Reweighting the OHS and GHS to improve data quality: representativeness, household counts, and small households

Paper submitted for the 2021 IUSSP International Population Conference

Amy Thornton Martin Wittenberg

December 2021

Abstract

The October Household Surveys (OHS) (1994-9) and the General Household Surveys (GHS) (2002-present) collected by StatsSA comprise South Africa's only nationally-representative time series with information on both people and households for (almost) every year of the post-apartheid period. However, the quality of these data has been compromised in three ways by how the survey weights have been calibrated. We document these problems and their implications in detail; and then use cross-entropy estimation to recalibrate the survey weights for a stacked version of these surveys between 1995 and 2011 to address these weaknesses. The first of these is that the weight calibration procedure breaks with sampling practise by calibrating person and household weights separately. This creates conceptual problems because the data is not properly representative of the population. It also creates statistical problems, including that a series of total population and household counts cannot be reliably extracted from the series, which is typically a first-order output for such a time series. Secondly, issues with the benchmarks StatsSA use mean the series of household counts extracted from the GHS is probably too low. Thirdly, no compensation is made by the survey weights for the chronic undersampling of small households over the entire period. Our new weights make headway in resolving these issues. Our weights yield consistent counts of people and households benchmarked on both person and household auxiliary information for the first time; and, benchmarked counts of one-, two-, and three-person households. Work is ongoing to improve the weights.

1 Introduction

The October Household Surveys (OHS) (1993-1999) and General Household Surveys (GHS) (2002-present) collected by Statistics South Africa (StatsSA) are among the country's best data sources for information about people and households. These data comprise South Africa's only nationally-representative time-series that covers (almost) every year of the post-apartheid period with comprehensive information about both people and households. Both the OHS and GHS were specifically designed to assist with evidence-based policy-making for the new democratic era. Therefore, both surveys collect a common set of basic demographic information and include dedicated modules to monitor several of the State's social policies, such as access to basic services (e.g. water, electricity) and social grants (e.g. pension, child support grant). This makes the OHS-GHS series one of the best sources of information about the progress of welfare over the post-apartheid period, and the only nationally-representative source for the first ten years of this period. This positioning in South Africa's data infrastructure, means the importance of the OHS and the GHS to both researchers and policy-makers in South Africa cannot be overstated.

However, the quality of these data is compromised by the way the survey weights are calibrated. Survey weights are critical to the accuracy of estimates and the degree to which data analysis can be relevant for the national population. These properties in the OHS and GHS are undermined by three main weak points: weighting practise breaking with sampling practise; the quality of the household benchmarks; and, the undersampling of small households. A contribution of this paper is, firstly, to document these weaknesses in detail and discuss their implications. This is information about the quality of the OHS and GHS of which any researcher aiming to use these data should be aware. The weights currently released with these data statistically and conceptually undermine analyses combining person- and household-level information; lead to an underestimation of household counts of the order of about 5%; and, make weighted analysis of trends in small households unreliable. Our second contribution is to construct a new set of survey weights using cross-entropy estimation that overcome these weaknesses, although work on the weights is ongoing. These weights are made available with this article.

There are three main weaknesses with the existing weights in the OHS and GHS, and probably the most important relates to StatsSA weighting the household and person data separately, in a process that is at odds with the how the survey data were collected. The OHS and the GHS are released with two weights, a household weight and a person weight, but these weights are not linked in any way whatsoever even though they apply to the same sample. Instead, they are calibrated on mutually exclusive auxiliary information in totally separate procedures. The result is one weight that yields a representative person universe, and another that yields a representative household universe, but not one that is representative of both at the same time. This creates statistical problems because one can now extract two different estimates for the same statistic. But, also creates conceptual problems, since people are effectively detached from their own households, having the effect of restricting the number of research questions the data can be used to answer coherently.

The second weakness of the calibration is inaccuracy in the series of household auxiliary information StatsSA use to calibrate the GHS household weight. These issues stem from choice of data sources for benchmarks and inconsistent treatment of the worker hostel sub-population in the benchmarks. The main effect of these weaknesses is a series of household counts that is too low; a factor which is particularly concerning for policymakers for whom household counts are important for planning purposes. On average, this undercount is of the order of about 650 000 households per year, or 5%. The final third weakness

with the survey weighting is that no compensation has been made for the chronic undersampling of small households. Since the household is the unit that is sampled, systematically missing certain types of households could lead to bias in a broad range of statistics. The serious undersampling of small households was first documented by Kerr & Wittenberg (2015) in the OHS era and we show this has continued in the GHS era. The gap in the total number of single-person households in the country as measured by the censuses versus the OHS-GHS widens over time, so that by 2011 the GHS underestimates the count by about 800 000 households, or 25%. Single-person households were the household type that grew the fastest between the 1996 and 2011 censuses; however, inadequacies with the survey weights mean the OHS-GHS data cannot be used to study this important constituency reliably.

Taken together, these three weaknesses with the survey weights represent a considerable cost to the country's research agenda; to the quality of information feeding into policy-making; and, to the public funds spent on the collection of these data. Fortunately, these costs are not insurmountable. The nature of the calibration process, as well as, the fact that it is undertaken after a survey is collected means it can be remedied relatively painlessly compared to other influences on data quality, like survey design or sampling errors. In this paper, we use cross-entropy estimation to reweight a stacked version of the OHS and GHS between 1995 and 2011 to address each of the three weaknesses. In the first case, we combine the information StatsSA use in two separate calibration processes, into one procedure. It was not apparent at the outset that this could practically be done because, as far as we know, StatsSA have never released survey weights calibrated using both auxiliary person and household information at the same time. This also makes this type of reweighting slightly different to work by (Branson & Wittenberg 2014) and Kerr et al. (2019) who reweight StatsSA labour market data on person auxiliary information only.[1] To target the second two weaknesses, we improve the quality of the series of household count benchmarks; and explicitly constrain the calibration on the number of one-, two-, and three-person households.

The new cross-entropy weights cohere with sampling practise, thereby restoring the mutual representativeness of people and households and recovering the number of questions these data can be coherently employed to answer. An immediate output is a set of survey weights we plan to eventually make freely available on data repository site, DataFirst. Advantages of these new weights include a benchmarked series of total household counts extending back into the OHS era for the first time; and a benchmarked series of single-, two-, and three-person households, for the first time for the GHS. The series of person and household counts we present here are also the first time such a complete series of counts for the post-apartheid period has been presented that are benchmarked on both person and household auxiliary information. The reweighting goes a long way towards improving the data quality of the OHS and the GHS, but is neither a panacea for data quality issues in these data, nor without its own challenges and limitations. The work described here is our progress so far on an ongoing project to improve the survey weights in the OHS and the GHS.

In the next section we introduce the OHS and GHS data in more detail and then describe what a weighting practise consistent with sampling practise would look like in Section 3. Section 4 then details weaknesses with the current weighting practise in the OHS and GHS. Our method for recalibrating the weights is described in Section 5. The performance of our new weight on selected statistics is presented in Section 6. We discuss one of the limits of our weights in Section 7. Section 8 concludes.

[1]Kerr et al. (2019) release their weight in the Post-Apartheid Labour Market Series (PALMS). Groundwork for the PALMS weight comes from the series constructed by Branson & Wittenberg (2014)

2 The OHS and GHS data

2.1 The surveys

The OHS is the only large nationally representative annual household survey undertaken by StatsSA in the period 1994-1999.[2] The OHS covered a wide range of topics from core socio-demographic information to detailed labour market outcomes. After 1999, the OHS was conceptually split into the Labour Force Surveys (LFSs) and General Household Surveys (GHSs). The former ran bi-annually and focused on the labour market outcomes; whilst the GHS - which only launched in 2002 - focused on the socio-demographic outcomes and continued to run annually. The OHS and the GHS collect both person and household information and have modules dedicated to monitoring government social policies, such as access to social grants and basic household services. Both the OHS and the GHS are cross-sectional and survey approximately 30 000 dwelling units based on about 3 000 Primary Sampling Units drawn from the Master Sample of enumerator areas used during the most recent census at the time. Exceptions are that the 1996 and 1998 October Household Surveys only surveyed about 16 000 and 20 000 dwelling units, respectively, due to budget constraints. A stratified, two-stage cluster sampling design is employed in each case, stratified at the provincial level.[3]

Between these two surveys then, we have large samples of nationally representative cross-sectional data on individuals and households for every year in the period 1994-present, with the exception of 2000 and 2001. In terms of time coverage, this represents South Africa's most comprehensive nationally-representative time-series for the post-apartheid period. We combine 15 releases of the OHS (1995-9) and GHS (2002-11) surveys (StatsSA 2010-2013, 2011-2018*b*) into what we call the OHS-GHS series by merging person and household information; stacking the cross-sections by year; cleaning and harmonising a subset of variables; and, keeping both the household weight from the household files and the person weight from the person files. The result is a data set of about 1.5 million people and 385 thousand households spanning the period 1995-2011, with the exception of 2000 and 2001. Our reasons for choosing this time period are explained when we describe our method in Section 5.

3 Calibration: why it matters and why it is complex

Survey weights are an essential aspect of data quality and it is standard practise for national household surveys to be released with calibrated weights. For example, attention is usually dedicated to achieving a consistent and reliable series of population and household counts over time in this type of time-series cross-sectional survey data, and survey weights are a key part of curating these series. Survey weights are scaling factors assigned to each unit of observation to make the sample representative of the surveyed population so that statistics estimated from the sample satisfactorily approximate what they would have been had they been estimated from a census (Deaton 1997). Ensuring that the sample is representative of the population, or sampling frame, has received a lot of scholarly attention in recent decades and is the

[2]OHS data exist for 1993, but we exclude this year because the 1993 survey had a different sampling frame to the later surveys in that it excluded what were then the independent homelands of Transkei, Bophuthatswana, Venda, and, Ciskei.

[3]The 2004 Master Sample used for GHS 2005-7 was stratified at the district council level, although StatsSA caution that the data is not representative at this level and more recently released versions of the GHS for 2002-2007 do not include a district council variable (DataFirst 2015).

topic of an extensive academic and practical literature (Deaton 1997, Deville 2000, Lavallée & Beaumont 2015, Deville & Särndal 1992).

There are usually three steps to this process: (1) assigning the design weights which account for probability of a unit being selected for sampling, dependent on survey design; (2) adjusting the weights to compensate for units that were selected to be surveyed but were not surveyed for some reason (e.g. non-response; non-coverage); and (3) tuning the weights so that estimates from the weighted sample reflect known population totals, often taken from a census, in a process known as calibration (Lavallée & Beaumont 2015). The statistical problem of calibration is that unnecessary randomness is introduced into estimation by the process of sampling and sampling design. The aim is to improve the quality of inferences in case a poor sample is selected by weighting the sample to make it as representative as possible of the true population. In the event that the true population is the country population, this process strengthens the degree to which the data is 'nationally-representative', a valued feature of the OHS-GHS data.

Calibrating weights is not as objective a process as obtaining the design weights which depend entirely on the survey design (Deaton 1997). By contrast, calibration is essentially a modelling problem and the modelling can be of better or worse quality, depending on the standard of the auxiliary information available and the modelling decisions made by the survey statistician. Modelling choices include specification of the distance function; choice of constraints from the auxiliary data; as well as, the level of disaggregation of the constraints. On the one hand, the more constraints that are used (e.g. age, region, race) and the more detailed (e.g. finer age bands, more disaggregated regions), the closer the sample should get to resembling the true population and the more precise your estimates. However, too many constraints relative to raw sample size can undermine the computing of the calibration and introduce other small sample biases as the sample is cut into finer and finer cells (Deville & Särndal 1992). The calibration process can fail if too many constraints are being imposed on the data; or, if the requirements of the constraints are too onerous for the calibration to meet at once.

Furthermore, if the external source providing the known population totals is of poor quality or the sampling errors from non-coverage or non-response are large and difficult to overcome, the calibration will be undermined or introduce unintended distortion elsewhere. In these cases, calibration can create as many problems as it tries to solve (Smith 1991). The survey statistician thus faces a trade-off between improvement in the accuracy of a core set of estimates; the potential distortion caused by the calibration; and, the practical feasibility of the calibration procedure, itself. In the following section, we discuss three areas in which the calibration of the survey weights in the OHS-GHS series is compromised, undermining the quality of the OHS-GHS series overall and the extent to which it is representative of the population. These are: weighting practise breaking with sampling practise; the quality of the household benchmarks; and, the undersampling of small households.

4 Problems with the survey weight calibration in the OHS-GHS series

4.1 Breaking with sampling practise: the two-weight system

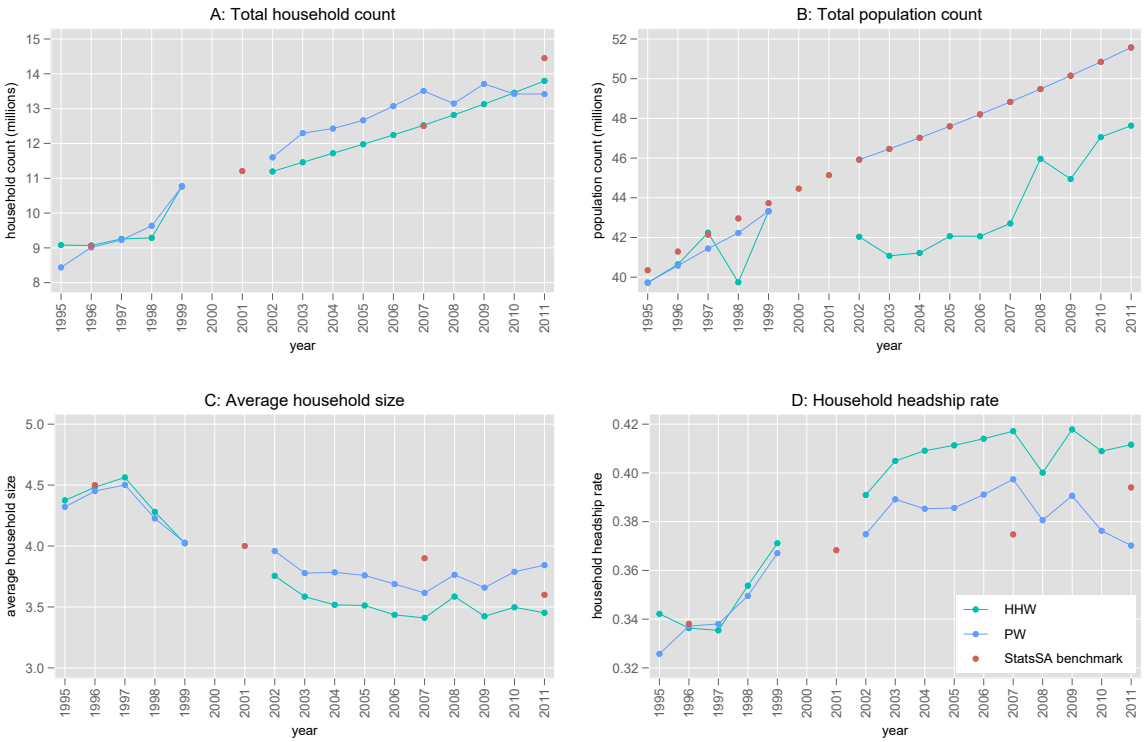
Probably the most fundamental problem with the weights in the OHS-GHS is that weighting practise breaks with sampling practise. This should never be the case because weighting procedure should be informed by sampling procedure which dictates the relationship between different units of analysis, such as people and households. Sampling practise for both the OHS and the GHS is that, if a household is selected for sampling, every person in that household is surveyed. This produces two implications for the surveys weights. Firstly, survey weights should be 'integrated': members of the same household should have the same weight. Secondly, there should also be no distinction between person and household weights because the chance a person will be sampled equals the chance their household is sampled. In other words, there should be one weight in the data, equal within households. In cases where household and person data is released in separate files, like the OHS and the GHS, the weight for a given household in the household file should equal the weight for a member of that household in the person file. This is almost never the case in the OHS or the GHS.

When calibrating the survey weights for the OHS, StatsSA calibrated the person weights in such a way that people within a household had different weights, i.e. weights were not integrated (Branson & Wittenberg 2014). This raised the problem of what the household weight should be and resulted in the release of a separate household weight (and for most years of the OHS it is unclear how the final household weight was calibrated[4]). The survey weights for the GHS have undergone a series of updates, during which StatsSA reweight the entire historical series. As of the most recent update in 2017, the procedure is to weight the person and household information in separate procedures. The person file includes a person weight calibrated on person demographic information from the Mid-Year Population Estimates (MYPE). The household file includes a household weight benchmarked using household headship rates from the 2001 census 10% samples, the 2007 Community Survey, and the 2011 census 10% sample. In other words, the information used to calibrate one weight, never enters the calibration of the other weight, effectively totally de-linking a representative person universe from a representative household universe.

This creates two overarching problems: noise is introduced into estimation, and and an incoherent conceptual framework is embedded in the data set. The first point relates to the fact that one can reach a different estimate for the same statistic, e.g. total household counts, depending on whether one uses the person or household weight, even though one is using the same sample. To illustrate this, we present Figure 1 which uses the OHS-GHS series to plot trends in some key demographic statistics using both weights. Looking at Panel A plotting total household counts, the data essentially provides two answers to the question 'how many households are there in South Africa in 2005?'. In this case, the household-weighted series has clearly been benchmarked so it might be clear that researchers are meant to choose the answer weighted by the household weight. The opposite is true for the counting up the total population in Panel B. However, this choice gets much murkier when considering statistics that are combinations of both person and household information, such as those plotted in Panels C and D.

[4]For example, metadata from years 1997-9 describes designing the household weight, but only post-stratifying the person weight (StatsSA 1997, 1998, 1999). In 1994, though, the weight of the household head was used as the household weight (Wittenberg 2008).

Figure 1: Trends in selected statistics in the OHS-GHS series



Notes: own calculations using the OHS-GHS series. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. StatsSA benchmark = Panel A: sourced from StatsSA Headship Model based on the census 10% samples for years 1996, 2001, and 2011 and the 2007 Community Survey; Panel B: StatsSA Mid-Year Population Estimates; Panels C + D: own calculations using the census 10% samples for 1996, 2001, and 2011, and the 2007 Community Survey and using the same definitions for the person and household sample as the StatsSA Headship Model. In Panel C, household size is the person-weighted population count divided by the household-weighted household count. In Panel D, the headship rate is the household-weighted number of households divided by the person-weighted count of the population aged 15+ years.

In Panel C, we plot average household size which could be described as a household attribute, but is defined by the count of people per household. In Panel D, we plot headship rates. Being a household head is a person attribute, but because there is one head per household in this data, the count of heads must tally with the count of households. It is much less obvious which series provides the 'correct' answer in Panels C and D; relatedly, neither series adequately meets census benchmarks. If researchers want to weight headship rates, for example, they are forced to confront the absurd-sounding choice of whether this is more of a person outcome or a household outcome? This question makes no sense because these outcomes cannot be neatly divided into person-only and household-only; they are inseparably related to each other. It is far from obvious how one is meant to mediate between these two weights in these cases.

This brings into sharp focus the second problem, which is that conceptually researchers have to choose between weights that either yield a representative person universe or a representative household universe, but not both at the same time. Estimates from the OHS or the GHS are only reliable for analysis that is person-level only, household-level only, or in-sample. But, weighted estimates from these data are not reliable and analysis is not conceptually coherent for any type of analysis that requires consistency

between the person and household level. Such consistency is a prerequisite for an extensive number of research questions and some of the most important for tracking the progress of the post-apartheid project. For example, this would include any type of welfare analysis using a statistic like per capita household income, which uses household information to draw a person-level conclusion. It also hampers extrapolation from individual receipt of wages or social grants (which the GHS is particularly good at collecting) to household-level welfare - an exercise that has been especially important for planning emergency relief for the Covid-19 pandemic. Analysis making the opposite inference is also affected; for example, how the roll-out of basic services to households (e.g. electricity) impacts individual welfare. Essentially, our ability to understand how the social connections captured by the household mediate a plethora of critical social outcomes - education, physical and mental health, employment - is undermined. The effect of compartmentalising people and households is therefore to reduce the number of research questions these data can be used to answer consistently and coherently. Since this procedure is embedded in more than 20-years-worth of data, such an erosion of data quality counts as a considerable cost both to the country's research agenda and to the efficiency of public spending.

4.2 Issues with the series of household count benchmarks

Besides the problems created by the two-weight system, there is another set of problems relating to the series of benchmarks StatsSA use for the household weights which are likely resulting in household counts that are too low. Household counts are not benchmarked in the OHS era, but the GHS household weight is based on benchmarks using the 2001 census, the 2007 Community Survey, and the 2011 census. We again present the series of total household counts in Figure 2, this time with annotations about corrections we have made in our Cross-Entropy Weight (CEW) Benchmark series.

4.2.1 The distorting effect of the Community Survey 2007

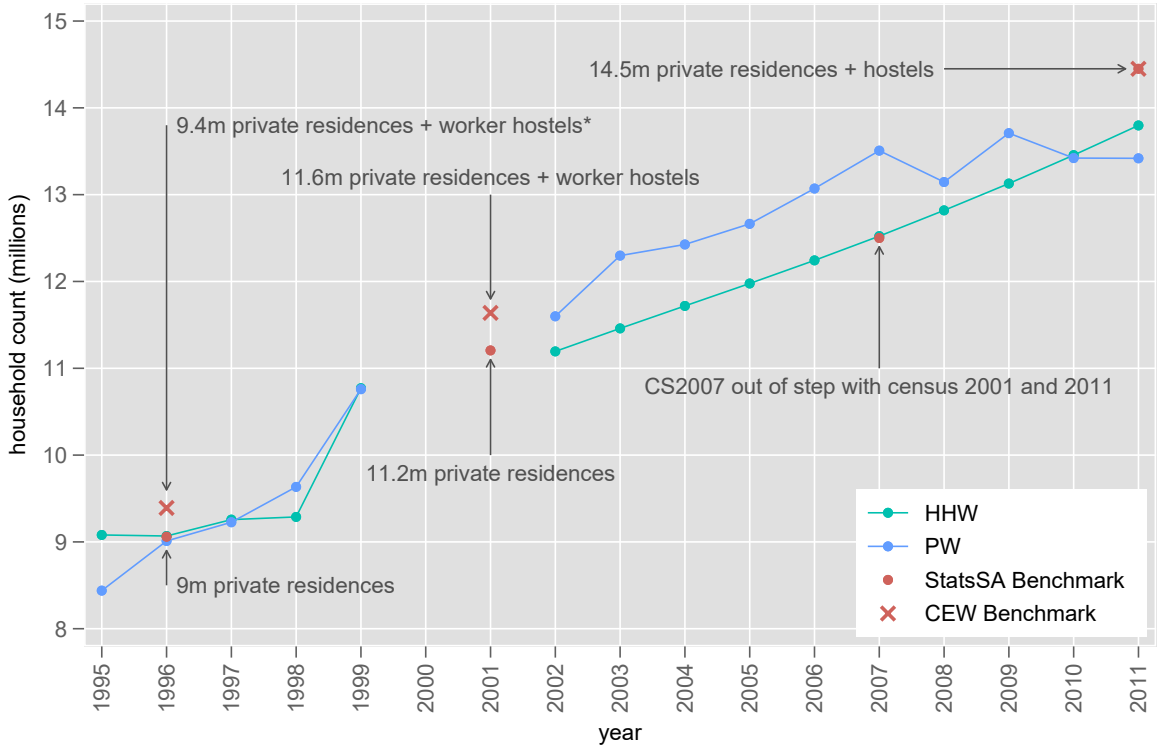
StatsSA include the 2007 Community Survey (CS2007) as a household count benchmark for the GHS household weight, even though it is not a census, but a larger-than-usual household survey. It is clear from Figure 2 that the CS2007 is out of step with the rest of the census and is distorting the household-weighted trend by pulling it lower than it should be. The CS2007 is also out of step with the census on a number of other statistics, especially single-person households. We omit this benchmark when constructing the CEW Benchmark series.

4.2.2 Inconsistent treatment of worker hostels

Another reason the GHS series of household counts is too low is because worker hostels were omitted from the household count for the 2001 benchmark, but included for 2011. The OHS and the GHS survey the 'population in households' which is the population that can sensibly be separated into distinct households. This includes people living in private residential units and people living as households in worker hostels.[5] The StatsSA benchmark count for 2001 only counted private residential households (11.2 million) and

[5]This excludes two other sub-populations: people living in institutions (e.g. hospitals, prisons, military barracks), and people living in collective living quarters that cannot be clearly delineated into separate households (e.g. student residences, residential motels, homes for the aged).

Figure 2: Inspecting the series of benchmarks for household counts



Notes: own calculations using the OHS-GHS series. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. StatsSA benchmark = Panel A: sourced from StatsSA Headship Model based on the census 10% samples for years 1996, 2001, and 2011 and the 2007 Community Survey. CEW Benchmark = own calculations using the census 10% samples for years 1996, 2001, and 2011. * households in worker hostels are not observable in the 1996 census 10% sample, this estimate is based on the 9 million private residences plus an inflation for worker hostels based on a projection from census 2001.

excluded the additional 400 000 households in worker hostels. In 2011 though, the benchmark is a count of 14.5 million households in private residential units and worker hostels. StatsSA have not benchmarked household counts in the OHS era, although the household weights in the 1996 OHS meet the census estimate. Note, however, that the 1996 census did not sample households within worker hostels so this census estimate is for private residences only. For our CEW Benchmark series, we construct a total household count estimate for 1996 that is consistent with the rest of the series in its inclusion of both private residences and worker hostels. Our estimate is based on back-projecting the share of worker hostels in the household count using census 2001, following a similar procedure by Machemedze et al. (2007). More detail on this inflation is in Section 5.

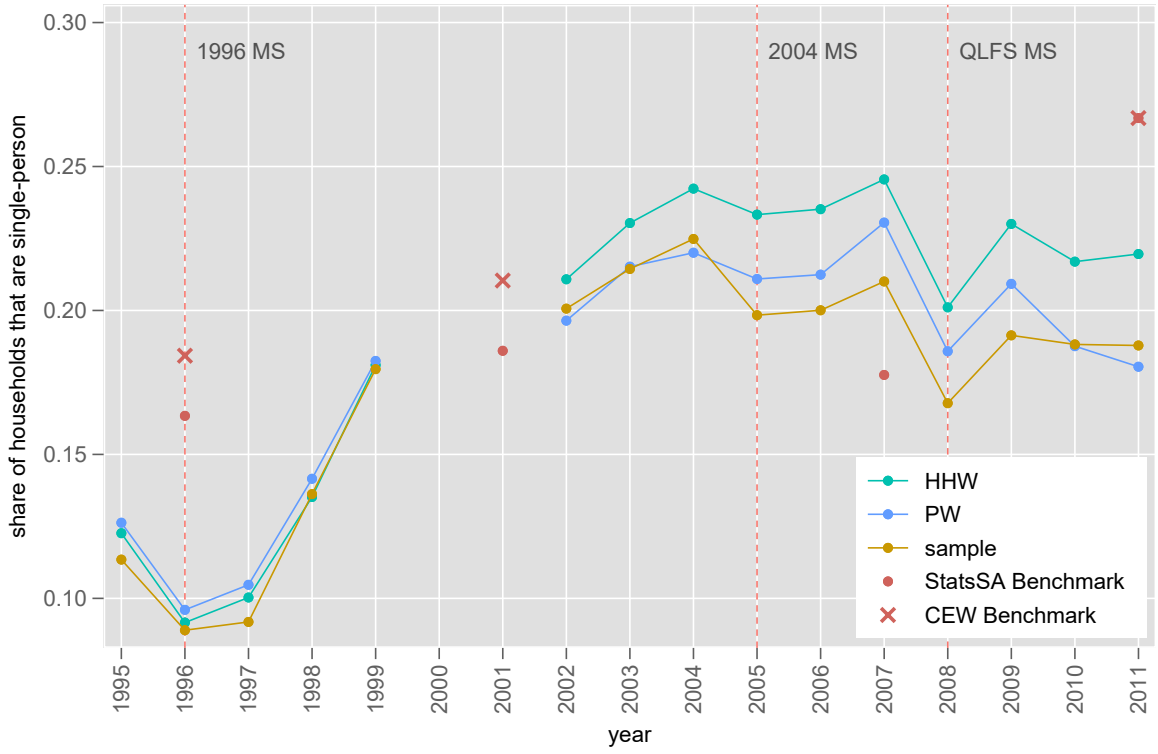
4.3 Undersampling of small households

Because the household is the unit that is sampled, attention should be paid to whether certain household types are being systematically missed. Small households have been chronically undersampled in the OHS and the GHS, and no compensation has been made in the calibration of the survey weights to account for

this. Figure 3 charts the share of households that are single-person in the OHS-GHS series, along with the StatsSA Benchmark and our updated CEW Benchmark. Change in Master Sample (MS) is indicated by red vertical lines. The gold sample line shows how single-person households have been underrepresented in the unweighted sample.

This problem originally arose in the 1994-1998 OHS owing to fieldwork design as documented by Kerr & Wittenberg (2015). Fieldworkers were instructed to only sample one household per dwelling unit with probability of selection proportional to size, meaning that smaller households (e.g. domestic workers living on an employer's property or backyard shack dwellers) were systematically missed. Single-person households occurred at a rate of 16% in the 1996 census and 18% when inflating for worker hostels; but, only 9% in the OHS 1996 household sample meaning they were occurring at half the rate they should have by the latter benchmark. From OHS 1999, fieldworker practise was improved by instructing fieldworkers to sample every household on a dwelling unit, amongst other changes (Kerr & Wittenberg 2015). In accordance with this change, the incidence of single-person households increases and moves closer to the benchmark's trend in 1999.

Figure 3: The share of households that are single-person in the OHS-GHS series



Notes: own calculations using the OHS-GHS series. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. sample = unweighted estimates in the OHS-GHS series. StatsSA Benchmark = own calculations using the census 10% household samples for the 1996, 2001, and 2011 census and the 2007 Community Survey. CEW Benchmark = own calculations using the census 10% household samples for the 1996, 2001 and 2011 census, and including an inflation of the 1996 estimate to account for worker hostels. Vertical lines correspond to change in Master Sample (MS).

The GHS period began with healthy levels of small-household sampling, particularly 2002-2004, when

sampling was still based on the Master Sample drawn up from the 1996 census. However, with the change to the 2004 Master Sample in 2005, the incidence of single-person households lost ground and the gap between the sample and the CEW Benchmark widened substantially over time. This drop-off happened again with the onset of the QLFS Master Sample in 2008, so that by 2011, single-person households occurred in the GHS at a rate of about 18%, when they occurred at a rate of 27% in the census. Each new Master Sample in the GHS seems to adjust the series further downwards; although, the trend is increasing within each Master Sample (with the exception of the QLFS Master Sample which seems flat). Both the 2004 and QLFS Master Samples are based on the 2001 census (as is the sampling for CS2007). The 2008 GHS stands out for having particularly poor sampling of single-person households related to a higher-than-usual rate of dwelling units being missed in the collection of this survey, in general (StatsSA 2008).

Unfortunately, the StatsSA household weight (HHW) does not assist in correcting this series and instead mirrors the trend traced by the unweighted sample. Looking at the HHW series in isolation, one would struggle to identify that the incidence of single-person households was consistently increasing as in the CEW Benchmark series. The HHW series moves incongruously with the CEW Benchmark trend and instead overestimates the share of single-person households in the early GHS, only to seriously underestimate it later on, and muddy the direction of the trend in the process.

In this paper, our goal is to recalibrate the weights and account for each of these three problems. Firstly, we overcome the two-weight system by consolidating the information StatsSA use in two separate procedures, into one procedure. Secondly, we update the series of household benchmarks to be the CEW Benchmark series motivated in Figure 2. This series omits the distorting effect of CS2007 and consistently includes worker hostels in all years. Thirdly, we explicitly constrain the calibration on the number of one-, two-, and three-person households in order to mitigate the impact of the undersampling of small households. These changes are consolidated into a single cross-entropy calibration procedure, which we describe next.

5 Method

5.1 The calibration procedure

Calibrated weights can be solved for by minimising a distance function between estimated population totals using the design weights and known population totals from an auxiliary data source (Deville 2000, Deville & Särndal 1992). The distance function can be defined in various ways, such as an a linear function (e.g. generalised regression, iterative raking); a multiplicative function (e.g. calibration to marginal totals in the CALMAR algorithm); or a minimum entropy measure. We use cross-entropy to calibrate the weights, meaning we use an entropy measure as our distance function. This technique gives identical solutions to iterative raking and, because it is run using a Stata program, compiles very quickly. This technique and the applied program we use, *maxentropy* from Wittenberg (2010), has been used to calibrate the National Income Dynamics Study (Wittenberg 2009) and the Post-Apartheid Labour Market Series (Kerr et al. 2019).

Given moment constraints and a normalisation restriction, cross-entropy minimises the information

($\mathbf{I}(\mathbf{p}, \mathbf{q})$) needed to make the new distribution of weights ($\mathbf{p}$) resemble as much as possible information we already have about what it should look like, which could be a distribution of existing StatsSA calibrated or design weights ($\mathbf{q}$). This can be formalised as:

$$\mathbf{I}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \ln(\frac{p_i}{q_i}) \tag{1}$$

where $\mathbf{I}(\mathbf{p}, \mathbf{q})$ is minimised subject to:

$$y_j = \sum_{i=1}^{n} X_{ji} p_i, j = 1, ..., J \tag{2}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{3}$$

where there are J population moments; y_i is the population mean of the random variable X_j; and equation 3 is the normalisation restriction. This constrained optimisation problem can be solved by maximising the unconstrained dual cross-entropy objective function:

$$L(\lambda) = \sum_{j=1}^{J} \lambda_j y_j - ln[\Omega(\lambda)] = \mathbf{M}(\lambda) \tag{4}$$

where $\Omega(\lambda)$ is given by:

$$\Omega(\tilde{\lambda}) = \sum_{i=1}^{n} q_i \exp(\mathbf{x}_i \tilde{\lambda}) \tag{5}$$

Golan et al. (1997) show that this function behaves much like maximum likelihood. The function $\mathbf{M}$ can be characterised as an expected log likelihood where $\mathbf{p}(\lambda)$ is the exponent and the parameter is λ. The λ output acts as a measure of how informative the constraints were; or, how different the new distribution of weights, $\mathbf{p}$, is relative to the underlying design weight distribution, $\mathbf{q}$.

5.2 The practical application

The sample for each year is obtained from the publicly accessible data set downloaded from South African data repository DataFirst's website (StatsSA 2010-2013, 2011-2018*b*). We use the *maxentropy* command from Wittenberg (2010) in Stata 16 to practically carry out the estimation. Table 1 summarises the sources and breakdown of various inputs for the calibration procedure. The distribution of prior information, $\mathbf{q}$, comes from the design weights in the case of the GHS, and the calibrated weights in the case of the OHS. Design weights are not publicly released for either the OHS or the GHS. The design weights for the GHS period 2002-2011 were received in private correspondence between Martin Wittenberg and StatsSA. As a result of only having weights for this period, we only run the recalibration up until 2011.[6]

The constraint matrix consists of 115 constraints, which enter as proportions of the total person pop-

[6]Although we plan to continue investigating, it appears there are no design weights still on record for the OHS at StatsSA.

Table 1: Breakdown and sources of constraints

	OHS	**GHS**
q	sampling weight	design weight
63 age-sex-race person	Back projections from DataFirst	MYPE
8 province	Back projections from DataFirst	MYPE
32 age-sex-race HH head	census 10% samples + 1996 inflated for WH	
12 race-hh size HH head	census 10% samples + 1996 inflated for WH	

Notes: MYPE = StatsSA Mid-Year Population Estimates. WH = worker hostels. hhsize = one-, two-, and three-person households. Headship constraints are interpolated between census benchmarks.

ulation: 63 individual age-sex-race categories; 8 province categories; 32 age-sex-race household headship categories; and, 12 race-household size household headship categories (being one-, two- and three-person households).[7] The individual-level and province population moments come from StatsSA (2018*a*,*c*) MYPE, which are publicly available for the period 2002 to present on the StatsSA website (i.e. the GHS period). Population estimates for the OHS period came from back projections of the population from DataFirst (2018). This was necessary as StatsSA do not provide disaggregated-enough population estimates for our purposes for these years.[8]

The household moments come from our adjustment to household headship rates calculated by StatsSA for their household weight model, the StatsSA Headship Model, also received by private correspondence between StatsSA and Martin Wittenberg. StatsSA calculate these headship rates using the ten percent samples of the three censuses (1996, 2001 and 2011), as well as, the 2007 Community Survey. For reasons explained previously in Section 4, we omit the 2007 Community Survey. We also update the benchmark for 2001 to include worker hostels. This could not be done for 1996 because the census was not set up to collect households living in hostels. Following Machemedze et al. (2007), we use 2001 to inform us about what the levels of worker hostels might have been in 1996. The household headship constraints are based on 32 age-sex-race categories for household heads. We calculate the share of each category who were living in worker hostels in 2001 and use this proportion to inflate the constraints for 1996 accordingly. The biggest revision was to male Black African heads aged 0-34 years whose headship rate increased by 8.4%. To the extent that the share of people living in worker hostels declined between 1996 and 2001, this will be an underestimate. Once benchmarks for 1996, 2001 and 2011 have been calculated, we interpolate headship rates for intervening years.

Effectively, constraints enter as shares the new weights must reflect when adding up to the overall person total. The total household count is then inserted into the constraint matrix via the benchmarked share of the total population who are household heads. This series of cross-entropy weights is then incorporated into the OHS-GHS series. There are now three weights in the OHS-GHS series: the StatsSA person weight (PW); the StatsSA household weight (HHW); and the new cross-entropy weight (CEW).

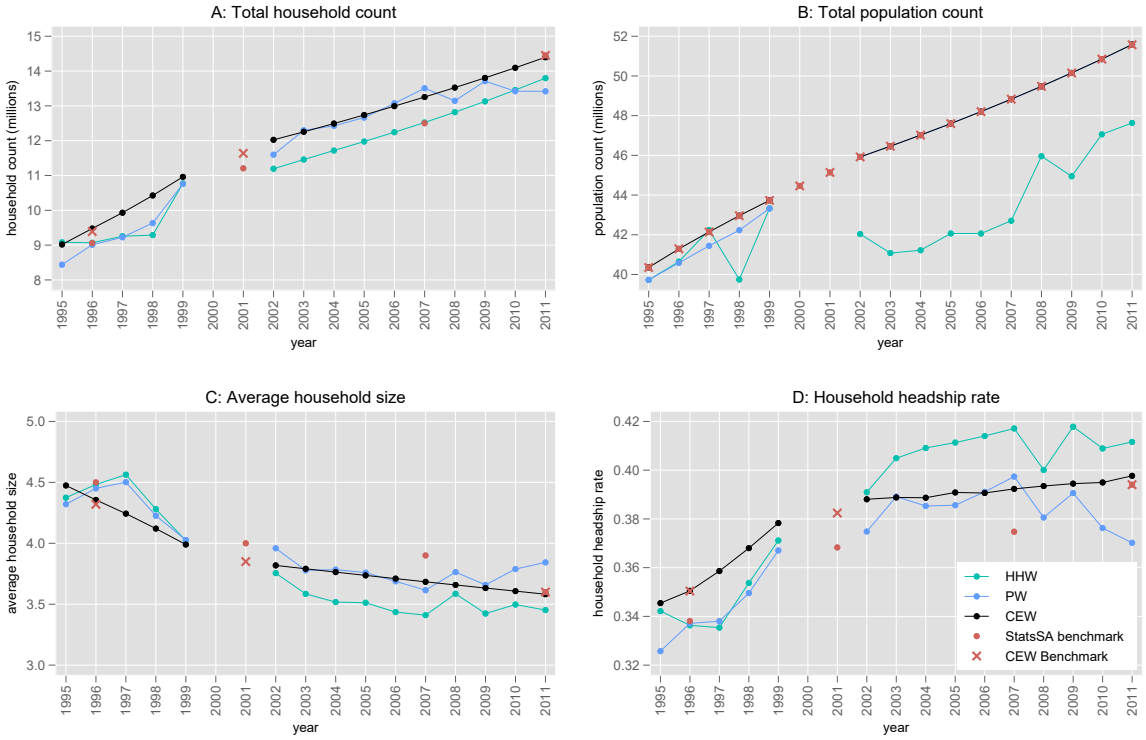
[7]The 64th individual and 9th province category are excluded as these sets are mutually exclusive. There are eight age categories for individuals in intervals of ten years, beginning with 0 - 9 years and ending with those aged 80 years and above. There were four age categories for the household heads. These were 0 - 34 years; 35 - 49 years; 50 - 64 years; and 65 years and older. For the race-household size categories there were four race groups by three household sizes, being one-, two, and three-person, leading to 12 categories.

[8]The population was back-projected by applying an exponential model of population growth to StatsSA MYPE from 2002 onwards and using population growth rates from the Actuarial Society of South Africa (ASSA).

6 Results

There are a number of ways to test the validity of the new weights, and not all of these are reported in this paper for reasons of space. For the purposes of this paper, we limit our description to how the CEW performs on producing series of key statistics, like counts of people and households. Figure 4 is a repeat of Figure 1, but with the addition of the CEW and its benchmark series. The numbers behind these figures for person and household counts are reported in Table 2 for a more detailed comparison. Quite clearly, the CEW outperforms the StatsSA weights in several aspects: it meets its benchmarks; it produces a smooth trend simultaneously for people and households; it produces a smooth trend for 'combination' variables in Panels C and D, implying consistency between people and households. The CEW series of household counts differs from the HHW series by being higher throughout the time period; and also benchmarked in the OHS era.

Figure 4: Performance of the CEW on key demographic statistics



Notes: own calculations using the OHS-GHS series. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. CEW = cross-entropy weight. StatsSA benchmark = Panel A: sourced from StatsSA Headship Model based on the census 10% household samples for years 1996, 2001, and 2011 and the 2007 Community Survey; Panel B: StatsSA Mid-Year Population Estimates; Panels C + D: own calculations using the census 10% samples for 1996, 2001, and 2011, and the 2007 Community Survey and using the same definitions for the person and household sample as the StatsSA Headship Model. CEW Benchmark = Mid-Year Population Estimates (2002-2011) and back-projections from DataFirst (1995-2001) in Panel B. Panels A, C and D: own calculations using the census 10% samples for 1996, 2001, and 2011, inflating the 1996 and 2001 estimate for worker hostels. Average household size benchmark = the person-weighted population count divided by the household-weighted household count. Headship rate benchmark = the household-weighted number of households divided by the person-weighted count of the population aged 15+ years.

Table 2: Counts of households and population in the OHS-GHS and its benchmarks

	Household counts (millions)					Population counts (millions)			
year	**PW**	**HHW**	**CEW**	**StatsSA Bm**	**CEW Bm**	**PW**	**HHW**	**CEW**	**MYPE**
1995	8.44	9.08	9.02			39.73	39.72	40.35	40.4
1996	9.01	9.07	9.48	9.06	9.39	40.58	40.64	41.29	41.3
1997	9.23	9.26	9.93			41.44	42.24	42.14	42.1
1998	9.63	9.29	10.43			42.23	39.74	42.96	43.0
1999	10.76	10.77	10.96			43.33	43.32	43.73	43.7
2000									44.5
2001				11.21	11.64				45.1
2002	11.60	11.19	12.03			45.92	42.04	45.92	45.9
2003	12.30	11.46	12.26			46.46	41.08	46.46	46.5
2004	12.43	11.72	12.49			47.02	41.22	47.02	47.0
2005	12.66	11.98	12.74			47.60	42.07	47.60	47.6
2006	13.07	12.24	12.99			48.20	42.06	48.20	48.2
2007	13.51	12.52	13.25	12.50		48.83	42.71	48.83	48.8
2008	13.15	12.82	13.53			49.48	45.96	49.48	49.5
2009	13.71	13.13	13.81			50.15	44.95	50.15	50.2
2010	13.42	13.46	14.09			50.85	47.06	50.85	50.9
2011	13.42	13.80	14.40	14.45	14.45	51.57	47.63	51.59	51.6

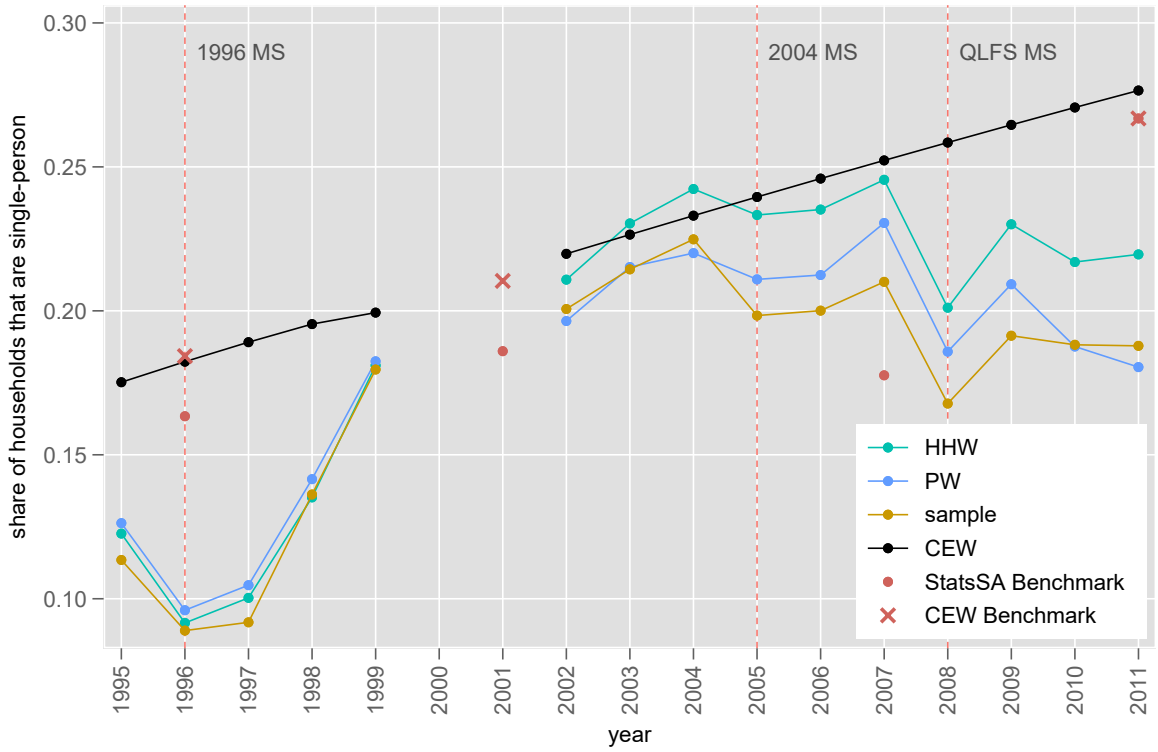
Notes: own calculations using the OHS-GHS series. PW = StatsSA person weights. HHW = StatsSA household weight. CEW = cross-entropy weight. StatsSA Bm (benchmark) = household counts from the StatsSA Headship Model based on the census 10% household files for 1996, 2001 and 2011, and the 2007 Community Survey. CEW Bm (benchmark) = household counts using the census 10% household files for 1996, 2001 and 2011 and including the population living in worker hostels in 1996 and 2001. MYPE = StatsSA Mid-Year Population Estimates.

Figure 5 reproduces Figure 3 for single-person households, this time including the CEW trend. Numbers for this figure are reported in Table 3. The CEW yields a much more reliable series that coheres with the benchmarked trend; i.e. one that is clearly increasing over time. Table 3 reports that by 2011, the HHW is underestimating the count of single-person households by about 800 000 households, or 25%.

6.1 Robustness on other variables

So far, we have compared the weights' performance on variables relating to person and household counts. These were variables on which the weights were explicitly constrained, meaning the weights should produce well-behaving trends unless something went wrong with the calibration. A good robustness check, then, is to test out the weights on statistics with which they were not constrained. Figure 6 presents the share of adults who are married or cohabiting in Panel A and the number of households with electricity as their main source of energy in Panel B. None of the weighted series perfectly matches the census estimates; however, we would argue that the CEW is performing best compared to the HHW and PW in both these examples. Panel A shows that married people appear to be oversampled in the OHS-GHS series. It is plausible that enumerators are more likely to find one partner of a married couple at home to enumerate, compared to non-married people. Whilst all three weights overestimate the share of married people versus the census, the CEW is the least biased. CEW does not appear to perform categorically better than the HHW or the PW in Panel B, but seems to produce a slightly smoother trend. Overestimation of the number in 1996 on the part of the CEW could be related to the inclusion of worker hostels

Figure 5: The performance of the CEW in plotting the share of households that are single-person in the OHS-GHS series



Notes: own calculations using the OHS-GHS series. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. CEW = cross-entropy weight. sample = unweighted estimates in the OHS-GHS series. StatsSA Benchmark = own calculations using the census 10% household samples for the 1996, 2001, and 2011 census and the 2007 Community Survey. CEW Benchmark = own calculations using the census 10% household samples for the 1996, 2001 and 2011 census, and including an inflation of the 1996 estimate to account for worker hostels. Vertical lines correspond to change in Master Sample (MS).

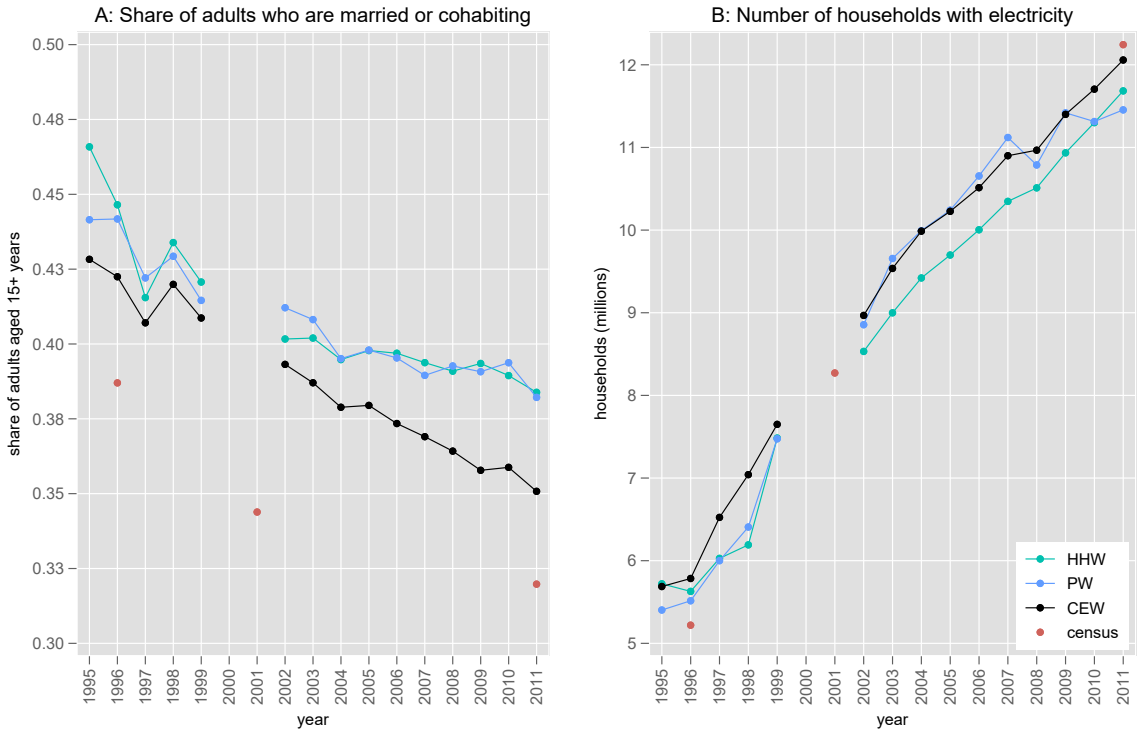
in the CEW estimate, but not in the census estimate.

Table 3: Counts of single-person households in the OHS-GHS and its benchmarks (millions)

year	PW	HHW	CEW	Sample*	StatsSA Bm	CEW Bm
1995	1.07	1.11	1.58	3 370		
1996	0.87	0.83	1.73	1 416	1.48	1.67
1997	0.97	0.93	1.88	2 737		
1998	1.36	1.25	2.04	2 582		
1999	1.96	1.95	2.19	4 694		
2000						
2001					2.08	2.48
2002	2.28	2.36	2.64	5 261		
2003	2.65	2.64	2.78	5 655		
2004	2.73	2.84	2.91	5 889		
2005	2.67	2.79	3.05	5 575		
2006	2.78	2.88	3.20	5 602		
2007	3.11	3.07	3.34	6 142	2.22	
2008	2.44	2.58	3.50	4 083		
2009	2.87	3.02	3.65	4 842		
2010	2.52	2.92	3.81	4 811		
2011	2.42	3.03	3.98	4 714	3.86	3.86

Notes: own calculations using the OHS-GHS series. PW = StatsSA person weights. HHW = StatsSA household weight. CEW = cross-entropy weight. Census + CS2007 = census 10% household samples for 1996, 2001, and 2011 and the Community Survey in 2007 - note the StatsSA weights are not benchmarked on this series. CEW Benchmark = census 10% household samples for 1996, 2001, and 2011, with the 1996 estimate inflated to account for worker hostels. *Counts are as is and not scaled to be in millions.

Figure 6: The performance of the CEW in plotting non-constraining variables



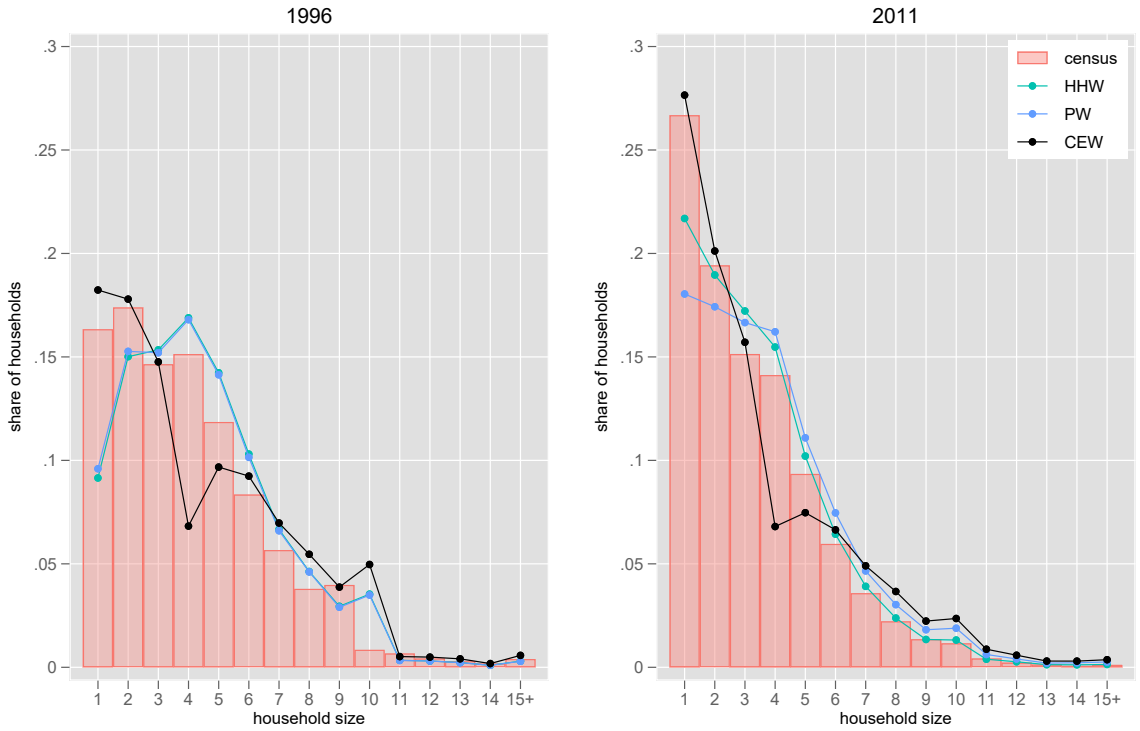
Notes: own calculations using the OHS-GHS series and census 10% samples. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. CEW = cross-entropy weight. We use the census person weight in Panel A and the census household weight in Panel B.

7 Limitation: distorting the distribution of household size

Previously, we explained how calibration can cause new problems. If the restrictions placed on the calibration are too onerous, this can cause the calibration procedure to break, or cause distortion in other ways. Here, we present one instance of distortion caused by the CEW calibration of which we are aware. In Figure 7, we compare the distribution of household size in the census 10% household files to the OHS-GHS, differently weighted. Clearly, the HHW and PW underestimate the share of one- and two-person households relative to the census in both years: in 1996, for example, the HHW underestimates the share of one-person households by about 7 percentage points. The HHW and PW then start to slightly overestimate the share of households sized four to about eight, before settling down after a heap at households of size ten. The underestimation of small households is corrected in the CEW distribution, which specially constrained on the number of one-, two-, and three-person households. However, this appears to have come at the cost of offsetting the share of four- and five-person households, which are now severely underestimated in both years.[9] The CEW then seems to overestimate the shares of larger households slightly more than the HHW and PW. The CEW clearly outperformed the HHW and PW in trends previously presented. But, in the case of the distribution of household size, all weights produce sub-par distributions, for different reasons. The choice is between an underestimation of small or medium-sized households. This is the case for now, at least, while we continue to work on the weights.

[9] In a previous version of the CEW, where we only constrained on one- and two-person households, the result was to seriously underestimate the share of three- and four-person households.

Figure 7: Distribution of household size in 1996 and 2011



Notes: own calculations using the OHS-GHS series and census 10% household samples for 1996 and 2011. HHW = StatsSA household weight. PW = StatsSA person weight. All StatsSA weights are from the latest 2017 reweighting. CEW = cross-entropy weight. No adjustment has been made to the 1996 census for worker hostels.

8 Conclusion

In this paper, we have documented key weaknesses of the survey weights StatsSA release with the OHS and the GHS and discussed their implications for different types of research. This is information about the quality of the OHS and GHS of which any researcher aiming to use these data should be aware. The most critical weakness of the weights is that the calibration of people is separated from the calibration of households, which compartmentalises a representative person universe from one that is representative of the households in which the same people live. A weakness that is of particular relevance to policy-makers is that the existing weights are probably underestimating total household counts by the order of about 5%. A final weakness is that the existing weights do not compensate for the chronic undersampling of small households, which are becoming a more and more important constituency over time. We presented a new series of survey weights which target the the key weaknesses with the existing weights. Our new cross-entropy weight is mutually representative of people and households at the same time and is able to produce a smooth series of person and household counts that closely agree with benchmarks, including for one-, two-, and three-person households.

When assessing the success of our cross-entropy weight, it is important to understand the limits of what we set out to achieve and of what we could possibly achieve with reweighting. Reweighting a data

set is not a panacea for every aspect of data quality. When facing deep underlying sampling problems, like low or non-coverage, the power of reweighting to solve certain types of problems is limited. For example, we showed that although the cross-entropy weight was less biased than the StatsSA weights, it could not completely overcome the undersampling of not-married people in the OHS-GHS series. Similarly, we discovered that the cross-entropy weight compensated for our correction of the shares of small households, by distorting and underestimating the shares of medium-sized households. This is a weak point for our cross-entropy weight that we will continue to work on. Nonetheless, we are able to show that our new cross-entropy weight performs better that the two StatsSA weights on a number of key outcomes; and provides the first series of benchmarked total household counts extending back into the OHS era.

References

Branson, N. & Wittenberg, M. (2014), 'Reweighting South African national household survey data to create a consistent series over time: a cross-entropy estimation approach', *South African Journal of Economics* **82**(1), 19–38.

DataFirst (2015), South Africa - General Household Survey 2005 - Metadata, Technical report, University of Cape Town, Cape Town.

DataFirst (2018), Projected Population Distribution: 1990-2001, Mimeograph, University of Cape Town, Cape Town.

Deaton, A. (1997), *The analysis of household surveys: A microeconometric approach to development policy*, The Johns Hopkins University Press, United States.

Deville, J.-C. (2000), Generalized calibration and application to weighting for non-response, *in* 'COMPSTAT', Springer, pp. 65–76.

Deville, J.-C. & Särndal, C.-E. (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association* **87**(418), 376–382.

Golan, A., Judge, G. & Miller, D. (1997), The maximum entropy approach to estimation and inference, *in* 'Applying Maximum Entropy to Econometric Problems', Emerald Group Publishing Limited, pp. 3–24.

Kerr, A., Lam, D. & Wittenberg, M. (2019), 'Post-Apartheid Labour Market Series: 1993-2019 [dataset]', University of Cape Town: DataFirst [producer and distributor]. Version 3.3.

Kerr, A. & Wittenberg, M. (2015), 'Sampling methodology and fieldwork changes in the October Household Surveys and Labour Force Surveys', *Development Southern Africa* **32**(5), 603–612.

Lavallée, P. & Beaumont, J.-F. (2015), 'Why we should put some weight on weights', *Survey Methods: Insights from the Field* . Special Issue 'Weighting: Practical Issues and 'How to' Approach'. Invited Article.

Machemedze, T., Kerr, A. & Wittenberg, M. (2007), Recalibrating the OHSs to adjust for sampling changes, DataFirst Technical Paper No. 28, DataFirst, University of Cape Town, South Africa.

Smith, T. M. (1991), 'Post-stratification', *Journal of the Royal Statistical Society: Series D (The Statistician)* **40**(3), 315–323.

StatsSA (1997), October Household Survey 1997: Metadata, Technical report, Statistics South Africa (StatsSA), Government of South Africa, Pretoria, South Africa.

StatsSA (1998), October Household Survey 1998: Metadata, Technical report, Statistics South Africa (StatsSA), Government of South Africa, Pretoria, South Africa.

StatsSA (1999), October Household Survey 1999: Metadata, Technical report, Statistics South Africa (StatsSA), Government of South Africa, Pretoria, South Africa.

StatsSA (2008), General Household Survey 2008 - Statistical Release P0318, Technical report, Statistics South Africa (StatsSA), Government of South Africa, Pretoria, South Africa.

StatsSA (2010-2013), 'October Household Survey: 1995-1999 [datasets]', Pretoria: Statistics South Africa (StatsSA) [producer]. University of Cape Town: DataFirst [distributor]. Version 1.1 [1996-1999]; Version 1.2 [1995].

StatsSA (2011-2018*b*), 'General Household Survey: 2002-2015 [datasets]', Pretoria: Statistics South Africa (StatsSA) [producer]. University of Cape Town: DataFirst [distributor]. Version 1 [2013]; Version 1.1 [2011, 2014]; Version 1.2 [2002, 2015]; Version 1.3 [2003-2009]; Version 2.1 [2010]; Version 2 [2012].

StatsSA (2018*a*), Country projection by population group, sex and age (2002-2017) [Excel file], 'Additional download' related to publication P0301-Mid-year population estimates, Statistics South Africa (StatsSA), Government of South Africa, Pretoria, South Africa. Retrieved from http://www.statssa.gov.za/?page$_id = 1854PPN = P0302$.

StatsSA (2018*c*), Provincial projection by sex and age (2002-2017) [Excel file], 'Additional download' related to publication P0301-Mid-year population estimates, Statistics South Africa (StatsSA), Government of South Africa, Pretoria, South Africa. Retrieved from http://www.statssa.gov.za/?page$_id = 1854PPN = P0302$.

Wittenberg, M. (2008), October Household Survey 1994, Technical report, DataFirst, University of Cape Town.
URL: *https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/407/download/5258*

Wittenberg, M. (2009), Weights: Report on NIDS Wave 1, NIDS Technical Paper no. 2, National Income Dynamics Study (NIDS), University of Cape Town, South Africa.

Wittenberg, M. (2010), 'An introduction to maximum entropy and minimum cross-entropy estimation using Stata', *Stata Journal* **10**(3), 315.