Guide to version 1 of the Post-Apartheid Socio-Economic Series

Andrew Kerr
School of Economics
University of Cape Town
andrew.kerr@uct.ac.za

Amy Thornton
SALDRU & DataFirst
University of Cape Town
amy.thornton@uct.ac.za

Martin Wittenberg*
DataFirst &
School of Economics
University of Cape Town

November 2025

*Acknowledgment

We acknowledge the special contribution of the late Prof. Martin Wittenberg who sadly passed away during this project. Prof. Wittenberg conceptualised and provided technical guidance for this project, although his role as a teacher, researcher and mentor more broadly was highly valued and is sorely missed.

Contents

1 Introduction

The Post-Apartheid Socio-Economic Series (PASES) is a stacked series of 28 nationally-representative cross-sectional household surveys covering every year between 1995-2024, with the exception of 2000 and 2001. The source data consists of the October Household Surveys (OHSs) collected every year by Statistics South Africa (StatsSA) between 1995-1999[1] and the General Household Surveys (GHSs) collected every year by StatsSA commencing in 2002 and continuing until the most recent publicly available survey, which is currently 2024 (StatsSA 2010-2013, 2011-2025). These surveys include both household and person information and the series currently consists of just over 640 000 households and their 2.4 million members. A subset of variables from the surveys have been included and harmonised across time. Along with some new and derived variables, there are currently 119 variables in the data set covering social, demographic and economic outcomes.

The construction of this series is based on the experience of creating the Post-Apartheid Labour Market Series (PALMS) which is a stacked and harmonised dataset of StatsSA's labour market surveys by Kerr et al. (2025) for DataFirst. PALMS demonstrated the usefulness of combining otherwise separate surveys and paying attention to particular data quality issues when stacking cross-sections over time. The strength of PASES is its coverage of a variety of socio-economic and demographic outcomes in one dataset at both the person and household level. For example, there are specific modules or question sets in the source survey questionnaires on person demographics, household structure, education, health, welfare and social grants, basic household services, labour market activity, amongst others.

Changes to the survey questionnaire over time as well as problems with how the source surveys have been weighted have presented the main harmonisation and data quality challenges. Additionally, the breadth of topics covered in general socio-economic surveys makes fully harmonising all these subject areas at once a highly effort- and time-intensive task. Instead it is more practical to harmonise the surveys incrementally by subject area. The focus of this first version of PASES has been on the labour market variables, mainly being labour market status and the harmonising, weighting and imputing of earnings information. We also consolidate a general set of variables on demographic and household characteristics, geography, education, household income, social grants, fertility, and transport. In future versions, we hope to incorporate a wider range of variables.

The data are released in two files, a main file and an extra income variables file available on DataFirst along with the do files to create the data. If you use PASES, please cite the data and this guide if you use it. The data can be cited as follows: **Kerr, A., Thornton, A. and Wittenberg, M. (2025) Post-Apartheid Socio-Economic Series [dataset]. Version 1. Cape Town: DataFirst [producer and distributor]**

The next few sections describe core contributions in terms of harmonisation and derived variables, as well as important information and guidance regarding the survey weights. A specific contribution of PASES is a more reliable survey weight until 2011, but unfortunately we do not currently provide this weight for the full series due to data quality issues related to the 2022 Census (Moultrie & Dorrington 2024, Budlender & Thornton forthcoming, Thornton 2025). After describing these key aspects, we provide a description of how PASES was constructed and a full variable list with definitions.

[1] StatsSA collected the OHS from 1993 but the series uses the OHS from 1995. Sampling issues distinct to the earlier OHS's made reweighting the surveys particularly difficult.

2 Description of selected variables

2.1 Earnings

Most researchers use the StatsSA labour market surveys to investigate earnings, but in recent years warnings have been issued about the quality of the imputed earnings data in the Quartlerly Labour Force Surveys (QLFS) beginning in 2010 Quarter 1 and particularly from 2012 Quarter 3 (Kerr & Wittenberg 2017, 2019*a*,*b*, 2021, Kerr 2025). The earnings information in the GHS is unimputed and Kerr (2025) shows that the earnings in the GHS appear reliable and cohere with trends in (the more trustworthy years of) the labour market surveys and South African tax data. The GHS also provides earnings information for 2008 and 2009, two years for which StatsSA has not provided earnings information in the QLFS (via the Labour Market Dynamics Series data).

Respondents can provide earnings information in the OHS and GHS either as a point estimate or as an earnings bracket. About a third of earners in PASES provide bracket responses, often high earners. There are different ways to deal with this mixed data type (point estimate vs brackets) in the earnings data, the merits of which are discussed by Wittenberg (2017). Two different approaches are incorporated in PASES: midpoints and reweighting. In the main dataset, we provide a cleaned nominal and real earnings variable with midpoints for bracket responders, called earningsMP and realearningsMP respectively. The real variable is pegged to December 2023 as the base to be comparable with PALMS v4. Using midpoints for bracket responders has been shown to reproduce the global mean of the distribution and have little practical effect on coefficient estimates (Wittenberg 2017, Von Fintel 2007). This is probably most useful for the types of analysis for which PASES might be used, especially considering its coverage of other non-labour market related topics.

However, midpoints are less well-suited to analysis of the earnings distribution itself (Wittenberg 2017). Imputing midpoints tends to create artificial spikes at the midpoint values in the earnings distribution and this clustering can shift percentile values with implications for measurement of earnings inequality, for example. Wittenberg (2017) shows that a reweighting approach outperforms the midpoint approach in this case. This reweighting approach is incorporated in PALMS in the form of a bracketweight to be used in conjunction with the earnings point estimates. The bracketweight is constructed by multiplying the inverse probability of a bracket response in a given bracket in a given year by the person weight for that observation (Wittenberg et al. 2008). If there are relatively more bracket responders in a given bracket, the bracketweight will relatively up-weight point estimate responders in that bracket interval. Reweighting also circumvents the problem faced by midpoint imputation of how to code earners in the top open-bracket category.[2]

In case users are interested in measuring earnings inequality or other analysis of the earnings distribution and in order to facilitate comparability with PALMS, we also include the reweighting approach in PASES. A nominal and real earnings variable of point estimates only, called earnings and realearnings respectively, plus a bracketweight is provided in a separate data file called pasesv1_extraincome.dta. We save these variables in a separate dataset in order to remind users that the variables in this dataset are a different approach to dealing with bracket responders and the two approaches - midpoints and reweighting - should not be mixed up. To reiterate, if using the earnings variable with midpoints, users must apply

[2]We multiply the open bracket by 1.5.

the regular weights in the main dataset. If using the point estimates only in the extra income dataset, the bracketweight must be applied. Choice of approach is guided by one's research question. If one is investigating earnings as a main outcome of interest, the reweighting approach may be more appropriate. However, if one is interested in how other outcomes cohere with earnings, researchers must be aware that using point estimates without midpoints drops about a third of earners in the dataset and the impact of this one one's sample may not be appropriate.

Note that midpoints and the bracketweight adjust for bracket responders, but not other kinds of non-response such as refusals or otherwise missing answers. To account for this other type of non-response, we have included an additional set of imputed earnings variables discussed later on in the extra income file.

Note also that no earnings point estimates were collected in the 1996 OHS. Earnings were only collected in brackets and we do not provide these brackets. Users can merge them in from the OHS 1996 if they are interested.

2.1.1 Outliers

A flag for earnings outliers called outlier is included in the dataset and is analogous to the variable of the same name in PALMS. An individual's earnings are flagged as an outlier if the studentised residual from a Mincerian earnings regression had an absolute value of greater than 5. The earnings regression was a regression of log earnings on an interaction of gender and year, an interaction of population group and year, years of education, age, age squared, province, and metro. Studentised residuals are residuals normalised against their standard deviation from a regression in which that observation has been left out. Outliers make a material difference to a time series of mean earnings. Outliers are imputed for in the imputed earnings variables in the in the extra income variables file.

2.1.2 Imputed earnings

Above we discussed two approaches to dealing with bracket responders, midpoints or reweighting, but neither of these approaches deal with otherwise missing data. There is a significant increase over time in the share of employed who have missing earnings information either because they replied "Don't Know" or "Refused" to the earnings questions or because it is otherwise unspecified or missing. Rates of missingness are around 7-10% in the early 2000s but climb over 20% in 2016. In 2024, 27% of the employed had missing earnings information. These rates are comparable to the 2018-2020 QLFSs (Kerr 2025).

Multiple imputation is the correct way to deal with missing data. The main advantage of *multiply* imputing as opposed to single imputation is to improve the accuracy of standard errors by more correctly describing the loss of precision that comes with imputing data (Van Buuren 2018). PALMS version 2 was released with a set of multiple earnings imputations and in the ten years since these imputations have been available, they have hardly been used according to PALMS citations. In practice, using multiply imputed data is complicated and somewhat limited in its applications.

We have therefore taken the pragmatic decision to provide only one set of imputations for missing earnings. The imputations are provided in the extra income dataset in the form of two variables depending

on the preferred approach to bracket responders. The first variable is our real earnings variable with midpoints now also with missing data imputed, called imputed_realearningsMP. This variable is now completely non-missing for all employed in the GHS period and should be used with the usual survey weights. The second variable is the real earnings variable (point estimates only) with the same set of imputations for missing earnings, called imputed_realearnings. This variable is now only missing bracket responders and should be used with the bracketweight.

We imputed earnings for any employed person who had neither a point estimate nor bracket response, or who was estimated to be an outlier for the GHS years (2002 onwards). Earnings was imputed per year using predictive mean matching, a semi-parametric method that uses a nearest neighbour approach similar to hot deck imputation and which is usually the preferred method for imputing continuous variables (Van Buuren 2018). Predictors for the imputation were gender, education, population group, a quadratic on age and urban location.

People reporting zero earnings did not have earnings imputed, following the approach in PALMS. Imputing for zero-earners make a material difference to the Labour Force Surveys' earnings distribution (Vermaak 2012) but in general they are not well-understood. Kerr & Wittenberg (2019*b*) suggest they could come from a different data generating process (e.g. unpaid family workers, people waiting for jobs to start) compared to regular earners and there is a high chance they could be measurement error. Together these are good grounds to avoid trying to impute their earnings.

There are some known weaknesses of the imputation. First, the imputation assumes the earnings data are missing at random (MAR) when in all likelihood the data are missing not at random (MNAR) according to Rubin's typology of missing data. Without any more information about the nature of the missingness mechanism, we assume the data is MAR in order to move forward with the imputation. Second, predictors for the imputation model are relatively limited compared to what is available in the labour market datasets. We do not have information such as occupation or industry for the full series which could be expected to be important for the quality of the imputations.

2.1.3 Extra earnings variables

Additional earnings variables are included in a separate data file called pasesv1_extraincome.dta. This file includes earnings brackets for the GHS era for bracket responders; a point estimate earnings variable (nominal and real versions); the bracketweight; and, earnings variables with imputations for missing data.[3] These variables are saved in a separate dataset to remind users that they represent a departure from the earnings variable in the main dataset in some form. These variables can be merged back into the main file by opening the main file and entering the following code into Stata:

merge 1:1 year hhid personnum using pasesv1_extraincome.dta

The files should merge perfectly since the extra income variables file includes all observations who are in the main file, regardless of whether or not they have non-missing information for one of the extra income variables.

[3]This file also contains additional household income variables discussed later on in Section 2.6.

2.2 Labour market status

We include our own derived labour market status variables in the data set along with the official StatsSA variables. The derived variables are empstat1 and empstat2 and correspond to narrow and broad definitions of labour market status, respectively. We import these variables from PALMS for the OHS era and construct them for the GHS period. However, users should be aware that there are occasional breaks in these two time series because some survey years are missing variables we use to construct empstat1 and empstat2.

In years where we have all the variables we need, respondents are classified as employed if in the last seven days they worked for pay; ran their own business; did unpaid work; or, had a job or business they could return to in the case they had been absent from work. The paid work question for GHS 2002-2008 specifically excluded domestic workers, who were asked about separately. These survey years also include additional questions capturing different types of unpaid work such as subsistence farmers; people engaged in constructing or repairing their own homes; and people catching wildlife for sale or home consumption. Respondents who answered 'yes' to these questions or being a domestic worker were also classified as employed. The variables used to construct empstat1 and empstat2 are included in the data set should users wish to adjust these categorisations.

Respondents were classified as narrowly unemployed if they searched for work in the past seven days. Respondents were classified as broadly unemployed if they either searched for work in the past seven days or said that they would accept suitable work if offered. Everyone else aged 15 years and over is classified as not economically active. The variables are missing for those younger than 15 years. Fifteen years and older is used as the age restriction because the GHS economic activity modules were asked of people aged 15 years and older. Users should know this is based on questionnaire design as opposed to representing our idea of 'working age'.

In terms of missing variables, the most problematic years are 2009 and 2010 which omit both questions used to ask about unemployment: job search and wanting to work if offered. We cannot distinguish the unemployed from the not economically active in this case. Both empstat1 and empstat2 only report employed labour market status for these years because it is the only category that can be reliably identified. GHS 2019 also omitted a question about wanting work if offered, which means we cannot identify discouraged work seekers from the not economically active. The empstat2 variable for 2019 only reports employed labour market status as we cannot distinguish between the broadly unemployed and the not economically active.

Note that the StatsSA labour market status variables, ssa_empstat1 and ssa_empstat2, are also not available for all years. These variables are available starting in 1999-2008, 2014-2018, and 2021-2024 but omitted in between and value codes are not consistent over time.

2.3 Cross-entropy weight

Another important value-add from PASES is that we provide a newly calibrated survey weight, cewgt, in addition to the original StatsSA person and household weights. Reweighting the data was a requirement because the current way weights are released in the OHS and GHS is inconsistent with sampling practice (Thornton & Wittenberg 2022*b*). The StatsSA sampling practise implies the data should include a single weight integrated at the household-level. Instead, both the OHS and GHS are released with separate

person and household weights that are never equal to each other and based on mutually exclusive calibration models (i.e. benchmarks).[4] Separating the weighting process in this way results in a conceptual compartmentalising of the person and household universe. In other words, this raises the question of which weight to use for analysis that uses both person and household information, such as 'how many *people* live in *households* receiving social grants?'.

The new cewgt survey weight is consistent with sampling practise in the source data. It uses all the same information StatsSA use to benchmark their separate weights plus additional household size benchmarks combined in a single calibration procedure (Thornton & Wittenberg 2022*b*). These benchmarks include both person demographics and household headship rates meaning the cewgt is simultaneously calibrated to total person and household counts. Additionally, we include benchmarks for one-, two-, and three-person households since small households have been chronically undersampled almost throughout the series (Thornton & Wittenberg 2022*b*). Since the household is the sampling unit, missing certain types of households could bias a range of other outcomes in the data set. We also treat worker hostels as households throughout our benchmark series, whereas they were left out of the 1996 and 2001 census benchmark for the StatsSA household count. Household count estimates using the new cross-entropy weight are therefore slightly higher than the StatsSA weights especially at the beginning of the series.

The cewgt is provided for the period 1995-2011, and previously released on DataFirst by Thornton & Wittenberg (2022*a*). The weights stop in 2011 because the household benchmarks are derived from the census 10% samples and 2011 is the last year with reliable census information. Unfortunately, we cannot use the latest 2022 Census to update these benchmarks for two reasons. This is firstly because of substantial data quality problems raised by other researchers (Moultrie & Dorrington 2024, Budlender & Thornton forthcoming). A second problem is that even if the Census 2022 met quality criteria, its location in the aftershock of the Covid-19 pandemic presents a modeling challenge (Thornton 2025).

The alternative of maintaining the existing model of household change based on the 1996, 2001 and 2011 censuses and projecting it forward is a problem for two reasons. Firstly, the divergence between the cewgt-weighted vs StatsSA-weighted household size distribution becomes excessive as time progresses further away from 2011. The cewgt procedure produces an estimate of 32% of households being single-person in 2019 (before the complication of Covid-19 lockdowns becomes a factor) compared to 23% according to the StatSA household weight.[5] This is likely happening because the underlying model of household formation has changed since 2011 and only a census can update this model with sufficient accuracy. Without a census to anchor expectations we are not comfortable maintaining a benchmark scheme that produces such different results. Secondly, the population lockdowns triggered by the Covid-19 pandemic will have caused a structural break in household change. Currently, we know very little about the nature of this change in a way that is nationally-representative for households, other than that it happened.[6] This is enough to know that we cannot assume that household change over this period continued uninterrupted in the same way it was occurring prior to the pandemic (Thornton 2025).

A key purpose of creating the cewgt is to combine person and household information in one calibration

[4]There are some cases in the OHS where the weight for the household head matches the weight for the household and was possibly the result of the same calibration procedure although no information is provided in the metadata. An additional problem in the OHS is that person weights are not integrated within the same household.

[5]And 19% according to the StatsSA person weight.

[6]We know that South Africans adapted their living arrangements as part of livelihood strategies during the lockdown from the NIDS-CRAM survey, a rapid telephone survey that was close to being nationally representative of people (Posel & Casale 2021, Eyal & Njozela 2025). This survey did not purport to be nationally representative of households.

procedure. If we cannot do this, users may as well use the StatsSA weights beyond 2011 for lack of a better alternative. However, users are warned that the problem with this weighting scheme described above remains an issue (Thornton & Wittenberg 2022*b*) as well as questions around how to model households over Covid with sparse auxiliary information (Thornton 2025). In particular, the trend in the total number of households according to the GHS household weight increases in an uninterrupted way over 2020 and 2021, which is highly implausible especially since the same household weight also registers a drop in average household size.

In selecting a weight to use, researchers should be guided by their research question. Our recommendations regarding weight choice are as follows: if studying any outcome in the time-period before 2011-inclusive, use the cewgt. If studying a person-level outcome post-2011 or spanning the full series, then use the StatsSA person weight. If studying a household-level outcome post-2011 or spanning the full series, use the StatsSA household weight. If studying outcomes post-2011 that require consistency at both the person and household-level the choice is not clear-cut and we recommend researchers check the sensitivity of their results to using the StatsSA person and household weight.

2.4 Household relationships

We are able to provide household relationships for the full period in the form of the 'relationship to the household head' variable called relhead. There is one household head per household and this individual is self-identified. All other household members are classified by their relationship to this person: parent, spouse, child, sibling, other relative, or other non-relative. Co-resident spouses (including cases where neither spouse is a household head) can also be identified using an indicator variable for a co-resident spouse, spouseinhh, and the spouse's person number in that household, spouseno. Similarly, co-resident parents and children can be identified using indicator variables called motherinhh and fatherinhh; and the parent's person number, motherno and fatherno.

The spouseno variable has poorer coverage of the married sample in the OHS than the GHS. This is particularly the case in 1997 where only about 6 000 spouse numbers were captured for 30 000 married people and in 1999 where spouse numbers were not captured at all. The spouseno variable in the data set includes our fix for the OHS era. We use the 'relationship to the household head' variable and capture the person numbers of household heads and spouses of the head to populate the spouseno variable. This fix will miss married people where neither spouse is a household head. We cross-referenced our fixed variable with observed spouseno where it exists. Our fixed variable very closely matches the observed numbers and improves this variable's coverage of the married sample.

2.5 Social grants

2.5.1 Receipt

There are up to ten variables recording individual social grant receipt available depending on the year. The core set are the old age pension, disability, child support, care dependency, and foster care grants which are available for almost the full series. Grant in aid and social relief are added from the onset of the GHS in 2002. War veteran grant receipt is added from 2009 but halts in 2022 when it is absorbed

into the old age pension. Specifically, instead of asking whether someone received the old age pension or not, respondents are asked whether they received the old age pension *or* the war veteran grant or not. The grants have the same Rand value (for pensioners over the age of 75) and possibly by this point war veterans were too few in number to warrant a separate question but we have not investigated this properly. We have not yet investigated whether one one can receive both grants at the same time.

The South African government issued new grants during and just after the Covid-19 lockdowns. From 2020 a question is asked about receipt of the social relief of distress grant. And from 2023, there is also information about the top-up for the child support grant. It is not entirely clear from how this particular question was asked whether child support top up receipt implied child support grant receipt or not (with implications for the calculation of grant income). More specifically, we still need to investigate whether respondents who received the top up always also received the child support grant, or whether they only indicated they received the top up (and not the child support grant on its own) and receipt of the child support grant was implied.

Note that 2002 was an unusual year in that all grants were asked about at the household-level. Grant receipt is surveyed at the person-level in all other years.

2.5.2 Income

Variables recording the Rand income per grant are included in the data set with the exception of the social relief grant and an important caveat described here. From 2009, the Rand amount per grant was reported in the survey questionnaire so we can accurately provide grant income variables per grant from this year onwards. Prior to this year, this was not the case and we could only find out information about the values for some grants and not others. In particular, we could find information about the year-to-year values of the old age pension, disability, foster care, and child support grant pre-2009. We could not find information on the values of the care dependency, war veteran, grant in aid, or social relief grant. Looking at the post-2009 data, we noticed that the values of three of the "missing" grants were the same as one of the available ones. The war veteran and care dependency grants were always the same as the old age pension; and grant-in-aid was always the same as the child support grant. We therefore assume that these grants are pegged to each other and continue this assumption backwards to impute values for social grant income pre-2009.

Users who are making use of the grant income variables, must please take note of this assumption we have made. More importantly, if users are aware of either sources of the values of the missing grants or how to verify our pegging assumption, please do let us know. To reiterate, we have imputed the grant income values for the care dependency, war veteran, and grant in aid grants pre-2009 by assuming they are pegged to the old age pension for the first two and child support grant for the latter. This is based purely on our observation of the pattern post-2009.

2.6 Household income

An advantage of the GHS data is that it covers a number of main sources of household income. From 2009, the GHS questionnaire became more specific and detailed in how it asks about household income sources and the values of these sources. It is therefore relatively direct to create household income variables from

2009 onwards, but extending the series backwards sometimes requires judgement calls. Here we document derived variables we have made and any assumptions we made in order to create a more complete series. Much more detailed information can be found in the do file to create these variables called '5. PASES other hh income.do'. The do files to create PASES are also available for download from DataFirst.

2.6.1 Income sources

As part of the construction of a household income variable described in the next subsection, we created a set of dummies that indicated whether a household received income from a certain source. The dummy variables are indicators for whether the household received monthly income from labour market earnings (hh_earnings), social grants (hh_grants), remittances (hh_rem), or private pensions (hh_pensions). The hh_earnings dummy was created using the labour market status variables. If any household member was employed, the household was coded as receiving income from labour market earnings. Similarly, the hh_grant dummy was constructed using the social grant receipt indicators from Section 2.5. If any household member received any one of the social grants[7] surveyed in the series, the household was coded as receiving income from social grants.

Creating the remittance and private pension series was in some ways easier and harder. From 2009 onwards, the household questionnaire directly asked whether the household received income from various income sources, including remittances and private pensions.[8] However, this information was not provided in such a straightforward way in the years before 2009 and we use two other variables to infer whether households received income from these sources. The first is the 'main source of household income' question. Throughout the series, households are asked what is their main source of income and 'remittances' and 'pensions and grants' are possible answer categories. This only indicates remittance receipt in the case that remittances are the main source of income, but many households could be receiving remittances as a secondary or lesser income source. Lumping pensions and grants together as an answer category is quite unhelpful.

We also rely on a second question asked prior to 2009 only of not-employed people aged 15 years and older. This question was "how does [the respondent] support themselves?" to which the respondent could reply "supported by persons not in the household" or "Savings or money previously earned" amongst other options. We treat these responses as remittance[9] and private pension income, respectively. In the case of private pensions, we use the 'not-employed support' question to construct the private pension

[7]Except the social relief of distress grant because we have no Rand values for this grant. This grant has very low coverage though. Just over 1000 people receive this grant in the entire series. If any user has information about where to find Rand values for this grant, we would be very grateful.

[8]Note that 2009 itself is an outlier for the private pension series. In the 2009 GHS, the number of people answering they receive private pension income is much higher than in other years. We considered whether respondents were confusing this pension question with the Old Age Pension social grant since this was the first year of the new questionnaire. Removing recipients of the Old Age Pension social grant reduced the number but enough to bring it in line with the rest of the time series.

[9]Some years of the OHS asked questions that allowed us to test how well the 'not-employed support' question captured remittance income. The same question about supporting oneself was asked of not-employed people as well as a question elsewhere in the survey asking whether an individual received remittance income (asked of everyone, including the employed, over the age of 15 years). We tried to cross-reference these questions to quality check our assumptions. Good news was that it appeared to rarely be the case in the OHS that employed people said they received remittance income. This means the skip code in the GHS (only asking not-employed people) hopefully isn't too distorting. However, we did not find a perfect overlap between the not-employed support question and the individual remittance question restricted to not-employed. There are many possible reasons for this including difference in wording ("support" vs. "income") or other skip codes we might have missed. Detailed notes are in the do file mentioned at the beginning of this section.

receipt dummy before 2009. We cannot use the 'main income' question because it mixes private pensions and grants in the same answer code.

In the case of remittances, we experimented with using only the 'main income' question or only the 'not-employed support' question to create the pre-2009 series and settled on using both, which runs the risk of slightly overestimating household remittance receipt in that year.[10] Users should be aware that these judgment calls are embedded in the remittance and private pension series prior to 2009. More detailed information on this series especially can be found in the do file referenced at the beginning of this section.

2.6.2 Rand values

We provide real and nominal versions of two household income variables from 2010, with and without missing earnings data imputed. Our first household income variable in the main dataset is the sum of four components: household labour market earnings with midpoints for bracket responders[11]; household social grant income; household remittance income; and, household private pension income. The nominal version of this variable is called hhinc and the real version is called realhhinc. By 2024, 18% of households are missing the hhinc variable because missingness levels in the earnings data climb significantly over time. For this reason, we also provide a version of household income created in the same way as hhinc but using our imputed earnings variable, imputed_earningsMP. This is an almost complete-case version of household income and is provided in the extra income dataset as hhinc_imputed and realhhinc_imputed.

We created the components for the household income variables using two pieces of information. We used the receipt indicators from the section above to determine if a household received that component. If they did not, we allocated a zero Rand value to that component for that household. If they did, we turn to the income information we have for that component. The GHS started directly collecting Rand value information about household remittances in 2009 and private pensions from 2010. Person-level social grant Rand values are imputed based on the grant values (See Section 2.5) and summed at the household level. Earnings were summed at the household level to create household earnings. The household earnings and grants components extend back beyond 2009.[12] These component variables are also provided in the dataset in their real format as: real household grant income (realhhgrantinc0), real household earnings (realhhearnings0)[13], real household remittance income (realhhreminc0), and real household private pension income (realhhpensioninc0). The component variable names have a zero suffix as a reminder that they are set to zero for households that did not receive income from this income source.[14]

[10] Relying only on the 'main income' question was a clear underestimate when observing the trends. On the other hand, relying only on the 'not-employed support' question missed around 400-500 households a year who responded that remittances were a main source of income. Potentially, these were employed people who were receiving remittances although we have not investigated whether that is the case. In order not to miss these households, we decided to assign households to receiving remittance income if either a not-employed household member indicated they were supported by people outside the household or the household head indicated that remittances were the main source of household income. Again, detailed notes are in the do file mentioned at the beginning of the section.

[11] Excluding outliers

[12] OHS 1997 and 1998 technically collected all the Rand values for the components, but there were questionnaire differences and levels and trends for the OHS did not always appear consistent with the GHS so we limited the hhinc variables to starting in 2010. If users wish to use these components for the OHS era, we suggest taking a careful look at the quality of the components.

[13] The imputed version of this variable is in the extraincome dataset

[14] The grant component is set to missing if anyone in the household has missing grant receipt information (since Rand

These components were then summed and households that had missing information for any component, are missing for household income. This means if a household member was employed, for example, but had no earnings information (neither point nor bracket), the household income variable is missing. It turns out missingness levels are very low for remittances, private pensions and grants. Only 2 053 households have missing remittance, private pension, or grant information in the 2010-2024 period.[15] The driver of missing household income information then is almost entirely missing earnings information, often clustering at the household level. This motivated us to provide our version of household income that uses imputed earnings. Our hhinc_imputed and realhhinc_imputed provide a close-to complete-case household income variable but users must be aware that this is due to single imputation of the earnings data. These imputed household income variables are only missing 2 053 households of the over 300 000 in the 2010-2024 period which amounts to 0.66% of households.[16]

StatsSA also provide a derived variable of household income in the GHSs released from 2009 which according to the GHS metadata (StatsSA 2010) is the sum of household earnings (point estimates and brackets), household remittance income, and grant income and has a maximum cut-off of R20 000. In practice, this variable ranges well above R20 000 and also sums private pension income although we cannot fully recover how this variable is constructed. Specifically, the StatsSA variable is non-missing for many more households than our hhinc and realhhinc variables in our main dataset. For example, for the 20 927 households in the 2023 GHS sample, the StatsSA household income variable (totmhinc) is non-missing for 20 545 housheolds, whilst our hhinc variable is non-missing for 17 403. This difference is entirely explained by treatment of missing data since 17 413 households in the 2023 GHS have complete earnings data in the event they include an employed person.[17] Summary statistics of the StatsSA totmhinc variable and our hhinc variable are relatively similar for these 17 403 households (approximately R9 800). However, for the remaining households, the mean of totmhinc is about R3 500 higher and notably pulls up the global average for totmhinc. We do not know how these values for totmhinc were calculated.

PASES is a work-in-progress. Please email amy.thornton@uct.ac.za or andrew.kerr@uct.ac.za with any feedback or ideas about how to improve the data set or if you pick up any errors.

References

Budlender, J. & Thornton, A. (forthcoming), Irregular imputation and implausible households in the 2022 south african census, Saldru working paper.

Eyal, K. & Njozela, L. (2025), 'Navigating crisis: Household recomposition and welfare patterns during covid-19 in south africa', *Development Southern Africa* pp. 1–20.

values are imputed). The earnings component is set to missing if any employed household member has missing earnings information.

[15]Only 52 are missing remittance and private pension information, the rest is grants. Possibly there is some imputation of these values since the response rate is so high but we have no evidence of this and there is no indication of this in the metadata.

[16]This is 11 742 people out of just over a million, or 1% of the person sample.

[17]The rest of the discrepancy to our 17 403 households is earnings outliers which we set to missing in constructing our household income variable.

Kerr, A. (2025), 'Earnings and earnings inequality in south africa: Evidence from household survey and administrative tax microdata from 1993 to 2020', *Review of Income and Wealth* **71**(1), e12695.

Kerr, A., Lam, D. & Wittenberg, M. (2025), 'Post-Apartheid Labour Market Series: 1993-2024 [dataset]', University of Cape Town: DataFirst [producer and distributor]. Version 4.

Kerr, A. & Wittenberg, M. (2017), Public sector wages and employment in south africa, REDI and SALDRU Working Paper No. 214, SALDRU, UCT, University of Cape Town, South Africa.

Kerr, A. & Wittenberg, M. (2019*a*), Earnings and employment microdata in south africa, SA-TIED Working paper 2019/47, UNU-WIDER, Helsinki.

Kerr, A. & Wittenberg, M. (2019*b*), A guide to palms version 3.3, Technical guide, DataFirst, Cape Town.
URL: *https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/434/download/10286*

Kerr, A. & Wittenberg, M. (2021), 'Union wage premia and wage inequality in south africa', *Economic Modelling* **97**, 255–271.

Moultrie, T. A. & Dorrington, R. E. (2024), 'Problems and concerns with the 2022 south african census', *South African Journal of Science* **120**(7-8), 1–6.

Posel, D. & Casale, D. (2021), 'Moving during times of crisis: Migration, living arrangements and covid-19 in south africa', *Scientific African* **13**, e00926.

StatsSA (2010), General Household Survey 2009: Metadata, Report p0318.

StatsSA (2010-2013), 'October Household Survey: 1995-1999 [datasets]', Pretoria: Statistics South Africa (StatsSA) [producer]. University of Cape Town: DataFirst [distributor]. Version 1.1 [1996-1997; 1999]; Version 1.2 [1995; 1998].

StatsSA (2011-2025), 'General Household Survey: 2002-2024 [datasets]', Pretoria: Statistics South Africa (StatsSA) [producer]. University of Cape Town: DataFirst [distributor]. Version 1 [2017-2024]; Version 1.1 [2013 person; 2014; 2016]; Version 1.2 [2011; 2013 household; 2015]; Version 1.3 [2002]; Version 1.4 [2003-2009]; Version 2.2 [2010]; Version 2.1 [2012].

Thornton, A. (2025), Household information in south african census and survey data: household change, covid-19 and census 2022, Saldru seminar presentation.

Thornton, A. & Wittenberg, M. (2022*a*), 'OHS-GHS Cross Entropy Weights 1995-2011 [dataset]', University of Cape Town: DataFirst [producer and distributor]. Version 1.

Thornton, A. & Wittenberg, M. (2022*b*), 'Reweighting the OHS and GHS to improve data quality: representativeness, household counts, and small households', *South African Journal of Economics* **90**(4), 513–534.

Van Buuren, S. (2018), *Flexible imputation of missing data*, CRC press.

Vermaak, C. (2012), 'Tracking poverty with coarse data: evidence from south africa', *The Journal of Economic Inequality* **10**, 239–265.

Von Fintel, D. (2007), 'Dealing with earnings bracket responses in household surveys–how sharp are midpoint imputations?', *South African Journal of Economics* **75**(2), 293–312.

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

Wittenberg, M. (2017), 'Wages and wage inequality in south africa 1994–2011: part 1–wage measurement and trends', *South African Journal of Economics* **85**(2), 279–297.

Wittenberg, M. et al. (2008), 'Nonparametric estimation when income is reported in bands and at points', *Cape Town: Economic Research Southern Africa Working Paper* (94).

A Recreating PASES

The idea and construction of PASES is based on Martin Wittenberg and Andrew Kerr's experience in constructing the PALMS data set. Both data sets have been created in Stata and most of the code for PASES was originally adapted from the PALMS do files by Andrew Kerr and then consolidated and extended by Amy Thornton. The reweighting procedure was formulated by Martin Wittenberg, who also wrote the Stata command we use (Wittenberg 2010), and implemented by Amy Thornton. The code for PALMS itself is based on David Lam's efforts to stack and harmonise the Labour Force Surveys.

We release the files that create PASES on DataFirst. First we harmonise and stack the OHS and GHS, create several derived variables, and then we reweight the entire series. There are seven main Stata do files, one of which is a master do file that includes detailed comments about the steps and tasks of each do file. There are three main steps to the harmonisation: renaming, merging and appending (and cleaning). We first create a set of do files that consistently rename variables to have the same name across years. This is achieved using an Excel spreadsheet called 'OHS GHS master codebook' which captures how the name of the same variable varies survey to survey. The Excel spreadsheet is saved in a folder called 'Renaming' in the 'Harmonisation' folder. This is the process that originally comes from David Lam.

The OHS and GHS both have person and household files for each year and some OHS surveys also had separate files for workers or children. The second step is to individually merge these person, household and sometimes other files together for each year. This is done using 27 separate do files, one for each year, because of idiosyncratic cleaning required. These merging do files are found in the 'Merging' folder under the 'Harmonisation' folder and are executed by the master do file. Surveys are then appended together and cleaned. Cleaning decisions are recorded in the do files.

There are separate do files for the cleaning of the earnings data and the other household income data, which in addition to basic harmonising also create derived variables. The earnings do file saves some of these derived earnings variables in the separate extra earnings variables data file. The weighting do file is the last substantive step run by the master do file and assumes that users have a cleaned and appended PASES data set missing only a set of cross-entropy weights. The weighting do file attaches a set of weights provided with the release and saved in the 'Weighting' folder. The final do file executes some final tidying up before saving the data set.

If users wish to recreate the cross-entropy weights and run the full calibration procedure themselves, the weighting do file is set up for them to do so. The weighting do file includes the code both to calibrate and to attach the weights but the calibration procedure is commented out. Notes in the introduction of the weighting do file explain how to adapt the do file to run the full procedure. Running the calibration requires an additional input in the form of the benchmarks for the weighting procedure. These are saved in an Excel spreadsheet called 'Constraint Matrix' also in the 'Weighting' folder. Investigating and setting up these benchmarks from various StatsSA and other sources was a project of its own and their collection and calculation is described by Thornton & Wittenberg (2022*b*). Like PASES, the weights are a work-in-progress and any feedback is welcome. Please contact amy.thornton@uct.ac.za.

B Variable list and description

A list and description of all the variables in the data set grouped by topic.

Survey and Weighting

hhid household identifier. This is the original StatsSA household identifier in all years except 1996 where there was no household identifier variable. For 1996, we constructed a household identifier using magisterial district, enumeration area, and visiting point numbers. hhid is unique within years, not across years. hhid and year form a unique household identifier across years.

personnum person identifier within a household. personnum and hhid are a unique person identifier within years, but not across years. year, hhid and personnum form a unique person identifier across years.

year survey year. Range 1995-2024, omitting 2001 and 2000.

ea enumeration area number. This variable reflects the original StatsSA enumeration number in the source data.

stratum stratum number. This variable reflects the original StatsSA enumeration number in the source data. In order to set up the full time series for complex survey data in Stata, stratum should be combined with year after checking for duplicates that could arise from different stratum numbering systems across years. Note that GHS 2005-2007 did not include a stratum variable in the public release data sets. We have created this variable according to the description of the strata in the metadata, that is by province and urban-rural geography type.

pweight original StatsSA person weight.

hweight original StatsSA household weight.

bracketweight DataFirst weight to be used when conducting analysis of earnings variables that exclude bracket respondents. This weight was constructed by multiplying the inverse probability of a bracket response by the pweight for an individual. This variable is in the extra income data file.

cewgt DataFirst cross-entropy weight. Benchmarks for this weight are the StatsSA mid-year population estimates and household benchmarks calculated using the three ten percent censuses. Currently this weight is available from 1995-2011. Extension of the weights beyond 2011 has been compromised by the quality of the 2022 Census.

interviewmonth approximate interview month derived by DataFirst from sampling information in the documentation released by StatsSA with the survey releases. Between 2002-2008, all respondents were sampled in July. From 2009-2024, sampling happened over three months from July-September. From 2013 onwards, respondents were sampled over the full course of the year and divided into four 'rotation' groups that roughly correspond to four quarters which is available in the 'quarter' variable in this dataset. Note that each GHS is still meant to be a single nationally representative snapshot of the population which has implications for survey weighting. As far as we understand each quarter cannot be treated as nationally-representative.

proxyresp dummy variable indicating whether the survey respondent was the person themselves, or a proxy answering the questions on their behalf. Available 1999-2007. This question was asked at the beginning of the Economic Activity module.

proxyrespno the person number of the survey respondent for (usually) the person questionnaire. Available for 2009 onwards with the exception of 2015 and 2019-2021. The question exists in the 2015 questionnaire but is missing in the 2019-2021 questionnaires. Over time, there are slight changes to the wording of this question and its placement in the questionnaire. Sometimes this variable is missing for people younger than 15 years since they did not get asked the Economic Activity module.

quarter quarter of the year in which the respondent was surveyed. This variable was introduced when the GHS transitioned from being collected once a year to being collected over the course of the year in four installments from 2013.

Demography

gender gender of individual. 1 = male; 2 = female.

age age of individual in years.

popgroup population group of individual. 1 = African/Black; 2 = Coloured; 3 = Indian/Asian; 4 = White.

Household Relationships

relhead relationship to household head. Household head is self-appointed by the household members. Categorical variable ranging 1-9.

fatheralive respondent's father is alive. 1 = Yes; 2 = No; 3 = Don't Know.

motheralive respondent's mother is alive. 1 = Yes; 2 = No; 3 = Don't Know.

fatherno father's personnum if co-residing in the same household. Omitted in OHS 1997-9.

motherno mother's personnum if co-residing in the same household. Omitted in OHS 1997-9.

marstat marital status. 1 = married; 2 = living together; 3 = divorced or separated; 4 = widowed; 5 = never married. Note that years GHS 2002-4 omitted 'living together' as an answer category. In the original source data, 'never married' is split into 'single, never cohabited' and 'single, have cohabited' from GHS 2009 onwards. The latter category is inconsistently labelled over time and also very small. We combine these two categories.

spouseno personnum of spouse if married and spouse is co-residing in the same household. Omitted in OHS 1999 and poorly captured in OHS 1997. We have re-populated this variable for the OHS era (including 1999) using the relhead variable.

spouseinhh indicator for whether spouse co-residing in same household. Only available from GHS 2002 onwards.

fatherinhh father co-residing in same household. Omitted in OHS 1997-9. For OHS 1995 and 1996, we coded Yes in cases where there is a valid fatherno; coded to missing if father dead or respondent does not know if father alive according to fatheralive; coded No if father is alive according to fatheralive and there is no valid fatherno.

motherinhh mother co-residing in same household. Omitted in OHS 1997-9. For OHS 1995 and 1996, we coded Yes in cases where there is a valid motherno; coded to missing if mother dead or respondent does not know if mother alive according to motheralive; coded No if mother is alive according to motheralive and there is no valid motherno.

hhsize household size. This is a derived variable constructed by counting the number of individuals in a household.

Geography

province province of the household. Range = 1-9.

metro household located in a metropolitan area. 1 = Metro; 2 = Non-metro. Only available from GHS 2002 onwards.

urban household located in an urban or rural area. 1 = Urban; 0 = Rural. Derived variable created using StatsSA's 'geotype' variable. Number of location types vary frequently across years but we collapse to the lowest common denominator which is a binary of urban-rural location.

Education

yrseduc derived years of education variable. Variable constructed using the various 'highest years of education' variables across the series.

eduohs95 respondent's highest level of education variable for the OHS 1995. Range = 0-15.

eduohs96 respondent's highest level of education for OHS 1996. Range = 0-19.

eduohs97_98 respondent's highest level of education for OHS 1997-8. Range = 0-13. 99 = unspecified

eduohs99 respondent's highest level of education for OHS 1999. Range = 0-22.

edughs02_03 respondent's highest level of education for GHS years 2002-2003. Valid range = 0-22; 99 = unspecified.

edughs04_08 respondent's highest level of education for GHS years 2004-2008. Valid range 0-26. 99 = unspecified.

edughs09_16 respondent's highest level of education for GHS years 2004-2008. Valid range 0-31; 98 = no schooling; 99 = unspecified.

edughs17_18 respondent's highest level of education for GHS 2017-18. Valid range 0-29; 98 = no schooling; 99 unspecified.

edughs19_24 respondent's highest level of education for GHS years 2019-2024. Range 0-29; 98 = no schooling; 99 unspecified.

Household Income

hhinc total monthly household income variable created by summing four main sources of income: earnings with midpoints for bracket respondents, social grants, private pensions and remittances. Variable is set to missing if any one of the sources are missing. Available from 2010 onwards.

realhhinc real version of hhinc with December 2023 as the base.

hhinc_imputed version of hhinc that uses imputed earnings for missing earnings data, available from 2010 onwards. This variable is in the extraincome.dta file.

realhhinc_imputed a real version of hhinc_imputed available from 2010 onwards. This variable is in the extraincome.dta file.

hh_rem household receives monthly income from remittances. Only available from 2009 onwards.

hh_pension household receives monthly income from private pensions. Only available from 2010 onwards.

hh_grants at least one member of the household receives monthly income from social grants.

hh_earnings at least one member of the household is employed.

main_inc_source0208 main source of income for the household for GHS 2002-2008. Range = 1-6.

main_inc_source09plus main source of income for the household for GHS 2009 onwards. Range = 1-8.

grant_income derived variable that is the sum of an individual's monthly grant income.

realhhgrantinc0 total monthly real grant income for the household with December 2023 as the base. This variable was constructed as part of the process of creating hhinc. This variable is set to zero for households with no grant recipients.

realhhreminc0 total monthly real remittance income for the household with December 2023 as the base. This variable was constructed as part of the process of creating hhinc. This variable is set to zero for households not receiving remittances.

realhhpensioninc0 total monthly real private pension income for the household with December 2023 as the base. This variable was constructed as part of the process of creating hhinc. This variable is set to zero for households with no private pension recipients.

realhhearnings0 total monthly real labour market earnings income for the household with December 2023 as the base. This variable was constructed as part of the process of creating hhinc. This variable is set to zero for households no employed members; and missing for households where employed members had missing earnings information.

realhhearnings0_imputed a version of realhhearnings0 that uses imputed earnings from imputed_realearningsMP to deal with high levels of missing earnings data. This variable is in the extraincome.dta file.

Social Grants

oap_g respondent receives the Old Age Pension. 1 = Yes; 0 = No. Available from OHS 1997 onwards. Note from 2022 onwards, warvet_g falls away and is merged with oap_g. We have not verified whether these two grants could be held simultaneously.

dis_g respondent receives disability grant. 1 = Yes; 0 = No. Available from OHS 1997 onwards.

cs_g respondent receives child support grant on behalf of a child in that household. 1 = Yes; 0 = No. Available from OHS 1997 onwards.

caredep_g respondent receives a care dependency grant. 1 = Yes; 0 = No. Available from OHS 1997 onwards.

foster_g respondent receives a foster care grant to foster a child in that household. 1 = Yes; 0 = No. Available from OHS 1997 onwards.

gia respondent receives grant-in-aid. 1 = Yes; 0 = No. Available from GHS 2002 onwards.

soc_rel recipient receives social relief grant. 1 = Yes; 0 = No. Available from GHS 2002 onwards.

warvet_g respondent receives a war veteran grant. 1 = Yes; 0 = No. Available from GHS 2009 onwards. Note from 2022 onwards, warvet_g falls away and is merged with oap_g. We have not verified whether these two grants could be held simultaneously.

srd_g respondent receives social relief of distress grant. 1 = Yes; 0 = No. Available from GHS 2020 onwards.

cstopup_g respondent receives the child support grant top-up for orphans in that household. 1 = Yes; 2 = No. Available from 2023. Question wording ambiguous: unclear whether recipients of this grant indicated both cs_g receipt *and* cstopup_g receipt or if indicating cstopup_g receipt *implied* cs_g receipt.

oap_g_inc monthly Rand income from old age pension grant.

dis_g_inc monthly Rand income from the disability grant.

foster_g_inc monthly Rand income from the foster care grant.

warvet_g_inc monthly Rand income from war veteran grant.

cs_g_inc monthly Rand income from child support grant.

gia_g_inc monthly Rand income from grant in aid grant. Amounts assumed pegged to the child support grant for the period 2002-2008.

caredep_g_inc monthly Rand income from the care dependency grant. Amounts assumed pegged to the old age pension grant for the period 2002-2008.

cstopup_g_inc monthly Rand income from the child support to up for orphans grant.

srd_g_inc monthly Rand income from the social relief of distress grant.

Labour Market Status

empstat1 derived narrow labour market status variable. 0 = not economically active; 1 = employed; 2 = unemployed. Non-working-age population set to missing. See notes about variable construction and problem years in Section 2.2.

empstat2 derived broad labour market status variable. 0 = not economically active; 1 = employed; 2 = unemployed. Non-working-age population set to missing. See notes about variable construction and problem years in Section 2.2.

emp_business the respondent worked in their own business in the past seven days. 1 = Yes; 2 = No. This variable was used to construct the empstat1 and empstat2 variables.

emp_wage respondent worked for pay in the past seven days 1 = Yes; 2 = No. This variable was used to construct the empstat1 and empstat2 variables.

emp_domes respondent worked as a domestic worker in the past seven days for pay. 1 = Yes; 2 = No. The emp_wage question for GHS 2002-2008 specifically excluded domestic workers. For this set of survey years, the GHS included additional questions about employment such as emp_domes that we use to derive the empstat1 and empstat2 variables.

emp_nopay respondent worked for no pay in the last seven days. 1 = Yes; 2 = No. This variable was used to construct the empstat1 and empstat2 variables.

emp_farm respondent worked on the household's plot, farm, food garden, cattle post or kraal or helped in growing farm produce or in looking after animals for the household in the past seven days. 1 = Yes, 2 = No. Only available GHS 2002-8 and used to derive the empstat1 and empstat2 variables for those years.

emp_const respondent worked on construction or major repair work on his/her own home, plot, cattle post or business or those of the household in the past seven days. 1 = Yes, 2 = No. Only available GHS 2002-8 and used to derive the empstat1 and empstat2 variables for those years.

emp_catch respondent worked to catch fish, prawns, shells, wild animals or other food for sale or household food in the past seven days. 1 = Yes, 2 = No. Only available GHS 2002-8 and used to derive the empstat1 and empstat2 variables for those years.

emp_return respondent has a job, business or other economic activity to return to even if they did no work in the past seven days. 1 = Yes; 2 = No.

want_work respondent would accept a suitable job offer. Asked of respondents who did not work in the past seven days. 1 = Yes; 2 = No. This variable was used to construct the empstat1 and empstat2 variables.

searching respondent has searched for work in the past four weeks. Asked of respondents who did not work in the past seven days. 1 = Yes; 2 = No. This variable was used to construct the empstat1 and empstat2 variables.

ssa_empstat1 StatsSA official narrow employment status. Only available 1999-2008; 2014-2018; 2021-2024 and value codes are not consistent over time. Users should be aware of this because we have not harmonised this variable to save on information loss. For 1999-2008, the codes are 0 = Not economically active; 1 = Employed; 2 = Unemployed. For 2014-2018 and 2021-2024, the codes are 1 = Employed; 2 = Unemployed; 8 = Not Applicable; 9 = Unspecified (i.e. no code for Not economically active).

ssa_empstat2 StatsSA official broad employment status. Only available 1999-2008; 2014-2018; 2021-2024 and value codes are not consistent over time. Users should be aware of this because we have not harmonised this variable to save on information loss. For 1999-2008, the value codes are 0 = Not economically active; 1 = Employed; 2 = Unemployed. For 2014-2018 and 2021-2024, the codes are 1 = Employed; 2 = Unemployed; 3 = Not Economically Active; 8 = Not Applicable.

sector respondent works for or owns a business that is in the formal or informal sector. 1 = formal sector; 2 = informal sector; 3 = Don't Know. Only available from GHS 2010 onwards with a break in 2020 and 2021.

Earnings

earningsMP nominal monthly earnings variable with bracket midpoints for bracket responders.

realearningsMP real monthly earnings variable with bracket midpoints for bracket responders. Base is December 2023.

earnings_period period over which earnings are paid for the GHS 2002 onwards. 1 = per week; 2 = per month; 3 = annually; 8 = not applicable; 9 = unspecified.

outlier A flag if the studentised residual from an OLS log earnings regression (with independent variables gender, year, popgroup, yrseduc, age, age squared, province, metro and interactions between gender and year and popgroup and year) is more than 5. 1 = Outlier.

earnings nominal monthly earnings variable. Should be used in conjunction with the bracketweight to account for bracket responders. Available in the extra income variables file.

realearnings real earnings variable with base December 2023. This variable should be used in conjunction with the bracketweight to account for bracket responders. Available in the extra income variables file.

imputed_earningsMP earningsMP with imputations for those with missing earnings information. This variable is in the extraincome.dta file. It is complete-case for the employed in the GHS period.

imputed_realearningsMP realearningsMP with imputations for those with missing earnings information. This variable is in the extraincome.dta file. It is complete-case for the employed in the GHS period.

imputed_earnings earnings with imputations for those with missing earnings information. This variable is missing for bracket responders and should be used in conjunction with the bracketweight to account for this. This variable is in the extraincome.dta file.

imputed_realearnings realearnings with imputations for those with missing earnings information. This variable is missing for bracket responders and should be used in conjunction with the bracketweight to account for this. This variable is in the extraincome.dta file.

earnings_bracket97_18 the GHS earnings bracket for respondents who either did not know their earnings amount or refused to give a point estimate for GHS 2002-2018. Responses could be based on weekly, monthly, or annual earnings periods although the brackets are equivalent between these earnings periods. Valid range = 1-16. This variable is in the extraincome.dta file.

earnings_bracket19_21 the GHS earnings bracket for respondents who either did not know their earnings amount or refused to give a point estimate for GHS 2019-2021. Responses could be based on weekly, monthly, or annual earnings periods although the brackets are equivalent between these earnings periods. Valid range = 1-15. This variable is in the extraincome.dta file.

earnings_bracket22_24 the GHS earnings bracket for respondents who either did not know their earnings amount or refused to give a point estimate for GHS 2022-2024. Responses could be based on weekly, monthly, or annual earnings periods although the brackets are equivalent between these earnings periods. Valid range = 1-20. This variable is in the extraincome.dta file.

Fertility

pregnant the respondent has been pregnant in the past 12 months. Asked of all female household members between the ages of 12 and 50 years. 1 = Yes; 2 = No; 3 = Don't Know. Available from 2009-2019, and again from 2021-2024.

pregstatus current status of the pregnancy if respondent answered yes to being pregnant in the past 12 months. Valid range = 1-5. Available from 2009-2019, and again in 2023-2024.

Transport

own_car household owns a motor vehicle. Available from 2009.

commutemode mode of transport respondent takes to work. Available from 2009. From 2024, bicycles and motorcycles are separated into separate categories and 'own car/private vehicle' is separated from 'company vehicle/staff transport'. We collapse these categories back into the the pre-2024 labeling for consistency since the change only exists for one year so far.

commutetime duration of respondent's work commute in minutes. Available from 2009.

hhbustrips total number of bus trips take by household members in the last week. Available from 2009.

hhbuscost total household spend on bus fares in last week. Available from 2009.

hhbusdist distance to the nearest bus stop. Available from 2009.

hhtraintrips total number of train trips take by household members in the last week. Available from 2009.

hhtraincost total household spend on train fares in last week. Available from 2009.

hhtraindist distance to the nearest train station. Available from 2009.

hhtaxitrips total number of taxi trips take by household members in the last week. Available from 2009.

hhtaxicost total household spend on taxi fares in last week. Available from 2009.

hhtaxidist distance to the nearest taxi stop. Available from 2009.