SOUTHMOD

Technical note

Exploring the quality of income data in two South African household surveys which underpin SAMOD

Helen Barnes, David McLennan, Michael Noble, and Gemma Wright

March 2020

University of Essex

UNITED NATIONS UNIVERSITY
UNU-WIDER

SOUTHMOD technical note

Exploring the quality of income data in two South African household surveys which underpin SAMOD

Helen Barnes, David McLennan, Michael Noble, and Gemma Wright

The technical note is related to the WIDER Working Paper 173/2018 'Assessing the quality of the income data used in SAMOD, a South African tax-benefit microsimulation model'.

Introduction

This technical note provides additional information to complement two recent working papers that use SAMOD, a tax-benefit microsimulation model for South Africa. The first paper focuses on the quality of the income data in SAMOD's underpinning datasets for the purpose of simulating personal income tax (PIT) (Wright et al. 2018). The second working paper explores the quality of income data in household surveys from Tanzania and Zambia and applies one of the income imputation techniques to a South African dataset (McLennan et al. forthcoming).

SAMOD has been underpinned by several different household surveys over the past decade, including earlier versions of the National Income Dynamics Study (NIDS) and the Living Conditions Survey (LCS) 2008/09 (e.g. Wright et al. 2016). The current version of SAMOD (Version 6.6) is underpinned by NIDS Wave 4 Version 1.1 (SALDRU 2014) and the LCS 2014/15 (Statistics South Africa 2017).

This technical tote has three parts. In Part 1, information is provided about the preparation of the LCS 2014/15 as an underpinning dataset for SAMOD, including data cleaning and imputations. Part 2 provides additional information about the process of comparing simulated PIT generated using recent versions of NIDS and the LCS, and with administrative data on PIT (to be read alongside Wright et al. 2018). In Part 3, details are provided about the process of generating artificially missing income data in NIDS in order to test one of the multiple imputation methods that had been applied to missing and implausible income data in household surveys from Tanzania and Zambia (supplementing McLennan et al. forthcoming).

Part 1: Preparing the LCS 2014/15 as an underpinning dataset for SAMOD: data cleaning and imputations

This section summarises the imputations/adjustments that were undertaken on the LCS 2014/15 dataset. The assessment of market income data is reported in the main working paper (Wright et al. 2018).

Grant outliers: The LCS uses COICOP codes for income and expenditure. COICOP values for reported income from old-age grants and disability grants that were too low were set to the minimum value (ZAR100) in the variables *boa* and *bdi* respectively. There were a small number of cases where the value for disability grant was above the maximum grant amount (ZAR1,410 in April 2015). These were capped at ZAR1,480.5 (ZAR1,410 x 1.05). Similarly, there were a small number of cases where the value for old-age grant was above the maximum grant amount (ZAR1,410 in April 2015 for under 75s and ZAR1,430 for 75 and over). These were capped at ZAR1,480.5 (ZAR1,410 x 1.05) and ZAR1,501.5 (ZAR1,430 x 1.05) respectively.

Unemployment Insurance Fund (UIF) contributor: In terms of creating a flag for UIF contributors (*lfc*), the COICOP codes for UIF are only reported at household level so it was not possible to know who made the contributions. As a proxy, all people in employment in the household were flagged as UIF contributors, and a flag was assigned to the head of household (HH) in cases where there was not anyone in employment but a COICOP amount was recorded.

Expenditure: Various expenditure variables were assigned to the HH where only a household-level amount was provided: income tax (*tin*), non-taxable income (*ynt*), and other income (*yot*). Some expenditures were assigned proportionally to all earners (*xishl* – own expenditure on medical insurance; *xhl* – expenditure on health care; *xpp* – pension contributions) or to all employees (*xishler* – employer contribution to medical insurance). Pension contributions (*xpp*) were also capped at 50 per cent of own earnings. Lump sums were assigned to the oldest head under 65 and, if none, then the oldest other person under 65 on the assumption that older heads of household either would not be in employment or would have retired a long time ago.

Expenditure relating to value-added tax (VAT): All expenditure relating to VAT was assigned to the HH. Additionally, VAT was removed from standard-rated items to enable the simulation of VAT on the model.

Spouse ID (idpartner): In situations when the spouse ID was not recorded for an individual but the relationship to HH variable for that individual was recorded as husband/wife/partner of head, the ID number of the HH was assigned to the spouse ID variable for that individual. In addition, the husband/wife/partner's ID was assigned to the spouse ID variable (if not recorded) for heads of household.

Mother, father and parent ID (idmother, idfather and idparent): An idmother variable was created using first, the biological mother ID variable (Q115MOTHNO). However, 9,972 children did not have an ID number for their biological mother in Q115MOTHNO. Some of these were the son/daughter/stepchild/adopted child of the head according to Q16RELATION. These children were therefore given the ID number of the HH, where the head was female (210 additional children). In total there were 9,762 under 18s without an idmother, of which 7,313 had a mother who is alive but not living in the household.

A similar process was followed for idfather, using the biological father ID variable (Q112FATHNO) in the first instance and then giving the ID number of the HH, where the head was male (458 additional children). In total there were 22,946 under 18s without a father ID, of which 17,658 had a father who is alive but not living in the household.

To create idparent, which is used in the model for identifying primary caregivers, idmother was prioritized, and if there was not an idmother then idfather was used. There were 8,576 under 18s in total who had neither a mother nor father ID; these 'loose children' had to be assigned a primary caregiver. Where the child was the grandchild/great grandchild of the head (according to

Q16RELATION), the ID number of HH was used for idparent. This accounted for 6,369 of the loose children. For the remaining children, the ID number of the oldest person in the household was used for idparent, provided the oldest person was not the child themselves.

Part 2: Simulating personal income tax using two surveys that underpin SAMOD and comparing results with administrative data

As mentioned above, SAMOD is currently underpinned by two recent datasets: NIDS Wave 4 Version 1.1, and the LCS 2014/15. This has enabled PIT estimates to be generated for tax year 2015/16 and compared to administrative sources from both the South African Revenue Service (SARS) and the National Treasury (NT). The results of this exercise were used to examine the quality of the income data in both datasets. This enabled conclusions to be drawn as regards both the unit missing and item missing/item implausible data and the role that administrative data might play in enhancing the quality of survey data for the purposes of tax-benefit microsimulation. The working paper (Wright et al. 2018) presents the main findings but the concepts and techniques used are elaborated in this technical note.

Accrual and cash flow

A key challenge when comparing PIT estimates is to ensure that the simulated data is comparable to the administrative data. SAMOD generates estimates of the amount of PIT due for a particular tax year, given the incomes of individuals for that year. That is, the model generates estimates of liability for tax on the current year basis. In accountancy terms this is known as an 'accrual' basis — the requirement to pay tax is calculated, but no consideration is taken into account as to when the tax will actually be due to be paid (for example, the actual due date of the tax may be the following year). Similarly, on an accrual basis, any arrears of tax that had been due in previous years are not taken into account.

In contrast, it is an important requirement for SARS and the NT to be able to plan for the funding of key services etc., and therefore to estimate the amount of income from PIT that will be received in practice in a particular year. This is good business practice and is often referred to in accountancy terms as a 'cash flow' basis. Accordingly, administrative data is typically presented on a cash flow basis. This will mean that, for example, arrears of PIT and fines will be included in the estimates, as will any refunds of tax due. The estimates will not include tax liabilities for the current year which are not due until subsequent years. Also, the estimates may allow for a certain degree of non-punctual payment. The consequences of a cash flow basis of reporting are particularly relevant for non-PAYE income, as for PAYE on employment income the accrued tax and collected tax should be (more or less) congruent.

Timepoint

In terms of the timepoint, the analysis was undertaken for tax year 2015/16 as this was the nearest tax year to the date of the datasets and therefore least vulnerable to the vagaries of uprating.

Taxable income groups

Another issue when comparing simulated results with administrative data is that, in the main, income taxes are not reported by income tax bands but rather by a range of so-called 'taxable income groups' for reporting purposes (e.g. NT 2015). Using SAMOD it is possible to harmonize

information on taxes and taxpayers to match these taxable income groups for the purpose of comparing with published administrative data.

Nine taxable income groups were constructed:

1) taxable income of ZAR0-70,000
2) taxable income of ZAR70,001-150,000
3) taxable income of ZAR150,001-250,000
4) taxable income of ZAR250,001-350,000
5) taxable income of ZAR350,001-500,000
6) taxable income of ZAR500,001-750,000
7) taxable income of ZAR750,001-1,000,000
8) taxable income of ZAR1,000,001-1,500,000
9) taxable income of ZAR1,500,001+

Part 3: Using NIDS to test the imputation methods applied to income data in household surveys from Tanzania and Zambia

In the working paper that explored the quality of income data in household surveys in Tanzania and Zambia (McLennan et al. 2018), a decision was made to test one of the imputation methods — predictive mean matching (PMM) — on one of the datasets underpinning SAMOD. The rationale was that income data has been more routinely analysed in the South African surveys than in other developing countries in sub-Saharan Africa. In particular, income data in NIDS has been extensively tested by the local and international academic community. For this reason the NIDS dataset was selected for testing.

The methodology adopted, as explicated in the working paper, was to select the variable to impute (in this case waged income (or *yemwg* in the EUROMOD terminology) and then create a new variable to which was assigned the natural log of *yemwg*. This new variable was named *L_yemwg_miss*. The next step was to generate a random mumber *(random)* for each case in the dataset that had a positive value for *L_yemwg_miss*. Deciles of the random numbers were created (*n_random*). Ten versions of the dataset were generated: in each of the datasets a different ten per cent of *L_yemwg_miss* cases were set to missing, using the decile flag *n_random*. So, for example, in dataset 1 *yemwg* was set to missing for decile 1 of *n_random*.

The following syntax illustrates the process by which the ten missing data datasets were created:

```
/*****************************************************************/
/* set random 10% to missing on pay_LS and then create 10 datasets */
/* each with a different 10% of L_yemwg_miss set to missing       */
/*****************************************************************/

set seed 123456789
gen random = runiform() if yemwg_miss != .
xtile n_random = random if yemwg_miss != . ,nq(10)

forvalues x=1(1)10 {
preserve
replace L_yemwg_miss=. if n_random==`x'
gen miss_flag=0
replace miss_flag=1 if L_yemwg_miss==.
save  "$work\Base File `x'.dta", replace
restore
}
```

Next, for each dataset the missing values for *yemwg* were imputed using PMM. The actual model underpinning the imputation was selected as a result of a model-fitting process. Moreover, the number of nearest neighbours (*knn*) selected was 3 rather than the 5 selected for Tanzania and Zambia. Again, this number was selected after sensitivity testing. The syntax is as follows:

```
forvalues x=1(1)10 {
use  "$work\Base File `x'.dta", clear
mi set wide
mi register imputed L_yemwg_miss

mi impute pmm L_yemwg_miss L_dag  hh_eq_expenditure i.deh_i  i.urban_rural
i.loc_i   i.dgn_i   i.deh_i   i.drc   b_overall_house_cond   b_lighting_fuel
b_heating_fuel b_toilet_type ///
b_water_source  b_walls  b_flooring    [pweight  =  dwt],  add(50)  knn(3)
rseed(1234) noisily

keep if miss_flag==1

keep hhid pid *yemwg*  n_random dwt

save "$work\Base imputed File `x'.dta", replace
}
```

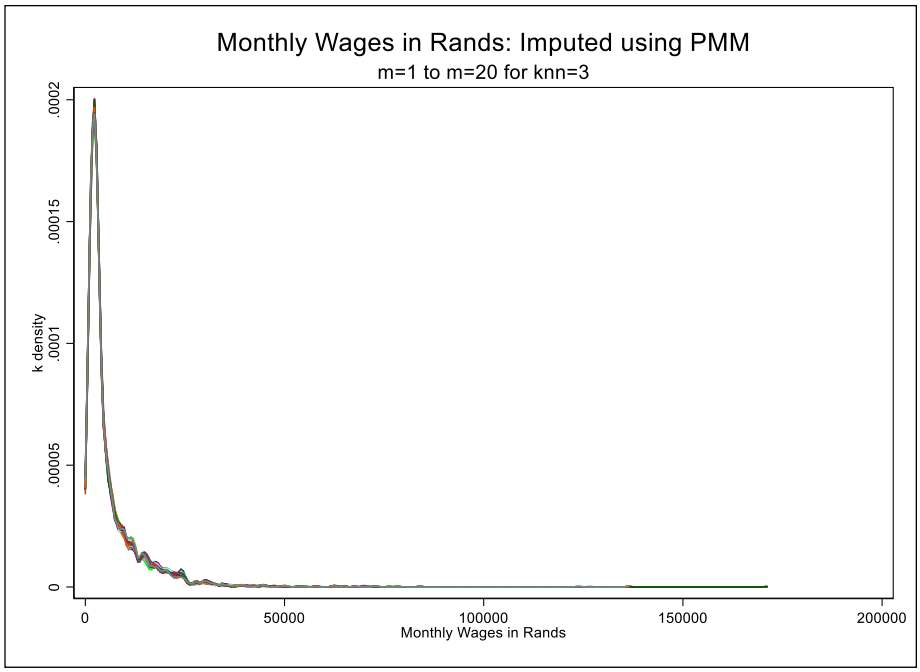
Each of the ten imputed datasets that are saved in the syntax shown above contained only the ten per cent of cases that were imputed in that imputation. Each dataset contained the following variables: household identifier (*hhid*); the individual identifier (*pid*); the variable containing the decile number of the random numbers (*n_random*); the weight (*dwt*); and finally all variables containing **yemwg**. These comprised the original waged income *yemwg*, the variable for imputation (*L_yemwg_miss*) and also each of the 50 imputed values for the missing cases. These are generated in the form *_m_L_yemwg_miss* where *m* is the number of the imputation ranging from 1 to 50.

Finally, the ten imputed datasets were appended, the anti-log was taken of each of the *_m_L_yemwg_miss* variables, and a mean imputed waged income variable was generated per case. The resultant file has all the cases where there is waged income and contains the original waged income as well as the wholly imputed waged income. This allows the original waged income and imputed waged income to be compared per case.

As reported in the working paper, scatters were created comparing original waged income with imputed waged income. A new underpinning dataset for SAMOD, containing wholly imputed waged income, was also created and the model output was compared.

In addition, various diagnostics were performed. Figure 1 shows a kernel density plot for the first 20 imputations using the wholly imputed dataset.

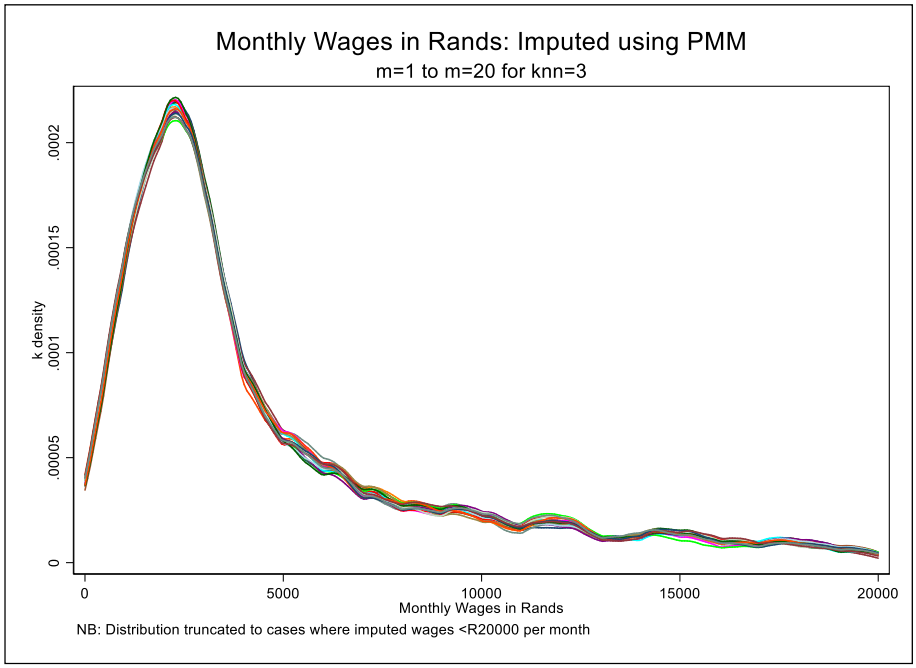
Figure 1: Kernel density plot of the first 20 imputations of monthly wages



Source: authors' calculations using imputed monthly wage data generated using PMM, for NIDS Wave 4 Version 1.1.

Because of the long tail it is difficult to examine the congruence of the 20 imputations. However, limiting the plot to cases where the imputed wages are less than ZAR20,000 per month illustrates the position more clearly (see Figure 2).

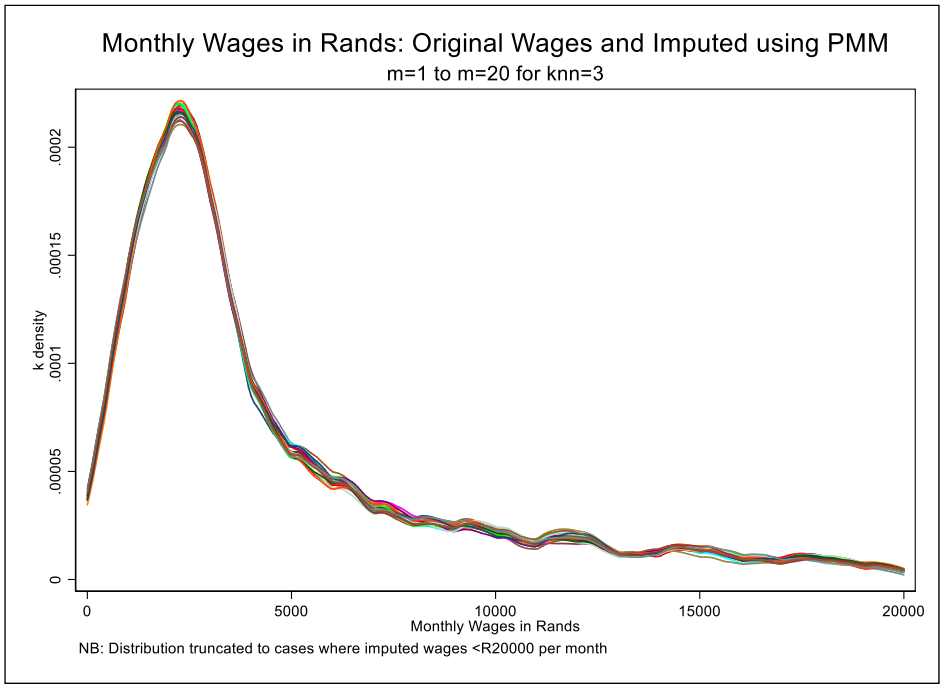
Figure 2: Kernel density plot of the first 20 imputations of monthly wages (truncated)



Source: authors' calculations using imputed monthly wage data generated using PMM, for NIDS Wave 4 Version 1.1.

If actual recorded wages are added in (again restricting to wages of less than ZAR20,000 per month), the following picture emerges (Figure 3).

Figure 3: Kernel density plot of the first 20 imputations of monthly wages (truncated) and the original reported wage data



Source: authors' calculations using imputed monthly wage data generated using PMM, for NIDS Wave 4 Version 1.1.

Although it is not possible to distinguish the addition of the original wages, the figure does demonstrate how congruent the imputations are with the original dataset.

Summary

This technical note has set out several data processes that have been undertaken using the income data in dataset(s) that underpin SAMOD. Part 1 describes various data-cleaning steps that were undertaken when preparing the LCS 2014/15 as an underpinning dataset for SAMOD. Part 2 elaborates on the process of comparing simulated estimates of PIT between two different datasets and with administative data sources. Part 3 describes a method for introducing artificial missing data in order to explore multiple imputation techniques for missing or implausible income data.

The processes of data-cleaning, imputation and validation are necessary steps for assessing and strengthening tax-benefit microsimulation models. These are themselves iterative processes, with specific issues identifiable in different country contexts and for different datasets for the same country. Findings will be shared with the custodians of the survey datasets as well as NT and SARS with a view to exploring ways in which to strengthen the estimation of eligibility for taxes and benefits.

References

McLennan, D., M. Noble, G. Wright, H. Barnes, and F. Masekesa (forthcoming). 'Exploring the quality of income data in two African household surveys for the purpose of tax-benefit microsimulation modelling: imputing employment income in Tanzania and Zambia'. WIDER Working Paper. Helsinki: UNU-WIDER.

National Treasury (2015). *Budget Review 2015*. Pretoria: National Treasury South Africa.

SALDRU (Southern Africa Labour and Development Research Unit) (2014). National Income Dynamics Study 2014–2015. Wave 4 [dataset]. Version 1.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016. Pretoria: Department of Planning Monitoring and Evaluation [commissioner], 2014.

Statistics South Africa (Stats SA) (2017) Living Conditions of Households in South Africa: An analysis of household expenditure and income data using the LCS 2014/2015, Statistical Release P0310 (2015). Pretoria: Statistics South Africa.

Wright, G., M. Noble, H. Barnes, D. McLennan, and M. Mpike (2016). 'SAMOD, a South African tax-benefit microsimulation model: recent developments'. WIDER Working Paper 2016/115. Helsinki: UNU-WIDER

Wright, G., H. Barnes, M. Noble, D. McLennan, and F. Masekesa (2018). 'Assessing the quality of the income data used in SAMOD, a South African tax-benefit microsimulation model'. WIDER Working Paper 2018/173. Helsinki: UNU-WIDER