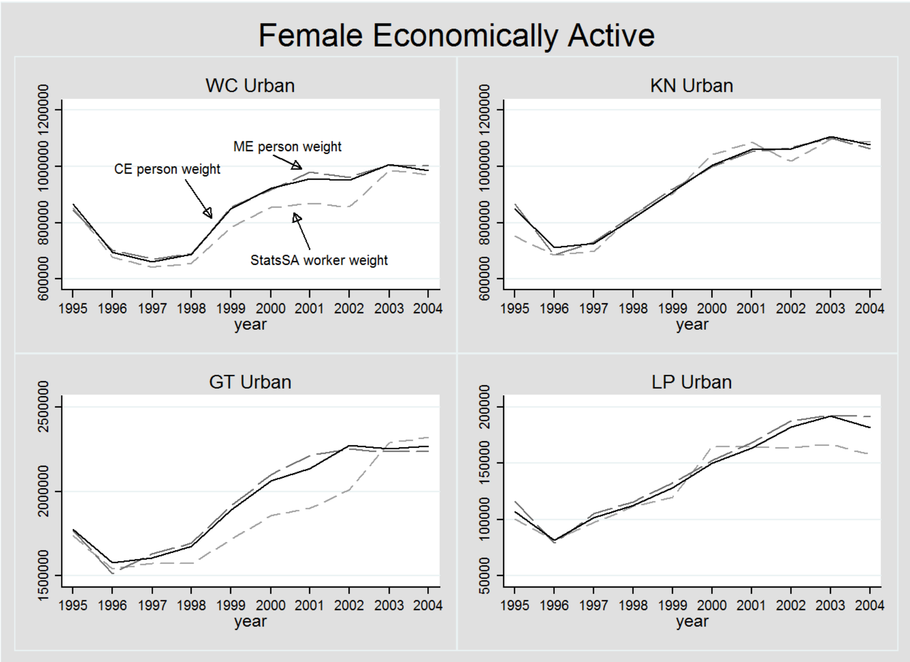


Figure 13 and 14 display the trends in the number of females[23] economically active in the population and at the province level for four large provinces, using the three different sets of weights. Figure 13 illustrates that the large increase in economic activity between 1997 and 2000 is not reduced at the population level through the use

[23] See Appendix E for male economically active

or the entropy weights. In fact, the overall increase between 1997 and 2000 is increased slightly when the cross entropy weights are used.

Figure 14: The trend in female economic activity by province



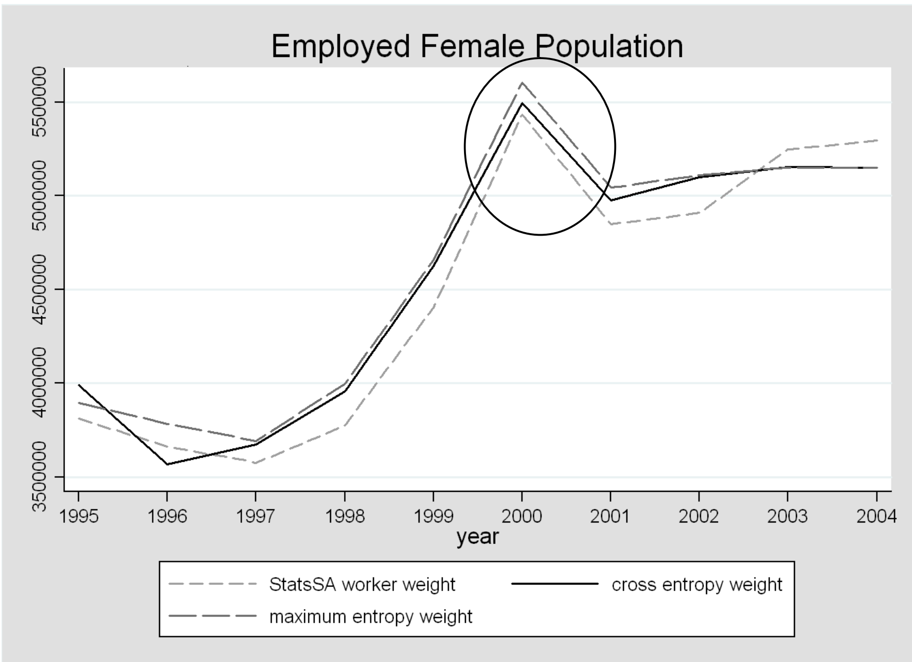
Notes: *The entropy weighted trend in the number of females who are economically active is smoother. This illustrates that while aggregate trends may not be significantly affected by the new weights, analyses at a less aggregated level are likely to see changes.*

Examining Figure 14, however, there are some noticeable shifts in the trend at the province level when the entropy weights are used. The entropy weighted trends are far smoother than when the original StatsSA weight is used. In Gauteng and the Western Cape, the increase in female economic activity starts and levels off sooner when the entropy weights are used. In the Limpopo province, the increase continues until 2003. Thus while the entropy weights have no major effect at the aggregate level, analyses at the province level are likely to be affected.

What is noticeable in both Figure 13 and 14 is that the upward shift in 2003 (a consequence of the change in sampling frame) can be smoothed by using consistent weights. The female population trend in economic activity between 2000 and 2004 has a smooth moderate gradient. This trend appears realistic. This points towards the conclusion that the LFS's have a similar approach to measuring economic activity; when weighted with consistently calculated weights the series is consistent.

Figure 15 shows the trend in the number of females employed between 1995 and 2004. The figure shows that the large number found employed in 2000 is not reduced when the entropy weights are used. This signals that the 2000 LFS measured employment differently (a point noted in the literature). If the 2000 point is removed, the CE-weights create a consistent trend. A 'growth shape' curve is evident; from 1996 the number of employed females increases at an increasing rate and from 2001 onwards, the growth slows. 1995 is still a clear outlier.

Figure 15: The number of females employed over time



Notes: *The 2000 spike in female employment is not diminished by the entropy weights. This signals that employment was measured differently in 2000 and the spike is not a result of a shift in the survey weights.*

Analysis of the proportion of the population in each state tells a similar story. Although there are small changes, the entropy weights have no significant effect in creating a more consistent trend in the labour market variables between 1995 and 2004. In other words, the large inconsistencies in the labour market variables are not a result of shifts in the weights.

The insignificant changes observed in the trends in employment status indicate that the increase in economic activity and the high employment levels found in 1995 and 2000 are unlikely to be a function of incorrect weights caused by post-stratification errors. Therefore by default, these results give further importance to the argument that the shifts are a function of the increased effort over time to find economic activity, in other words, are either real or a result of measurement error.

6. Conclusion

OHS and LFS 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. In this paper ten years of national household survey data between 1995 and 2004 are re-weighted to a consistent series of benchmarks from the ASSA model and Census data.

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 aggregate trends. 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 and the Census data. When comparing different years of the LFS and OHS as a time series, results will be more realistic if the benchmarks are consistent over time and if the post-stratification is based on a consistent post-stratification adjustment in each year. In other words, working with data calibrated in a similar manner on 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 and household weights are designed to show far more internal consistency. This is important for analyses where both person and household level variables are used. Up until 2003 the StatsSA household weights were not adjusted and as a result the variable in the household files produce erratic trends over time and should not be used as a series.

Finally, some variables will be affected by the weights. This is illustrated in Figure 14, where the use of the entropy weights at the province level affects the trend in economically active females quite noticeably. If, for instance, the spike in employment in 2000 was the result of the 2000 StatsSA weights over representing provinces that had high levels of employment, then by adjusting the weights to meet a series of consistent aggregates this spike would be reduced. The fact that the aggregate employment status analysis is not significantly affected just signals that this variable is not sensitive to the weights, which is reassuring.

The following two extensions would benefit the analysis. While the ASSA model and Census data produce consistent aggregates over time, these data are themselves imperfect measures of the true population. Thus just as the StatsSA mid-year estimates introduce error through their inaccuracies, any other benchmark used will introduce a certain level of error. The accuracy of the weights could be further improved by allowing for measurement error in the aggregate data. The generalised cross entropy framework allows for this extension. Second, while the household weights produced trends which were consistent with the person level data, restricting the person weights to be common within household would be more theoretically sound.

7. References

Ardington, C., Leibbrandt, M., Lam, D., & 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.ProvOutput, Version 051129* .

Bethlehem, J., & Wouter, J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics, Volume 3* , 141-153.

Bhorat, H., & 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., & Wittenberg, M. (2007). The Measurement if Economic Status in South Africa using Cohort Analysis, 1994-2004. *South African Journal of Economics, Volume 72:2* , 313-326(14).

Branson, N., & Wittenberg, M. (2007). The Measurement of Employment Status in South Africa using Cohort Analysis, 1994-2004. *South African Journal of Economics, Volume 75:2* .

Burger, R., & Yu, D. (2006). Wage trends in post-apartheid South Africa: Constructing an earnings series from household survey data. *Stellenbosch Economic Working Papers: 10/06* .

Casale, D., Muller, C., & 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, Volume 72:5* , 978-1002.

Casale, D., Muller, C., & Posel, D. (2004). Two million net new jobs: a reconsideration of the rise in employment in South Africa, 1995-2003. *African Development and Poverty Reduction: The Macro-Micro Linkage DPRU& TIPS* .

Cronje, M., & Budlender, D. (2004). Comparing Census 1996 and Census 2001. *South African Jounral of Demography, Volume 9:1* , 67-89.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric approach to development policy.* Published for the World Bank The John Hopkins University Press: Baltimore and London.

Deville, J.-C., & Sarndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *journal of the American Statistical Association, Volume 87:418* , 376-382.

Deville, J.-C., Sarndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of American Statistical Association, Volume 88:423* , 1013-1020.

Dorrington, R., & Kramer, S. (2007). The 2004 mid-year estimates: Method, Reliability and Implication. *Unpublished* .

Fraser, I. (2000). An Application of Maximum Entropy Estimation: The Demand for Meat in the United Kingdom. *Applied Economics, Volume 32:1* , 45-59.

Golan, A., Judge, G., & Miller, D. (1996). *Maximum Entropy Economics, Robust Estimation with Limited Data.* West Sussex, England: John Wiley and Sons Ltd.

Holt, D., & Smith, T. (1979). Post Stratification. *Journal of the Royal Statistical Society, Volume 142* , 33-46.

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

Keswell, M., & Poswell, L. (2004). Returns to education in South Africa: A retrospective sensitivity analysis of the avaiable evidence. *The South African Journal of Economics, Volume 72: 4* , 834-860.

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

Lemaitre, G., & Dufour, J. (1987). An Integrated Method for Weighting Persons and Families. *Econometrica 48* , 1333-1346.

Little, R. (1993). Post-stratification: a modeler's Perspective. *Journal of the American Statistical Association, Volume 88* , 1001-1012.

Merz, J. (1994). Microdata Adjustment by the Minimum Information Loss Principle. *FFB Discussion Paper, No 10* .

Merz, J., & Stolze, H. (2006). Representative Time Use Data and Calibration of the American Time Use Studies 1965-1999. *FFB Discussion Paper No 54* .

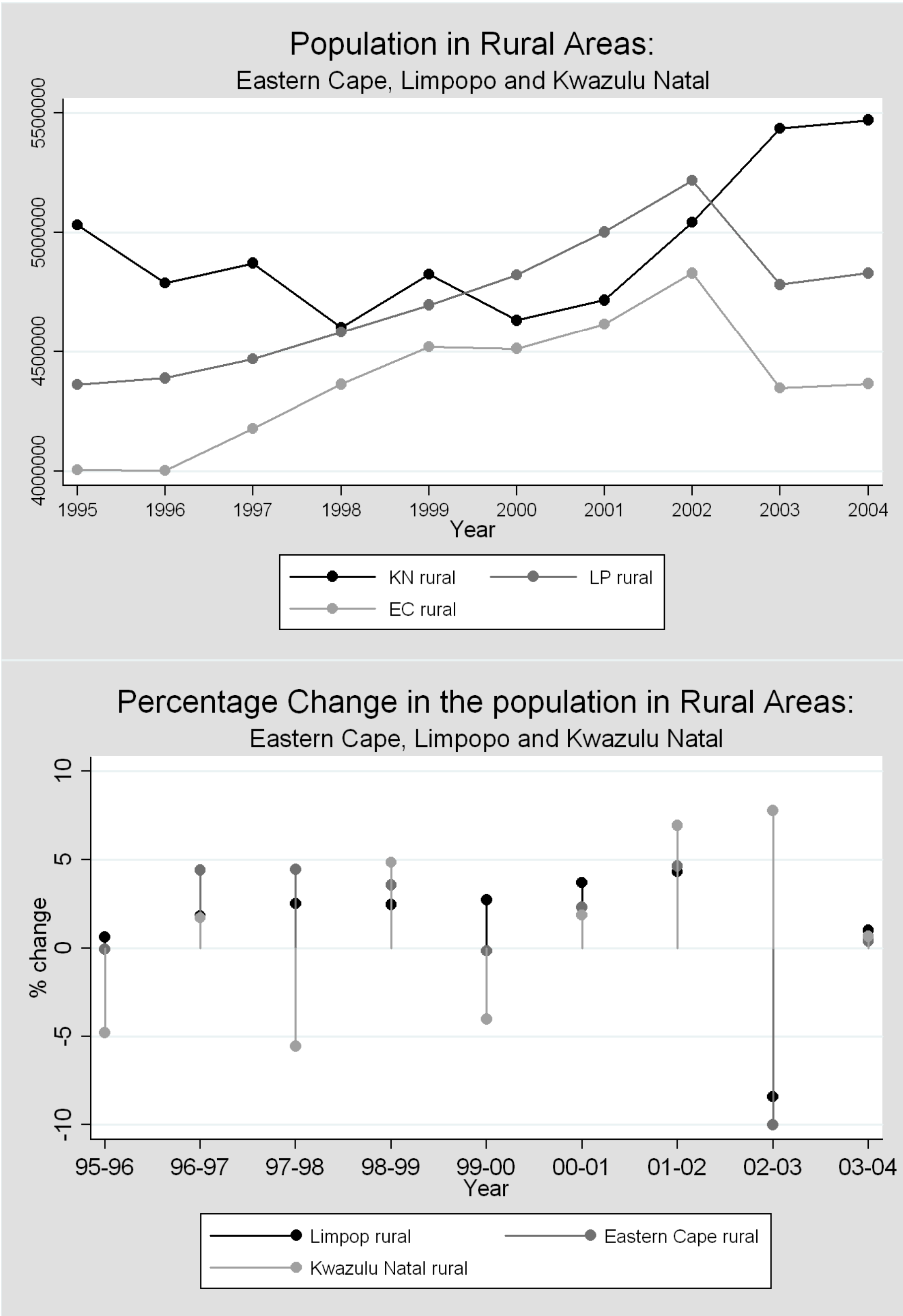
Muller, C. (2003). Measuring South Africa's Informal Sector: An Analysis of National Household Surveys. *Development Policy Resarch Unit, Working Paper 03/71* .

Neethling, A., & Galpin, J. (2006). Weighting of Household Survey Dara: A Comparison of Various Calibration, Intergrated amd Cosmetic Estimators. *South African Statistics Jounal, Vol 40:2* , 123-150.

Posel, D., & Casale, D. (2003). What has been hapening to internal labour migration in South Africa, 1993-1999. *The South African Jounal of Economics, Volume 71:3* , 455-479.
Robilliard, A., & Robinson, S. (2003). Reconciling Household Surveys and National Accounts Data using a Cross Entropy Estimation Method. *Review of Income and Wealth, Volume 49:3* .
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, Volume 40:3* , 315-323.
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, Volume 35:3* , 130-137.
Zhang, L.-C. (2000). Post-stratification and Calibration-A Synthesis. *The American Statistician, Volume 54:3* , 178-184.

Appendix A

Figure A.1: Population in Rural Areas



Notes: *Large positive changes in Kwazulu Natal observed between 2001 and 2003. Large negative population changes in Limpopo and the Eastern Cape. The figure illustrates that unconditional usage of the OHS and LFS as a series can result in erroneous conclusions.*

Appendix B

Table b.1: Restriction values

			Restriction values									
			1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Stratum	1	WC urban	3539316	3652860	3759760	3865689	3970256	4073144	4174290	4267739	4341031	4408106
	2	WC rural	456215	456094	454483	452398	449830	446783	443287	439007	446787	453936
	3	EC urban	2297616	2365387	2409723	2450618	2487964	2521744	2552627	2586911	2594082	2602719
	4	EC rural	4054037	4097419	4096732	4088919	4074171	4052831	4026309	4005891	4018251	4032888
	5	NC urban	535361	580784	610649	638887	665305	689768	712258	733770	738317	742602
	6	NC rural	262695	247724	225872	204933	185066	166390	148997	133429	134575	135678
	7	FS urban	1832503	1888013	1942293	1992941	2039466	2081555	2119534	2156189	2156197	2155429
	8	FS rural	900369	864193	827234	789798	752049	714210	676685	641302	642075	642618
	9	KN urban	3698899	3812432	3932023	4047751	4158528	4263517	4363007	4458347	4491334	4517478
	10	KN rural	4997321	5033117	5070505	5098579	5116527	5123936	5121791	5114227	5154083	5186116
	11	NW urban	1118821	1192861	1257659	1323339	1389490	1455770	1522211	1588023	1600014	1610129
	12	NW rural	2210521	2225079	2212670	2195954	2174732	2149024	2119443	2087493	2105307	2120682
	13	GT urban	7079933	7316076	7598172	7878565	8155240	8426603	8692848	8910939	9084371	9221557
	14	GT rural	222048	226270	231679	236839	241697	246216	250411	253132	258119	262080
	15	MP urban	1117633	1144872	1181342	1216782	1250933	1283629	1315050	1344928	1358765	1370825
	16	MP rural	1772383	1783189	1806616	1827058	1844264	1858138	1869090	1877453	1897348	1914772
	17	LP urban	517597	549559	580233	611954	644709	678506	713491	750156	760048	770198
	*	LP rural	4369609	4446430	4496079	4541343	4582084	4618350	4651101	4686686	4751907	4818836
			40982880	*41882359*	*42693725*	*43462350*	*44182312*	*44850113*	*45472429*	*46035621*	*46532612*	*46966647*
Sex	18	male	0.4856	0.4853	0.4850	0.4848	0.4846	0.4844	0.4843	0.4842	0.4841	0.4840
	*	female	0.5144	0.5147	0.5150	0.5152	0.5154	0.5156	0.5157	0.5158	0.5159	0.5160

*reference category
source: ASSA 2003 , Census 1996 & 2001

Table B.1 continued

			Constraint values									
			1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Age Group	19	age 0-4	5063796	5039164	5123569	5189283	5238755	5272192	5287237	5245938	5211475	5184446
	20	age 5-9	4855288	4845834	4856881	4865236	4870194	4872996	4877406	4943605	5001985	5050253
	21	age 10-14	4811282	4879486	4910864	4921330	4917230	4908146	4901908	4890365	4874018	4857624
	22	age 15-19	4201287	4306400	4418825	4553112	4690447	4808169	4889759	4934355	4948847	4941035
	23	age 20-24	3919076	4050191	4113711	4157958	4196719	4249269	4327426	4428836	4546954	4666841
	24	age 25-29	3429162	3538698	3627188	3730953	3837762	3932242	4005345	4048203	4068200	4080557
	25	age 30-34	3139952	3232844	3278756	3310758	3336618	3368143	3413205	3471696	3538385	3602350
	26	age 35-39	2661468	2772038	2855379	2931777	2998169	3050573	3088525	3105944	3105027	3094639
	27	age 40-44	2137810	2236835	2322844	2407684	2489301	2564864	2633045	2689889	2734556	2764482
	28	age 45-49	1664093	1752752	1827586	1899070	1968194	2036785	2106781	2175182	2238672	2294659
	29	age 50-54	1286946	1319592	1365250	1424668	1492663	1562171	1629667	1693033	1752133	1807249
	30	age 55-59	1145935	1167438	1175466	1175333	1174717	1182724	1205705	1244524	1295474	1353612
	31	age 60-64	896731	915610	940529	969977	999529	1022978	1038390	1044266	1043200	1041575
	32	age 65-69	704384	726017	740356	749733	757473	767132	781876	802450	827115	851901
	33	age 70-74	469050	476608	493606	516930	542117	564264	581001	591823	598883	604934
	34	age 75-79	359487	358030	353479	347439	341956	341245	347303	359851	376711	394724
	35	age 80-84	191662	205702	214062	220587	225834	228366	227658	224848	221164	218003
	*	age > 85	45472	59120	75374	90525	104637	117854	130192	140813	149813	157762
			40982880	*41882359*	*42693725*	*43462350*	*44182312*	*44850113*	*45472429*	*46035621*	*46532612*	*46966647*
Share of population in Single person HH	36	single	0.0331	0.0351	0.0373	0.0396	0.0421	0.0447	0.0475	0.0503	0.0505	0.0505
	*	other	0.9669	0.9649	0.9627	0.9604	0.9579	0.9553	0.9525	0.9497	0.9495	0.9495

*reference category
source: ASSA 2003 , Census 1996 & 2001

Table b.2: Difference between restrictions and initial values

Category	Restriction No	Description	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
			Difference: Restrictions-Initial									
Stratum	1	WC urban	149941	136808	150930	171483	281959	302709	374606	423936	149423	160843
	2	WC rural	-38035	14438	35212	59608	-10266	5068	-14703	-52321	-132288	-137296
	3	EC urban	120619	61144	150636	217952	248585	135614	139876	226026	425447	430307
	4	EC rural	51372	96728	-80108	-273901	-444646	-459133	-589204	-823181	-328812	-331436
	5	NC urban	-54996	-8134	-50017	35598	53692	64501	83119	144536	183872	192438
	6	NC rural	29070	-3671	30144	-60201	-91562	-84180	-100280	-170058	-132362	-131693
	7	FS urban	299803	81440	-42340	143516	57351	56005	158888	143651	292721	264763
	8	FS rural	-141589	37345	120393	-101794	-76963	-67008	-204266	-234103	-235646	-221155
	9	KN urban	489611	184202	204985	-96866	2966	-148083	-95162	174715	119341	88246
	10	KN rural	-33412	244582	199172	497198	292530	493567	404808	70473	-280996	-284050
	11	NW urban	-179126	21166	31658	141749	79730	34733	241841	290293	244846	227438
	12	NW rural	231977	42381	6093	-130958	-97200	-21631	-189244	-313218	-347466	-346749
	13	GT urban	302224	187683	274213	474791	678219	806455	921395	980254	-11976	-54004
	14	GT rural	-174518	8743	46122	31739	-32941	-53001	32049	-2592	-148858	-134771
	15	MP urban	260552	52181	78655	78331	44767	17184	86948	88551	62776	44200
	16	MP rural	-110822	76208	46604	38866	55027	64411	-27891	-61271	-68214	-59188
	17	LP urban	46700	8230	20450	-1709	30720	-69164	-6754	74962	111507	115699
	*	LP rural	8331	58347	27821	-40000	-112676	-204468	-350355	-531416	-28112	-9278
		Total	*1257701*	*1299821*	*1250624*	*1185404*	*959295*	*873580*	*865668*	*429237*	*-124796*	*-185686*
Sex	18	male	0.0055	0.0047	0.0031	0.0021	0.0009	0.0039	0.0037	0.0028	0.0081	0.0075
	*	female	-0.0055	-0.0047	-0.0031	-0.0021	-0.0009	-0.0039	-0.0037	-0.0028	-0.0081	-0.0075
Share of population	19	single	0.0063	0.0138	0.0140	0.0073	-0.0034	-0.0061	-0.0025	-0.0017	-0.0102	-0.0135
in Single person HH	*	other										

Table b.2 continued

Category	Restriction No	Description	Difference: Constraint-Initial 1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Age Group	20	age 0-4	996868	564261	344280	861596	812190	-620285	-953429	-1333475	871608	758884
	21	age 5-9	-80826	260771	190940	-386278	-503041	285250	370407	491546	125306	133084
	22	age 10-14	6856	42804	203003	-189693	-310790	133804	147977	105049	-475951	-388381
	23	age 15-19	59500	75572	98574	138285	183109	181103	181763	143113	-421095	-452251
	24	age 20-24	-124797	-160795	46902	-146628	-200749	92242	150353	213566	86314	241349
	25	age 25-29	44975	217010	39447	139135	156994	108673	116294	90957	140550	-95705
	26	age 30-34	97887	108205	104254	78427	28096	36987	65095	74037	126310	145022
	27	age 35-39	63764	89242	101818	170878	171297	130200	132015	86988	-21696	-188213
	28	age 40-44	97434	117336	68287	233868	268602	108583	121703	59125	-80450	-111404
	29	age 45-49	75992	69842	49657	207421	239235	100818	105964	100335	-34139	-20540
	30	age 50-54	97054	46578	20343	153373	196902	86792	96404	82704	-103718	-39247
	31	age 55-59	90216	85962	63059	49346	26140	66718	81384	97371	8285	84365
	32	age 60-64	10009	-73334	25825	24476	35750	99605	124625	111268	-125509	-114205
	33	age 65-69	-80398	-81347	-60673	-84898	-69713	46854	46972	69560	-11115	53110
	34	age 70-74	-77148	-25473	-4229	-27677	-26822	16434	15614	192	-114303	-115605
	35	age 75-79	47176	-20393	-35657	9171	-15782	22022	51159	63250	-34893	-19169
	36	age 80-84	-11446	46751	64176	7643	10409	-7709	22955	14508	-9276	-13697
	*	age > 85	-55412	-63171	-69382	-53038	-42530	-14511	-11587	-40858	-51023	-43086

Appendix C

Missing Values

	year	initial sample size	age	sex	age*sex	strata	total observations dropped	percentage dropped	final sample size
OHS	1995	130787	0	0	0	0	0	0.00%	130787
	1996	72889	0	0	0	0	0	0.00%	72889
	1997	140015	0	0	0	0	0	0.00%	140015
	1998	82263	0	0	0	2	2	0.00%	82261
	1999	106650	184	44	2	0	226	0.21%	106424
LFS	2000	105371	117	13	1	0	129	0.12%	105242
	2001	106439	139	0	0	0	139	0.13%	106300
	2002	102480	118	28	0	0	146	0.14%	102334
	2003	98748	52	1	0	0	53	0.05%	98695
	2004	98256	72	19	9	0	82	0.08%	98174

Appendix D

Table D.1: ME-weight regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
StatsSA Person Weight	0.0220***	0.00619***	0.0279***	0.00276***	0.00129***	0.00495***	0.00823***	0.0117***	0.0113***	0.00291***
	(0.00073)	(0.00094)	(0.0011)	(0.00050)	(0.00022)	(0.00031)	(0.00031)	(0.00033)	(0.00061)	(0.00033)
WC rural	-144.5***	-34.25***	-111.5***	-54.47***	-307.1***	-352.0***	-336.3***	-343.1***	-381.1***	-381.7***
	(0.59)	(2.80)	(0.91)	(0.94)	(0.29)	(0.44)	(0.44)	(0.45)	(0.79)	(0.85)
EC urban	-120.5***	-246.2***	73.55***	77.37***	-68.69***	-120.4***	-98.52***	-103.2***	-115.4***	-135.2***
	(0.36)	(1.36)	(0.62)	(0.54)	(0.21)	(0.32)	(0.32)	(0.33)	(0.59)	(0.62)
EC rural	-62.75***	-251.8***	-4.883***	22.98***	-26.90***	-69.22***	-46.26***	-56.53***	-78.61***	-85.78***
	(0.33)	(1.21)	(0.49)	(0.45)	(0.19)	(0.30)	(0.30)	(0.30)	(0.53)	(0.57)
NC urban	-200.5***	-440.9***	-171.3***	-290.1***	-272.6***	-319.9***	-290.6***	-290.1***	-325.3***	-352.5***
	(0.49)	(1.79)	(0.70)	(0.62)	(0.27)	(0.40)	(0.40)	(0.41)	(0.74)	(0.75)
NC rural	-161.6***	-308.1***	-162.9***	-361.9***	-352.3***	-417.5***	-401.1***	-429.8***	-465.5***	-479.3***
	(0.71)	(2.93)	(1.07)	(0.83)	(0.36)	(0.55)	(0.56)	(0.54)	(0.97)	(1.04)
FS urban	-98.57***	-122.5***	-54.17***	-39.99***	-92.99***	-128.5***	-119.2***	-95.13***	-129.7***	-142.0***
	(0.39)	(1.53)	(0.58)	(0.53)	(0.22)	(0.34)	(0.33)	(0.35)	(0.62)	(0.65)
FS rural	-138.8***	-164.2***	-43.14***	-160.5***	-233.0***	-287.0***	-283.7***	-276.8***	-317.0***	-328.4***
	(0.46)	(1.96)	(0.79)	(0.67)	(0.27)	(0.41)	(0.41)	(0.43)	(0.76)	(0.81)
KN urban	-0.828**	-55.26***	92.90***	144.0***	103.7***	3.200***	48.30***	54.74***	73.68***	28.55***
	(0.35)	(1.30)	(0.55)	(0.49)	(0.21)	(0.30)	(0.30)	(0.31)	(0.56)	(0.58)
KN rural	-17.53***	58.23***	-1.276***	197.8***	-0.523***	-90.99***	-47.52***	-16.91***	-36.08***	-49.96***
	(0.32)	(1.27)	(0.47)	(0.47)	(0.19)	(0.28)	(0.28)	(0.29)	(0.51)	(0.55)
NW urban	-121.2***	-138.4***	-20.55***	-41.66***	-168.2***	-220.6***	-172.9***	-154.3***	-171.8***	-180.3***
	(0.43)	(1.77)	(0.70)	(0.62)	(0.24)	(0.35)	(0.36)	(0.37)	(0.66)	(0.70)
NW rural	67.18***	-105.8***	-67.06***	-99.31***	-119.5***	-156.5***	-146.0***	-160.3***	-179.1***	-183.1***
	(0.42)	(1.47)	(0.55)	(0.50)	(0.21)	(0.33)	(0.33)	(0.33)	(0.59)	(0.64)
GT urban	264.3***	-86.24***	116.0***	330.7***	99.26***	105.5***	133.3***	223.8***	204.5***	206.9***
	(0.39)	(1.14)	(0.48)	(0.46)	(0.18)	(0.27)	(0.27)	(0.29)	(0.51)	(0.53)
GT rural	-170.9***	52.75***	234.6***	576.2***	199.4***	23.80***	165.2***	301.1***	185.7***	216.1***
	(0.74)	(4.07)	(1.97)	(1.95)	(0.69)	(0.93)	(1.02)	(1.14)	(1.91)	(2.03)
MP urban	-116.8***	-65.24***	-59.23***	-95.62***	-215.0***	-231.0***	-189.7***	-198.5***	-242.9***	-253.3***
	(0.45)	(1.87)	(0.68)	(0.61)	(0.23)	(0.36)	(0.37)	(0.37)	(0.65)	(0.69)

Table D.1 continued

MP rural	-104.9***	-232.0***	-94.45***	-155.4***	-156.1***	-171.3***	-144.8***	-150.3***	-168.9***	-170.1***
	(0.39)	(1.48)	(0.56)	(0.51)	(0.22)	(0.34)	(0.34)	(0.35)	(0.62)	(0.66)
LP urban	-162.1***	-300.2***	-5.522***	46.89***	-258.5***	-302.5***	-261.1***	-243.7***	-284.0***	-305.5***
	(0.54)	(2.11)	(0.97)	(0.90)	(0.28)	(0.41)	(0.42)	(0.43)	(0.76)	(0.79)
LP rural	71.42***	-240.5***	-1.171**	-21.67***	-34.82***	-78.08***	-46.18***	-42.25***	-61.26***	-66.63***
	(0.35)	(1.20)	(0.48)	(0.43)	(0.19)	(0.29)	(0.28)	(0.29)	(0.52)	(0.55)
age 5-9	-78.45***	-86.34***	-43.50***	-96.93***	-100.3***	-101.8***	-116.9***	-113.5***	-104.1***	-95.05***
	(0.30)	(1.05)	(0.44)	(0.41)	(0.17)	(0.26)	(0.27)	(0.29)	(0.49)	(0.53)
age 10-14	-96.44***	-107.7***	-50.05***	-104.6***	-113.5***	-132.1***	-160.0***	-180.5***	-188.6***	-198.3***
	(0.30)	(1.04)	(0.44)	(0.41)	(0.17)	(0.26)	(0.27)	(0.29)	(0.48)	(0.51)
age 15-19	-113.1***	-116.7***	-54.67***	-115.2***	-112.4***	-116.3***	-130.9***	-147.5***	-158.6***	-153.8***
	(0.31)	(1.07)	(0.45)	(0.42)	(0.17)	(0.26)	(0.27)	(0.29)	(0.48)	(0.52)
age 20-24	-81.60***	-98.04***	-21.72***	-67.43***	-86.19***	-97.87***	-94.68***	-119.6***	-131.2***	-129.3***
	(0.32)	(1.10)	(0.47)	(0.44)	(0.18)	(0.27)	(0.28)	(0.30)	(0.50)	(0.54)
age 25-29	-82.19***	-31.51***	2.064***	-35.80***	-56.73***	-59.13***	-83.55***	-83.82***	-73.34***	-93.83***
	(0.33)	(1.18)	(0.50)	(0.46)	(0.19)	(0.28)	(0.29)	(0.31)	(0.53)	(0.57)
age 30-34	-69.64***	-34.27***	21.83***	-3.682***	-63.26***	-76.71***	-94.66***	-93.36***	-118.5***	-115.7***
	(0.35)	(1.22)	(0.52)	(0.49)	(0.19)	(0.29)	(0.30)	(0.32)	(0.54)	(0.58)
age 35-39	-105.2***	-38.31***	3.992***	-26.51***	-77.50***	-83.56***	-99.36***	-84.27***	-123.1***	-118.8***
	(0.35)	(1.27)	(0.54)	(0.50)	(0.20)	(0.30)	(0.31)	(0.33)	(0.56)	(0.60)
age 40-44	-98.90***	-86.99***	6.834***	-53.37***	-73.89***	-80.44***	-98.86***	-135.8***	-149.1***	-151.3***
	(0.38)	(1.33)	(0.58)	(0.53)	(0.21)	(0.31)	(0.32)	(0.33)	(0.57)	(0.61)
age 45-49	-135.4***	-92.72***	-28.09***	-72.67***	-104.0***	-105.1***	-123.4***	-136.0***	-134.2***	-127.6***
	(0.40)	(1.44)	(0.61)	(0.56)	(0.22)	(0.33)	(0.34)	(0.36)	(0.61)	(0.65)
age 50-54	-122.3***	-97.15***	-26.96***	-134.1***	-106.9***	-112.4***	-123.6***	-166.9***	-183.5***	-187.5***
	(0.44)	(1.59)	(0.68)	(0.60)	(0.24)	(0.36)	(0.37)	(0.38)	(0.64)	(0.68)
age 55-59	-131.3***	-68.45***	-41.48***	-126.5***	-109.2***	-93.26***	-138.3***	-163.6***	-169.2***	-174.7***
	(0.45)	(1.69)	(0.70)	(0.64)	(0.26)	(0.41)	(0.41)	(0.42)	(0.72)	(0.76)
age 60-64	-146.3***	-156.1***	-61.96***	-144.4***	-108.6***	-150.0***	-150.7***	-191.3***	-229.6***	-231.5***
	(0.48)	(1.74)	(0.74)	(0.69)	(0.28)	(0.41)	(0.43)	(0.44)	(0.74)	(0.79)
age 65-69	-175.8***	-157.4***	-94.71***	-192.0***	-142.9***	-124.7***	-155.5***	-166.3***	-189.0***	-184.5***
	(0.50)	(1.90)	(0.77)	(0.72)	(0.30)	(0.47)	(0.47)	(0.49)	(0.84)	(0.90)
age 70-74	-161.3***	-121.8***	-70.52***	-172.1***	-118.7***	-145.8***	-158.0***	-188.0***	-200.1***	-205.6***

Table D.1 continued

	(0.61)	(2.36)	(0.96)	(0.86)	(0.36)	(0.52)	(0.53)	(0.55)	(0.95)	(1.01)
age 75-79	-134.2***	-126.2***	-100.4***	-125.7***	-117.9***	-113.9***	-119.2***	-165.0***	-185.2***	-205.1***
	(0.72)	(2.62)	(1.05)	(1.08)	(0.45)	(0.68)	(0.69)	(0.69)	(1.18)	(1.22)
age 80-84	-58.83***	99.62***	36.36***	-121.9***	-78.77***	-164.6***	-150.3***	-216.2***	-170.2***	-197.1***
	(1.08)	(4.06)	(1.64)	(1.34)	(0.56)	(0.77)	(0.81)	(0.81)	(1.53)	(1.61)
age 85+	-291.2***	-383.9***	-191.2***	-302.8***	-212.5***	-197.8***	-209.7***	-210.7***	-262.1***	-264.6***
	(1.20)	(4.35)	(1.67)	(1.61)	(0.66)	(0.98)	(0.96)	(0.99)	(1.58)	(1.67)
male	13.55***	49.99***	28.83***	32.76***	13.24***	14.40***	17.94***	7.442***	9.742***	10.90***
	(0.14)	(0.50)	(0.21)	(0.20)	(0.078)	(0.12)	(0.12)	(0.12)	(0.22)	(0.23)
Constant	420.9***	769.5***	312.8***	561.6***	554.1***	611.3***	597.8***	625.1***	674.8***	695.9***
	(0.42)	(1.39)	(0.58)	(0.50)	(0.21)	(0.35)	(0.35)	(0.38)	(0.63)	(0.60)
Observations	130787	72889	140013	82261	106424	105242	106300	102334	98695	98174
R-squared	0.96	0.79	0.82	0.97	0.99	0.98	0.98	0.99	0.96	0.96

WC urban is the omitted category
Standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1

Table D.2: CE-weight regression

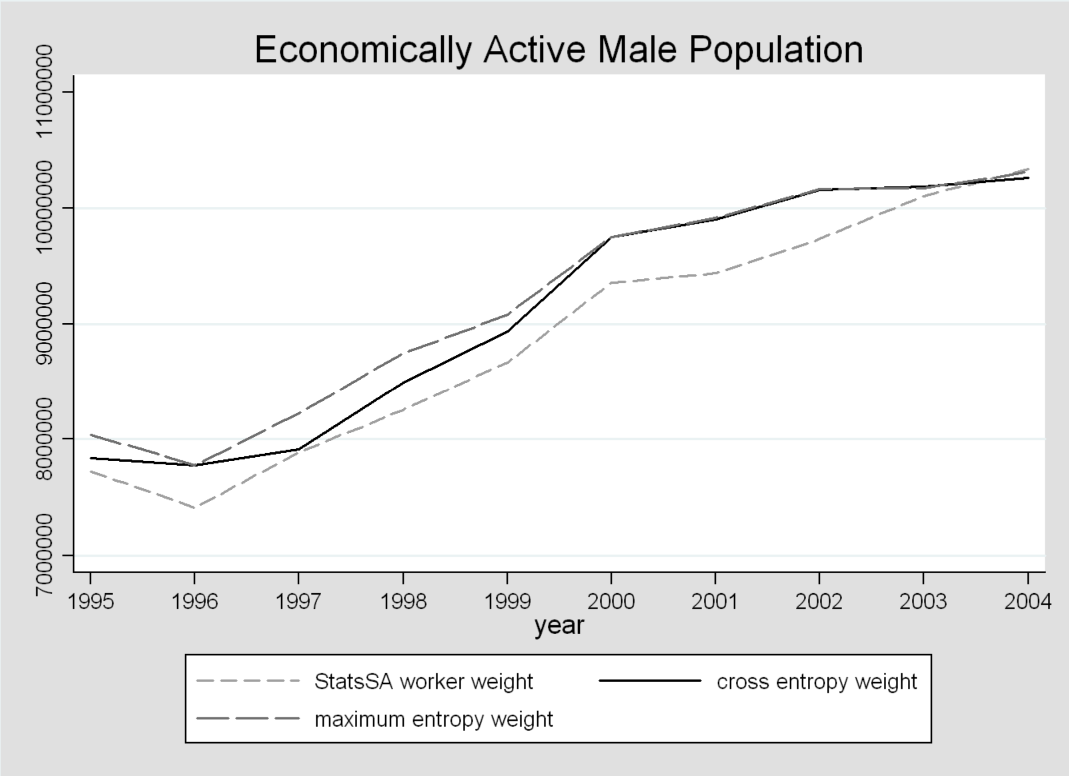
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
StatsSA Person Weight	1.018***	1.050***	1.055***	1.015***	1.008***	0.995***	0.966***	0.910***	0.990***	0.985***
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
WC rural	-28.96***	-3.935	9.223***	38.63***	-35.60***	-39.04***	-60.32***	-95.73***	-73.31***	-82.91***
	[0.87]	[3.27]	[0.87]	[1.86]	[0.98]	[0.83]	[0.99]	[1.25]	[1.26]	[1.61]
EC urban	-0.800	-3.110*	8.077***	28.53***	7.338***	-18.10***	-25.65***	-23.27***	50.64***	47.84***
	[0.53]	[1.59]	[0.59]	[1.06]	[0.73]	[0.61]	[0.72]	[0.91]	[0.94]	[1.18]
EC rural	-9.941***	-3.964***	-18.46***	-57.96***	-82.88***	-90.77***	-109.1***	-136.2***	-57.85***	-60.61***
	[0.48]	[1.41]	[0.47]	[0.89]	[0.65]	[0.55]	[0.66]	[0.83]	[0.84]	[1.07]
NC urban	-25.11***	-10.19***	-14.69***	-6.364***	-15.57***	-21.74***	-30.17***	-34.22***	32.82***	29.18***
	[0.72]	[2.09]	[0.67]	[1.21]	[0.91]	[0.75]	[0.90]	[1.13]	[1.17]	[1.42]
NC rural	8.504***	-19.46***	16.19***	-60.43***	-94.07***	-98.60***	-124.4***	-182.4***	-109.2***	-117.6***
	[1.05]	[3.42]	[1.03]	[1.62]	[1.23]	[1.03]	[1.25]	[1.51]	[1.55]	[1.97]
FS urban	26.64***	4.869***	-15.82***	11.79***	-22.51***	-29.16***	-20.20***	-29.73***	36.35***	28.61***
	[0.57]	[1.79]	[0.55]	[1.05]	[0.76]	[0.64]	[0.75]	[0.96]	[0.98]	[1.24]
FS rural	-45.20***	5.105**	29.92***	-65.43***	-57.38***	-63.10***	-116.5***	-148.4***	-107.7***	-107.5***
	[0.68]	[2.29]	[0.76]	[1.32]	[0.92]	[0.78]	[0.92]	[1.20]	[1.22]	[1.54]
KN urban	31.28***	8.161***	3.881***	-40.51***	-34.45***	-57.46***	-53.35***	-20.56***	-1.557*	-8.511***
	[0.51]	[1.52]	[0.52]	[0.96]	[0.70]	[0.57]	[0.69]	[0.86]	[0.88]	[1.10]
KN rural	-16.62***	8.463***	-0.193	43.64***	-6.708***	2.159***	-10.61***	-41.41***	-47.11***	-49.23***
	[0.47]	[1.48]	[0.45]	[0.91]	[0.63]	[0.52]	[0.63]	[0.80]	[0.81]	[1.03]
NW urban	-47.83***	-10.63***	-4.315***	28.38***	-14.90***	-33.01***	0.751	-1.453	37.20***	30.83***
	[0.64]	[2.07]	[0.67]	[1.21]	[0.80]	[0.67]	[0.80]	[1.02]	[1.05]	[1.32]
NW rural	28.16***	-10.71***	-8.459***	-45.48***	-49.18***	-43.68***	-79.06***	-108.3***	-81.40***	-85.29***
	[0.63]	[1.72]	[0.53]	[0.98]	[0.73]	[0.62]	[0.74]	[0.93]	[0.95]	[1.20]
GT urban	7.018***	-6.992***	-3.425***	23.75***	13.50***	22.09***	26.71***	49.69***	-17.76***	-21.78***
	[0.57]	[1.33]	[0.46]	[0.90]	[0.60]	[0.51]	[0.61]	[0.79]	[0.82]	[1.01]
GT rural	-148.0***	0.378	89.58***	117.1***	-128.8***	-158.5***	45.64***	-22.04***	-445.4***	-415.8***
	[1.10]	[4.75]	[1.89]	[3.83]	[2.35]	[1.74]	[2.30]	[3.18]	[3.04]	[3.84]
MP urban	39.87***	6.270***	7.735***	5.627***	-22.88***	-36.43***	-29.97***	-45.43***	-7.085***	-14.21***
	[0.66]	[2.19]	[0.65]	[1.20]	[0.80]	[0.68]	[0.83]	[1.03]	[1.04]	[1.30]

Table D.2 continued

MP rural	-27.95***	5.161***	-1.825***	-12.75***	-23.23***	-27.79***	-53.59***	-70.78***	-34.18***	-35.18***
	[0.57]	[1.73]	[0.54]	[1.00]	[0.74]	[0.64]	[0.77]	[0.97]	[0.99]	[1.26]
LP urban	4.718***	-6.769***	-1.499	-24.91***	-21.61***	-63.61***	-54.68***	-44.10***	17.09***	14.13***
	[0.80]	[2.46]	[0.93]	[1.78]	[0.94]	[0.78]	[0.94]	[1.20]	[1.22]	[1.50]
LP rural	-15.33***	-9.559***	-10.65***	-26.42***	-44.78***	-59.22***	-78.09***	-98.64***	-22.30***	-22.43***
	[0.52]	[1.40]	[0.46]	[0.85]	[0.63]	[0.54]	[0.64]	[0.81]	[0.82]	[1.04]
Constant	8.487***	-7.585***	-4.007***	15.05***	30.37***	42.01***	60.72***	92.70***	24.21***	29.04***
	[0.50]	[1.34]	[0.45]	[0.78]	[0.57]	[0.49]	[0.56]	[0.71]	[0.80]	[0.85]
Observations	130787	72889	140013	82261	106424	105242	106300	102334	98695	98174
R-squared	0.95	0.94	0.93	0.96	0.96	0.98	0.97	0.95	0.95	0.97

WC urban is the omitted category
Standard errors in parentheses
***p<0.01, **p<0.05, *p<0.1

Appendix E



Economically Active Male Population
11000000
10000000
9000000
8000000
7000000
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
year
StatsSA worker weight
cross entropy weight
maximum entropy weight

Appendix F

```
****************************************************************************
*Entropy Weight Programme
*Uses OHS and LFS data, ASSA and Census Aggregates
****************************************************************************
clear
set type double, perm
set mem 200m
set more off
global filepath="specify path directory
local x="95" /*this is done for each year 1995-2004*/
*use national household survey person level data
	use "$filepath\datasets\\`x' strata.psu.person_merge.dta", clear
	keep if age`x'~=.
	sort hhid persno
*create proportionate weights*
	rename perswgt q1
	egen double qtot=sum(q1)
	gen double qrescale=q1/qtot
*create person stratum dummies*
	forvalues y=1/18{
		gen x`y'=0
		replace x`y'=1 if stratum==`y'
	}
*create person age group dummies
	gen agegrp`x'=1 if age`x'<5
	replace agegrp`x'=2 if age`x'>=5 & age`x'<10
	replace agegrp`x'=3 if age`x'>=10 & age`x'<15
	replace agegrp`x'=4 if age`x'>=15 & age`x'<20
	replace agegrp`x'=5 if age`x'>=20 & age`x'<25
	replace agegrp`x'=6 if age`x'>=25 & age`x'<30
	replace agegrp`x'=7 if age`x'>=30 & age`x'<35
	replace agegrp`x'=8 if age`x'>=35 & age`x'<40
	replace agegrp`x'=9 if age`x'>=40 & age`x'<45
	replace agegrp`x'=10 if age`x'>=45 & age`x'<50
	replace agegrp`x'=11 if age`x'>=50 & age`x'<55
	replace agegrp`x'=12 if age`x'>=55 & age`x'<60
	replace agegrp`x'=13 if age`x'>=60 & age`x'<65
	replace agegrp`x'=14 if age`x'>=65 & age`x'<70
	replace agegrp`x'=15 if age`x'>=70 & age`x'<75
	replace agegrp`x'=16 if age`x'>=75 & age`x'<80
	replace agegrp`x'=17 if age`x'>=80 & age`x'<85
	replace agegrp`x'=18 if age`x'>=85 & age`x'<.
	forvalues y=1/18{
		gen da`y'=0
	replace da`y'=1 if agegrp==`y'
	}
*create male dummy
	gen ds1=0
	replace ds1=1 if  gender`x'==1
*create single person dummy
	sort hhid`x'  persno`x'
	egen hhsize=count(persno`x'), by(hhid`x')
	gen single=0
	replace single=1 if hhsize==1
	gen cons=1
*strata restriction values from ASSA2003
	do "$filepath\y_input\y `x' matrix.do"
	svmat y
	egen double pop=sum(y1)
	replace pop=pop/2
	drop y1
	scalar yt=pop[1]
*create restriction matrix
	matrix A1=y[1..17,1]
```

```
        matrix A2=y[19..35, 1]
        matrix define A3=(0.485586523)
        matrix define A4=(0.0331)
        matrix A=A1\A2
        matrix y1=A/pop[1]
        matrix y=y1\A3\A4
*** Max entropy
sort hhid
cap program drop mymaxent2
program define mymaxent2
        version 9.2
        args todo b lnf
        tempname lambday
        tempvar omega xb
        mleval `xb' = `b'
        quietly{
                sort hhid
                gen double `omega'=$ML_y1*exp(-`xb')
                replace `omega'=sum(`omega')
                matrix `lambday'=`b'*y
        }
        scalar `lnf'=-(`lambday'[1,1])-ln(`omega'[_N])
end

****MAXIMUM ENTROPY****
#delimit ;
ml model d0 mymaxent2 (cons=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16
x17 da1 da2 da3 da4 da5 da6 da7 da8 da9 da10 da11 da12 da13 da14 da15 da16 da17 ds1
single, nocons);
#delimit cr
ml maximize


predict double p3
replace p3=exp(-p3)
egen double omega=sum(p3)
replace p3=p3/omega
drop omega
sort stratum
gen double me_wgt=p3*yt

****CROSS ENTROPY****

ml model d0 mymaxent2 (qrescale=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15
x16 x17 da1 /*
*/ da2 da3 da4 da5 da6 da7 da8 da9 da10 da11 da12 da13 da14 da15 da16 da17 ds1
single, nocons)
ml maximize

predict double p4
replace p4=qrescale*exp(-p4)
egen double omega=sum(p4)
replace p4=p4/omega
drop omega
sort stratum
gen double ce_wgt=p4*yt

drop x* da* ds*  qtot qrescale pop p3 p4 c

save "$filepath\pdata.entropy95.dta", replace
```

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