Technical Paper Series

PROVIDE PROJECT

The Provincial Decision-making Enabling Project

Technical Paper 2005:1

Creating a 2000 IES-LFS Database in Stata®

Elsenburg
February 2005

PROVIDE PROJECT

The Provincial Decision-making Enabling Project

Overview

The Provincial Decision-Making Enabling (PROVIDE) Project aims to facilitate policy design by supplying policymakers with provincial and national level quantitative policy information. The project entails the development of a series of databases (in the format of Social Accounting Matrices) for use in Computable General Equilibrium models.

The National and Provincial Departments of Agriculture are the stakeholders and funders of the PROVIDE Project. The research team is located at Elsenburg in the Western Cape.

PROVIDE Research Team

Project Leader:	Cecilia Punt
Senior Researchers:	Kalie Pauw
	Melt van Schoor
Junior Researchers:	Benedict Gilimani
	Lillian Rantho
Technical Expert:	Scott McDonald
Associate Researchers:	Lindsay Chant
	Christine Valente

PROVIDE Contact Details

Private Bag X1
Elsenburg, 7607
South Africa

ceciliap@elsenburg.com
+27-21-8085191
+27-21-8085210

For the original project proposal and a more detailed description of the project, please visit www.elsenburg.com/provide

Creating an IES-LFS 2000 Database in Stata® [1]

Abstract

The Income and Expenditure Survey of 2000 is the most recent comprehensive source of information on income and expenditure patterns of South African households. This data is used to compile various household-related sub-matrices for a series of Social Accounting Matrices for South Africa. By linking the Labour Force Survey of September 2000, which contains detailed employment data on occupation codes, activity codes and wages of workers, with the Income and Expenditure Survey of 2000 various factor-related sub-matrices can also be extracted. This paper discusses the steps followed to extract data, correct problems and errors where appropriate or necessary, merge various files, and create Stata format data files that can be used to compile the relevant sub-matrices. The focus remains highly technical throughout.

[1] The main author of this paper is Kalie Pauw, Senior Researcher of the PROVIDE Project. The Stata® software (referred to throughout as Stata) is a registered trademark of the Stata Corporation (StataCorp, 2001).

Table of Contents

List of Figures

List of Tables

1. Introduction

The Income and Expenditure Survey of 2000 (IES 2000) conducted by Statistics South Africa contains detailed information on income and expenditure of households. When merged with the Labour Force Survey of September 2000 (LFS 2000:2), the data can be used for the compilation of various sub-matrices of South African Social Accounting Matrices (SAMs). Up until recently most of the existing post-1995 South African SAMs, including the PROVIDE national and provincial SAMs, relied on the IES 1995 and the October Household Survey (OHS) of 1995 for data on income, expenditure and employment patterns of households. The use of the more recent 2000 dataset can be regarded as an improvement since changes in employment and income and expenditure patterns are likely to have taken place between 1995 and 2000. However, there are some concerns about the quality of the IES 2000 data in particular. Correcting these errors, which range from simple computing errors to more serious inconsistencies in the data, has proven to be quite an elaborate process, especially since there is no single correct way of treating such problems. However, as argued by Van der Berg *et al*. (2003a), the IES 2000 is the most recent available data and one should attempt to work with it.

This paper is organised as follows. Section 2 gives a brief overview of the data files and sample design of the IES 2000 and LFS 2000:2. A section is also devoted to the theory of sampling and weighting. Next, the merging of the IES 2000 and LFS 2000:2 is discussed, and finally some of the data problems, specifically in the IES 2000, are discussed in some detail. Section 3 describes the Stata do-files that were used to read in the data from the original ASCII-format data files and create person- and household-level IES 2000 and LFS 2000:2 Stata-format data files. Section 4 explains some of the final data adjustments made to prepare the dataset for extraction of various household- and factor-related SAM sub-matrices. Some concluding remarks are made in section 5.

2. An overview of IES 2000 and the LFS 2000:2

2.1. Data files

2.1.1. IES 2000

The IES is conducted by Statistics South Africa every five years. It measures the detailed income and expenditure of households. These surveys were originally designed and are still used to determine weights for the South African Consumer Price Index (CPI).[2] The IES is,

[2] Because of this objective of the IES consumer goods and services are not grouped according to the Standard Industrial Classification (SIC) codes. The SIC classification is used for activities and commodities in the

however, also useful to show the earning and spending capacity and expenditure patterns of South African households (SSA, 2002a). The survey is also based on the same sample of households interviewed for the twice-yearly LFS (SSA, 2002b), which contains more detailed information pertaining to employment activities of household members.[3] This proves to be quite useful as the IES 2000 and the LFS 2000:2 can be merged to form a comprehensive dataset that combines the detailed household income and expenditure data of the IES 2000 with employment data in the LFS 2000:2.[4]

The metadata file published with the IES 2000 provides a description of the data, the sample design, the sampling weights and the variables contained in the dataset (SSA, 2002a). The raw data are published in four ASCII text files with one line of given length per record or observation. Each line represents a household or a person, depending on whether it is a person- or household-level file. The first file, *person.txt*, contains person-level data of all members in the household. The maximum household size allowed for is 25 members. This file contains categorical and continuous variables for gender, age, race, work status and income from employment of each household member. The inclusion of labour income data at a person-level is a new addition to the IES 2000. Previously in the IES 1995 employment information was not available at a person-level, which meant that data on occupation codes, industry codes and labour income had to be extracted from the OHS 1995. Since this data is now available in the IES 2000 there is the option of using only the IES dataset for all the sub-matrices, including the factor-related sub-matrices. However, as argued in section 2.3, there a number of reasons why it was decided to rather use the LFS 2000:2 data for all factor-related information.

A second file, renamed *domworker.txt*, contains information on domestic workers employed by households. In the PROVIDE SAM domestic work is regarded as a service purchased by the household and supplied by an activity called domestic services, and hence this information should also be extracted from the raw data. A third file, renamed *homegrown.txt*, contains information on home production for home consumption (HPHC) of farm produce and livestock at the household level. This information is included in the income and expenditure sides of the applicable households and takes into account the market value of goods produced, the amount consumed, and the value of excess production sold. Input costs are also accounted for. Finally, *general.txt* contains all the general income and expenditure

PROVIDE SAMs. As a result expenditure items from the IES have to be mapped to the correct commodity groups. This mapping is based on the mapping used by (McDonald and Punt, 2001) for their SAM for the Western Cape province.

[3] The LFS replaced the OHS, which was conducted annually until 1999.

[4] Since the LFS 2000:2 and the IES 2000 were conducted at around the same time in 2000 Statistics South Africa suggests that the September edition of the LFS be merged with the IES 2000.

data, including income and expenditure summary tables. This file is the largest of all the data files and contains the bulk of the information collected for the IES 2000.

2.1.2. LFS 2000:2

The LFS 2000:2 also comes with a metadata file explaining the sampling framework and a list of the files that are contained in the LFS 2000:2 dataset. The sample design of the LFS 2000:2 is the same as that of the IES 2000. Data files include *person.txt*, *worker.txt* and *house.txt*.[5] The file *person.txt*, as its namesake in the IES 2000, contains all the person-level information of household members, while *worker.txt* contains employment data of all household members of working age (15 – 65). Finally, *house.txt* contains general household variables. A fourth data file, *stratum_psu.txt* contains variables identifying the primary sampling units (PSUs) and the strata used in the survey (see section 2.2). When merged with the IES 2000 only data contained in *person.txt* and *worker.txt* are used.

2.2. Sampling and weighting[6]

2.2.1. Survey design

The design of a survey has important implications for the way in which data analysis should be undertaken. Often budgets and time constraints dictate the sampling and data collection methods used, and ingenious ways have to be sought to reduce data collection costs without jeopardising the quality and 'representativity' of the data. Ideally the sampling design should match the type of survey being conducted. Deaton (1997:17) suggests that each different application of a survey mandates a different survey design – "*precision for one variable is imprecision for another*". However, given budgetary constraints "*it makes no sense to design a survey for each*". The IES 2000, for example, was designed specifically for calculations of the CPI, but understandably so, has become a general-purpose household survey with a range of applications.

A typical households survey selects households randomly from a list of all households in the population known as the sampling frame. The sampling frame is often the most recent Census. In the case of the IES 2000 and LFS 2000:2 the South African Population Census of 1996 was used as sampling frame (SSA, 1998). A Census contains a list of all households and household members. The most common way of choosing representative households from the sample frame is based on a two-stage selection process. At the first stage clusters or groups of households are selected randomly from the population. These clusters are often based on existing geographical boundaries. Next, the census data are used to compile a list of all

[5] To avoid confusion these files were renamed *lfsperson.txt*, *lfsworker.txt* and *lfshouse.txt*.
[6] This section draws mainly on Deaton (1997) unless otherwise cited.

households in each selected cluster. The second stage then involves drawing households from each sampled cluster to enter into the survey. Often this stage of the selection process is informed by prior knowledge about households, which implies that stratification comes into play. Clustering and stratification are discussed in more detail in the following section.

2.2.2. Clustering and stratification

In the two-stage sample design clusters are first selected randomly from a list of clusters covering the entire population. Next, households are selected from each of the sampled clusters. This generates a final sample in which households are not randomly distributed over space, but are grouped geographically. The most important reason for clustering is the cost-effectiveness of this approach. With clustering it also becomes more feasible to gather village-level information on, for example, schools, clinics and (local) government services. The Census of 1996 forms the basis for clustering in the IES 2000 sample. The 3,000 primary sampling units (PSUs) in the IES 2000 are drawn randomly from the list of census enumeration areas (EAs) (SSA, 2002a).

Before households are drawn from the list of random clusters, it has to be decided whether prior knowledge about households should be used to influence the selection process. Often surveys are required to generate statistics for population sub-groups, e.g. by geographical area, race or standard of living. Stratification is a method used to ensure that observations from each of these groups are adequately represented in the final sample by "*effectively* [converting] *a sample from one population into a sample from many populations*" (Deaton, 1997:13). Household income and expenditure surveys "*nearly always*" distinguish between rural and urban areas, and sometimes further stratification by geographical region, race and income group are added. Such stratification is also known as explicit stratification.

Stratification can also be done implicitly by means of a systematic sampling process. A list of households are ranked or sorted according to some household characteristic. A random starting point is selected and thereafter every j^{th} observation is selected into the sample, with the value of j depending on the size of the clusters and the total number of households that will eventually be included in the sample. If, for example, households are sorted according to income, selection of every j^{th} observation will ensure that the final sample will contain observations from across the entire income spectrum. Such a survey is then said to be implicitly stratified by income.

The IES 2000 is explicitly stratified by the nine provinces and by location (urban and rural) (SSA, 2002a), giving 18 explicit strata in total. Each PSU was also implicitly stratified firstly by magisterial district or district council, and thereafter by average household income

(in the case of urban areas or hostels) or EA (presumably in the case of rural areas).[7] This basically means that all urban households or households living in hostels are first sorted by magisterial district and then by their average household income. The household income data come from the Census of 1996. Rural households are sorted by magisterial district and then by EA. Ten households are then selected randomly from each of the stratified PSUs.

The way in which the two-stage sampling process is designed ensures that each household has an equal chance of selection into the final sample. If each cluster is selected randomly, with probability of selection proportionate to the size of the cluster, and if the same number of household is selected from each cluster, then the design is 'self-weighting', i.e. each household has the same chance of being included in the final sample. The IES 2000 is an example of a self-weighted sample design. The 3,000 randomly selected clusters were selected with probability proportionate to their size, while 10 households were selected from each sample. In theory each of the 30,000 households in the sample all had an equal chance of being included in the sample.

2.2.3. The 'design effect'

When a sample is stratified, say, along rural-urban lines, there are essentially two independent surveys that are being conducted. This ensures that the final combined survey is representative of households from both sectors in the population. The overall variance of an estimate, say income, will then be the weighted sum of the variance of rural income and urban income. The covariance or between-sector variance is zero because the two samples are independent. However, if the overall sample were a single random survey the covariance would come into play. More importantly, if the means of rural and urban incomes, say, were very different, the overall variability would be greater. The conclusion from this is that stratification enhances 'precision', where the term precision refers to the variability of an estimator (Deaton, 1997:14).

Clustering, on the other hand, reduces precision. This can be explained as follows. Generally speaking, households within clusters are more similar in terms of their characteristics and behaviour than households of different clusters. Thus, by sampling several households from the same cluster there is potentially less information content in the survey. The precision of an estimate therefore depends on the correlation between the observations in the cluster. The sample design therefore affects the precision, with stratification improving it, but clustering working against it.

Kish (1965, cited in Deaton, 1997) came up with the concept of 'design effect' – also known as *deff*. It is defined as the ratio of the variance of an estimate to the ratio of the

[7] Statistics South Africa is not entirely clear on this.

variance had the sample been a simple random one. Stratification typically reduces *deff* below one, while clustering increases it above one. Deaton (1997:15) suggests that most surveys have a *deff* of more than one, which proves that "*in survey design the practical convenience and cost considerations of clustering usually predominate over the search for variance-reduction*".

2.2.4. Unequal selection probabilities

Although surveys such as the IES 2000 are usually designed to be self-weighting, the probabilities of inclusion differ between observations. The possibilities of non-cooperation and non-contact cannot be taken into account when designing a survey. In some cases it also costs more to sample certain households. In such instance households that are costly to interview may be excluded on purpose, which affects the probability of inclusion of those observations. Since each sampled observation or household is representative of a number of other non-sampled households, it is necessary to adjust the weight of each observation to account for over- or under-representation of certain types of representative households. Deaton (1997:15) explains as follows:[8]

> *"The rule here is to weight according to the reciprocals of sampling probabilities because households with low (high) probabilities of selection stand proxy for large (small) numbers of households in the population."*

Differences in probabilities of selection are either a result of design (in the case of surveys that were not designed to be self-weighting) or accidental (for example when households refuse to cooperate). In the case of accidental differences in selection probabilities it is necessary to add weights to the survey ex-post. However, as Deaton warns, it is very difficult to find those factors or characteristics that sufficiently explain non-response. A good example is the apparent low response rate for White households in the IES 2000. Whether the race explains this low response rate or whether it is as a result of a combination of factors such as race, income and location is impossible to say. The difficulty in explaining the source(s) of over- or under-representation suggests that there is a real threat that the ex-post weighting adjustments could sometimes be incorrect.

2.2.5. Weights in Stata

When specifying the weight option in a Stata command line, Stata attaches a weight to each observation. This weight can alter the 'importance' of each observation in the estimation of the moments of an observation. The Stata reference manual (StataCorp, 2001) discusses four types of weights that can be implemented in Stata:

[8] See section 2.2.5 (inverse probability weights) for a discussion of the practical implementation in Stata.

- Frequency weights (*fweight*): These are integer weights that indicate duplicated observations. If for a given observation *fweight* = *n* it implies that there are *n* other identical observations in the population that are represented by this single sampled observation.
- Sampling weights or inverse probability weights (*pweight*): These weights denote the inverse of the probability that a certain observation is selected to enter into the sample. Sampling weights are typically associated with survey data. Although, as discussed in the previous section, survey designs are often quite complicated, a simple example shows how inverse probability weights are related to normal frequency weights. Suppose there are $N = 200$ households in the population from which the sample is drawn. If 10 households are selected randomly, the probability of selection is $P = 10/N = 0.05$. The inverse of this is $1/P = 20$, which is simply a frequency weight. Usually the interpretation of sampling weights is not as straightforward. The calculation of the probability of selection is more complicated when the sampling design involves clustering and stratification.
- Analytic weights (*aweight*): These weights are inversely proportional to the variance of an observation. Thus, if the variance of the jth observation were σ^2 / w_j, w_j would represent the weight attached to that observation. Typically the observations represent averages, and the weight is the number of elements that gave rise to these averages.
- Importance weights (*iweight*): This type of weight has no formal statistical definition. Each observation's weight indicates the relative 'importance' of that observation. Since they are rarely used with survey data it will not be discussed any further.

The weights provided with the IES 2000 are inverse probability weights and are based on the Census of 1996. This in itself is problematic since the Census of 2000 (SSA, 2003a) revealed some biases in the Census of 1996.[9] A family of commands specifically designed to handle the complexity of such sample designs exists in Stata (*svy*-commands). The following section gives a more detailed overview of survey estimation.

[9] A number of other sets of weights are also available, although at this stage none of these have been officially approved by Statistics South Africa, and hence the original household weights of IES 2000 and the person weights of LFS 2000:2 are used at this stage by the PROVIDE Project.

2.2.6. Survey estimation in Stata

Complex surveys typically have three characteristics: (1) the survey weights are inverse probability weights; (2) the sample is drawn from clusters rather than from the entire population; and (3) the data are stratified. Sampling weights, whether added to the data ex-post or designed beforehand, have to be used to adjust for differing selection probabilities between observations. Failure to use weights will result in biased estimates. When the sample is drawn from clusters, observations are not independent. Many statistical estimators assume independence and use of these estimators without making the correct adjustments will result in standard errors being too small. Finally, since stratification can reduce estimates of standard errors, it is also necessary to adjust for it.

Consider the following example.[10] Suppose we wish to estimate the average total income (variable *totinc*) of South African households. We can use the confidence interval (command *ci*) to show the mean, standard error and the 95% confidence interval. In the Stata output table below 'unweighted' data are used. This effectively means that the mean is the sample mean, which is at its best a crude estimate of the population mean.

```
. ci totinc

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      totinc |   26177     39186.44    638.5181        37934.91    40437.97
```

If we use weights Stata will compute a more accurate estimate of the population mean. Since *pweight* does not work with the *ci* command, we allow Stata to choose the type of weight.[11] However, if we do wish to use the *pweight* option, we have to make use of the *svymean* command. Initially the *svyset pweight wgtselect* option is set, i.e. clustering and stratification is ignored. The output of these two examples are listed below:

```
. ci totinc [weight = wgtselect]

(analytic weights assumed)

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      totinc |   26177     42793.12    653.4643        41512.29    44073.95


. svymean totinc

Survey mean estimation

pweight:  wgtselect                              Number of obs    =     26177
Strata:   <one>                                  Number of strata =         1
```

[10] The *ies2000h.dta* database is used for the example (see section 3). The weight variable *wgtselect* is used. (The current version of the *ies2000h.dta* has changed slightly since these examples were run – KP 15/02/2005).

[11] Alternatively, we can specify frequency weights (*fweight*), but then the truncated version of the weight, *fwgtselect*, has to be used since *fweight* only allows integer weights. This will give similar means and standard deviations (see section 2.2.5).

```
PSU:      <observations>                        Number of PSUs   =     26177
                                                Population size  =  11221840

------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
---------+--------------------------------------------------------------------
  totinc |   42793.12    725.4612    41371.18    44215.06    1.232494
------------------------------------------------------------------------------
```

The *ci* command with the *aweight* option gives the exact same point estimates as *svymean* with *pweight*. However, the standard errors are different since *aweight* uses a different formula for the standard error. The standard error of the point estimate, which should not be confused with the standard deviation of a variable, as well as the 95% confidence interval of the estimate, is slightly larger when *pweight* is used.

As argued before it is also important to specify clustering and stratification if applicable. The IES 2000 data used in this example made use of clusters (PSUs) and stratification along provincial and rural/urban lines (variable *provloc*). In the two examples that follow *svyset psu psuno* is specified, and thereafter *svyset strata provloc* is added.[12] In each instance the point estimates are shown. Notice the effect on the standard error, confidence interval and *deff* (see section 2.2.3).

```
. svyset psu psuno

. svymean totinc

Survey mean estimation

pweight:  wgtselect                             Number of obs    =     26177
Strata:   <one>                                 Number of strata =         1
PSU:      psuno                                 Number of PSUs   =      2956
                                                Population size  =  11221840

------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
---------+--------------------------------------------------------------------
  totinc |   42793.12    1042.246    40749.52    44836.72    2.543879
------------------------------------------------------------------------------

. svyset strata provloc

. svymean totinc

Survey mean estimation

pweight:  wgtselect                             Number of obs    =     26177
Strata:   provloc                               Number of strata =        18
PSU:      psuno                                 Number of PSUs   =      3327
                                                Population size  =  11221840

------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
---------+--------------------------------------------------------------------
  totinc |   42793.12    975.1044    40881.25    44704.99    2.226684
------------------------------------------------------------------------------
```

[12] Variable *psuno* is supplied by Statistics South Africa. Variable *provloc*, which represents the strata in the survey, was created by grouping variables for province (*prov*) and *location*.

When the cluster option is activated the standard error increases and the confidence interval widens compared to the previous example where clustering was ignored. Also, *deff* increases substantially due the effect of clustering on the precision, i.e. the variance increases.[13] When stratification is also taken into account the standard deviation declines and the confidence interval becomes narrower in line with expectations (see section 2.2.3). However, *deff* is still substantially higher than one.

In conclusion it can be said that the *svy*-commands are useful and indeed important to use when the distribution of a variable is of concern. Income distribution data, for example, will only be reliable when weights, clustering and stratification are specified. Test statistics will also be more accurate. However, if the only concern is finding the means or total income or expenditure (mean multiplied by the number of observations), normal analytic or frequency weights will suffice.

2.3. Merging the IES 2000 and the LFS 2000:2

2.3.1. Overview

The IES 2000, unlike its predecessor, the IES 1995, contains enough information on employment activities of household members to determine their occupation codes, industry codes and wages or salaries. Employment data also appears in the LFS 2000:2 in somewhat more detail. Therefore, depending on the information requirements, it may be unnecessary to merge the two files. However, recently education data, which is only available in the LFS 2000:2, was required for the formation of new household groups for the PROVIDE SAM. As a consequence it was necessary to merge these files, and hence the LFS 2000:2 employment data became available within the IES 2000 in any event. Furthermore, since the LFS 2000:2 is designed specifically to gather information on employment and related activities of the population, the quality of the data is arguably better. For example, the IES 2000 only asks a single question to determine a person's occupation or industry code. In contrast, occupation and industry codes in the LFS 2000:2 are based on a series of questions. Consequently there are fewer 'unspecified' factors and industries in the LFS 2000:2 (see section 2.3.2).

Various researchers have encountered difficulties when merging the IES 2000 and LFS 2000:2 data files. Van der Berg *et al.* (2003a) find that when merging these datasets there are a substantial number of observations for which age, gender and race variables do not match.

[13] Incidentally, *deff* will equal one if none of *pweight*, *psu* and *strata* were specified, since the variance is then simply equal to the sample variance as computed before in *ci totinc*. When only *pweight* is specified *deff* increases to 1.23, which indicates that weighting (in this instance) increases the variability. A tabulation of average weights by income deciles will reveal that the weights attached to high-income households is higher than for low-income households. Thus, when weights are specified the inequality in the distribution of income will increase since more weight is now attached to high-income households in the sample.

One hypothesis is that this is due to a mismatch of individuals within households rather than a mismatch of households. This can occur when individuals in the LFS 2000:2 do not have the same unique identification numbers (person numbers) as in the IES 2000.[14] In order to avoid these types of problems the LFS 2000:2 is used throughout as the main source of demographic data, while the IES 2000 is only used for household income and expenditure data (excluding wage or salary income data). More problematic are the "*irreconcilable differences*" between the LFS 2000:2 and the IES 2000 weights (Van der Berg *et al.*, 2003a). As a rule of thumb the LFS 2000:2 person weights were used when working with person-level data, while the IES 2000 household weights were used when working with household-level income and expenditure data. Below we make some comparisons of the demographic and labour income data of the IES 2000 and LFS 2000:2.

2.3.2. Comparing IES 2000 and LFS 2000:2 data

When merging the IES 2000 with the LFS 2000:2 there are 416 observations unique to the IES that are not in the LFS, and 1 626 observations in the LFS not in the IES. Just over 98% of the observations appear in both. As mentioned previously we use the LFS 2000:2 as the main source of demographic information of persons. However, for those 416 observations that are unique to the IES 2000 demographic data is of course not available in the LFS 2000:2. For these observations the IES data is used. This prevents the loss of a substantial number of records. As explained in detail in section 4.2.1, variables are 'created' for factors (*mergefact*), labour income (*mergeinclabp*), activities (*mergeact*), gender (*mergegender*), age (*mergeage*), province (*mergeprov*), location (*mergeloc*), race (*mergerace*) and person weights (*mergepwgt*) by using the LFS 2000:2 data as the basis and substituting missing data points by IES 2000 data points. These *merge-* variables are therefore in some sense 'combined' IES 2000 and LFS 2000:2 variables.

The choice between the LFS and IES factor income data is, however, not a straightforward one. Initial explorations revealed large numbers of outliers in the LFS data. Comparison of the two sources revealed a fair degree of correlation for the majority of the observations, but in many cases the difference was substantial. This required some further explorations and eventually a new 'combined' factor income variable was created that contained more of the IES data points than *mergefact*. In Figure 1 the average wage and number of observations falling within each factor group are compared. Four income variables are compared, namely the original LFS labour income (*inclabp_lfsorig*), the original IES labour income (*inclabp_iesorig*), and two versions of the new 'combined' income variable, *inclabp_old* and *inclabp_new*. In this section the comparison of the original labour income variables are discussed. From the figure it is clear that the average wages reported in the LFS are generally

[14] This hypothesis is later shown to be wrong.

higher than those reported in the IES. For certain occupation groups, such as clerks, plant and machine operators, elementary occupations, and unspecified occupations, the average LFS wage is more than twice that of the IES (also see Table 1).

Closer inspection revealed that large outliers in the LFS data are often the cause of the large income differences. This also explains the large difference between the overall IES and LFS labour income (R32,405 compared to R53,091). In at least 8 observations the LFS figure was exactly 1000 times higher than the IES figure, which clearly points to data capturing errors. In many other instances the figure from the one survey was exactly 10, 12 or 100 times the figure from the other survey. This discovery necessitated looking at records with large differences individually. A new person-level labour income variable, *inclabp_new*, which essentially selects the 'more appropriate' of the two reported labour income figures, was created. Section 7.1 in the appendix explains how this was done. This variable was later renamed *inclabp_old* when *inclabp_new* was scaled to match the household-level *inclab* variable. Section 4.2.4 explains how the scaling was done to create *inclabp_new*. The discussion of *inclabp_old* and *inclabp_new* and Figure 1 is also continued in section 4.2.4.

Figure 1: Comparing LFS 2000:2 and IES 2000 average wages and employment figures

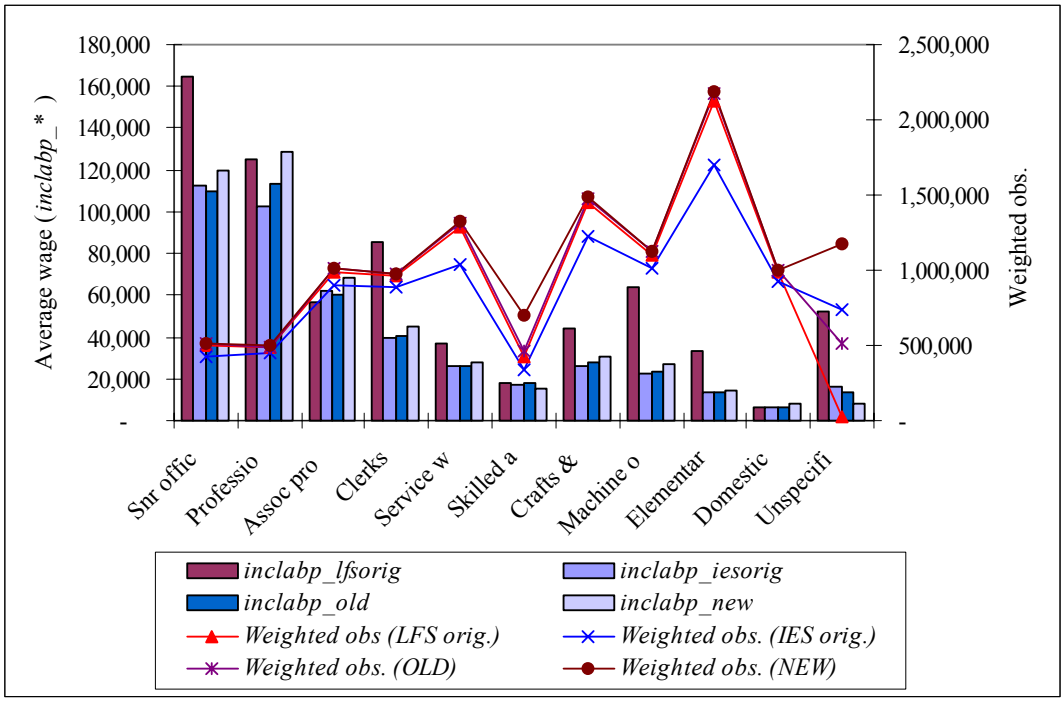


Table 1: Percentage differences (employment and wages)

	IES 2000 No. of obs.	LFS 2000:2 No. of obs.	Percentage under-/overreported	Mean (inclabp) (IES 2000)	Mean (w_inclabp) (LFS 2000:2)	Percentage under-/overreported
Legislators senior officials and managers	420,402	501,689	19%	112,787	165,009	46%
Professionals	449,222	481,781	7%	102,199	125,365	23%
Technicians and associate professionals	905,578	988,889	9%	62,146	56,831	-9%
Clerks	893,638	960,147	7%	39,350	85,415	117%
Service workers & shop market sales workers	1,038,507	1,289,362	24%	25,977	36,532	41%
Skilled agricultural and fishery workers	340,695	428,772	26%	16,751	17,810	6%
Craft and related trades workers	1,225,808	1,445,966	18%	25,852	43,746	69%
Plant and machine operators and assemblers	1,011,376	1,099,325	9%	22,855	64,158	181%
Elementary occupations	1,705,561	2,121,789	24%	13,442	33,358	148%
Domestic Workers	923,499	981,741	6%	6,302	6,258	-1%
Unspecified	737,041	22,946	-97%	15,937	51,858	225%
Total	9,651,327	10,322,407	7%	32,405	53,091	64%

Also reported in Figure 1 and Table 1 are employment figures reported in the IES and LFS. The LFS reports higher employment figures for almost all occupation groups. The patterns of employment in the LFS and IES are, however, similar. The unspecified category shows the largest difference. This is probably due to the fact that the LFS uses two separate questions to determine the occupation code, while the IES has only one question. Table 2 is a cross-tabulation of the two occupation code variables. Only 45 persons report their occupation code as 'unspecified' in the LFS, compared to the 1,798 unspecified workers in the IES. Many of these workers (1,253) reported no income in the LFS, which is why they are classified as 'not applicable' in the LFS.

As for the remainder of the occupation codes there is a fairly high correlation between the LFS and IES. In Table 2 those observations on the diagonal of the cross-tabulation show those persons who report the same occupation code in both the LFS and the IES. The percentage 'correctly categorised' observations is fairly high. The 'correctly categorised' row is defined as the number of observations on the diagonal divided by the column total (IES number of observations), while the column is defined as the number of observations on the diagonal divided by the row total (LFS number of observations). Section 4.2.4 continues the discussion of the new employment figures that are used together with *inclabp_new* and *inclabp_old*.

Table 2: Cross-tabulation of employment data

IES 2000 ➔ LFS 2000:2 ➓	Not applicable	Snr officials	Professionals	Assoc profess	Clerks	Service workers	Skilled agric	Crafts & trade	Machine operators	Elementary	Domestic workers	Unspecified	**Total**	*"Correctly" categorised*
Not applicable	77,472	15	14	51	40	58	84	89	84	164	65	1,253	79,389	97.60%
Snr officials	196	706	2	2	3		2	1	1	2		6	921	76.70%
Professionals	74	1	759	7	1	3	1			2		14	862	88.10%
Assoc profess	195	2	13	1,860	3	8		2	1	2		46	2,132	87.20%
Clerks	123			5	1,777	4	1	2	1	3		57	1,973	90.10%
Service workers	608	2	3	5	2	2,255	2	1	1	8	1	65	2,953	76.40%
Skilled agric	306	1		1	1	6	805	1		16	1	24	1,162	69.30%
Crafts & trade	527	3		1	3	2		2,826	5	12		63	3,442	82.10%
Machine operators	203	4		1	3		2	10	2,494	23		64	2,804	88.90%
Elementary	1,000		1	2	6	8	31	9	15	4,643	20	135	5,870	79.10%
Domestic workers	165					1	7	1	1	17	2,309	46	2,547	90.70%
Unspecified	10	1	2	1		1		1	1	2	1	25	45	55.60%
Total	80,879	735	794	1,936	1,839	2,346	935	2,943	2,604	4,894	2,397	1,798	104,100	
'Correctly' categorised	95.80%	96.10%	95.60%	96.10%	96.60%	96.10%	86.10%	96.00%	95.80%	94.90%	96.30%	1.40%		

Next comparisons between person- and household-level data for province, age, gender, location and race are made. The province variable was a perfect match for all observations and needs no further investigation. As far as age is concerned, we find that 1,214 observations report different ages. A further 2,091 observations are missing in either the LFS or the IES data, and hence cannot be compared. The remaining 102,359 of the observations (approximately 97% of the sample) report the same age. Furthermore, in 278 cases age only differs by one year, which could be accepted as an actual birthday taking place between the surveys or a minor reporting error. This suggests that the merge is fairly accurate on account of the age variable.

In Table 3 gender, location and race are cross-tabulated. The accuracy rates are very high for all these variables, and differs very little between the person- and household-level variables. This suggests that the mismatched person numbers is, after all, not such a big factor (see footnote 14). The location variable is, however, a bit worrying, with a substantial number of households (1,415) reporting rural in the LFS and urban in the IES, giving an accuracy rate of only 91.1%. This perhaps points at a definitional difference of urban and rural between the two surveys.

Table 3: Cross-tabulating gender, location and race

Gender	Person-level		Household-level	
LFS↓ IES→	Male	Female	Male	Female
Male	48799	424	15785	79
Female	414	53960	46	10196
Accuracy	99.2%	99.2%	99.7%	99.2%

Location	Household-level	
LFS↓ IES→	Urban	Rural
Urban	14466	383
Rural	1415	9854
Accuracy	91.1%	96.3%

Race	Person-level			
LFS↓ IES→	African	Coloured	Asian	White
African	83896	60	2	12
Coloured	93	11462	12	7
Asian	3	0	2040	4
White	36	6	0	5914
Accuracy	99.8%	99.4%	99.3%	99.6%
	Household-level			
LFS↓ IES→	African	Coloured	Asian	White
African	20719	15	1	5
Coloured	33	2689	2	1
Asian	1	0	523	2
White	6	2	0	2098
Accuracy	99.8%	99.4%	99.4%	99.6%

2.4. IES 2000 data problems

2.4.1. Literature review

Most of the data problems in the IES 2000 dataset can be ascribed to accounting and coding errors (Poswell, 2003). There are also a few inconsistencies when compared to the IES 1995 dataset. This restricts the confidence with which conclusions can be drawn about changes in income or expenditure over time. Van der Berg *et al*. (2003a) point out some of the specific problems that they have encountered in the IES 2000 dataset:

- When compared to IES 1995 the 2000 results indicate that income in South Africa has been declining strongly in real terms between 1995 and 2000. This contradicts national accounts and demographic statistics also compiled by Statistics South Africa. Statistics South Africa has since admitted that the surveys are incomparable.[15, 16] When building a

[15] This inconsistency is also pointed out by Simkins (2003). Simkins looks at the components of income and finds that there are large and inexplicable drops in some of these components, particularly net profits from business (half the 1995-level in nominal terms), occupational perquisites (down 42%) and

SAM one is more concerned about working with a database that reports the correct expenditure pattern rather than the correct absolute expenditure levels, since external control totals from the National Accounts (South African Reserve Bank) are used to scale expenditure up or down in any event. However, working with the correct expenditure levels is important when analysing poverty, since results (poverty lines and poverty measures) have to be realistic.

- Van der Berg *et al*. (2003a) find that there is evidence of "*sloppy work in the both gathering and the management of data*". Some of this 'sloppiness' is dealt with when expenditure totals are recalculated and expenditure items are mapped to new expenditure categories for use in the SAM. Some obvious inconsistencies are also dealt with in various ways where possible, although it is impossible to correct all errors resulting from miscoding, misrepresentation by respondents and misinterpretation of questions. Section 3 contains more detail in this regard.
- Van der Berg *et al*. (2003a) estimate that about 25% of the records are unusable for many purposes, mainly because total expenditure (including savings) and total income differ by more than 30%.[17] For the purpose of the SAM it is assumed that the larger of income or expenditure is the correct welfare measure, and all income or expenditure items are scaled up accordingly without changing the relative income or expenditure patterns of the households. Van der Berg *et al*. (2003a) are also concerned about the food expenditure reporting. In about 350 observations food expenditure was reported as zero (see section 3.2.7, which explains how food expenditure was imputed for households reporting zero food expenditure).[18]

While these problems are disheartening and are cause for concern about the usefulness of the dataset as a whole, the IES 2000 dataset remains, as pointed out by Van der Berg *et al*. (2003), the most recent available data and one should attempt to work with it. In sections 2.4.2 and 2.4.3 some further research is conducted, first into aggregate income and

unspecified income (down "*sharply*"). He finds that the share of income attributable to the household as a whole drops from 15.7% in 1995 to 9.9% in 2000, "*suggesting worse income component measurement in 2000 than in 1995*" (2003: 5).

[16] This also does not say that the 1995 levels were correct. Some believe that 1995 had some inconsistencies of its own. A recent study by Hoogeveen and Özler (2004) does in fact compare the two surveys after making various adjustments.

[17] It is unclear why 30% or more is regarded by the authors as too large a difference.

[18] The questionnaire asks respondents to give the actual value of expenditure or goods received in kind during the last 30 days. It does not ask the value of consumption. This may be an explanation for the occurrence of zero food expenditure.

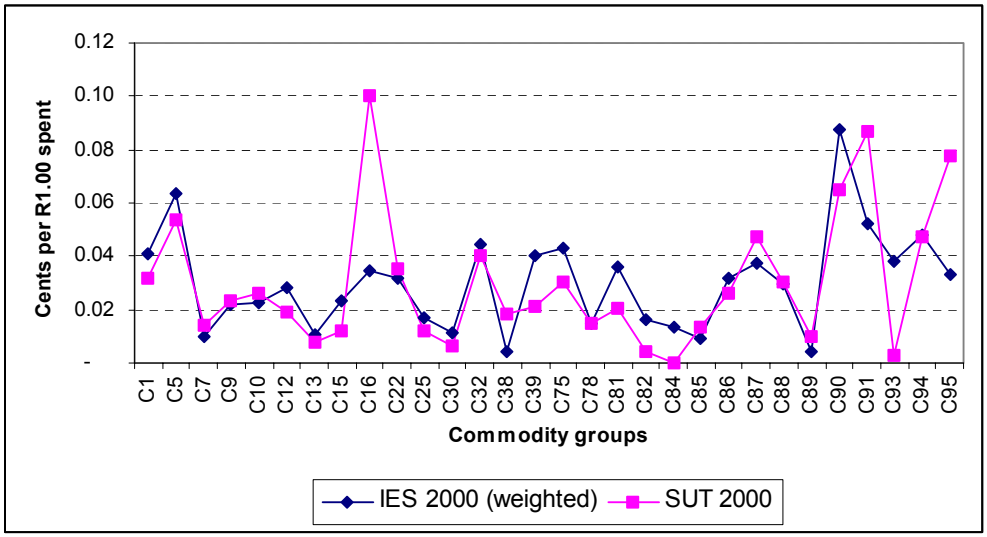
expenditure patterns reported in the IES 2000 compared to those found in other data sources, and second into the income and expenditure patterns by income or expenditure deciles within the IES 2000 database.

2.4.2. Comparing income and expenditure patterns with other data sources

The IES 2000 data is used to derive household expenditure accounts of the PROVIDE SAM. The commodity accounts are labelled *C1* to *C96,* denoting 96 household commodities or services that can be purchased by households. Additional expenditure accounts include household transfers (*hhtrans*), household income tax (*hhinctax*) and local taxes (*hhlocaltax*), household savings (*hhsav*) and other expenditures (*hhother*). All these additional expenditure accounts are also mapped directly from the IES 2000 expenditure data (see section 3.2.8 for more on the mapping of these accounts). A full listing and description of the accounts appears in the appendix (section 7.2, Table 11).

The Supply and Use Tables for 2000 (SUT 2000) (SSA, 2003b) contains an estimate of national household expenditure on 95 commodity groups (*C96* – domestic services – is not included in the SUT 2000 as a commodity/service). Although these expenditure estimates are based on data from the IES 1995, the Bureau of Market Research and the SARB, it provides a useful benchmark against which to compare the IES 2000 commodity expenditure estimates. Relative expenditures are compared, i.e. expenditure on each commodity group is divided by total expenditure to derive the expenditure share. Figure 2 only includes those commodity categories for which expenditure exceeded 1% of total expenditure, leaving 30 commodity groups. The figure shows that the relative expenditures reported by IES 2000 and SUT 2000 are similar to some extent, but large differences remain for some commodity groups (see Table 11 for account descriptions). Different interpretations between Statistics South Africa and PROVIDE of how expenditure categories should be mapped to the 95 commodity groups, as well as changes in the expenditure patterns between 1995 and 2000 may explain some of these differences.

Figure 2: Comparing patterns of expenditure from IES 2000 and SUT 2000



Note: Analytic weights assumed (variable *weight*)

The IES 2000 income and expenditure patterns can also be compared with the National Accounts for 2000 estimates published by the South African Reserve Bank (SARB 2000). The 2000 data (current prices) in the September 2002 bulletin is used in this analysis (SARB, 2002). The SARB uses a more aggregated commodity grouping to report final consumption expenditure by households. The ten commodity groups are listed in Table 4. The 95 commodity classes in the IES 2000 and SUT 2000 are mapped to these ten commodity groups.[19] Once mapped, the IES/SUT 2000 expenditure patterns can be compared to those reported by the SARB 2000. The comparison is shown in Table 4 and graphically in Figure 3.

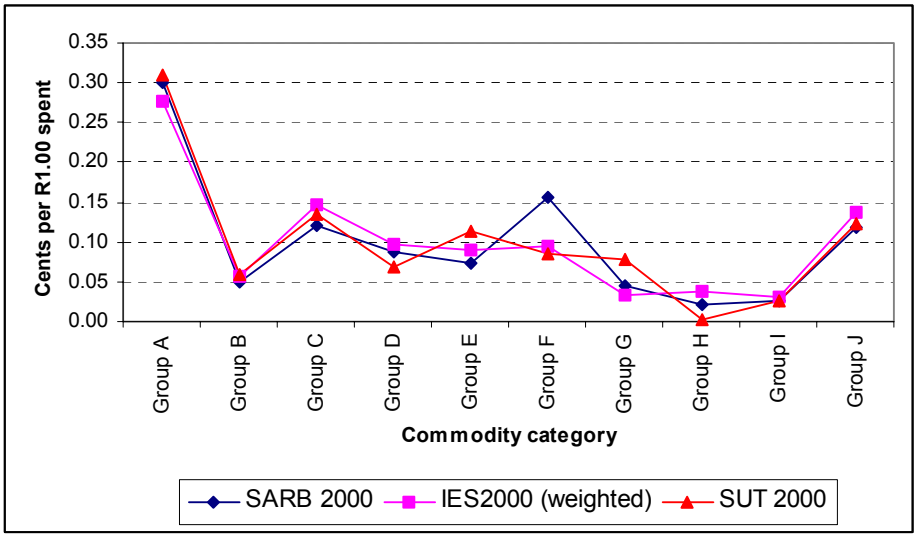
Table 4: Comparing patterns of expenditure from IES 2000, SUT 2000 and SARB 2000

		Cents per R1.00 spent		
Group	**Description**	**SARB 2000**	**IES2000 (weighted)**	**SUT 2000**
Group A	Food, beverages and tobacco	0.30	0.28	0.31
Group B	Clothing and footwear	0.05	0.06	0.06
Group C	Housing, water & electricity, other fuels	0.12	0.15	0.14
Group D	Furnishings, household equipment and routine maintenance	0.09	0.10	0.07
Group E	Health	0.07	0.09	0.11
Group F	Transport	0.16	0.09	0.08
Group G	Recreation, entertainment and culture	0.05	0.03	0.08
Group H	Education	0.02	0.04	0.00
Group I	Hotels, cafes and restaurants	0.03	0.03	0.03
Group J	Miscellaneous and services	0.12	0.14	0.12
	Total	1.00	1.00	1.00

Note: Analytic weights assumed (variable *weight*)

[19] The mapping file is available from the author on request.

Figure 3: Comparing patterns of expenditure from IES 2000, SUT 2000 and SARB 2000



Note: Analytic weights assumed (variable *weight*)

The IES 2000 and SUT 2000 expenditure patterns are more similar since the commodity groups are now more aggregated. As far as comparisons with SARB 2000 are concerned it is clear that the general expenditure pattern is similar to patterns in SUT/IES 2000. The largest difference occurs in transport (group F). The absolute expenditure levels, which are not shown in the figure, do however differ quite substantially.

The SARB also reports on household income from various sources. Current income is made up of (1) compensation of employees, (2) income from property, (3) current transfers from government, (4) current transfers from incorporated business enterprises and (5) transfers from the rest of the world. The income categories used for the PROVIDE SAM that are mapped from the IES 2000 income data include income from labour (*inclab*), income from GOS (*incgos*), income from household transfers (*inctrans*), income from incorporated business enterprises (*inccorp*), transfers from government (*incgov*), other income (*incother*) and income from the sale of home produce and livestock (*inchphc*). Unfortunately the mapping between these two sets of variables is not straightforward, and hence comparisons are difficult to make (see section 3.2.8 for more on the mapping of income sources in the IES 2000).

The SARB income from property is defined as dividend receipts, interest receipts net of interest payments, rent receipts net of maintenance, mortgage interest and consumption of fixed capital, and net profits of non-incorporated business enterprises. In the IES 2000 dividend income and interest receipts were mapped to *inccorp*. To avoid confusion it was decided to create a new income from property category, which is made up of the SARB-defined income from property plus transfers from incorporated business enterprises. In the case of IES 2000 this category is made up of *inccorp* plus *incgos*.

SARB 2000 does not make provision for inter-household transfers because, in theory, net (national) domestic transfers in any economy should be zero. Hence, it was decided to form a single transfer income category. In the case of SARB 2000 this is made up of net transfers from abroad and from government. In the case of IES 2000 it is made up of *inctrans* plus *incgov* minus *hhtrans* (household transfer expenditure). These two income sources are not strictly comparable, although it does give a general idea of the extent of total transfer income. Other income (*incother*) and income from the sale of home production (*inchphc*) in the IES 2000 was netted out by increasing the other income categories *pro rata*.

The relative contribution of each income source is shown in Table 5. The percentages reported in columns two and four show the relative contribution of the three income sources to total current income. Income tax and savings are expressed as percentages of current income.

Table 5: Comparing IES 2000 and SARB 2000 income and expenditure patterns

	SARB 2000 (R millions)	**Percentages**	**IES 2000 (R millions)**	**Percentages**
Income from labour (*)	421,168	64.7%	326,862	*77.6%*
Income from property (**)	194,377	29.9%	55,632	*13.2%*
Income from transfers (***)	35,238	5.4%	38,931	*9.2%*
Current income	650,783	100.0%	421,424	*100.0%*
Minus Income tax	90,296	13.9%	38,555	*9.1%*
Disposable income	560,487		382,869	
Minus Consumption	558,425		340,036	
Savings	*2,062*	*0.3%*	42,833	*10.2%*

Notes:

- IES 2000 figures are weighted and multiplied by the number of households (11 million) to obtain an estimate of the national totals. Data was also adjusted so that total income and expenditure matches (see section 4).
- (*) SARB 2000 income from labour quoted directly from source; IES 2000 labour income includes income from the sale of home produce.
- (**) SARB 2000: income from property plus transfers from incorporated business enterprises; IES 2000: *incgos* plus *inccorp*.
- (***) SARB 2000: net transfers from general government plus net transfers from the rest of the world; IES 2000: *inctrans* plus *incgov* minus *hhtrans*.
- Note: Analytic weights assumed (variable *weight*) for IES 2000 data

From Table 5 it is clear that the income patterns differ quite substantially between the two data sources. In both data sources labour is the most important source of income, but the relative contribution of labour is much higher in IES 2000. Notably, the definition of income from labour is fairly broad in IES 2000, as it includes non-monetary forms of remuneration such as food, housing and clothes. As far as the other income sources are concerned differences are most likely as a result of definitional differences. However, Simkins (2003) also notes a large drop in income from net profits in the IES 2000 data compared to IES 1995,

which may explain partly the large gap between SARB and IES 2000 as far as income from property is concerned (see footnote 15).

As far as income from transfers is concerned, the crude assumption that income from transfers from other households minus transfer payments to other households is equal to net transfers from the rest of the world presents a possible explanation for the large difference here between the SARB 2000 and IES 2000 data. Note that transfers from the rest of the world are not included in the IES 2000. Net inter-household transfers, which, according to the assumption made earlier is equal to net income from the rest of the world, equals R930 on average per household in the IES 2000, which is 2.8% of household income. The comparative figures in the SARB 2000 data indicate that net foreign transfers only contributed 0.02% to total household income.

Income tax is hugely underreported in the IES 2000 data. According to the IES 1995 figures the average income tax rate was 8.6% in 1995. This has dropped to 7.5% in 2000.[20] In the SARB 2000 the comparative average tax rates, calculated here as tax on income and wealth divided by current income, was 13.0% in 1995 and 13.8% in 2000. Although there have been some reductions in marginal income tax rates between 1995 and 2000, improved tax collections and a broadening of the tax base would have counteracted the impact of tax relief on average tax rates. This suggests that the SARB 2000 is more likely to be correct and that households underreported income tax in the IES 2000. Section 2.4.3 extends the investigation into taxes by looking at tax rates by expenditure deciles.

The IES 2000 questionnaire asks respondents to indicate the amount of tax paid in the last 12 months, be it provisional payments or PAYE and SITE deductions from salary. The financial year-end is at the end of February, while the IES 2000 was conducted around September/October. At this stage only about six months' worth of tax payments would have appeared on the salary slips. It is likely that many respondents simply failed to include tax payments made for the last six months of the previous tax year ending February 2000. It is further unclear why Statistics South Africa chose to include income tax as a household-level variable while clearly working individuals – who are required to provide other wage and salary information in any case – would have been able to give a better indication of their tax payments for the last 12 months. Furthermore, an unexpectedly large number of households failed to report any income tax, causing average rates to be lower than expected.

The reported savings in the IES 2000 appears to be very high. The savings variable in the IES 2000 is defined fairly broadly and is made up of various items including the capital

[20] These figures vary slightly depending on the way average income tax is calculated. The figure here is the average of all household tax rates rather than the rate calculated by dividing average income tax by average income (as was the case in Table 5).

component of mortgage repayments, investments in pension schemes and shares, funds deposited into savings accounts and 'stokvels', and all other forms of investment. In the SARB 2000 data it appears to be a residual item that balances current expenditure and current income, rather than an observed 'expenditure' by households.

2.4.3. Income and expenditure patterns by deciles

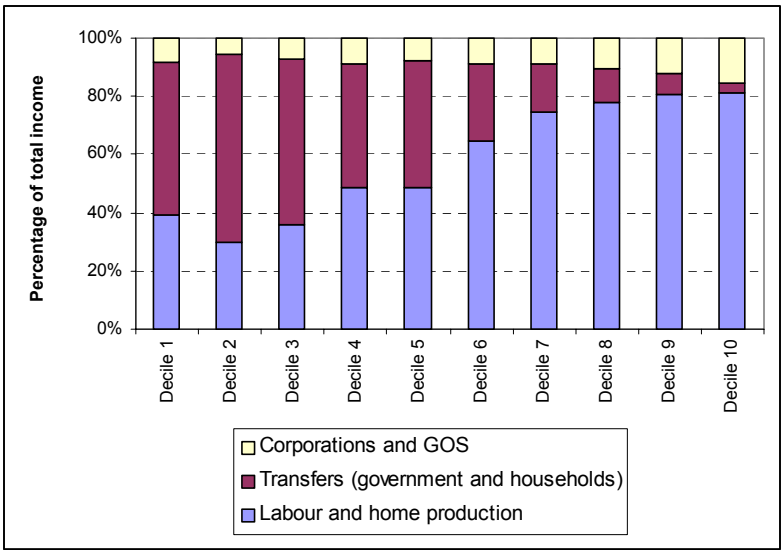
This section compares income and expenditure patterns between income or expenditure deciles. The objective is to see whether income and expenditure patterns behave according to expectations. As far as income is concerned, economic literature recognises the fact that sources of income usually differ between households in different income deciles. Typically low-income households rely more on financial support from government transfers, while middle- to high-income households earn a greater share of income from labour and from the ownership of capital.

Expenditure patterns across expenditure deciles are also considered. The food budget share provides a useful check to test the validity of Engel's Law, which states that low-income households spend a larger proportion of their income on food. Income tax payments were previously shown to be problematic due to the apparent underreporting in IES 2000, thus also calling for further explorations at the expenditure decile level. Finally, household savings are often strongly determined by household income. An analysis across expenditure deciles provides a useful check on the validity of the results.

The IES 2000 sources of income are grouped into three income groups for simplicity reasons. These are (1) income from corporations and GOS, (2) income from labour and the sale of home produce, and (3) income from government and household transfers. Figure 4 shows that income from labour and the sale of home produce and livestock, although by far the most important source of income at a national level, only contributes between 30% and 50% of total income in the low-income deciles (one to five). These income groups depend much more on transfer income from government and other households. Income from corporations and GOS are not a particularly important source of income for these low-income households.

As we move to the higher income groups, labour income becomes a very important source of income. Transfer income drops rapidly in relative terms (although not necessarily in absolute terms), while income from corporations and GOS increases steadily and causes labour income's share to drop slightly in the tenth decile. From the evidence presented it can be said that income patterns appear to be consistent and realistic across income deciles.

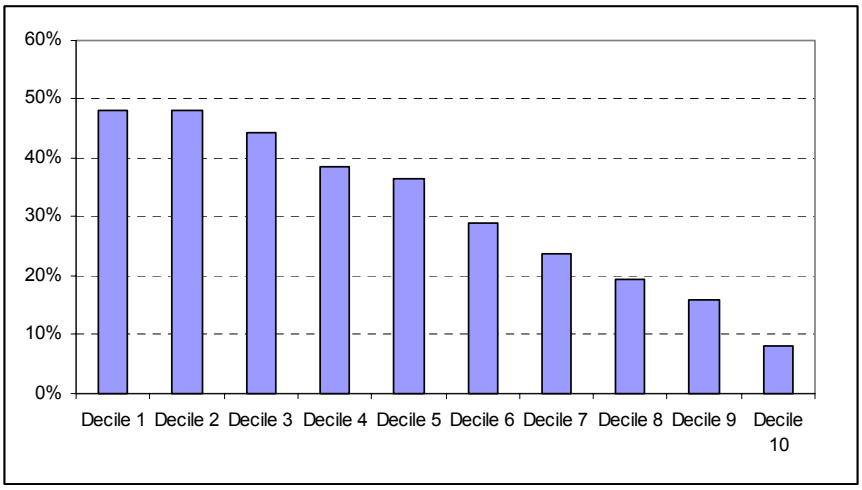
Figure 4: Relative income sources by income deciles



Note: Analytic weights assumed (variable *weight*)

Figure 5 graphs the food budget share by expenditure deciles. The expenditure deciles are based on adult equivalent *per capita* expenditure levels (see footnote 31). The food budget share stays the same between deciles one and two, and thereafter food becomes a normal good as the food budget share declines. In general the pattern of food expenditure is in line with expectations.

Figure 5: Food budget share by expenditure deciles using adult equivalent scales



Note: Analytic weights assumed (variable *weight*)

Previously it was shown that the average national tax rate as reported in the IES 2000 is significantly lower than that of the SARB 2000. By looking at the average reported tax rates within deciles it can be established whether (1) the pattern of tax expenditure is consistent with increasing household income levels, and (2) the extent of the underreporting. Using SARB 2000 data and the tax expenditure pattern of IES 2000, the 'expected tax rate' was

calculated and compared with the actual tax rate reported. In columns 1 and 2 of Table 6 the reported total expenditure and tax expenditure are listed.

The IES 2000 total expenditure of R415,526 million (2000 prices, including current expenditure, income tax and savings) is about 1.57 times lower than total current income of households before tax, consumption and transfers reported in SARB 2000 (R651,675 million, column 4). Unfortunately the SARB 2000 does not have data on the distribution of this income between deciles. Thus, assuming that the income distribution reported in the IES 2000 is correct, each decile's total expenditure can be increased 1.57 times so that the total income adds up to R651,675 million. Similarly, according to the SARB 2000 data total income tax was R90,296 million in 2000, almost three times as much as reported in IES 2000. The entries in column 5 are calculated by multiplying each decile's reported tax expenditure by a factor of 2.89. The 'expected' decile-specific average tax rates are now calculated (column 6) and compared with the reported tax rates (column 3). Given the way in which the expected rate is calculated the expected rate is exactly 1.84 (2.89 divided by 1.57) times the reported rate.

Table 6: Tax rates reported in IES 2000 (R millions)

	Total expenditure (IES 2000, pre-adjustment) (1)	Total tax (IES 2000, pre-adjustment) (2)	Tax rate (IES 2000, pre-adjustment) (3)	Total expenditure (SARB 2000) (4)	Total tax (SARB 2000) (5)	Expected tax rate (SARB 2000) (6)	Total expenditure (IES 2000, post-adjustment) (7)	Total tax (IES 2000, post-adjustment) (8)	Tax rate (IES 2000, post-adjustment) (9)
Decile 1	3,132	0	0.01%	4,912	1	0.02%	3,132	0	0.01%
Decile 2	5,955	3	0.05%	9,340	8	0.08%	5,955	3	0.05%
Decile 3	8,175	9	0.11%	12,821	25	0.19%	8,175	9	0.11%
Decile 4	11,006	22	0.20%	17,260	65	0.38%	11,006	22	0.20%
Decile 5	14,527	87	0.60%	22,782	251	1.10%	14,688	249	1.69%
Decile 6	19,641	289	1.47%	30,804	835	2.71%	20,199	847	4.19%
Decile 7	27,399	771	2.82%	42,970	2,228	5.19%	28,134	1,507	5.36%
Decile 8	40,616	1,522	3.75%	63,699	4,397	6.90%	41,656	2,562	6.15%
Decile 9	71,692	4,664	6.51%	112,435	13,473	11.98%	73,494	6,466	8.80%
Decile 10	213,384	23,890	11.20%	334,653	69,013	20.62%	222,452	32,958	14.82%
Total	**415,527**	**31,257**	**7.52%**	**651,675**	**90,296**	**13.86%**	**428,892**	**44,622**	**10.40%**

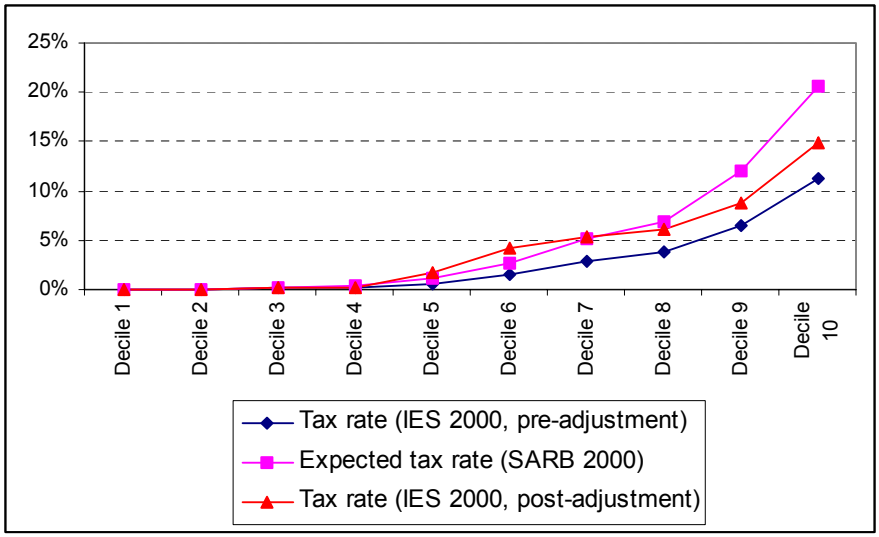
Possible reasons for the under-reporting of tax were given in section 2.4.2. In short, households could simply have understated the amount of tax paid relative to the income or expenditure level reported. Also, many households reported zero taxation, either in error or because they did not want to disclose the information. In the latter case the variable should have been coded as missing. When many households report zero taxation, the average within deciles is dragged down. The large number of households that report zero taxation suggests that this second reason is perhaps an important reason for the low average tax rates within deciles.

SARS (2004) report that in 2001/02 there were 3.5 million registered personal income taxpayers. Given income levels and the income tax brackets it is fair to assume that the majority of taxpaying households are in deciles five to ten (6.6 million households). Roughly 40-50% of all households are expected to be taxpaying households.[21] The IES 1995 data showed that 49% of households reported zero tax. Despite increases in the number of taxpayers between 1995 and 2000 the IES 2000 data reports that 77% of households (20,087 out of 26,224) paid no tax at all. If we only focus our attention on deciles five to ten the number reduces to 62% (9748 out of 15734), which is still considerably higher than our expected range of 40-50% (see footnote 20).

Two options now exist for adjusting the tax rates. The first is to multiply each household's tax expenditure by 1.84, the factor of under-reporting calculated earlier. This will ensure that the reported tax rate increases, but since the total expenditure also has to be increased at the same time, the adjusted rate will still be lower than the actual rate. Also, if the tax expenditure reported by households is simply multiplied by a fixed factor, those households that initially reported zero tax will still be recorded as paying no tax. A second option is to estimate the tax rate of households using an econometrically estimated model. Given the shortcomings of the first option, this approach was taken. The assumption was made that all households above the median of total expenditure were expected to also report income tax. For all households falling within this category and reporting zero tax, a tax expenditure value was imputed. Figure 6 compares this 'post-adjustment' tax rate with the 'pre-adjustment' and 'expected' tax rates (also see Table 6, columns 7 – 9). Section 3.2.7 explains in detail how the model was estimated.

[21] According to the LFS 2000:2 about 7831 individuals in the database earned income from salaries and wages sufficient to be eligible for tax, i.e. they earned more than R22000 (SARS tax tables for 2000/01). These 7831 individuals live in 6258 households, i.e. for every 100 taxable employees, about 80 households are expected to become taxpaying households. Thus, if the 3.5 million taxpayers all live in deciles five to ten, we expect about 2.8 million households to be taxpayers out of 6.6 million eligible households. This gives a rate of about 42%, or, say, 40-50% to allow for some flexibility and to include those households that pay tax on non-labour income.

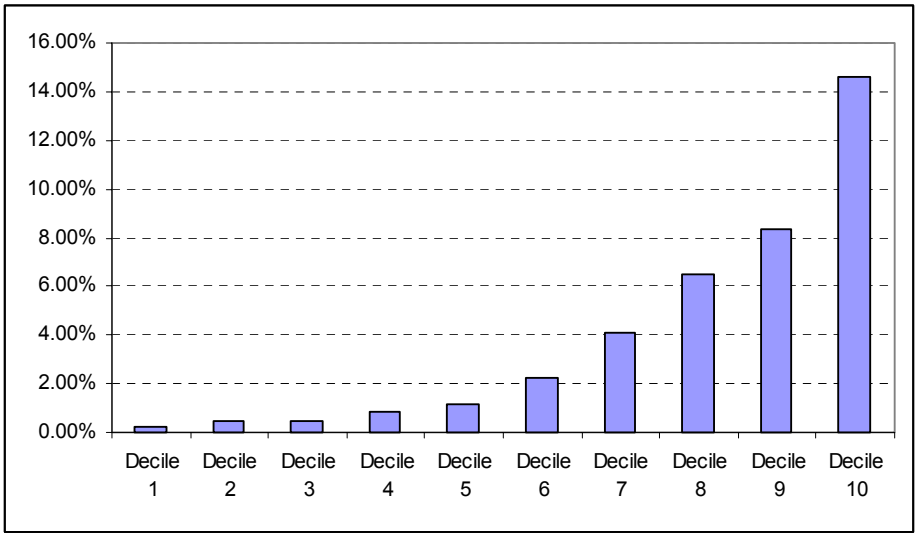
Figure 6: Average tax rates by expenditure deciles



Note: Analytic weights assumed (variable *weight*)

Finally, savings rates by expenditure decile are analysed. As shown in Figure 7 the savings rate increases exponentially as we move along the income scale. The highest income group contributes 74% to the pool of savings.

Figure 7: Savings rate by expenditure deciles



Note: Analytic weights assumed (variable *weight*)

3. Stata do-files to extract and reorganise data (*ies2000.do*)

A series of do-files were created to automate the process of extracting data, merging files, creating variables and making the necessary changes to variables. The discussion below describes in detail the functions of each do-file. Figure 8 and Figure 9 summarise the discussion and also show how the do-files are structured. The master do-file, *ies2000.do*, calls

up four sub-do-files, namely *readin.do*, *ies2000h.do* and *ies2000p.do* and *lfs2000_2.do*. Once these do-files have been run the 'original versions' of the IES 2000 (person- and household-level) and the LFS 2000:2 are saved (*ies2000h_orig.dta*, *ies2000p_orig.dta* and *lfs2000_2_orig.dta*). The second part of *ies2000.do* runs *adjustments.do*, which makes adjustments to the data so that incomes and expenditures match (see section 4.2), and *print.do*, which prints data tables for use in various sub-matrices of the SAM (see section 4.3). Do-files *ies2000h.do* and *ies2000h.do* also contain further sub-routines, which are discussed below. The Stata output of the first part of *ies2000.do* is saved in a log-file called *ies2000.log*.

Figure 8: Do-file structure of *ies2000.do*

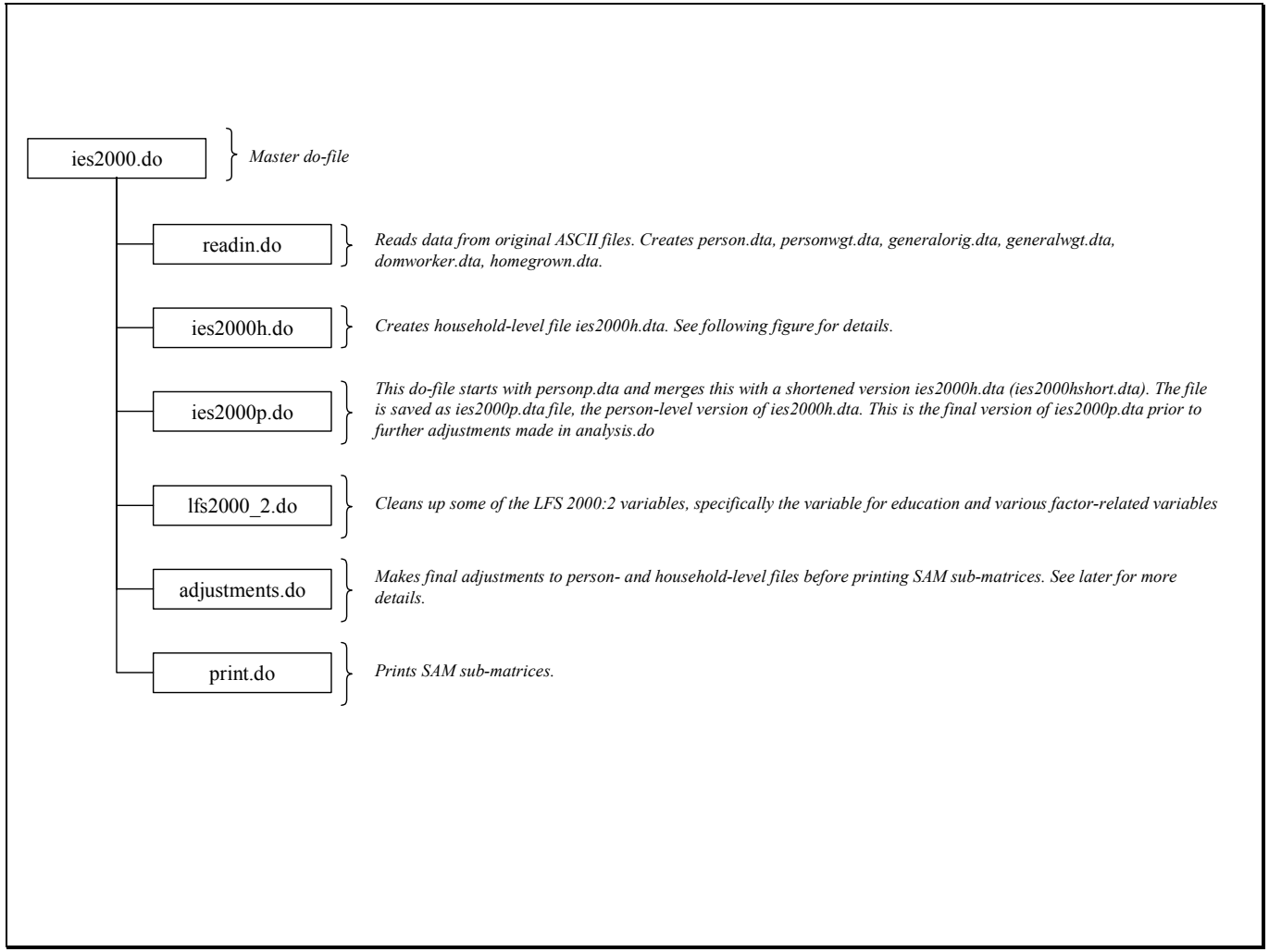
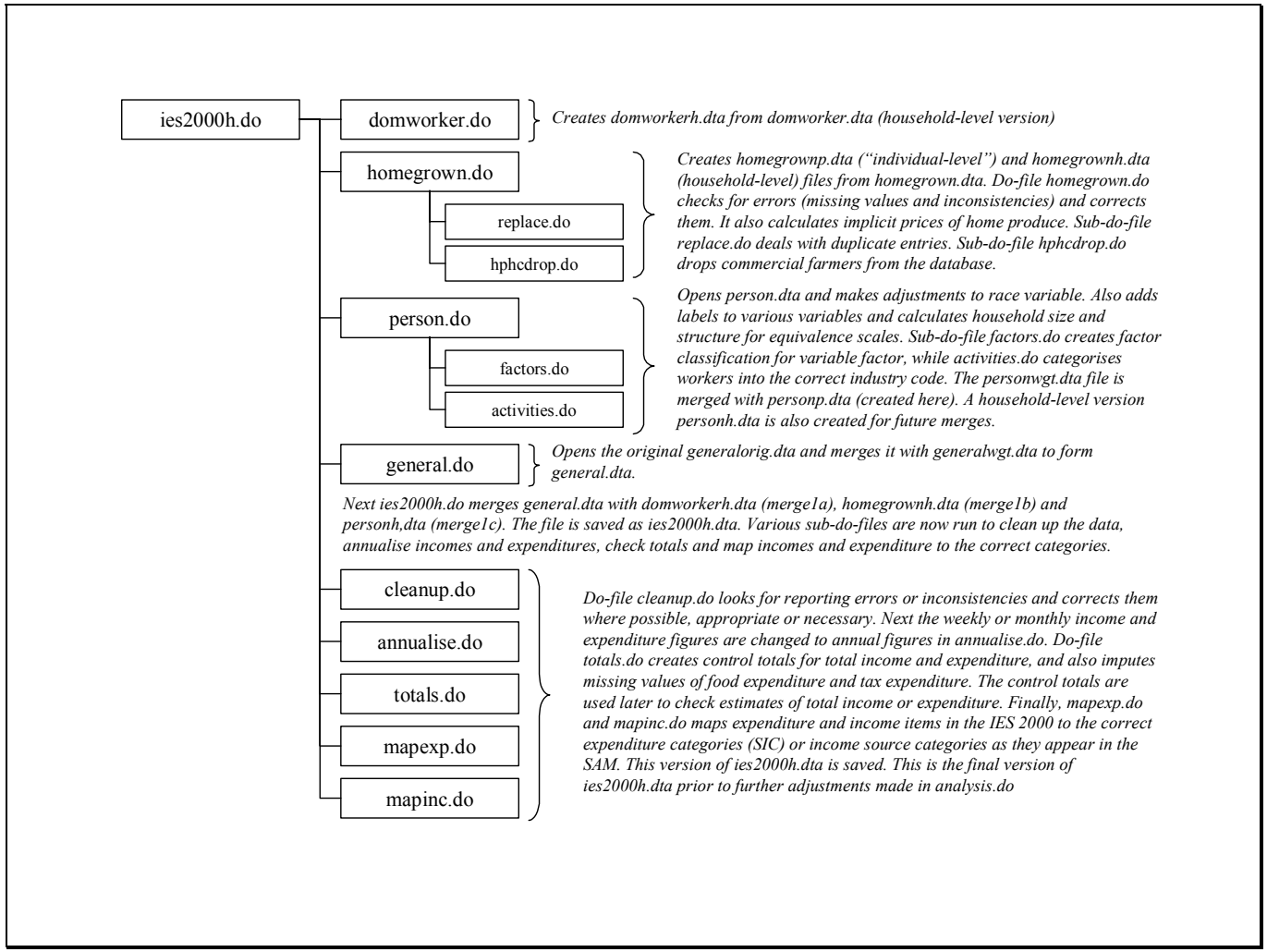


Figure 9: Do-file structure of *ies2000h.do*



3.1. Reading in the data (*readin.do*)

The raw IES 2000 data is supplied by Statistics South Africa in a series of ASCII text files. These fixed-width files are read into Stata using dictionary files specifying the location (column number) and length of each variable as it appears in each row of the ASCII files. The do-file *readin.do* calls up all the dictionary files. The IES 2000 ASCII files are converted to Stata files and saved as *person.dta*, *personwgt.dta*, *generalorig.dta*, *generalwgt.dta*, *domworker.dta* and *homegrown.dta*. [22, 23] These files are merged at a later stage to form person- and household-level IES 2000 files. The LFS 2000:2 ASCII files are also converted to Stata files and saved as *lfsperson.dta*, which is merged with the data from *worker.txt* to form *lfs2000_2p.dta*, and *lfshouse.dta*, which is merged with the data from *stratum_psu.txt* to form *lfs2000_2h.dta*. Finally, *lfs2000_2p.dta* and *lfs2000_2h* are merged to form a file called *lfs2000_2.dta*, which contains person- and household-level LFS 2000:2 data.

3.2. Forming a household-level IES 2000 dataset (*ies2000h.do*)

The main aim of do-file *ies2000h.do* is to create the household-level file *ies2000h.dta*. It starts by merging *general.dta* with *domworkerh.dta*, *homegrownh.dta* and *personh.dta*. Four do-files are called up within *ies2000h.do* in order to create or prepare these data files for merging.

3.2.1. Domestic workers (domworker.do)

Unlike the other household-level data files, the original file *domworker.dta* does not necessarily only contain a single entry per household. If a certain household has more than one domestic worker a new entry with the same household identification number (variable *hhid*) is added to the database. It is therefore necessary to create a household-level version of this file where each entry or observation reports the total expense for all domestic workers employed by the household. This avoids double counting when merging files. The following command adds up domestic worker expenses for observations with the same *hhid* number.[24]

```
for var P*: by hhid, sort: egen Xh = sum(X)
```

[22] To save computing time this do-file can be skipped by placing an asterisk at the beginning of the command line *do readin.do*, provided that the various **.dta* files already exist in the relevant folder.

[23] Originally only four ASCII files are supplied with the IES 2000 data. Two new files (*personwgt.dta* and *generalwgt.dta*) were obtained from Ingrid Woolard (HSRC). These files contain newly released person – level and household-level weights for the IES 2000. At present they are not 'official' yet and cannot be used. Also note that *general.txt* is now read in and saved in Stata as *generalorig.dta*.

[24] Note that *P** refers to all the variables starting with *P-*, i.e. the expenditure variables in *domworker.dta*.

Once this is done the household identifier (*hhid*) together with the household-level expenditure variables are saved as a new file named *domworkerh.dta*, the *–h* referring to the fact that this is now a household-level database.

3.2.2. Home production for home consumption (homegrown.do)

The next do-file (*homegrown.do*) starts with the original *homegrown.dta* database. Various modifications have to be done. As was the case with *domworker.dta* this file also allows for multiple observations with the same *hhid* when a household produces multiple products or keeps more than one type of livestock. Consequently it is necessary to create household-level income or expenditure variables before exporting some of the information to the household-level IES 2000 database. However, before these new variables can be created there are various problems in *homegrown.dta* that need to be addressed first. Apart from containing missing values, there were also numerous duplicate entries and various types of inconsistencies. The fact that some commercial farmers also reported under this section is problematic.[25]

A further problem, and perhaps a more serious error on the part of Statistics South Africa, is the treatment of the value of sales of livestock and produce. In the IES 2000 this is regarded as an expense and the value of sales is added to the expenditure side in the summary expenditure tables. Arguably the value of sales could be regarded as the input cost of produce sold, but an additional variable for value of inputs is also included in the database. This has led us to believe that their treatment of sales is incorrect. The correct procedure would have been to add the value of sales to the income side. Furthermore, the value of consumption of home produced goods should be added to the expenditure side. However, this is completely ignored by Statistics South Africa. The input cost of home production is correctly reported by Statistics South Africa on the expenditure side of the household account.[26]

The do-file is divided into seven parts.[27] After opening the *homegrown.dta* database and identifying multi-product households (parts 1 and 2) the occurrence of missing values is investigated in part 3. Missing values are usually problematic since they represent those cases where respondents failed to answer a certain question. However, in the case of the *homegrown.dta* database (and the IES 2000 database in general) various variables contain

[25] Commercial farmers typically have relatively large sales figures as they produce mainly for the market. This skews the data, since the majority of the respondents in *homegrown.dta* are small subsistence farmers or normal households producing goods for own consumption. The idea behind this section of the questionnaire was to capture information about these particular households and not commercial farmers.

[26] Statistics South Africa was unable to confirm or deny this author's belief that the treatment of these expenses and incomes is incorrect in the IES 2000.

[27] This section provides a brief summary of the functions of the do-file. A more detailed discussion appears in PROVIDE (____-b) (not published yet).

large numbers of missing values. These are in actual fact not true missing values, but rather values that are "uncoded", i.e. values that should have been coded as zeroes but were coded as missing because the specific section did not apply to the respondent.[28] Missing expenditure values only occurred when respondents indicated that they did not produce any goods for home consumption, nor did they keep any livestock. Thus, these missing values should rightfully have been coded as zeroes and are changed accordingly.

Next, a series of checks are performed (part 4) to see whether there are any coding inconsistencies in the data. The first type of inconsistency checked for is one where respondents indicate that they have no livestock but nevertheless report expenses (this error is later referred to as "miscoded" – see footnote 28 and section 3.2.5). The second inconsistency checked for is double counting of expenses. For some inexplicable reason many entries are clearly duplicate entries. All entries where the same expense appeared two or more times under the same *hhid* and the same produce or livestock code were identified and duplicate expenses were changed to zeroes. A sub-do-file *replace.do* was created to make the necessary changes (see Figure 8).

Part 5 of *homegrown.do* calculates the implicit prices of produce and livestock sold by the household. These prices are needed to value produce and livestock consumed by the household. As mentioned previously the value of home produce consumed was never captured on the expenditure side of the IES 2000. Where Hoogeveen and Özler (2004) used actual market prices for 2000 to value home consumption, the approach here makes use of actual sales data to calculate implicit prices. It is believed that this represents a more accurate valuation of produce and livestock in rural areas where general market prices seldom prevail due to separation from formal markets. The database contains data for quantity of sales and value of sales. The implicit price is the value divided by the quantity. Since the price is calculated for each observation, there is considerable variation in the implicit price. After closer inspection and correspondence with various specialists in the field it was decided to use the median price of produce and livestock sold to value home consumption. The study on home production for home consumption (PROVIDE, ____-b) gives a detailed analysis of the process followed to calculate implicit prices.

As mentioned in footnote 25 sales figures of commercial farmers should not have been included in the database. Thus, in part 6 of the do-file respondents that clearly had production and sales levels that are only possible when operating at a commercial scale were identified. The value of sales was changed to zero, but care was taken to keep home consumption figures

[28] This anomaly is also found in various other parts of the IES 2000 database. See section 3.2.5 for a more detailed discussion.

in the database, provided that these levels seemed realistic.[29] The entire part 6 is included as a separate do-file named *hphcdrop.do*.

Finally, in part 7, household-level variables were created for value of produce and livestock sold and consumed (*valprodcons*, *valprodsale*, *vallivecons*, *valliveprod*). These values, together with the household-level input costs (*P2205TOT*) are saved as *homegrownh.dta*, which is subsequently merged with the other household-level files.

3.2.3. Person-level data file (person.do)

The next do-file is *person.do*. This do-file opens *person.dta*, which contains all the information about each individual in each household, such as employment data and general demographic information. Variable *race* is slightly problematic since about 159 individuals report race as 'unspecified' (code 5 or 9). Since the SAM household and factor accounts are all disaggregated along racial lines, information about race is important. One option is to have a separate racial category labelled 'undefined', but this is not justifiable given that only 0.15% of the 104,153 individuals in *person.dta* do not specify their race. Another option is to drop observations with unspecified race from the sample, but this is also undesirable if it is possible to work around the problem.

Closer inspection revealed that some of the 'unspecified' individuals live in households where the head of the household did report his or her racial group. These individuals' race was changed to that of the head of the household. If the head of the household's race is unspecified, it is changed to that of the second household member (if available). After this adjustment 134 individuals remain unspecified. These people live in 39 households in which all members are unspecified. Unfortunately the whole process only 'saves' 25 individuals and 5 households.

Next, the do-file adds labels to variables and creates a few new ones, such as variable *region*, which maps the province variable (*prov*) to the four SAM regions. New variables are also created for the number of children (variable *K*), the number of adults (variable *A*), the total household size (variable *H*)[30], and the adult equivalent household size (variable *E*)[31].

---

[29] In some cases home *per capita* consumption levels were extremely high. One explanation for this is that own produce (such as maize) is possibly used for livestock feed, in which case it should have been reported as an input cost. Consumption levels were truncated at certain levels when they appeared unrealistically high.

[30] Although the original *person.dta* comes complete with a household size variable this variable appears to be incorrect. Consequently it is re-calculated here.

[31] The adult equivalence scale adjusts the actual household size to take into account differences in size and structure of households. The adjusted household size variable *E* is constructed using the formula $E = (A + \alpha K)^{\theta}$. May (1995, cited in Woolard and Leibbrandt, 2001) suggest that $\alpha = 0.5$ and $\theta = 0.9$ are plausible values for South Africa. Some sensitivity analysis around these values will be done at a later stage.

The *personwgt.dta* file, which contains alternative but unofficial person-level weights for IES 2000, is also merged with *person.dta* at this point.[32]

Next two sub-do-files are run within *person.do*. The first is do-file *factors.do*, which converts employment information contained in variable *jobcode* to create factor codes as used in the SAM (variable *factors*). Variable *jobcode* is based on answers given to the following question in the IES 2000 questionnaire: "*What kind of work did [the respondent] do in his/her main job during the past seven days?*" Statistics South Africa uses this information to give each respondent a four-digit occupation code based on the International Standard Classification of Occupations (ISCO 88). The meta-data file included with the LFS 2000:2 data shows how these codes can be used to derive an aggregated occupation code variable (*factors*) containing ten types of work and one category for 'unspecified or not adequately defined' (see Table 7). This aggregation differs only in one respect to the occupation codes used before in PROVIDE (2003b) in that 'domestic workers' have now replaced 'armed forces'.[33]

Table 7: Occupation codes (variable *factors*)

Factor code	Description
0	Not applicable/not working
1	Legislators, senior officials and managers
2	Professionals
3	Technical and associate professionals
4	Clerks
5	Service workers and shop and market sales workers
6	Skilled agricultural and fishery workers
7	Craft and related trades workers
8	Plant and machine operators and assemblers
9	Elementary Occupation
10	Domestic workers
11	Not adequately or elsewhere defined, unspecified

Source: (SSA, 2002a)

The second do-file in *person.do* is do-file *activities.do*. Individuals were asked to indicate in which industry they work. The answers were used to derive an industry code variable called *stccode*, based on the International Standard Industrial Classification (ISIC 1993) of all economic activities. Variable *stccode* was then used to group workers into 96 different

[32] Statistics South Africa has been unable to confirm whether these new weights can be used, hence the continued use of the old weights.

[33] Previously domestic workers were included under unskilled factors, but given that a separate industry for domestic services is also included in the SAM this distinction is useful. Armed forces used to be a separate group, but there were concerns about the representativity of this group as a separate factor account. Only 0.3% of African, 0.2% of Coloured and 0.2% of White workers were members of the armed forces. On aggregate only 0.2% of all workers were employed as members of the armed forces (IES 1995). On the other hand, 5.9% of African, 3.6% of Coloured, 0.4% of Asian and 0.1% of White workers are domestic workers, giving an aggregate of 5.0% of all workers (IES 2000).

industries or activities that are based on the same mapping of the commodity accounts (variable *activities*).[34]

Unfortunately the ISIC 93 codes used for variable *stccode* were not in all cases disaggregated enough in order to map factors to each of the 96 industry categories. In some cases, for example food production, the activity disaggregation went one step beyond the factor code disaggregation in variable *stccode* (see *activities.do*). The problem cannot be fixed in Stata. After the data has been extracted to a spreadsheet to form the factor-activity sub-matrix the Supply and Use Tables (SUT 2000) were used to find the relative value-added payments from activities to factors for those industries that are not disaggregated enough. The value-added payments are then allocated to the more disaggregated activity accounts in the ratios calculated from the Supply and Use Tables.

In order to obtain household-level labour income data the person-level labour income data has to be converted to household-level data. The following statement in Stata is used to achieve this.

```
for var P*: by hhid, sort: egen Xh = sum(X)
```

Only the observations relating to the head of the household is kept to create a household-level database that contains, among other things, total household-level income from labour, the race and gender of the head of the household, and information relating to the adult equivalence scales. This file is saved as *personh.dta* to distinguish it from the person-level *person.dta*.

3.2.4. General income and expenditure file (general.do)

Once *domworker.do*, *homegrown.do* and *person.do* has been run, the file *generalorig.dta*, which contains the bulk of the household-level income and expenditure data, is opened and merged with *generalwgt.dta*. The resulting file is saved as *general.dta*. The do-file programme now returns to *ies2000h.do* and merges *general.dta* with the household-level files *domworkerh.dta*, *homegrownh.dta* and *personh.dta*. The merge processes are done in succession. Variables *merge1a*, *merge1b* and *merge1c* show the merge results. Tabulating *merge1a* shows that there were 24,134 observations in *general.dta* not found in *domworkerh.dta*. It can be safely assumed that these households did not employ domestic

[34] The IES 2000 metadata file is somewhat confusing in this regard. It appears as if variables *stccode* and *jobcode* were meant to be used jointly to form a single occupation code variable based on ISCO 88. This is in fact how it was done in the LFS 2000:2 (see variable *Q41Occup*). In the LFS a second set of questions was then asked relating to the type of goods produced at the workplace. This information was then used to derive the activity code based on ISIC 93 (variable *Q42Indus* in LFS 2000:2). However, a comparison of the IES and LFS suggests that variable *stccode* is the same as the LFS industry code variable, while *jobcode* is the same as the LFS occupation code variable. It is therefore assumed that *stccode* (*activities*) and *jobcode* (*factors*) are correctly defined and coded in IES 2000.

workers and consequently did not answer this section. There were also 40 observations in *domworkerh.dta* for which no match could be found in *general.dta*.

```
  general & |
 domworkerh |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      24134       91.75       91.75
          2 |         40        0.15       91.90
          3 |       2131        8.10      100.00
------------+-----------------------------------
      Total |      26305      100.00
```

While 5 of these 40 observations report zero expenditure, the remaining 35 observations report expenditure ranging from R1,020 to R48,600, with an average of R10,195. The tabulation of *merge1b* shows 38 observations in *general.dta* not found in *homegrownh.dta*. One can again safely assume that these households did not partake in any home production for home consumption. However, 4 observations were found in *homegrownh.dta* that were not in *general.dta*. These households report zero expenditure on inputs, zero sales and very low consumption of own produce and livestock (output appears below).

```
  general & |
 homegrownh |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         38        0.14        0.14
          2 |          4        0.02        0.16
          3 |      26267       99.84      100.00
------------+-----------------------------------
      Total |      26309      100.00

             hhid  v~inputs  v~prodsale  v~prodcons  v~livesale v~livecons
 7353.  3.251e+12         0          0         248          0       3000
 7413.  4.061e+12         0          0           0          0          0
10924.  5.032e+12         0          0          45          0          0
11446.  5.072e+12         0          0          75          0          0
```

Finally, the merge between *general.dta* and *personh.dta* revealed that 46 observations were only found in *general.dta*. Whereas with the previous merges this was not a problem (one could simply assume that the relevant expenditures were zero) it is more problematic here since demographic information (race, gender, age, province) and employment data are now missing for 46 observations. This renders these 46 observations virtually unusable. Many of these 'mismatched' observations are dropped from the sample at a later stage.

```
  general & |
    personh |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         46        0.17        0.17
          3 |      26263       99.83      100.00
------------+-----------------------------------
      Total |      26309      100.00
```

3.2.5. Cleaning the data (cleanup.do)

After merging the datasets *cleanup.do* is run. As discussed in section 2.3 the IES 2000 dataset is plagued by numerous data problems. Do-file *cleanup.do* aims to rectify some of the minor ones, such as the simple adding-up problems. It also checks for consistency in the reported

totals and recalculates them where necessary. Before any of the actual 'cleaning up' can start the problem of missing values has to be investigated.

Usually missing values are coded in Stata as a dot (full stop). A large number of the variables in IES 2000, fortunately only on the expenditure side, contain very large numbers of missing values. Missing values in a Stata dataset create various problems. Any arithmetic operation on a missing value yields a missing value, which becomes problematic if, for example, total expenditure is to be calculated. Closer inspection revealed that large numbers of missing values only occurred in those variables that relate to optional questions. This created the suspicion that these are not true missing values, but rather a result of incorrect coding by Statistics South Africa. The following definitions are defined to clarify matters, i.e. observations that are coded with a full stop in the IES 2000 can fall into one of the following three categories:

- *Uncoded* – Some questions in the IES 2000 questionnaire were optional. Optional sections are preceded by a question that asks the respondent whether the expenses relating to that section are relevant to the household. If they answer no they may skip the section. In many instances Statistics South Africa coded expenses in these optional sections with missing values when the section was skipped. These are defined as uncoded observations and can be changed to zeroes.
- *Miscoded* – In some instances the preceding question to the optional sections was answered in the negative, but positive expenses were nevertheless reported in the optional section following the question. In these instances it is assumed that the original question was miscoded and should have been coded as 'yes'. Consequently the information content in the section is left as is.
- *(True) missing values* – The remaining missing values relate to respondents who should have answered a section given their response to the preceding question, but failed to do so. These are therefore true missing values. It can be argued that some of these missing values are a result of miscoding, i.e. that the preceding question should have been coded as 'no'. However, there is no basis on which such an assumption can be made, and consequently these values have to be treated as missing.

All variables coded with a full stop were systematically analysed to determine in which category they fall. Table 8 shows all the missing values (uncoded and true missing values) in the IES 2000 database, as well as those that were miscoded. The numbers of missing values reported in the original database is shown in column C. Only expense categories that

originally contained ten or more 'missing values' are included in the table. The table shows that the vast majority of these 'missing values' are in fact uncoded (column A). True missing values make up a very small number of the total number of observations (columns B and F). A number of observations were also miscoded, i.e. households that were expected to report zero expenditure (or a missing value if Statistics South Africa had been consistent in their treatment of optional sections) (column D). The sum of the 'missing values', miscoded observations and the remaining non-zero values (column E) gives the total number of observations in the database.

Table 8: Uncoded, miscoded and true missing values in IES 2000 [35]

	(A) Uncoded	(B) True missing	(C) Total "missing" (A + B)	(D) Miscoded	(E) Not missing, not miscoded	(F) Total obs. (C + D + E)
Monthly housing cost if rented						
P0303Q0101	17,485	5	17,490	1,018	7,755	26,263
P0303Q010101	17,541	10	17,551	962	7,750	26,263
P0303Q010102	17,562	12	17,574	941	7,748	26,263
P0303Q02	17,534	11	17,545	969	7,749	26,263
P0303Q03	17,452	49	17,501	1,051	7,711	26,263
P0303Q04	17,426	29	17,455	1,077	7,731	26,263
Monthly housing cost if owned						
P0303Q0501	8,006	9	8,015	1,309	16,939	26,263
P0303Q050101	8,075	14	8,089	1,240	16,934	26,263
P0303Q050102	8,075	13	8,088	1,240	16,935	26,263
P0303Q0502	8,033	11	8,044	1,282	16,937	26,263
P0303Q0503	8,018	12	8,030	1,297	16,936	26,263
Annual housing costs						
P0304Q0401	111	1	112	24,642	1,509	26,263
P0304Q0402	111	1	112	24,642	1,509	26,263
P0304Q0403	111	1	112	24,642	1,509	26,263
P0304Q0404	111	1	112	24,642	1,509	26,263
P0304Q0405	111	1	112	24,642	1,509	26,263
P0304Q05	109	-	109	24,644	1,510	26,263
Holiday expenses						
P03052Q0101	25,023	-	25,023	146	1,094	26,263
P03052Q0102	25,093	1	25,094	76	1,093	26,263
P0305~010401	26,032	-	26,032	49	182	26,263
P0305~010402	26,060	-	26,060	21	182	26,263
P03052Q02	26,060	-	26,060	21	182	26,263
Cigarettes						
P0702Q01	16,947	3	16,950	22	9,291	26,263
P0702Q02	16,947	3	16,950	22	9,291	26,263
P0702Q03	16,947	2	16,949	22	9,292	26,263
P0702Q04	16,947	2	16,949	22	9,292	26,263
P0702Q05	16,947	3	16,950	22	9,291	26,263
P0702total	16,947	2	16,949	22	9,292	26,263

[35] In this table observations were dropped when *merge1a* = 2 or *merge1b* = 2 or *merge1c* = 1. This leaves 26,263 observations.

Table 8 continued…

	Uncoded	True missing	Total "missing" (A + B)	Miscoded	Not missing, not miscoded	Total obs. (C + D + E)
Public transport						
P1504Q01	25,259	9	25,268	113	882	26,263
P1504Q02	25,263	9	25,272	109	882	26,263
P1504Q03	25,263	9	25,272	109	882	26,263
P1504Q04	25,262	9	25,271	110	882	26,263
P1504Q05	25,263	9	25,272	109	882	26,263
P1504Q06	25,263	10	25,273	109	881	26,263
P1504Q07	25,263	9	25,272	109	882	26,263
P1504Q08	25,263	9	25,272	109	882	26,263
P1504total	25,258	9	25,267	114	882	26,263
Cost of other sport/recreation goods						
P2003Q01	24,726	-	24,726	102	1,435	26,263
P2003Q02	24,736	1	24,737	92	1,434	26,263
P2003Q03	24,735	2	24,737	93	1,433	26,263
P2003Q0401	24,731	1	24,732	97	1,434	26,263
P2003Q05	24,734	-	24,734	94	1,435	26,263
P2003Q06	24,735	-	24,735	93	1,435	26,263
P2003Q07	24,735	1	24,736	93	1,434	26,263
P2003Total	21,405	-	21,405	3,423	1,435	26,263

From Table 8 it is clear that the numbers of true missing values are quite low. The only sections that contain more than ten missing values in certain variables are the two sections on monthly housing costs. Changing missing values to zeroes is justifiable given the small number of true missing values. It is better to rather lose information content of a few true missing variables by changing it to zeroes than lose the entire observation due to the adding-up restrictions in Stata. As explained previously missing values in expenditure categories from *domworkerh.dta* and *homegrownh.dta* were also changed to zero given that these were optional section in the questionnaire.

The only other anomaly in Table 8 is variable *P2003Total*. Answering this question was not optional, and hence the large number of uncoded observations is strange. However, as explained below, it was established that this was a result of a coding error. Since this is a sub-total it could simply be recalculated.

Another concern relates to questions 3.1 and 3.2 in the housing section. The questionnaire requests respondents to answer either the section labelled monthly housing cost *if rented* or the section labelled monthly housing cost *if owned*. In 21 cases households answered both sections. It does not seem highly unlikely that some households own property (excluding holiday homes) as well as rent property. The list below shows the reported values for each of the questions in these two sections. There appears to be no duplication (apart from record number 26015 – levy is reported twice). The error is small enough and is unlikely to affect results in any great deal, and is consequently ignored.

	P0303Q0..	P0303Q02	P0303Q03	P0303Q04	P0~050101	P0~050102	P030~0502	P030~0503
103.	24	0	0	0	5280	10788	0	72
25875.	12	0	0	0	0	4800	0	0
25895.	12	0	0	0	1029.6	842.4	0	0
25905.	12	0	0	0	5874	4806	0	0
25982.	24	0	0	0	240	264	0	0
26015.	6000	0	720	0	0	0	0	720
26031.	840	0	0	0	0	0	0	900
26078.	1236	0	0	0	679.8	556.2	7500	0
26102.	1188	0	0	0	1201.2	982.8	0	0
26153.	3600	0	0	0	3102	2538	2040	0
26180.	39600	0	0	0	2983.2	2440.8	0	0
26189.	1200	0	0	0	3828	3132	0	0
26198.	11760	0	0	0	2724	1980	0	0
26202.	12	0	0	0	6600	5400	0	0
26208.	6000	0	0	0	2400	12000	0	0
26220.	2400	0	0	0	9622.801	7873.2	0	0
26221.	1200	0	0	0	9622.801	7873.2	0	0
26222.	12	0	0	0	0	0	17496	0
26229.	10800	0	0	0	8400	0	0	11244
26250.	30600	0	0	0	16500	13500	0	600
26254.	6000	0	0	0	19800	16200	0	0

The remainder of do-file *cleanup.do* looks at some of the other problems in the database. The housing section (Part 3) of the questionnaire contains various problems. In section 3.3, question 1.1 should, by definition, be the total of questions 1.1.1 and 1.1.2. For 4525 out of the 26265 households interviewed this is not the case. Poswell (2003: 2) ascribes this to the "*wording problem*" in question 1.1.1. Fortunately when the expenditure categories used by PROVIDE are calculated, only the total monthly rent, i.e. the reported value in question 1.1, is taken into account. Hence this specific problem does not affect the work.

More problematic is question 5.1 of section 3.3. In this question homeowners with bonds are asked to provide a breakdown of their monthly instalments. The capital and interest parts should add up to the total monthly payment, but this is not the case for 1213 of the respondents. Since these individual components are used separately in the calculation of various expenditure categories, it is important to attempt to identify the problem. These 1213 observations are classified into four different 'types' of errors, as shown in Table 9.

Table 9: Four error types in housing section (monthly instalment on bond)

Error type	**No of obs.**	**Problem description**	**Action taken**
1	23	No total reported, only breakdown	Breakdown provided is nonsensical; hence sub-components are deleted (set to zero)
2	149	Assumed calculation errors (+/- R100)	Total monthly instalment is recalculated given the components
3	846	Only total reported, no breakdown	Capital-interest breakdown of 55/45 assumed given evidence of 'average' breakdown reported by other households.
4	195	Nonsensical reporting, incorrect capital	Same action as for error type 3.
TOT	**1213**		

The 23 observations of error type 1 are listed below. From the list it is quite clear that the reported capital component is often nonsensical. It appears as if respondents reported the

principal loan amount rather than the capital component of the monthly instalment. It was decided to set both the capital and interest components to zero.

	P030~0501	P0~050101	P0~050102
8321.	0	1250	0
9756.	0	6	0
10835.	0	6	0
11483.	0	0	10
13537.	0	0	30
14058.	0	1	0
14231.	0	98000	0
14242.	0	100000	0
14243.	0	130000	0
14244.	0	180000	0
14248.	0	70000	0
14253.	0	150000	0
17011.	0	350000	0
17673.	0	60000	0
19250.	0	75000	0
19255.	0	49000	0
19289.	0	1550	0
22106.	0	3	0
22246.	0	13	0
22401.	0	8	0
22486.	0	0	6
22623.	0	6	0
24638.	0	15000	0

The type 2 errors include those observations for which the calculated total differs with R100 or less from the actual reported total. Of the 1213 errors 149 households fall into this category. It is assumed that this relatively small error is a calculation error.[36] The total is therefore simply recalculated.

About 846 households only reported a total monthly instalment and provided no breakdown. Missing values are constructed or estimated given the available information on the average breakdown of those respondents that did provide all the information relating to repayment of bonds. Prior to the correction of errors a total of 816 households correctly reported their monthly instalments. An average capital payment of R1031 (55%) and an average interest payment of R839 (45%) were reported, giving a total of R1870 per month.[37] This breakdown will be assumed for the 846 households that only reported a total monthly repayment.

The remaining 195 households have a range of problems that are not always easily identifiable. However, the most important problem in this group of households is an

[36] The average monthly instalment reported by the 1897 bonded homeowners is R1691 (after correction of error types 1 – 4). The +/-R100 error boundary represents an error of less than 6%.

[37] The average repayment roughly relates to a principal loan amount of R133000, assuming a bond period of 20 years and an interest rate of 16% per annum. The specific capital-interest composition would be reached after about 16 years, irrespective of the principal amount. One would expect the average period lapsed to be closer to 10 years. However, given that most South African bonds have an 'access bond' facility whereby additional funds can be paid into the account to reduce the capital outstanding, the result is understandable. The principal also seems realistic given that these amounts were borrowed approximately 10-16 years prior to 2000 and the average house prices prior to 1990.

incorrectly reported capital amount. Respondents probably interpreted the question incorrectly and reported the principal loan amount rather than the capital component, the same problem as the one identified before. The summary below clearly shows how capital is grossly over-reported in these 195 households. Since many of the interest components were also nonsensical, it was decided to follow the same procedure as before by replacing the two components with estimated values for capital and interest, using the 55-45 split.

Variable	Obs	Mean	Std. Dev.	Min	Max
P0303Q0501	195	1621.282	984.4212	178	7000
P0303Q050101	195	55917.74	77056.19	0	550000
P0303Q050102	195	2002.79	9342.135	0	120000

After all corrections have been made a total of 1897 households correctly report a monthly instalment and the capital-interest breakdown. Incidentally, the capital-interest split remains fairly close to the initial estimate, with an average of 55.2% of the monthly instalment going towards capital and the remainder towards interest on the loan. The total instalment, however, is considerably lower than before. This is as a result of the incorrect capital components that have been deleted.

Variable	Obs	Mean	Std. Dev.	Min	Max
P0303Q0501	1897	1691.131	4730.005	3	190000
P0303Q050101	1897	934.128	4535.202	0	190000
P0303Q050102	1897	757.0033	885.5662	0	22500

The only other major change made in the housing section is made in question 6. Since all household expenditures in a SAM should be reported inclusive of VAT, the VAT component in this question should be added pro-rata to the other components that make up total payments for housing services. It is assumed that no VAT is payable on assessment rates and taxes, as this expenditure type is already a local or municipal tax on property. All other components, namely water, electricity, gas, sanitary services and refuse removal should be inclusive of VAT. VAT is therefore added pro-rata to these components. Some households (11 in total) reported only VAT. For these households VAT is added to the various components in six equal shares.[38]

Finally, a check is performed to see what the impact of the changes made has been on the reported total. Poswell (2003) points out that prior to any changes a discrepancy between calculated and reported housing expenditure totals in about 1.5% of households was seen. Having made the changes discussed above, it is found that this discrepancy has risen to almost 2% of households (520 observations). The average difference over all households, including those with zero difference, is about R6.43, as seen in the Stata output below.

[38] Previously (see PROVIDE, 2003a) the VAT component of households that reported only VAT was added to the variable for household indirect taxes. This is viewed as an unnecessary complication, hence the change in the procedure.

Poswell (2003) has no explanation for the discrepancy, apart from reporting or calculation errors. The figure for total housing expenditure will not be used in further calculations, hence no attempt is made to try and find the error.

Variable	Obs	Mean	Std. Dev.	Min	Max
P0303total	26265	354.3109	1534.388	0	190300
P0303total~K	26265	360.7422	1539.738	0	190300
diff	26265	6.431296	112.0424	-100	4544

Section 3.4 of Part 3 covers annual housing costs. These are costs that are typically made only once (if at all) in a given calendar year. Expense items in this category include alterations, improvements, and transfer costs. As seen in the Stata output table below, there is an average difference of R16.72 between the calculated and reported total. These differences occur in 171 of the observations. Closer inspection suggests that most of the differences are simply calculation errors. For some observations the total annual housing expenditure was reported as zero despite having reported expenses that supposedly make up the total. This explains many of the large differences. The total annual housing expenditure figure will not be used further, and hence no attempt is made to correct the discrepancies that exist.

Variable	Obs	Mean	Std. Dev.	Min	Max
P0304total~K	26265	561.9166	5673.867	0	266200
P0304total	26265	545.1938	5642.526	0	266200
diff	26265	16.72283	602.733	-2889	67400

Finally, section 3.5 of Part 3 covers expenditure on holiday accommodation. As shown in the Stata output tables below the calculated total differs only slightly from the reported total. The difference cannot be explained properly, but only occurs in 180 of the observations. At this point nothing is done about the problem.

Variable	Obs	Mean	Std. Dev.	Min	Max
P03052tota~K	26835	186.5814	2649.756	0	235000
P03052total	26835	186.6786	2649.828	0	235000
diff	26835	-.0971492	32.81482	-4000	3002

Variable	Obs	Mean	Std. Dev.	Min	Max
P03052tota~K	180	339.3667	1443.965	0	10002
P03052total	180	353.85	1461.898	0	10000
diff	180	-14.48333	401.5173	-4000	3002

Part 5 covers expenditure on food items. All totals were recalculated and compared to the reported totals. As shown in the Stata output table below, no substantial errors were found. The largest difference occurred in question 5.3 (*P0503tot**). The error is small enough to safely ignore and is likely to be a calculation error.

A more serious problem in the food expenditure section is that about 351 observations (1.3% of the dataset) report zero food expenditure. Given that respondents should also report

the value of free food received, it is highly improbable that food expenditure in a given month (in this case inflated to annual figures) can be zero.

Variable	Obs	Mean	Std. Dev.	Min	Max
P0501total	26835	149.0692	162.9203	0	11000
P0501total~K	26835	149.0815	163.7497	0	11330
P0502total	26835	140.9402	178.3022	0	7295
P0502total~K	26835	140.9402	178.3022	0	7295
P0503total	26835	20.03041	39.03968	0	1632
P0503total~K	26835	19.32327	36.3151	0	1632
P0504total	26835	23.93777	25.74955	0	471
P0504total~K	26835	23.93777	25.74955	0	471
P0505total	26835	54.24405	75.13046	0	1989
P0505total~K	26835	54.24405	75.13046	0	1989
P0506total	26835	61.56944	61.47751	0	1520
P0506total~K	26835	61.56944	61.47751	0	1520
P0507total	26835	28.06924	43.64455	0	1148
P0507total~K	26835	28.06924	43.64455	0	1148
P0508total	26835	24.4973	23.43113	0	2000
P0508total~K	26835	24.4973	23.43113	0	2000
P0509total	26835	9.757146	22.75167	0	623
P0509total~K	26835	9.757146	22.75167	0	623
P0510total	26835	19.12331	34.51434	0	2000
P0510total~K	26835	19.13956	35.45512	0	2116
P0511total	26835	5.829216	31.71888	0	1000
P0511total~K	26835	5.832793	31.83621	0	1096
P0512total	26835	27.53367	40.9759	0	2000
P0512total~K	26835	27.52055	40.39669	0	1648
P0513total	26835	31.94481	122.7268	0	3500
P0513total~K	26835	31.94481	122.7268	0	3500

Parts 6 to 11 cover beverages (alcoholic and non-alcoholic), tobacco products, personal care products, general household services and household fuel. As before, all reported totals were checked against the calculated totals. The discrepancy in question 6.1(2) (*P0601Q02**) appears to be another error. The total of question 6.2(2) was incorrectly reported as the total of 6.1(2). Again, since the reported totals will not be used nothing is done to rectify the error.

Variable	Obs	Mean	Std. Dev.	Min	Max
P0601Q01tot	26835	6.362102	26.69734	0	1840
P0601Q01to~K	26835	6.362102	26.69734	0	1840
P0601Q02tot	26835	23.75629	112.7319	0	8700
P0601Q02to~K	26835	20.76083	51.71083	0	3000
P0602Q01tot	26835	13.61394	82.6549	0	6000
P0602Q01to~K	26835	13.61394	82.6549	0	6000
P0602Q02tot	26835	23.75629	112.7319	0	8700
P0602Q02to~K	26835	23.75629	112.7319	0	8700
P0702total	26835	30.31276	82.5225	0	1680
P0702total~K	26835	30.31276	82.5225	0	1680
P0801total	26835	99.03335	143.3689	0	3655
P0801total~K	26835	99.03335	143.3689	0	3655
P0901total	26835	35.3466	39.63331	0	1040
P0901total~K	26835	35.3466	39.63331	0	1040
P1001total	26835	3.336464	61.85509	0	4784
P1001total~K	26835	3.336464	61.85509	0	4784
P1100total	26835	31.51127	129.9219	0	18500
P1100total~K	26835	31.51127	129.9219	0	18500

Parts 12 to 21 contain all the annual expenditures incurred by the household. This ranges from expenditure on clothing to various services purchased. Part 17 pertains to communication expenses. As was the case with VAT in the housing services section, VAT on

telephone services was added pro-rata to the components that were thought to make up telephone fees (question 17.1(1)). Thus, VAT was added pro-rata to telephone rental, private calls, calls made from cellular phones, cellular network charges (connection and rent) and Internet charges. For those households that only reported VAT and no other telephone expenses VAT was added in equal shares to the various components.

The Stata output table below shows that the calculated total of all expenditure totals match the reported total, except for Part 20.1 (*P2001Tot**). It appears a similar error to the one explained before was made here, with the total of Part 20.3 (*P2003Tot**) incorrectly reported here as the total of Part 20.1. The error is not amended since the components of the total are used individually. Another interesting observation can be made in Part 21.3 (*P2103Tot**). The figure reported here is for net income tax, i.e. income tax paid minus rebates. The range given shows that there are some households that 'paid' negative taxes, i.e. their rebates exceeded their payments. In such cases income tax is in actual fact an 'income' rather than an expense.

Variable	Obs	Mean	Std. Dev.	Min	Max
P1201total	26835	947.2283	1600.976	0	39160
P1201total~K	26835	947.2283	1600.976	0	39160
P1202total	26835	451.6619	756.3367	0	14800
P1202total~K	26835	451.6619	756.3367	0	14800
P1301total	26835	440.802	2193.456	0	94500
P1301total~K	26835	440.802	2193.456	0	94500
P1302total	26835	154.7855	673.6884	0	40000
P1302total~K	26835	154.7855	673.6884	0	40000
P1303total	26835	247.9006	972.7045	0	34600
P1303total~K	26835	247.9006	972.7045	0	34600
P1304total	26835	37.6171	224.8277	0	13400
P1304total~K	26835	37.6171	224.8277	0	13400
P1401total	26835	1031.47	4035.006	0	144000
P1401total~K	26835	1031.47	4035.006	0	144000
P1402total	26835	197.8958	1140.258	0	75200
P1402total~K	26835	197.8958	1140.258	0	75200
P15012total	26835	1617.553	7933.802	0	328930
P15012tota~K	26835	1617.553	7933.802	0	328930
P1502total	26835	641.0117	1314.854	0	40320
P1502total~K	26835	641.0117	1314.854	0	40320
P1504total	26835	47.24636	667.8682	0	42000
P1504total~K	26835	47.24636	667.8682	0	42000
P1601total	26835	203.2405	1345.858	0	45500
P1601total~K	26835	203.2405	1345.858	0	45500
P1701total	26835	687.5704	1999.454	0	72000
P1701total~K	26835	687.5704	1999.454	0	72000
P1801Q01Tot	26835	920.5109	3877.379	0	157150
P1801Q01To~K	26835	920.5109	3877.379	0	157150
P1801Q02Tot	26835	123.605	1412.967	0	100000
P1801Q02To~K	26835	123.605	1412.967	0	100000
P1901Q01Tot	26835	2.748873	26.65607	0	1680
P1901Q01To~K	26835	2.748873	26.65607	0	1680
P1901Q02Tot	26835	4.421576	34.92754	0	2040
P1901Q02To~K	26835	4.421576	34.92754	0	2040
P1901Q03Tot	26835	42.72256	331.2455	0	29050
P1901Q03To~K	26835	42.72256	331.2455	0	29050
P2001Total	26835	71.33564	601.3598	0	40000
P2001Total~K	26835	243.5663	1729.987	0	97000
P2003Total	26835	71.70225	621.7803	0	40000
P2003Total~K	26835	71.33564	601.3598	0	40000
P2004Total	26835	196.9735	1544.142	0	222000
P2004Total~K	26835	196.9917	1546.755	0	222488
P2101Total	26835	120.0666	2253.016	0	251800
P2101Total~K	26835	120.0666	2253.016	0	251800

```
    P2102Total |   26835    1152.961   4247.576          0     360000
  P2102Total~K |   26835    1152.961   4247.576          0     360000
    P2103Total |   26835    2603.309   27987.51     -20000    3000000
  P2103Total~K |   26835    2603.309   27987.51     -20000    3000000
    P2104Total |   26835    4470.326   39426.95          0    3105800
  P2104Total~K |   26835    4470.326   39426.95          0    3105800
    P2105Total |   26835     432.705   7858.462          0     874000
  P2105Total~K |   26835     432.705   7858.462          0     874000
```

3.2.6. Annualising and creating control totals (annualise.do and totals.do)

Some of the income and expenditures reported in the IES 2000 questionnaire is weekly or monthly, and should be converted to annual figures. This do-file simply changes all the weekly or monthly figures to annual figures. Do-file *totals.do* recalculates the income and expenditure sub-totals and also creates variables *totincCHECK* and *totexpCHECK*, which can be used to make sure that the mapping of income and expenditure in do-files *mapinc.do* and *mapexp.do* is done correctly. At the end of do-file *totals.do* food expenditure values that are missing or zero are imputed. A similar process if followed for missing or zero tax expenditure values when the total income level of the household creates the expectation that the household should have reported tax expenditure. A discussion of food and tax expenditure imputations follows in section 3.2.7.

3.2.7. Imputing 'missing' food and tax expenditure values

The total food expenditure variable of all households reporting zero food expenditure was changed to missing based on the assumption that each household has to at least report some expenditure on food. These 'missing' food expenditure values were then imputed by estimating a double-log Engel equation of the form

$$\ln(Y_i) = a + b.\ln(X_i) + c.\ln(H_i) + \varepsilon_i$$

where a, b and c are constants, Y_i is the food expenditure (*logfoodexp*), X_i the total household expenditure (*logtotexp*), H_i the household size (*logH*) and ε_i the error term. This double-log formulation ensures that the share of total expenditure spent on food declines as total expenditure increases, while larger households benefit from scale economies (see Van der Berg *et al.*, 2003b). The following regression results were obtained (sampling weights used):

```
      Source |       SS       df       MS              Number of obs =   25944
-------------+------------------------------           F(  2, 25941) =30994.49
       Model |   14165.074     2    7082.537           Prob > F      =  0.0000
    Residual |   5927.7668 25941  .228509572           R-squared     =  0.7050
-------------+------------------------------           Adj R-squared =  0.7050
       Total |  20092.8408 25943   .77449951           Root MSE      =  .47803

------------------------------------------------------------------------------
  logfoodexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logtotexp |   .5992633   .0025101   238.74   0.000     .5943434    .6041832
        logH |   .2185103   .0041317    52.89   0.000      .210412    .2266087
       _cons |   2.439185   .0247711    98.47   0.000     2.390632    2.487738
------------------------------------------------------------------------------
```

The model explains about 70.5% of the variation of food expenditure. All coefficients are significant, while the positive sign of all the coefficients makes economic sense. This model was now used to estimate the expected values of the approximately 350 households that failed to report any food expenditure (missing or zero values), but did report a value for total expenditure. The model suggests that a two-person household earning R250,000 per annum will spend about R1,909 per month on food, while a two-person household earning R60,000 per year will spend about R812 per month on food. These results appear to be realistic. The average food expenditure rose by R7,096 per annum to R7,162 per annum as a result of the imputation. The total expenditure levels were adjusted accordingly.

Section 2.4.3 looked in some detail at average tax rates per expenditure decile and found that tax was grossly underreported in the IES 2000. The tax imputation was done in a similar way. However, rather than fitting the regression model to the entire sample, only those households whose total expenditure was higher than the median total expenditure were included. The progressive tax system in South Africa is such that only about half of households are eligible for tax. A regression model was run with the following double-log function form:

$$\ln(T_i) = a + b.\ln(X_i) + c.\ln(W_i) + \varepsilon_i .$$

Parameter a is a constant term, while b and c are coefficients. The independent variable, T_i is the tax level (*logtax*), X_i the total expenditure (*logtotexp*) and W_i the number of working adults per household (*logW*). This variable was found to be more significant than the household size variable. Various other independent variables were also tested, but none of these were significant. These included race dummy variables, location dummy variables, sector of employment (formal/informal) dummy variables, occupation code dummy variables and so forth. The regression model results appear below:

```
      Source |       SS       df       MS              Number of obs =    5828
-------------+------------------------------           F(  2,  5825) = 2976.37
       Model |  8225.37271     2  4112.68636           Prob > F      =  0.0000
    Residual |  8048.85079  5825  1.38177696           R-squared     =  0.5054
-------------+------------------------------           Adj R-squared =  0.5053
       Total |  16274.2235  5827  2.79289918           Root MSE      =  1.1755

------------------------------------------------------------------------------
      logtax |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logtotexp |   1.420247   .0185721    76.47   0.000     1.383839    1.456656
        logW |  -.2130343   .0263952    -8.07   0.000    -.2647787     -.16129
       _cons |  -7.334131    .201661   -36.37   0.000    -7.729462   -6.938801
------------------------------------------------------------------------------
```

The regression results indicate that the model explains about 50.5% of the variation in reported tax expenditure. The positive coefficient for variable *logtotexp* shows that as total expenditure, which is a proxy for income, increases, more tax is paid. Furthermore, because the coefficient is greater than one, a higher marginal tax rate is payable as households moves

into a higher income bracket. The negative coefficient for variable *logW* suggests that less tax is payable if more family members contribute to the pool of household income. This makes economic sense as well. For example, if a single household member contributes R120,000 to the household income, her marginal tax rate will be much higher than two members contributing R60,000 each. Finally, the negative constant, which is equivalent to about -R1,532 can be interpreted as a tax rebate (the actual tax rebate was about R3,800 in 2000). The increase in the average reported tax rate was quite significant, increasing from R2,831 to R4,041. However, given the large number of households reporting zero tax where they were expected to have paid tax, this sharp increase was expected. The total expenditure levels were adjusted accordingly.

3.2.8. Mapping income and expenditure categories (mapexp.do and mapinc.do)

Since the IES 2000 was initially developed to provide information for the calculation of the CPI, the expenditure items are not grouped in a way that is consistent with the commodity categories used in the SAM. On the expenditure side each expenditure item has to be mapped to a specific commodity group (see *mapexp.do*). These commodity groups are based on the activity-based Standard Industrial Classification (SIC) codes of 1993. A total of 95 commodity groups are created based on the SIC codes, plus an additional commodity group denoting domestic services (*C1 – C96*) (see Table 11).

Apart from the 96 commodity groups created, additional expenditure categories are also created for inter-household transfers (variable *hhtrans*), the payment of income tax (variable *hhinctax*), the payment of local or provincial 'taxes' (variable *hhlocaltax*) and savings (variable *hhsav*). Some expenditure items, such as pocket money, costs relating to home production, gambling expenses and other losses could not be mapped successfully and were included in a variable called *hhother*. The value of home consumption plus input costs (variable *hhhphc*) is also created separately, but can be incorporated into *hhother* if this information is not needed separately. Variable *hhother* is netted out by allocating the expenses pro-rata to the other expense categories (see discussion later).

A few important assumptions were made in order to map or create commodity group *C89* (Financial Services Indirectly Measured, or FSIM). This is an expense category that cannot be directly mapped from the IES 2000 data since it is an implicit expense incurred by households that either borrow or invest money with a bank. A large part of the financial service industry's revenue comes from charging higher interest rates for loans than it pays for deposits made at the bank. Thus, when households pay interest on loans an implicit finance charge is included since the household pays a premium on the 'base' interest rate. Similarly, when a household receives interest payments from a bank it actually receives less than the 'base' interest rate. These implicit finance charges appear nowhere on the expenditure side of

the IES 2000. In SUT 2000 the national-level FSIM is derived from the System of National Accounts (SNA), but no breakdown is given as to how individual household budgets are affected by such charges (SSA, 2003b). In order to incorporate FSIM at a household level the following assumptions are thus made:

- *Expenditure side*: 10% of interest payments made will be regarded as FSIM expenses. These interest payments include the interest component of the monthly installment on a bond (variable *P0303Q050102*) and other interest on finance (variable *P2104Q0102*).
- *Income side*: Since interest received would have been higher if FSIM were zero, household-level interest receipts (variable *P2401Q05h*) are increased by 10%. This additional income is also added to the expenditure category FSIM.

The net effect of this change is that total income and expenditure are both increased by an amount equal to 10% of interest receipts. This represents an average increase in expenditure/income of less than 0.01%, so the effect is minimal.

Various income items are also mapped to a number of income groups or sources (see *mapinc.do*). These are income from labour (variable *inclab*), income from gross operating surplus (variable *incgos*), income from transfers from other households (variable *inctrans*), income from corporations (variable *inccorp*), transfers from government (variable *incgov*) and other income (variable *incother*), which is again netted out. Variable *inchphc* represents income from the sale of home produce or livestock. If this information is not required separately it is simply added to *incother*.

3.3. Forming a person-level IES 2000 dataset (*ies2000p.do*)

Very little remains to be done to form a person-level IES 2000 database. Do-file *ies2000p.do* starts with the newly formed *ies2000h.dta* and keeps whichever variables are relevant to the user. This shortened version of the file is saved as *ies2000hshort.dta*. It then opens *person.dta* and merges it with *ies2000short.dta*. The merge results are stored in variable *merge2*. As was the case previously when *general.dta* and *personh.dta* were merged, there are 46 observations in *ies2000h.dta* that do not have a matching *hhid* in *person.dta*. The data file is saved as *ies2000p.dta*.

3.4. Cleaning up education and factor data in the LFS 2000:2 (*lfs2000_2.do*)

The LFS 2000:2 data file, which was created in do-file *readin.do*, contains education, factor and activity data that needs to be cleaned up before it becomes usable. Sub-do-file *education.do* creates a variable called *education* that groups persons in the LFS 2000:2 into six education categories, namely (1) none or pre-primary, (2) primary, (3) lower secondary

(standard 8), (4) upper secondary (standard 10), (5) tertiary and (6) other, don't know or missing. These education categories are used later in the formation of representative household groups for the SAM (see PROVIDE, 2005).

Sub-do-file *factact.do* creates a wage or salary-income variable (*w_inclabp*), an occupation code variable (*w_fact*) and an activity variable (*w_activity*). Respondents had the choice of either specifying their exact weekly, monthly or annual wage or salary, or selecting an income category within a specific income band (discrete variable). Consequently it was necessary to adjust the data so that variable *w_inclabp* is a continuous variable showing the respondent's annual income from labour. Variable *w_fact* has the same occupation categories as listed in Table 7, while variable *w_activity* is similar to variable activities created in the IES 2000.

4. Further data analysis and adjustments

The second part of *ies2000.do* starts by running *adjustments.do*. This do-file can also be run independently from *ies2000.do*. Do-file *adjustments.do* adjusts the data in *ies2000h.dta* and *ies2000p.dta* (or *ieslfsmerge.dta*, a merged IES 2000 and LFS 2000:2 dataset) so that they are ready to use for the creation of various sub-matrices in a series of SAMs for South Africa. These SAMs are compiled from a variety of datasets, using sources of national data, such as the SARB data, to provide control totals. This is necessary so that the SAM represents a realistic picture, not only of patterns of income and expenditure of agents, but also of the overall income and expenditure levels.

The main objective of *adjustments.do* is to balance incomes and expenditure of individual households. By definition income and expenditure of each household group in the SAM should balance. Although the final balancing of a SAM is done during the SAM estimation process (PROVIDE, ____-a), balancing incomes and expenditures at an individual household-level at this stage in the process if also useful as it ensures that households either income or expenditure can be used to determine the household's level of welfare.[39] Before proceeding with the discussion *adjustments.do*, section 4.1 looks at the income and expenditure differences at the household level.

4.1. Income and expenditure differences

In theory total income and expenditure reported in the IES 2000 should be the same due to the way the questionnaire is set up and because of the economic identity *income* = *consumption* + *savings*. However, this is not true for the IES 2000. This can be due

[39] For example, if household groups were formed on the basis of income or expenditure, large differences between income and expenditure would cause households to be allocated to very different groups.

to various reasons, ranging from poor data capturing, inconsistent or erroneous reporting and deliberate misrepresentation by respondents. In the welfare literature total expenditure is often used as a more accurate measure of welfare. However, in the case of the IES 2000 there is no reason to believe that expenditure was captured or reported more accurately than income, since, as we show below, there appears to be no consistency in the way in which income is over- or underreported in the data.[40]

On average total income (*totinc*) and total expenditure (*totexp*) do not differ that much. The (unweighted) average *totinc* is R34,470, compared to the average *totexp* of R32,759. If households are grouped into those over-reporting income, those where income equals expenditure and those underreporting income, it can be seen than most households (16,590 out of 26,215) over-report income, with income exceeding expenditure, on average, by 36.2%. This is an interesting result given evidence that households usually tend to underreport income in these types of household surveys. Only 17 households report the exact same income and expenditure, while 9,608 households underreport income. For these households expenditure exceeds income by an average of 33.6%.

```
. sum totinc totexp

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
      totinc |   26215    34470.39   92908.21          0    5602178
      totexp |   26215    32759.18   84078.72         12    7568643

. sum totinc totexp if totinc > totexp

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
      totinc |   16590    34906.31   106543.5        300    5602178
      totexp |   16590    25634.41   59721.15         42    3751763

. sum totinc totexp if totinc == totexp

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
      totinc |      17    5098.941   4518.258       1440      18864
      totexp |      17    5098.941   4518.258       1440      18864

. sum totinc totexp if totinc < totexp

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
      totinc |    9608    33769.68   62846.27          0    1713000
      totexp |    9608    45110.37   113529.9         12    7568643
```

Below is the detailed summary statistics of variable *diff*, defined as *totinc* minus *totexp*. Variable *diff* ranges from –R5.86 million to R5.50 million, with a mean value of R1,711 (income is over-reported on average). Graphically the distribution of *diff* looks fairly symmetrical (see Figure 10), but bear in mind that the *x*-axis in the figure is truncated. In reality the distribution is skewed to the right. These absolute differences are, however,

[40] The data reported here comes from *ies2000h_orig.dta* after dropping mismatched observations from the various merges [`drop if merge1a == 2 | merge1b == 2 | merge1c == 1 | merge0b == 1`].

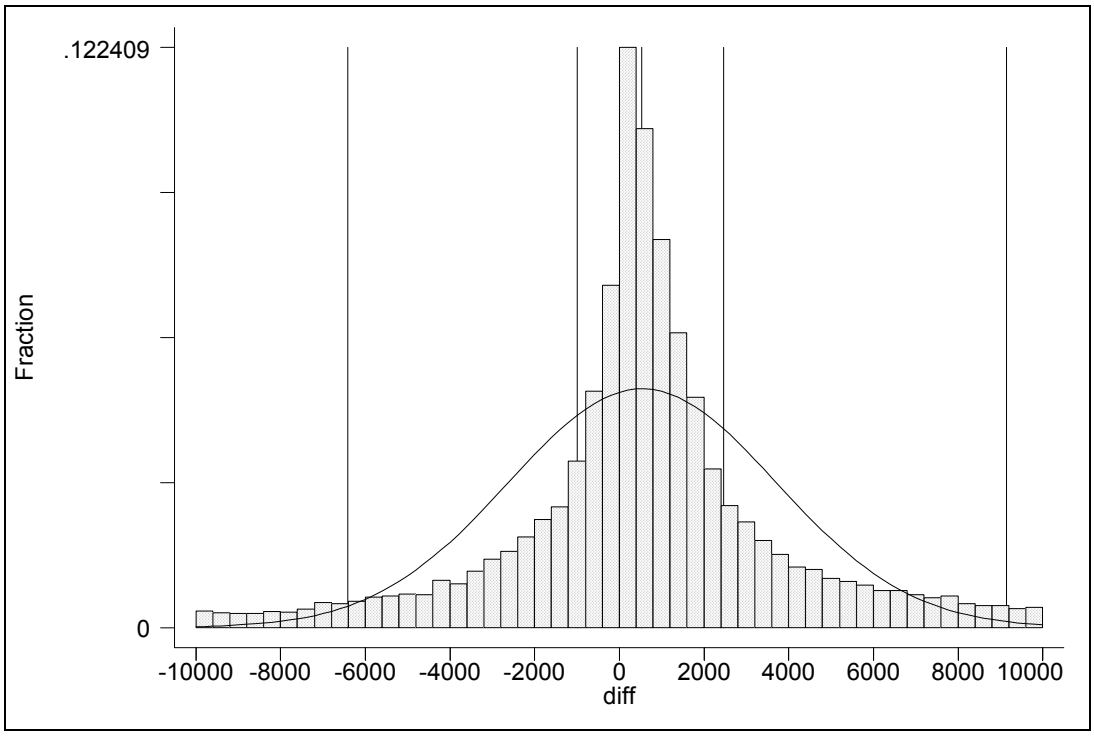
difficult to interpret. In order to evaluate the relative differences – i.e. the percentage by which income is over- or underreported with respect to total expenditure – variable *diffp* is created. This variable ranges from –100% (where *totinc* is zero and *totexp* positive) and a rather substantial 8,509.8%. This variable is also highly skewed to the right (see Figure 11).

```
. sum diff, detail

                              diff
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -68783.84       -5861002
 5%    -14883.78       -2383134
10%        -6417       -1777605       Obs               26215
25%    -987.1641        -909550       Sum of Wgt.       26215

50%          529                      Mean           1711.209
                        Largest       Std. Dev.      75058.32
75%         2468        2902067
90%     9140.105        3198446       Variance       5.63e+09
95%        19135        3695809       Skewness       7.326016
99%     78539.17        5495151       Kurtosis       3066.211
```

Figure 10: Distribution of the difference between income and expenditure (variable *diff*)



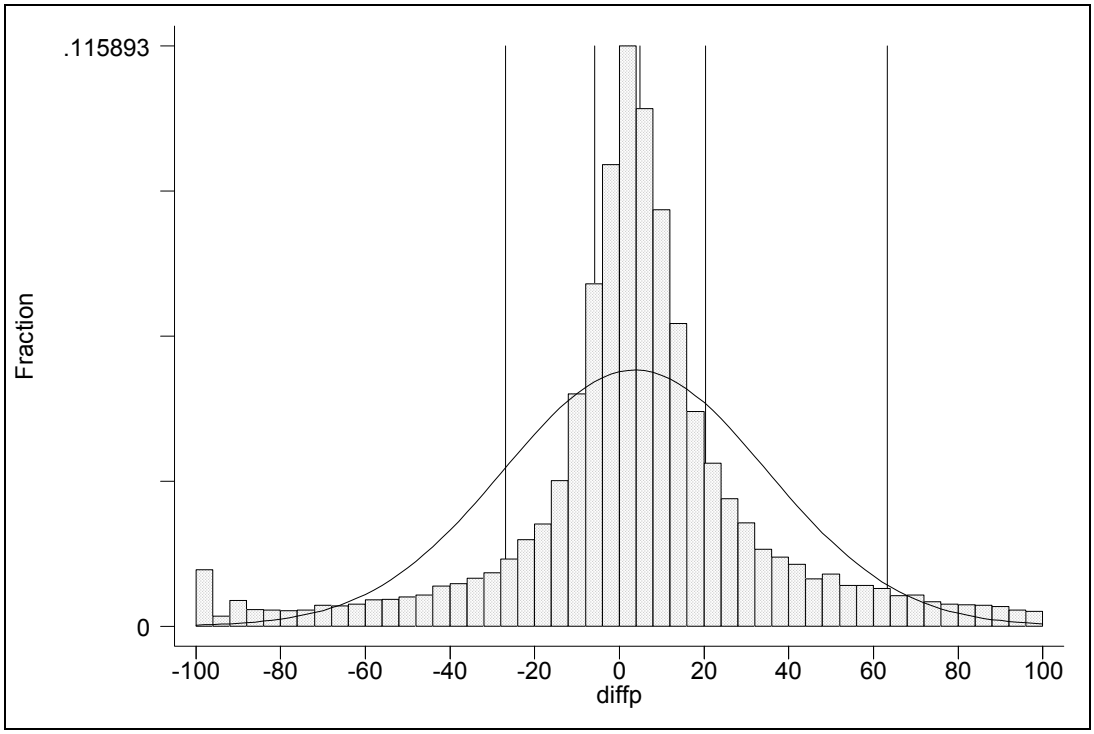
Note: Only values between –10000 and 10000 included in the graph. The vertical lines represent (from left to right) the 10th, 25th, 50th, 75th and 90th percentiles of variable *diff*.

```
                              diffp
-------------------------------------------------------------
      Percentiles      Smallest
 1%     -97.3604          -100
 5%    -52.92025          -100
10%    -26.82177          -100       Obs               26215
25%    -5.802562          -100       Sum of Wgt.       26215

50%     4.914149                     Mean           19.41118
                        Largest       Std. Dev.      112.9211
75%     20.33355       3045.161
```

```
90%      63.31995       3090.994       Variance       12751.18
95%      116.2348       5134.359       Skewness       26.65743
99%      336.1646       8509.819       Kurtosis       1499.697
```

Figure 11: Distribution of the relative income and expenditure difference (variable *diffp*)



Note: Only values between –100% and 100 included in the graph. The vertical lines represent (from left to right) the 10th, 25th, 50th, 75th and 90th percentiles of variable *diffp*.

The fact that income or expenditure is underreported is not necessarily the problem, as this is natural for most surveys of this kind. More problematic is the large average differences between the two. A simple experiment performed here ranks households first by expenditure (deciles) and then by income (deciles). Table 10 tabulates household income deciles against household expenditure deciles. If income and expenditure were exactly the same, or even within reasonable distances from each other, one would expect all households to lie on the diagonal of the matrix. The shaded band above and below the diagonal shows those households that move one group up or down. On average between 49.2% and 83.9% of households remain in the same deciles. If the bands above and below are included, the figures rise to between 81.6% and 94.8%.[41]

[41] These percentages are reported in the last two rows.

Table 10: Forming deciles using income and expenditure

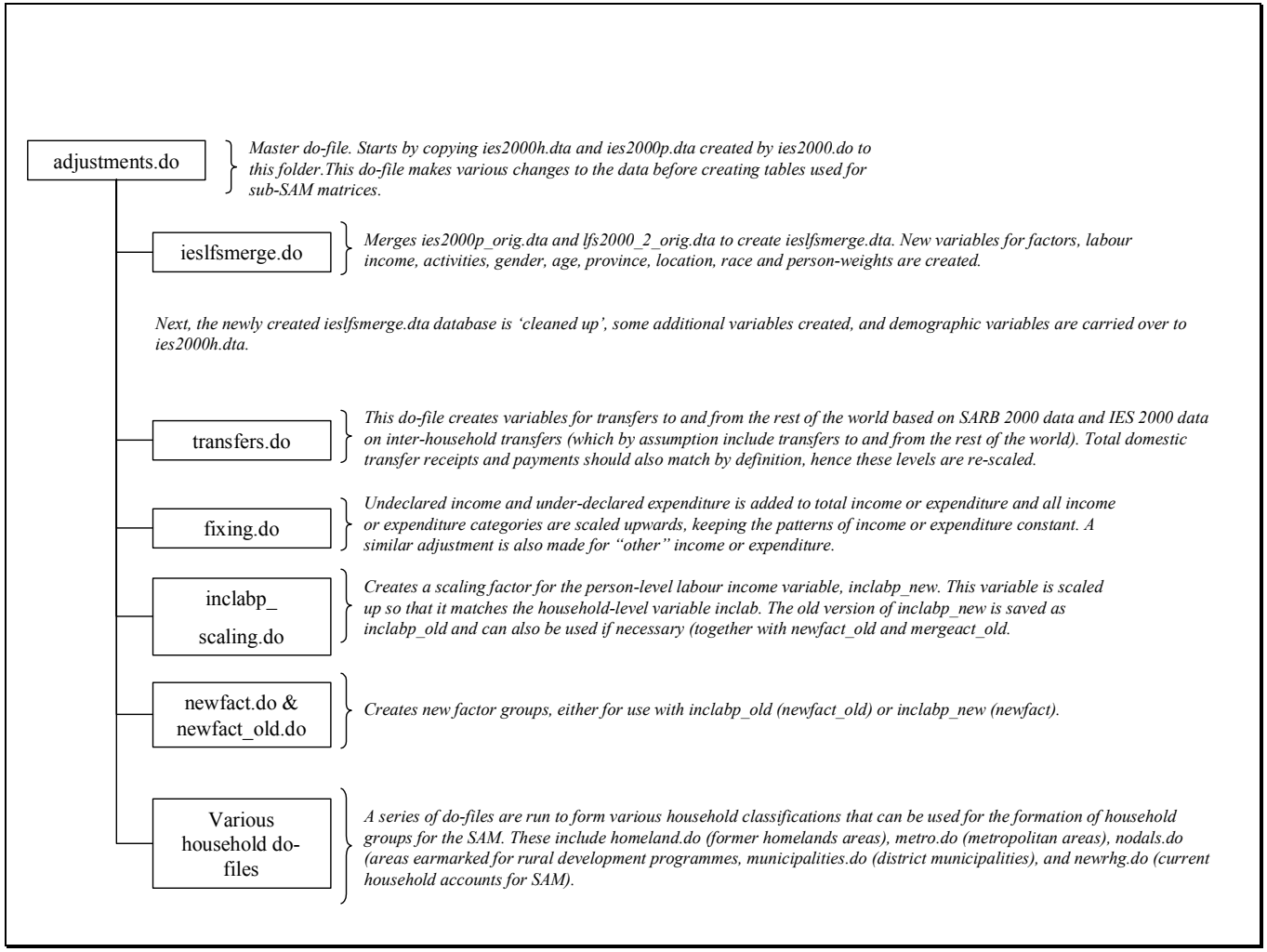
	Exp decile 1	Exp decile 2	Exp decile 3	Exp decile 4	Exp decile 5	Exp decile 6	Exp decile 7	Exp decile 8	Exp decile 9	Exp decile 10
Inc decile 1	1,791	312	132	84	84	54	59	46	41	45
Inc decile 2	525	1,460	499	137	66	42	22	18	13	5
Inc decile 3	165	510	1,291	376	95	51	24	17	19	3
Inc decile 4	59	147	422	1,330	373	92	39	18	14	10
Inc decile 5	48	101	153	436	1,387	402	77	44	14	1
Inc decile 6	14	55	69	137	380	1,415	412	60	19	15
Inc decile 7	14	18	28	66	140	387	1,490	394	61	24
Inc decile 8	5	7	17	37	57	125	385	1,613	344	33
Inc decile 9	1	11	8	12	27	44	98	359	1,800	285
Inc decile 10		1	3	5	13	9	16	52	297	2,200
Total obs.	**2,622**	**2,622**	**2,622**	**2,620**	**2,622**	**2,621**	**2,622**	**2,621**	**2,622**	**2,621**
Diagonal	*68.3%*	*55.7%*	*49.2%*	*50.8%*	*52.9%*	*54.0%*	*56.8%*	*61.5%*	*68.6%*	*83.9%*
Shaded band	*88.3%*	*87.0%*	*84.4%*	*81.8%*	*81.6%*	*84.1%*	*87.2%*	*90.3%*	*93.1%*	*94.8%*

Note: Analytic weights assumed (variable *weight*)

4.2. Adjusting the data (*adjustments.do*)

Do-file *adjustments.do* creates final person- and household level files from which the various household- and factor-related SAM sub-matrices are extracted. The person-level file is created by merging the IES 2000 and LFS 2000:2 files. This allows the user the option to either use the IES 2000 employment data or the LFS 2000:2 employment data for the factor-related sub-matrices. Some adjustments, which are discussed below, are also made to the household-level IES 2000 file. Figure 12 shows the structure of *adjustments.do* and its sub-do-files.

Figure 12: Do-file structure of *adjustments.do*



4.2.1. Merging the IES 2000 and LFS 2000:2 files (ieslfsmerge.do)

Do-file *ieslfsmerge.do* uses the original versions of the person-level IES 2000 (*ies2000p_orig.dta*) and LFS 2000:2 (*lfs2000_2_orig.dta*) created in part one of *ies2000.do* to form a merged file called *ieslfsmerge.dta*. Given some of the discrepancies between the IES 2000 and LFS 2000:2 in terms of demographic variables, the LFS 2000:2 variables are used. However, in some cases the demographic variables are missing in the LFS 2000:2 but not in the IES 2000. In such cases the IES 2000 variables are used to replace missing LFS variables. This do-file creates variables for factors (*mergefact*), labour income (*mergeinclabp*), activities (*mergeact*), gender (*mergegender*), age (*mergeage*), province (*mergeprov*), location (*mergeloc*), race (*mergerace*) and person-weights are created (*mergepwgt*).

4.2.2. Adjusting transfer variables (transfers.do)

Do-file *transfers.do* has two objectives. Firstly, it creates variables for transfers to and from the rest of the world (*rowtransinc* and *rowtransexp*), and secondly, it addresses the disparity between total transfer receipts and total transfer payments in IES 2000. The initial mean values of transfer expenditure (*hhtrans*) and transfer receipts (*inctrans*) are listed below.

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
hhtrans	26265	11041643.4	851.5219	5480.98	0	360000
inctrans	26265	11041643.4	1724.127	5838.838	0	396185

The IES 2000 does not distinguish between domestic and foreign transfer receipts or payments. We make the assumption that the transfer receipts and payments reported are inclusive of both domestic and foreign transfers. In order to separate out the foreign transfers from total transfers, we make use of SARB 2000 data, which reports on total foreign transfer receipts and payments to and from South Africa. These transfers make up 0.04% and 0.02% of current household income respectively. This information is used to estimate the share of foreign transfers in total transfer receipts and payments. The distribution of foreign transfer payments or receipts is weighted according to each household's share of total (national) transfer receipts or payments, i.e. foreign transfers follow the same distribution pattern as domestic transfers.[42] The total sum of transfer receipts is created so that it equals 0.04% of household income, while total transfer payments is created so that it equals 0.02% of household income (national level).

[42] This assumption is made due to the lack of information regarding the distribution of foreign transfers across households. An alternative approach would be to assume that foreign transfer activities are related to the income of the household, e.g. that a higher income household is more likely to receive income or make payments to household abroad. However, since there is no real basis for such an approach the former approach was opted for.

The remainder of total transfer incomes and receipts is assumed to be domestic transfers. The sum of variables *hhtransinc* and *hhtransexp* should in theory be the same, since all domestic transfer receipts should exactly offset domestic transfer payments.[43] In reality reported transfer payments are lower than transfer receipts, and hence the expenditure side has to be adjusted. The following equation is used:

```
. replace hhtransexp = hhtransexp + ((hhtransexp/sumhhtransexp)*sumhhtransnet)
```

After this adjustment the sum of net transfers (variable *sumhhtransnet*) is recalculated and equals zero. Note that some of these adjustments are made only once *fixing.do* has been run (see section 4.2.3). The final mean values of transfers to and from the rest of the world and inter-household transfers are listed below.

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
rowtransexp	26183	10977096.4	8.677201	52.98943	0	3234.902
hhtransexp	26183	10977096.4	1948.476	11884.91	0	725277.5
rowtransinc	26183	10977096.4	48.46138	701.238	0	94926.11
hhtransinc	26183	10977096.4	1948.476	6733.124	0	379948.7

The inter-household transfers sub-matrix reports on all net transfer flows between all RHGs in the matrix. The IES 2000, however, does not supply any information on 'to whom' and 'from whom' transfers were paid or received. The only information we have is the total value of transfer receipts and payments of each household (or household group). The mapping of transfers between specific households or household groups cannot be done in Stata. Certain assumptions have to be made about how these payments or receipts are distributed. In the appendix (section 7.1) we explain the assumptions and show how the cells of the inter-household transfers matrix may be populated using a simple MS Excel model.

4.2.3. Income and expenditure differences (fixing.do)

After running *transfers.do*, do-file *fixing.do* is run to close the gap between income and expenditure. This gap is removed by assuming that the larger of income or expenditure reported by each household is the correct welfare estimate. The under-declared figure is increased while each of the components that make up the under-declared figure is scaled upwards.

Do-file *fixing.do* starts with the income side and calculates each observation's undeclared income (variable *incundecl*) if income is less than expenditure and adds this to *incother*. Total income is increased by *incundecl*. In order to ensure that the components of income add up to the adjusted income figure, each component is scaled

[43] Because it is a 'zero sum game' national accounts ignore inter-household transfers. However, this data is usually included in a SAM and hence it cannot be ignored here.

upwards pro-rata. At the same time incother is also netted out so that the components now only include *inclab* (which is inclusive of *inchphc*), *incgos*, *inccorp*, *incgov*, *rowtransinc* and *hhtransinc*.

```
for var inclab incgos inccorp incgov rowtransinc hhtransinc:
      replace X = X + (incother*X/totincadj) if totincadj > 0
```

Some households report zero income but positive expenditure. In these cases *incother* is positive but and *totincadj* = *totinc* – *incother* is zero, and no information exists about the pattern of income of the household. For this reason the above Stata command contains an *if*-statement that prevents division by zero. For these households the average income pattern across all households is used to estimate the 'missing' income components.

A similar approach was followed for the expenditure side. Under-declared expenditure (*expundecl*) was added to *hhother*, and the net increase added to *totexp*. The components of *totexp*, namely *C1* (including *hhhphc)* to *C96*, *hhtransexp*, *rowtransexp*, *hhinctax*, *hhlocaltax* and *hhsav* are scaled upward using the Stata command below. For those households reporting zero *totexpadj* the expenditure components were estimated as before using the average expenditure pattern across all households.

```
for var C1 - C96 hhtransexp rowtransexp hhinctax hhlocaltax hhsav :
      replace X = X + (X/totexpadj)*(hhother) if totexpadj > 0 ;
```

This do-file also makes a final adjustment to domestic transfers (variables *hhtransinc* and *hhtransexp*) to ensure their weighted averages are equal in the final database. Finally, some checks are performed to make sure that all adjustments were done correctly, i.e. that the income and expenditure components add up to total income and expenditure respectively.

4.2.4. Scaling up the person-level factor income variables (inclabpscaling.do)

Due to the adjustments made in *fixing.do* the person-level *inclabp_new* variables of a household do not necessarily add up to the new household-level *inclab* variable any longer. Changes to *inclab* originate from three sources. Firstly, income from the sale of home produced agricultural products is now included under *inclab* since the current SAM does not separate out this type of income. Secondly, *inclab* was scaled up in cases where total income was less than total expenditure so that total income and expenditure matches. Thirdly, in some cases where total income was zero and total expenditure greater than zero, it was assumed that the true households income is equal to total expenditure. The components of total income were 'built up' on the assumption that the household's structure of income was the same as that of the 'average' household. Under

this assumption various households now earn income from labour although the person-level labour income variables for that household are all zero.

In do-file *inclabscaling.do* a scaling factor is created that either leaves the person-level *inclabp_new* variable in tact or scales it up or down so that the sum of the *inclabp_new* variables in each household equal the household-level *inclab* variable. For 4,946 households *inclab* remained zero before and after the adjustments. Variable *inclab* remained positive and unchanged for a further 7,428 households. For all these households the *inclabp_new* variables remained unchanged. For 11,399 households the *inclabp_new* variables were scaled upwards by an average factor of 1.43 due to changes made to *inclab*. A further 2,410 households initially reported zero income from labour but now had positive income figures. For these households the head of the household was assumed to have earned that income. The original *inclabp_new* variable was saved as *inclabp_old* and *inclabp_new* was scaled up to its new levels.

Figure 1 compares the new and old versions of person-level labour income. As expected the average income is now slightly higher for all of the occupation groups, except for farmers and unspecified workers. Many households that previously reported zero income from labour were now added to these two groups due to the adjustments made to *inclab* in *fixing.do*. Specifically the addition of income from home production to *inclab* explains the increase in the number of agricultural workers. The new workers added to this group obviously had a lower average wage than the rest of the agricultural workers, which explains why the average wage drops. Most of the other 'new' additions were allocated to the unspecified category, because these workers did not previously report income and never specified an occupation category. The average wage of unspecified workers drops for the same reason as the drop in agricultural wages.

4.2.5. Forming factor groups (newfact.do and newfact_old.do)

Do-file *newfact.do* creates a province-level occupation code variable called *newfact*. This variable is similar to *mergefact* but disaggregates workers further by race and province. It contains 88 different types of labour. The original occupation groups mapped from the LFS 2000:2 are (1) legislators, senior officials and managers; (2) professionals; (3) technical and associate professionals; (4) clerks; (5) service workers and shop and market sales workers; (6) skilled agricultural and fishery workers; (7) craft and related trades workers; (8) plant and machine operators and assemblers; (9) elementary occupations; (10) domestic workers; and (11) not adequately or elsewhere defined, unspecified. In some provinces certain of these province-race-labour sub-categories are not well represented, in which case aggregate groups are formed by merging an occupation group with another of similar skills level. Thus, high skilled are

made up of factor groups 1 and 2, skilled include groups 3 to 5, and semi- and unskilled include groups 6 to 11. A detailed description of these factor groups appears in PROVIDE (2005). Variable *newfact* is used in the formation of the factor-related sub-matrices.

Do-file *newfact_old.do* creates variable *newfact_old*, which is exactly the same as *newfact* except that it is meant to be used together with *inclabp_old* rather than *inclabp_new* (see Figure 1, which shows the difference between employment figures for *inclabp_new* and *inclabp_old* – variable *mergefact* and *mergefact_old*). The user therefore has a choice whether to use *inclabp_old*, *newfact_old* and *mergeact_old* to form factor-related sub-matrices, or whether to use *inclabp_new*, *newfact* and *mergeact*.

4.2.6. Forming variables for various possible household classifications

Next, a series of do-files are run to form various household classifications that can be used for the formation of household groups for the SAM. Do-file *homeland.do* indicates whether households live in former homeland areas. Magisterial district information was used to do the mapping. Do-file *metro.do* is similar to variable location, but disaggregates urban areas metropolitan areas and other urban areas. The metropolitan areas are those areas that were declared metropolitan municipalities by the Demarcation Board after 1994. Do-file *nodals.do* uses magisterial district information to indicate which households live in so-called nodal areas for the implementation of rural development programmes. These were areas identified by President Mbeki during his 2002 State of the Nation Address. Do-file *municipalities.do* indicates for every household the local district municipality in which they live.

Finally, do-file *newrhg.do* creates the current household groups used in the household-related SAM sub-matrices. Households are disaggregated by province, race, gender of the head of the household, location (former homeland areas), agricultural and non-agricultural households and education of the head of the household. A detailed description of these household groups, as well as the do-files listed above, is provided in PROVIDE (2005).

4.3. Printing SAM sub-matrices (*print.do*)

The final do-file, *print.do*, prints various tables to log-file called *print.log*. At present the do-file is set-up to use the LFS 2000:2 factor data where applicable. The first table, *SAMDATA1*, gives the total expenditure by household groups on all commodities (*C1 – C96*), transfers, taxes, savings and total expenditure. When transposed, part of this table becomes the commodities-households sub-matrix, while the other expense items are

used in various other SAM sub-matrices. Also included in *SAMDATA1* are all the income-side variables (income from labour, GOS, transfers, etc.) for each household group. All incomes and expenditures are weighted using the household weights from the IES 2000.

Table *SAMDATA2* is the household-factor sub-matrix, which cross-tabulates variables household groups and factor groups, with total income from labour in the cells. Here it is necessary to choose between the 'old' and 'new' factor income and occupation groups. This table represents the flow of resources from factors to households, which forms part of the functional income distribution. Table *SAMDATA3* cross-tabulates factors and (activities), and represents the value-added sub-matrix of the SAM. In addition to choosing the factor income and occupation groups, the 'old' or 'new' activity variable should also be specified.

5. Concluding remarks

This paper discussed in detail the IES 2000 and LFS 2000:2 datasets and the process followed to correct errors, make adjustments to the data and merge the two datasets. The final version of the IES 2000 created in Stata is perfectly balanced in the sense that total expenditures equal total receipts for every household. Although not too many problems were encountered when merging the IES 2000 and LFS 2000:2, one problem that remains is the incompatibility between the LFS and IES labour income data. The LFS data suggests much higher average wages than reported in the IES. This means that the levels and distribution of income from labour in the IES is different from that of the LFS. Some more work may have to be done to establish how large the effect is from switching from the IES to the LFS as the main source of factor-related data.

6. References

Deaton, A. (1997). *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy*. John Hopkins University Press: London.

Hoogeveen, J.G. and Özler, B. (2004): "Not Separate, Not Equal. Poverty and Inequality in Post-Apartheid South Africa." World Bank.

McDonald, S. and Punt, C. (2001): "A Social Accounting Matrix for the Western Cape." Department of Agriculture, Western Cape Government.

Poswell, L. (2003): "Comments on Income and Expenditure Survey 2000." Development Policy Research Unit.

PROVIDE (2003a). "Creating a 1995 IES Database in STATA," *PROVIDE Technical Paper Series,* 2003:1.

PROVIDE (2003b). "Creating a 1995 OHS and a Combined OHS-IES Database in STATA," *PROVIDE Technical Paper Series,* 2003:2.

PROVIDE (2005). "Forming Representative Household Groups in a SAM," *PROVIDE Technical Paper Series,* 2005:2.

PROVIDE (____-a). "Compiling a Social Accounting Matrix for South Africa: 2000," *PROVIDE Technical Paper Series,* Forthcoming.

PROVIDE (____-b). "The Economic Contribution of Home Production for Home Consumption," *PROVIDE Background Paper Series,* Forthcoming.
SARB (2002): *Quarterly Bulletin, September 2002,* Pretoria: South African Reserve Bank.
Simkins, C. (2003). "A Critical Assessment of the 1995 and 2000 Income and Expenditure Surveys as Sources of Information on Incomes," *Mimeo*.
SSA (1998): *The People of South Africa: Population Census 1996,* Pretoria: Statistics South Africa.
SSA (2002a): *Income and Expenditure Survey 2000,* Pretoria: Statistics South Africa.
SSA (2002b): *Labour Force Survey September 2000,* Pretoria: Statistics South Africa.
SSA (2003a): *Census 2001,* Pretoria: Statistics South Africa.
SSA (2003b): *Final supply and use tables, 2000: an input-output framework,* Pretoria: Statistics South Africa.
StataCorp (2001). *Statistical Software: Release 7.0*. Stata Corporation: College Station, TX.
Van der Berg, S., Nieftagodien, S. and Burger, R. (2003a). "Consumption patterns and living standards of the black population in perspective." *Paper presented at the Biennial Conference of the Economic Society of South Africa*, Somerset West.
Van der Berg, S., Nieftagodien, S. and Burger, R. (2003b). "Consumption Patterns of South Africa's Rising Black Middle-Class: Correcting for Measurement Errors." *Paper presented at the CSAE conference on Poverty Reduction, Growth and Human Development in Africa, Oxford, March 2004*,
Woolard, I. and Leibbrandt, M. (2001). "Measuring Poverty in South Africa." In *Fighting Poverty. Labour Markets and Inequality in South Africa*, edited by Bhorat, H., Leibbrandt, M., Maziya, M., Van der Berg, S. and Woolard, I. Cape Town: UCT Press.

7. Appendix

7.1. Wage and salary income from labour – data adjustments

Section 2.3.2 compared the LFS 2000:2 and IES 2000 labour income data. This section describes how the 'combined' labour income variable and its related factors and activities variables were created. The combined variable was subsequently adjusted (scaled upwards) so that total income from labour in the person-level file (*ieslfsmerge.dta*) matches the total income from labour (*inclab*) in the household-level file (*ies2000h.dta*). Although the person- and household-level labour income data did match originally by construction, some adjustments were made to the household-level variable *inclab* in do-file *fixing.do*. This necessitated the changes, and hence two person-level variables for income from labour exist, namely *inclabp_old* and *inclabp_new*.[44]

Initially, when only LFS data was used for the factor-related sub-matrices, there were various upper and lower outliers that caused total wages within the SAM sub-matrices to be biased. These outliers in the LFS were also cause for the large differences in average wages reported in the LFS and the IES. Because of this it was necessary to investigate the issue further. As a first adjustment it was assumed that if an

[44] Either one of these variables may be used to form sub-matrices. It doesn't make much of a difference since the patterns of income distribution remain largely the same.

individual reported zero income in the one survey and non-zero income in the other, the non-zero entry was assumed to be the correct one.

Next, variable *diffp*, defined as the percentage difference (in absolute terms) between the LFS wage (*inclabp_lfs*) and the IES wage (*inclabp_ies*), was constructed. If this difference was less than 30% the larger of the two income levels was chosen as the correct factor income. Initially there were about 24,700 workers (unweighted) in the LFS, and 23,221 in the IES. When considering all people reporting income from labour in a combined IES-LFS survey, there were almost 27,000 workers. For the majority of these (approximately 23,000 observations) the difference between IES and LFS wage income was less than 30%.

The remainder were evaluated record by record, following the basic rule of thumb that the larger of the two incomes is correct. However, in instances where it was reasonable to believe that the larger entry was incorrect, the smaller entry was select. For example, in many instances the one figure was exactly 10, 100, or 1000 times the other, which clearly suggests an error in the data capturing process. In such instances the more realistic figure was selected, given the average income of the factor group, the education level of the respondent, the total household income or expenditure (unadjusted), as well as the reported income levels of other members of the household. In all instances where it was still unclear whether to use the IES or LFS data, the one that would be more successful at closing the gap between total household income and expenditure was selected. Of the 4000 observations examined there were 95 cases where the LFS income was larger, but the IES was used, and 59 cases where the IES was larger, but the LFS was used. For the rest the larger of the two seemed more appropriate.

The new income variable created was initially saved as *inclabp_new*, but later renamed *inclabp_old* when *inclabp_new* was scaled upwards to match the household-level variable *inclab*. This is discussed in section 4.2.4.

7.2. Household expenditure accounts

Table 11: Commodity accounts and other expenditure categories

A/c name	Description	A/c name	Description	A/c name	Description
C1	Agricultural products	C35	Primary plastic products	C69	Wire and cable products
C2	Coal and lignite products	C36	Pesticides	C70	Accumulators
C3	Gold and uranium ore products	C37	Paints	C71	Lighting equipment
C4	Other mining products	C38	Pharmaceutical products	C72	Other electrical products
C5	Meat products	C39	Soap products	C73	Radio and television products
C6	Fish products	C40	Other chemical products	C74	Optical instruments
C7	Fruit and vegetables products	C41	Rubber tyres	C75	Motor vehicles
C8	Oils and fats products	C42	Other rubber products	C76	Motor vehicles parts
C9	Dairy products	C43	Plastic products	C77	Other transport products
C10	Grain mill products	C44	Glass products	C78	Furniture
C11	Animal feeds	C45	Ceramicware	C79	Jewellery
C12	Bakery products	C46	Ceramic products	C80	Other manufacturing
C13	Sugar products	C47	Cement	C81	Electricity
C14	Confectionaryproducts	C48	Other non-metallic products	C82	Water
C15	Other food products	C49	Iron and steel products	C83	Buildings
C16	Beverages and tobacco products	C50	Non-ferrous metals	C84	Other constructions
C17	Textile products	C51	Structural metal products	C85	Trade services
C18	Made-up textile products	C52	Treated metal products	C86	Accommodation
C19	Carpets	C53	General hardware products	C87	Transport services
C20	Other textile products	C54	Other fabricated metal products	C88	Communications
C21	Knitting mill products	C55	Engines	C89	FSIM
C22	Wearing apparel	C56	Pumps	C90	Insurance services
C23	Leather products	C57	Gears	C91	Real estate services
C24	Handbags	C59	General machinery	C92	Other business services
C25	Footwear	C58	Lifting equipment	C93	General Government services
C26	Wood products	C60	Agricultural machinery	C94	Health and social work
C27	Paper products	C61	Machine-tools	C95	Other services / activities
C28	Containers of paper	C62	Mining machinery	C96	Household domestic services
C29	Other paper products	C63	Food machinery	**Other expenditure categories**	
C30	Published and printed products	C64	Other special machinery	hhtrans	Inter-household transfers - payments
C31	Recorded media products	C65	Household appliances	hhinctax	Income tax paid
C32	Petroleum products	C66	Office machinery	hhlocaltax	Local taxes, levies and charges
C33	Basic chemical products	C67	Electric motors	hhsav	Net savings
C34	Fertilizers	C68	Electricity apparatus	hhother	Other expenditure

7.3. Creating an inter-household transfers matrix

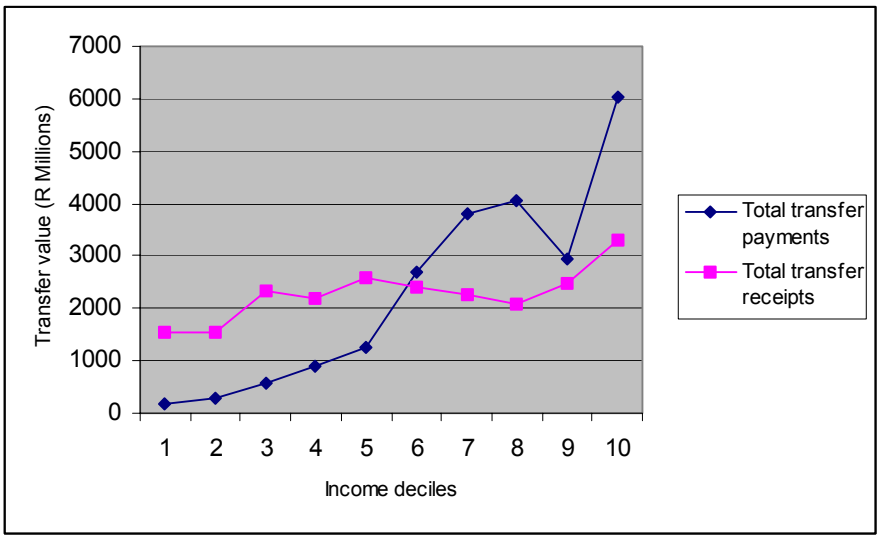
Data on inter-household transfers in the IES 2000 is problematic for two reasons:

- The national-level transfer payments by households are not equal to the national-level transfer receipts reported by households. In theory these two figures should be the same.

- There is no information that can be used to map incomes and receipts. The only information that can be gathered from the IES 2000 is the total amount of transfers received and the total amount of transfers paid during 2000. There is no information about where transfer receipts come from or to whom payments are made. In a SAM one aims to map these relationship between household groups

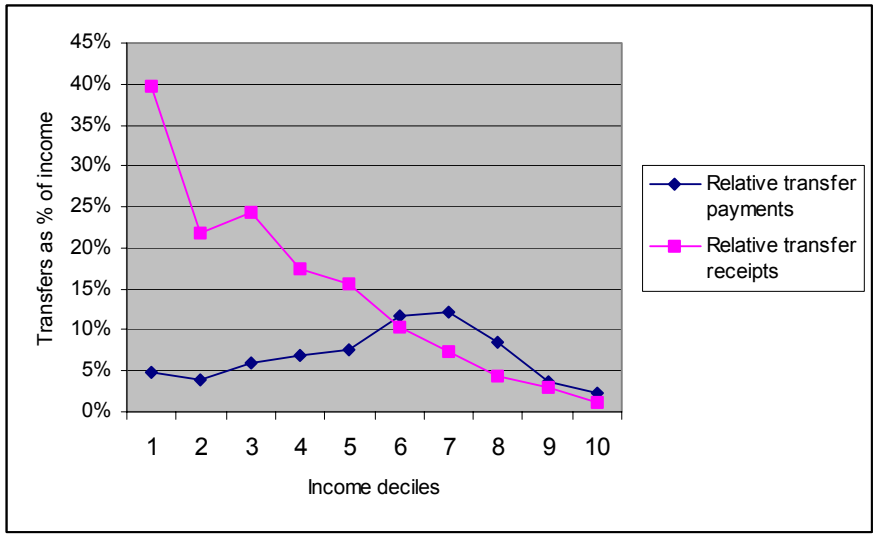
Section 4.2.1 elaborated on the process of correcting these problems. In this appendix we explain how the actual inter-household transfers sub-matrix is actually formed in MS Excel. As an example households are grouped into income deciles.[45] The only household group information that we can gather from the IES 2000 is the total transfer receipts and payments made by each household group. These values are shown graphically in Figure 13. As expected transfer payments increase as one moves to a higher income group, although there is a large dip in the 9th decile. There is no clear pattern as far as transfer receipts go. However, expressing transfer payments and receipts as a percentage of income shows a more interesting picture. The value of transfer receipts relative to the total household group income drops significantly as one moves to the higher incomes deciles (see Figure 14).

Figure 13: Total value of transfer payments and receipts by income decile



[45] As before 'income' is defined as the maximum of total income and total expenditure. Hence income and expenditure deciles are the same.

Figure 14: Transfer payments and receipts as a percentage of total income



The information collected from the IES 2000 is summarised in Table 12. These totals can be seen as the row and column totals of the inter-household transfers sub-matrix (matrix *T*) as shown in Table 13. Cell t_{ij} (in the ith row and j^{th} column) of matrix *T* is calculated as

$$t_{ij} = \frac{\sum_j t_{ij} . \sum_i t_{ij}}{\sum_i \sum_j t_{ij}}$$

where *i* and *j* denote the rows and columns respectively. It is easy to verify that summing the above expression over *j* gives the vector of column (expenditure) totals, while summing over *i* gives the vector of row (income) totals. The sum of all the cells is of course the total value of transfer incomes or payments.

The next step is to calculate the net receipts of each household group. This can be done by subtracting from matrix *T* its transpose, thus giving a symmetrical matrix $T^s = T - T'$ for which $t_{ij} = -t_{ij}$. Although one would expect to see all the positive entries above the diagonal (which implies that a richer household group transfers more money to poorer households than it receives from them) there are some household groups that have positive entries below the diagonal. Fortunately this only occurs at the higher end of the income distribution.[46] All diagonal entries of the net transfers matrix are zero ($t_{ij} = 0$ for $i = j$). The negative entries of net transfers are deleted, thus giving the final inter-household transfers matrix in Table 14. Note that the net transfers column

[46] See $t_{9,7}$, $t_{9,8}$ and $t_{10,8}$.

(last column) is the same as the net transfers column in Table 13. This same procedure can be applied to any number of household groups in the SAM.

Table 12: Adjusted transfers data extracted from IES 2000

	Transfer payments (R millions)	Transfer income (R millions)
Decile 1	189.51	1544.39
Decile 2	278.75	1526.48
Decile 3	571.50	2317.88
Decile 4	885.60	2202.95
Decile 5	1246.08	2576.17
Decile 6	2688.41	2408.68
Decile 7	3797.39	2247.75
Decile 8	4050.24	2088.48
Decile 9	2941.75	2482.87
Decile 10	6037.05	3290.61
Totals	*22686.27*	*22686.27*

Table 13: Inter-household transfers sub-matrix

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	*Totals*	*Net transfers*
Decile 1	12.90	18.98	38.91	60.29	84.83	183.02	258.51	275.72	200.26	410.98	*1544.39*	*1354.89*
Decile 2	12.75	18.76	38.45	59.59	83.84	180.89	255.51	272.53	197.94	406.21	*1526.48*	*1247.73*
Decile 3	19.36	28.48	58.39	90.48	127.31	274.68	387.98	413.82	300.56	616.81	*2317.88*	*1746.38*
Decile 4	18.40	27.07	55.50	86.00	121.00	261.06	368.74	393.30	285.66	586.23	*2202.95*	*1317.35*
Decile 5	21.52	31.65	64.90	100.57	141.50	305.29	431.22	459.93	334.06	685.55	*2576.17*	*1330.10*
Decile 6	20.12	29.60	60.68	94.03	132.30	285.44	403.18	430.03	312.34	640.98	*2408.68*	*-279.72*
Decile 7	18.78	27.62	56.62	87.75	123.46	266.37	376.24	401.30	291.47	598.15	*2247.75*	*-1549.64*
Decile 8	17.45	25.66	52.61	81.53	114.71	247.49	349.59	372.86	270.82	555.77	*2088.48*	*-1961.76*
Decile 9	20.74	30.51	62.55	96.92	136.38	294.23	415.60	443.27	321.96	660.72	*2482.87*	*-458.88*
Decile 10	27.49	40.43	82.89	128.46	180.74	389.95	550.81	587.48	426.70	875.67	*3290.61*	*-2746.44*
Totals	*189.51*	*278.75*	*571.50*	*885.60*	*1246.08*	*2688.41*	*3797.39*	*4050.24*	*2941.75*	*6037.05*	*22686.27*	*0.00*

Table 14: Final inter-household transfers matrix

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	*Totals*	*Net transfers*
Decile 1	0.00	6.22	19.54	41.89	63.31	162.90	239.73	258.28	179.52	383.49	*1354.89*	*1354.89*
Decile 2	0.00	0.00	9.97	32.52	52.19	151.30	227.89	246.86	167.43	365.78	*1253.95*	*1247.73*
Decile 3	0.00	0.00	0.00	34.99	62.42	214.00	331.36	361.21	238.01	533.92	*1775.90*	*1746.38*
Decile 4	0.00	0.00	0.00	0.00	20.43	167.03	281.00	311.77	188.73	457.77	*1426.74*	*1317.35*
Decile 5	0.00	0.00	0.00	0.00	0.00	172.99	307.76	345.22	197.68	504.80	*1528.45*	*1330.10*
Decile 6	0.00	0.00	0.00	0.00	0.00	0.00	136.82	182.54	18.11	251.03	*588.48*	*-279.72*
Decile 7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	51.71	0.00	47.34	*99.06*	*-1549.64*
Decile 8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	*0.00*	*-1961.76*
Decile 9	0.00	0.00	0.00	0.00	0.00	0.00	124.13	172.46	0.00	234.02	*530.61*	*-458.88*
Decile 10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.71	0.00	0.00	*31.71*	*-2746.44*
Totals	*0.00*	*6.22*	*29.52*	*109.40*	*198.35*	*868.21*	*1648.69*	*1961.76*	*989.49*	*2778.15*	*8589.79*	*0.00*

Technical Papers in this Series

Number	Title	Date
TP2003:1	Creating a 1995 IES Database in Stata	September 2003
TP2003:2	Creating a 1995 OHS and a combined OHS-IES Database in Stata	September 2003
TP2003:3	A Standard Computable General Equilibrium Model Version 3: Technical Documentation	September 2003
TP2004:1	SeeResults: A spreadsheet Application for the Analysis of CGE Model Results	November 2004
TP2004:2	The Organisation of Trade Data for inclusion in Social Accounting Matrix	December 2004
TP2005:1	Creating a 2000 IES-LFS Database in Stata	February 2005
TP2005:2	Forming Representative Household and Factor Groups in a SAM	March 2005

Other PROVIDE Publications

Background Paper Series
Working Papers
Research Reports