UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2021/134

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

David McLennan,[1] Michael Noble,[1] Gemma Wright,[1] Helen Barnes,[1] and Faith Masekesa[2]

August 2021

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: The quality of data on employment income is explored using Tanzanian and Zambian household survey datasets. The extent of missing and implausible income data is assessed and four different methods are applied to impute missing or implausible values. The four imputation methods are also applied to artificial missing data for Tanzania and Zambia, and—using one approach—for a South Africa dataset. Post-imputation results are assessed. It is argued that the treatment of missing data cannot be generalized, and that tax compliance should also be taken into account when assessing the validity of a microsimulation model's simulation of direct taxes.

Keywords: income imputation, microsimulation, missing data, Tanzania, Zambia

JEL classification: C63, C81, H24, D31

Acknowledgements: The results presented here are based on TAZMOD v1.8 and MicroZAMOD v2.0. TAZMOD and MicroZAMOD are developed, maintained, and managed by UNU-WIDER in collaboration with the EUROMOD team at ISER (University of Essex), SASPRI (Southern African Social Policy Research Insights), and local partners in Tanzania and Zambia: the University of Dar es Salaam and the Zambia Institute for Policy Analysis and Research (ZIPAR), respectively. We are indebted to the many people who have contributed to the development of SOUTHMOD, TAZMOD, and MicroZAMOD. Earlier versions of this paper were presented at the SOUTHMOD Workshop on 13 June 2018 in Helsinki, Finland, at a seminar on 6 September 2018 at ISER at the University of Essex, England, and at the WIDER Development Conference on 13 September 2018 in Helsinki, Finland. Thanks are extended to all who have commented on earlier versions. The results and their interpretation presented in this publication are solely the authors' responsibility.

Note: list of acronyms at the end of the paper (Appendix E)

Update (September 2021): The technical note mentioned in the text is available as WIDER Technical Note 2021/15 here: https://doi.org/10.35188/UNU-WIDER/WTN/2021-15.

[1] Southern African Social Policy Research Insights, Hove, UK, corresponding author: david.mclennan@saspri.org; [2] Southern African Social Policy Research Institute, Cape Town, South Africa

This study has been prepared within the former UNU-WIDER project The economics and politics of taxation and social protection and published within the current project SOUTHMOD – simulating tax and benefit policies for development (Phase 2), which is part of the Domestic Revenue Mobilization programme. The programme is financed through specific contributions by the Norwegian Agency for Development Cooperation (Norad).

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9267-074-0

https://doi.org/10.35188/UNU-WIDER/2021/074-0

Typescript prepared by Gary Smith.

United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

When undertaking tax-benefit microsimulation, it is important to have confidence in the quality of the income data that is contained within the underpinning dataset of the tax-benefit model. In developed countries, survey data on income are relied upon for a wide range of statistics, including poverty and inequality estimates. In consequence, income data are extensively scrutinized and a great deal of resources are employed to that end. The situation in developing countries is rather different. Although expenditure/consumption data in social surveys have been used extensively in developing countries, income data have been used much less frequently and have come under less rigorous scrutiny (e.g. Beegle et al. 2015), with some notable exceptions (e.g. Ferreira et al. 2016).

In Africa and elsewhere, a family of tax-benefit microsimulation models based on the EUROMOD software have been developed under the SOUTHMOD banner (Decoster et al. forthcoming; Sutherland and Figari 2013). In this paper an assessment is made of the income data underpinning two African SOUTHMOD tax-benefit microsimulation models: TAZMOD for Tanzania (Leyaro et al. 2017), and MicroZAMOD for Zambia (Nakamba-Kabaso et al. 2017). Both models are underpinned by nationally representative household survey datasets, and use the EUROMOD software (University of Essex 2017). During the process of building each of the country models, the developers identified that the underpinning survey datasets presented challenges regarding the quality of the income data. Although at the time of model development a number of steps were taken to address these data challenges, this paper explores the issues in more detail and demonstrates techniques that can be applied to further strengthen the income data for inclusion in the microsimulation model input dataset.

The focus of the paper is on practical approaches to data cleaning and imputation of income data rather than a general review of different approaches to imputation in general (for which see Graham 2009; Heeringa et al. 2017; Little and Rubin 2002) or multiple imputation specifically (Statacorp 2017; Rubin 1987, 1996). To emphasize the practical nature of this paper, it is followed by a Technical Note that gives detailed descriptions of the data cleaning and preparation steps undertaken in Tanzania and Zambia, as well as further details on the imputation methods employed (Technical Note forthcoming).

Household survey data on income can be problematic in a number of different ways, and although discussed in more detail below, the issues can be grouped into three categories. First, information on income may be missing for a particular individual within the data (referred to as 'item missing'). Second, the information on income may be present in the survey dataset but appear improbable, based on other information in the survey about that individual (referred to as 'item implausible'). Third, the survey may under-represent certain groups such as high-income earners (referred to as 'unit missing'). Although all three issues are important, this paper focuses on the first two issues only—that is, item missing and item implausible data.

TAZMOD and MicroZAMOD each use many different types of self-reported income data. However, for the purposes of this paper we focus on just one type of income: employee income from primary employment, which is the major or sole component of the variable '*yem*' in EUROMOD terminology.[1] This particular income category is used as a test case, in order to

[1] More specifically, in Tanzania the variable of main interest was the cash payment for primary activity (HBS 2011/12: question S12Q28A) for people aged 15 and over whose primary activity was 'working for pay' (HBS 2011/12: question S12Q10A). In Zambia the variable of main interest was 'How much is your regular gross monthly salary/wage including regular allowances and transport allowances, regular overtime, retention allowance, from the main job?'

highlight the importance of interrogating each of the different types of income data that are used by the models.[2]

The paper is structured as follows. Section 2 presents an overview of the income data that underpin the two country models, TAZMOD and MicroZAMOD, and highlights the main challenges posed by the income data. This section also includes a discussion of the key principles for consideration when addressing the issues of item missing and item implausible data, and provides a rationale for pursuing the various imputation techniques that are adopted in this paper. Section 3 outlines the data-checking and -cleaning steps that were undertaken on the Tanzania dataset; this includes reviewing and cleaning variables likely to be predictors of earned income and reviewing the income data in the context of other demographic and labour market variables in order to identify missing or implausible income values. Section 4 introduces the imputation techniques that were applied on both the Tanzanian and Zambian datasets. Section 5 presents the results for the Tanzanian dataset. Section 6 provides an account of the main findings for Zambia with an emphasis on how they differed from Tanzania, with further details included in Appendices B–D. In Section 7, one of the imputation techniques is tested on a South African dataset which has better-quality income data. Section 8 reflects on the main findings and includes recommendations for future work.

2 The Tanzania and Zambia datasets and key principles for dealing with missing data

This section introduces the Tanzania and Zambia survey datasets, and highlights the main challenges posed by the income data with reference to the two tax-benefit microsimulation models TAZMOD and MicroZAMOD. The section concludes with an overview of principles that should be taken into consideration when dealing with missing data.

2.1 Tanzania

The database underpinning TAZMOD v1.8 was drawn from the Tanzania Household Budget Survey (HBS) 2011/12, a cross-sectional survey of Mainland Tanzania that was conducted by the National Bureau of Statistics (NBS). The Main Report (NBS 2014a) and Technical Report (NBS 2014b) provide key information about the HBS.

The survey is representative at a national level for Mainland Tanzania, and at a sub-national level for Dar es Salaam, other urban areas, and rural areas. The questionnaire has a number of sections that cover aspects of income generation, including labour market participation, agricultural production, non-agricultural business conducted by individuals, as well as other sources of income, including state and private transfers.

The income data in the HBS have already been subjected to some cleaning by the NBS. So, for example, they write that:

(LCMS 2015: section 6, question 27), for people aged 15 and over who reported their main current economic activity status as being 'in wage employment' (LCMS 2015: section 5, question 1).

[2] Attempts were made to fit models to form the basis of imputations for the two other main sources of income in Tanzania: income from self-employment and income from agriculture. Suitable covariates could not be identified for them, and so instead they were each capped at the 99th percentile. It was also not possible to fit a model for income from secondary pay, but there were very few such cases in the Tanzanian dataset.

> data [were collected] on labour status, household businesses and individual income and hence data cleaning and validation dealt with these topics. The activities carried out in cleaning and validations of this form are:
>
> i) To check the logical flow of the questions and skipping patterns.
>
> ii) To check for the missing data.
>
> iii) To validate individual and households income. (NBS 2014b: 43)

Despite this, the NBS caution against the use of the income data for poverty measurement:

> The Tanzanian poverty estimates are based on aggregate household consumption as the key welfare indicator. As in many other parts of Sub-Saharan Africa, consumption is considered a more reliable indicator of welfare than income. First, consumption is typically less fluctuating than income and gives a better and steadier picture of long-term welfare. Second, individuals feel more comfortable answering questions related to consumption than to income. Third, *income measurement in countries with a large agricultural or informal sector is often highly inaccurate.* (NBS 2014b: 48; our emphasis)

Concerns about the quality of the HBS income data arose during the development of the TAZMOD model by the authors and colleagues. Early versions of the input dataset (which was derived from the HBS but did not incorporate any additional cleaning beyond that already undertaken by the NBS) generated very skewed income ranges. Moreover, the direct taxes generated by the model were far in excess of those reported by the Tanzania Revenue Authority (TRA) (Leyaro et al. 2017).

Preliminary investigations of the data revealed that income from employment was by far the largest contributor to the overestimation of personal income tax (PIT). Implausible employment incomes seemed, in part, to be generated by poorly recorded periodicity of receipt. For example, many of the highest and implausible incomes were as a result of the amount received being recorded as an hourly amount rather than, say, a monthly amount. This resulted in the multiplication of a relatively small sum by 40 (hours in a week) and then by 4.25 (weeks per month) in order to generate the requisite monthly amount. Several manual adjustments were made: a cap was applied to hourly and daily employment incomes above TZS5,000/US$2.22 and TZS100,000/US$44.44 respectively by recoding their periodicity to monthly;[3] employment income where periodicity is recorded as 'other'/'not stated' was set to missing; employment income was capped at the 99th percentile by occupational class; self-employed income and agricultural income were capped at the 99th percentile; and 'other income' was excluded if it related to a payment to a child, and was then capped at the 99th percentile.

Table 1 illustrates the impact of these initial outlier adjustments. Column 1 shows the total PIT simulated using TAZMOD for 2015 by income source before outlier adjustments. The PIT has been apportioned to the income sources according to their share of taxable income, as in practice these income streams are pooled for the purpose of calculating the PIT policy, having stripped out any income pertaining to the turnover tax. Column 2 shows the apportionment by income source after initial outlier adjustments by the model development team.

[3] Values suggested by Dr Leyaro and Dr Kisanga as plausible for work paid at an hourly or daily rate.

Table 1: PIT simulated in TAZMOD, tax year 2015/16, apportioned by share of income source, before and after initial outlier adjustments of income data—Tanzania

Source of taxable income	1 PIT, apportioned by share of taxable income—before outlier adjustments (TZS million)	2 PIT, apportioned by share of taxable income—after outlier adjustments (TZS million)	3 Contribution to PIT, apportioned by share of taxable income—before outlier adjustments (%)	4 Contribution to PIT, apportioned by share of taxable income–after outlier adjustments (%)
Employment	8,456,007	1,697,574	72.4	43.4
Self-employment	2,196,343	1,398,322	18.8	35.8
Agriculture	317,816	118,714	2.7	3.0
Other	710,935	695,454	6.1	17.8
Total	11,681,102	3,910,064	–	–

Note: annual amounts. PIT restricted to those working in the formal sector.

Source: authors' calculations using TAZMOD V1.8.

As can be seen, the outlier adjustments caused the amount of simulated PIT to reduce in size considerably, from TZS11,681 billion to TZS3,910 billion. Also, the contribution of employment income to PIT fell from 72 per cent to 43 per cent. A further consequence was that the proportional contribution to PIT from self-employment income rose from 19 per cent to 36 per cent. The simulated direct taxes using these two datasets in TAZMOD captured 493 per cent of reported direct taxes before outlier adjustments, and 167 per cent after outlier adjustments.

Further interrogation of the income data, including imputation where necessary, was thus regarded as essential to further enhance the quality of the model.

2.2 Zambia

MicroZAMOD v2.0 is underpinned by the Living Conditions Monitoring Survey (LCMS) 2010 and 2015, both of which were undertaken by the Republic of Zambia Central Statistical Office (CSO). The key results from these two waves of the LCMS, along with the methodologies, were published in the Living Conditions Monitoring Survey Report 2006–2010 (CSO 2012) and the 2015 Living Conditions Monitoring Survey Report (CSO 2016), respectively.

The LCMS 2015—the Zambia dataset used in this paper—is a household survey designed to be representative at national, provincial, and residence (urban/rural) levels. Enumeration was undertaken during the months of April and May 2015. The survey was designed to cover a representative sample of 12,260 non-institutionalized private households and achieved a response rate of 98 per cent. Non-responding households were systematically replaced, resulting in a final enumerated sample size of 12,251 households, containing 62,880 individuals.

The LCMS 2015 contained a series of questions on individual and household income sources, such as labour market participation, agricultural production, non-agricultural business conducted by individuals, as well as other sources of income, including state and private transfers. While the LCMS 2015 report (CSO 2016) refers to the application of imputation techniques to deal with missing or implausible values on consumption measures, there is no reference to any cleaning or imputation work undertaken by CSO on the various income sources.

In the same way as for Tanzania, concerns about the quality of the income data in the LCMS 2015 arose during the development of the MicroZAMOD model by the authors and colleagues. For

example, of those individuals aged 20–59 who reported their current economic status as 'employee' in the formal sector, 6 per cent reported a zero value for their regular gross monthly salary/wage, and a further 8 per cent had a missing value. For those individuals aged 20–59 who reported their current economic status as 'employee' in the informal sector, 16 per cent reported a zero value for their regular gross salary/wage, and a further 12 per cent had a missing value.

As reported in the Country Report (Nakamba-Kabaso et al. 2017), each income variable was assessed in terms of its distribution and the effects of any outliers. Where relevant, incomes were capped to minimize the effect of outliers. Two income categories were capped at the 99th percentile value (*ypr*, *yiyit*); one was capped at the 90th percentile value (*ypp*); four were capped at particular numeric values (*ys*e, *yiy*, *yot*, *yag*); and three were not capped at all as the distributions looked plausible (*yem*, *ytn*, *ypt*).

In addition to these missing and potentially implausible income values in the LCMS 2015 dataset, a comparison of the MicroZAMOD simulations of direct taxes with external validation statistics revealed a substantial disparity, with MicroZAMOD simulating only 31 per cent of the published figures (although the reported value is not strictly comparable as it also contains 'withholding tax'[4]). Table 2 shows the monetary values simulated in MicroZAMOD compared to the available external statistics.

The challenge with the Zambia data is therefore different from the Tanzania dataset, as it appears that income data may be under-reported in Zambia whereas it may be over-reported in Tanzania.

Table 2: Direct taxes simulated in MicroZAMOD, 2015—Zambia

Tax-benefit policy	1 MicroZAMOD 2015 (ZMW million)	2 External 2015 (ZMW million)	3 Per cent captured (1/2)
Turnover tax	698	10,005	31
Personal income tax	2,431		

Note: in column 1, outliers for certain components of income were adjusted, as described above, but no adjustments were made to employment income (*yem*). The figure in column 2 comprises PIT, turnover tax, and withholding tax. ZMW = Zambian Kwacha.

Source: authors, based on (column 1) MicroZAMOD V2.0 and (column B) Ministry of Finance (2016: 28, 30).

2.3 Key principles when dealing with missing data

There are many different methodological approaches available for dealing with the problem of missing data in social surveys (Graham 2009; Little and Rubin 2002). In order to choose the most appropriate method, it is necessary to identify the source(s) and type(s) of missing data; review the extent and pattern(s) of missingness in a dataset; and consider the analytical motivation(s) for dealing with the missing data problem (Heeringa et al. 2017). The remainder of this section addresses each of these issues.

[4] As elaborated by Nakamba-Kabaso et al. (2017), the Ministry of Finance publishes income tax totals for company tax (not relevant here), pay-as-you-earn (PAYE), and 'Other income tax—withholding tax' (which includes turnover tax and other income taxes). This means that turnover tax is combined with all other categories of withholding tax in the published data and so it is not possible to compare the simulated outputs with directly comparable categories of published figures for income tax. This is particularly relevant in 2015 as the Ministry of Finance notes that there was a particularly high amount of property transfer tax received that year, which is included within the withholding tax reported figure but was not simulated in MicroZAMOD: 'Withholding tax was also higher by 32.9 percent mainly boosted by higher than anticipated property transfer tax collections' (Ministry of Finance 2016: 29).

In terms of categorizing different types of missing data, it is possible to draw a basic distinction between 'item missing' and 'unit missing'.[5] Item missing data refers to instances in which a sample unit does participate in the survey, but there is a missing value on one or more variables in an otherwise complete survey record. Item missing data may be due to factors such as a respondent's refusal to answer a particular question (or subset of questions) or error in data capture by the interviewer. In contrast, unit missing data (or unit non-response) refers to those instances in which a sampled unit is not contained within the final survey dataset due to either refusal to participate or inability to participate. In such instances, no data are collected from these sample units. Unit missing data problems are typically dealt with by weighting the successfully enumerated survey cases appropriately to adjust for non-response bias (Groves et al. 2009; Lacerda et al. 2007). As previously stated, the focus of this paper is on item missing data, which includes cases with a missing value, as well as cases where the value has been set to missing as it is implausible.

The extent and pattern of missing data in the dataset also informs the choice of methodology for dealing with missing data. Where there is item missing data within a social survey dataset, it is necessary to assess the pattern of missingness and the extent to which the pattern of missingness is associated with other variables in the dataset. The relationship between the missing data pattern on any given variable(s) and the pattern of responses on other variables is referred to as the missing data mechanism. The missing data mechanism can take one of three forms, depending on the extent to which the pattern of missing data is conditional on the pattern of observed values across the dataset. The three missing data mechanisms can be summarized as follows:[6]

- Missing completely at random (MCAR), where the probability of a value being missing is random across the entire survey dataset.
- Missing at random (MAR), where the probability of a value being missing is random across different identifiable subsets of sample (e.g., sex, race, employment status), but where the probability of being missing is different *between* the identifiable subsets. In other words, missingness depends only on the observed values of the variables in the survey, not on the missing values themselves.
- Missing not at random (MNAR),[7] where the probability of a value being missing is not random across the entire sample or within identifiable subsets. In other words, missingness depends on the unobserved values of the variables with missing data (as well as potentially the observed values).

It is relatively straightforward to test whether the pattern of missingness is systematically different between different subsets of the survey dataset and, as such, to distinguish between MCAR and MAR as a starting point. However, the very nature of missing data means it may be difficult to make confident assumptions as to whether or not the pattern of missingness is partly (or indeed wholly) conditional on the unobserved values themselves. It may therefore be difficult to confidently discount the possibility of the missing data mechanism being MNAR.

---

[5] While the basic distinction between unit missing and item missing is a useful starting point for the analyses considered in this paper, other commentators have identified additional types of missing data that might be relevant in other country contexts. See, for example, Heeringa et al. (2017), who refer also to 'wave nonresponse' in relation to panel data and 'phase nonresponse' in relation to studies where survey enumerators ask for consent to match respondents' data with external data sources.

[6] For a more detailed discussion see, for example, Heeringa et al. (2017), Lacerda et al. (2007), and Little and Rubin (2002).

[7] Sometimes referred to as 'not missing at random' (NMAR).

When considering the possible missing data mechanism relating to missing income data, it is conceivable that while missing income data might be related in part to observable variables (e.g., employment status, occupation type), it may also be related in part to the true value of income, thereby meaning that it is actually a case of MNAR. Heeringa et al. (2017) note that most imputation methods assume the missing data mechanism is MAR and that approaches to dealing with data MNAR are far more complex than those designed to deal with data MAR. For the purpose of this paper, the assumption is made that the missing data mechanism is MAR, as MNAR cannot be ascertained because it is unobservable.[8]

Another determinant in choosing the most appropriate method for dealing with missing data is the analytical purpose that motivates the process. The methods can broadly be grouped into 'imputation' methods and 'non-imputation' methods (Graham 2009). If the purpose of the analysis is simply to calculate certain parameter estimates across the dataset, adjusting for the potential effect of missing data, then non-imputation methods may be adequate. However, if the purpose of the analysis is to replace missing values in the dataset with plausible values—as is the case here—then imputation methods are needed. A key feature of imputation methods is that they all result in a 'complete dataset' consisting of observed values plus imputed values, whereas non-imputation methods do not result in complete datasets.

The principle underpinning all imputation approaches is that the value of a missing data item can be predicted by drawing upon observed values in the dataset, with single imputation approaches resulting in a single complete dataset, and multiple imputation techniques resulting in a series of complete datasets. Irrespective of whether the single imputation or multiple imputation option is chosen, Heeringa et al. (2017) recommend that the exact methodology adopted should ideally be stochastic in nature, either through being based on random draws from selected observed cases (e.g. hotdeck imputation) or through being based on model parameters and error terms (i.e. regression models with in-built randomness).

An advantage of multiple imputation over single imputation is that multiple imputation enables the calculation of standard errors to qualify the robustness of the imputation process (Rubin 1986, 1987). As Ardington et al. state:

> any single imputation technique does not distinguish between observed and imputed values in the resultant data set and as such the variance of any estimates is understated. Multiple imputations generate a distribution of imputed values and a distribution of parameters of interest. This allows for the uncertainty due to imputation to be reflected in the standard errors of the estimates. Given such advantages, the imputation literature has a strong preference for running multiple imputations using a suitable multivariate technique. (Ardington et al. 2005: 4)

A further consideration is the extent to which an item missing data problem may also occur in variables other than income. For example, if certain key covariates of income are also subject to item missing data, then these may also need to be imputed, and this was indeed the case with the Tanzania and Zambia datasets. A particular family of multiple imputation techniques that is designed to deal with imputing multiple variables within the same imputation process is sequential regression multiple imputation (SRMI), sometimes referred to as multiple imputation using

[8] The analysis in Section 5 supports our assumption that the data are not MCAR. However, it is not possible to definitely state that the data are MNAR. It therefore remains an assumption that the data are MAR.

chained equations (MICE) (Azur et al. 2011; Raghunathan et al. 2001). This is discussed further in Section 4 with examples using the Tanzania dataset.

3 Data cleaning, income validation, and identification of implausible income: Tanzania

In this section, the internal data cleaning and income validation steps are summarized, using the Tanzania HBS dataset as an example. These necessary steps precede the multiple imputation routines reported in Section 4.

The extent of missing and implausible values was explored for a set of variables that were likely to be good predictors of employee income from primary employment: gender, age, level of education, labour market status, consumption, and lastly—the main variable of interest—employee income from primary employment. The demographic variables—gender, age, and level of education—were examined for all cases, whereas labour market status, consumption, and income were only examined if the individual was an employee according to the variable on primary activity (S12Q10A) and if the individual was aged 15 or over. The objective was to produce a dataset for imputation, comprising a subset of all cases in the HBS, of only those individuals aged 15 or over whose primary activity is 'working for pay' (i.e. employees in both formal and informal sectors[9]).

The HBS dataset had no missing values for *gender* and so no imputation was required for this variable. The NBS states that the *age* variable in the HBS had been checked for consistency between the year of birth and the age of the household member and that any inconsistencies had been corrected with reference to the original questionnaire (NBS 2014b). The outcome of the NBS data cleaning was checked using survey date and month and year of birth information. For the small number of cases (16) where age is coded as 98 (don't know) or 99 (not stated)—that is, missing data—age was either calculated from year of birth information or a manual imputation was made using other information in the survey, resulting in no missing values for age. A population pyramid suggests that age as cleaned is likely to be robust.

Regarding *level of education*, the question on highest grade completed was considered to be important for the purpose of imputation of employee income. One would not expect many of those currently in school to be working for pay (though there are 91 individuals where this is the case[10]). It was necessary to tidy inconsistencies between the question on current school attendance and highest grade completed (as the latter should not be completed if the individual is currently in school). A 'highest level of education' variable was then created with the following categories: 0 'No schooling';[11] 1 'Not completed primary'; 2 'Completed primary'; 3 'Completed secondary'; 4 'Completed tertiary'. In addition, the under-fives and those currently attending school were recoded as 'missing, not needing imputation'. The remaining cases with missing information for highest level of education were all recorded as not currently in school, though some had a value

[9] However, simulations of PIT in TAZMOD exclude the informal sector.

[10] Further investigations revealed that approximately three-quarters of this group are aged over 18 and so it is quite possible that many are incorrectly coded as in school. Additionally, over 90 per cent are receiving some pay for their primary activity, and over half are in the top three occupational classes ('Senior officials and managers', 'Professionals', and 'Technicians and associate professionals'). Again, this would suggest that many are incorrectly coded as in school.

[11] The 'no schooling' category includes individuals who have not had any schooling and those whose highest grade completed is adult education.

for the questions on whether the school is public or private and current grade. This inconsistency could either be that the question on current school attendance has been miscoded as they are actually in school, or that the current grade information instead relates to highest grade. A case-by-case judgement was made about how best to manually recode these cases in order to avoid any missing values for highest level of education.

With regard to *labour market status*, having assessed the range of questions in the HBS, the variables relating to primary activity and TASCO code[12] in section 12 of the survey appeared to be the most useful for the purpose of imputation of employee income. For all but 80 cases (aged 15 or over) where the primary activity is '4. Working for pay' (i.e. employees), there is an occupational TASCO code. We would not necessarily expect any of the other categories of primary activity to have an occupational code—except, perhaps, the apprentices—and therefore all cases (other than employees) missing a TASCO code were recoded as 'missing, not needing imputation'. There is no way of determining an appropriate occupational code for those who are working for pay but missing this information and therefore a decision was made to impute a value for these missing cases.

Turning next to *employee income from primary employment*, this was also explored for the group of interest, that is individuals who report their primary activity as '4. Working for pay' (i.e. employees). Of the 4,328 cases (aged 15 or over) who are working for pay, 99 individuals do not report a cash payment in the variable S12Q28A relating to their primary activity, referred to here as 'primary pay'. Of these, 45 are not planning to return to their primary activity and so their primary pay was recoded as 'missing, not needing imputation'. For the remaining 54 cases, primary pay was recoded as 'missing, needing imputation'. This group comprised 36 individuals with missing information for primary pay and 18 individuals with a value of zero. For all other cases reporting a primary pay and yet with a primary activity that is not 'working for pay', primary pay was recoded as 'missing, not needing imputation'.[13] Similarly, all remaining zero values for primary pay were recoded as 'missing, not needing imputation' as the individuals were either apprentices or working on the household farm.

Before investigating implausible values, the primary pay variable (in Tanzanian shillings, TZS) was converted to US dollars (US$) to help interpret results, then converted to a monthly amount using S12Q28B on periodicity of payment (having adjusted for number of months worked in the 12-month period using S12Q27).

For some cases there are very high values for primary pay where the time period is given as hourly or daily. Caps of TZS5,000/US$2.22 per hour and TZS100,000/US$44.44 per day were chosen and any primary pay above the threshold was recoded as 'missing, needing imputation'.[14] In addition, where the periodicity was unknown or 'other', primary pay was recoded as 'missing, needing imputation'.

For the purpose of further investigating the implausible values, primary pay in dollars was transformed to a natural logarithmic scale and boxplots were generated to give a visual

[12] Tanzania uses a four-digit TASCO (Tanzania Standard Classification of Occupations) code, the first digit of which maps to the International Standard Classification of Occupations (ISCO).

[13] Over half of these individuals are apprentices, and although it is quite likely that they are genuinely receiving pay for their work, they are a unique group in terms of pay which could distort the imputation process.

[14] These are the same thresholds described in Section 2, where previously periodicity was recoded as monthly for hourly and daily values above the respective thresholds (rather than being recoded as missing).

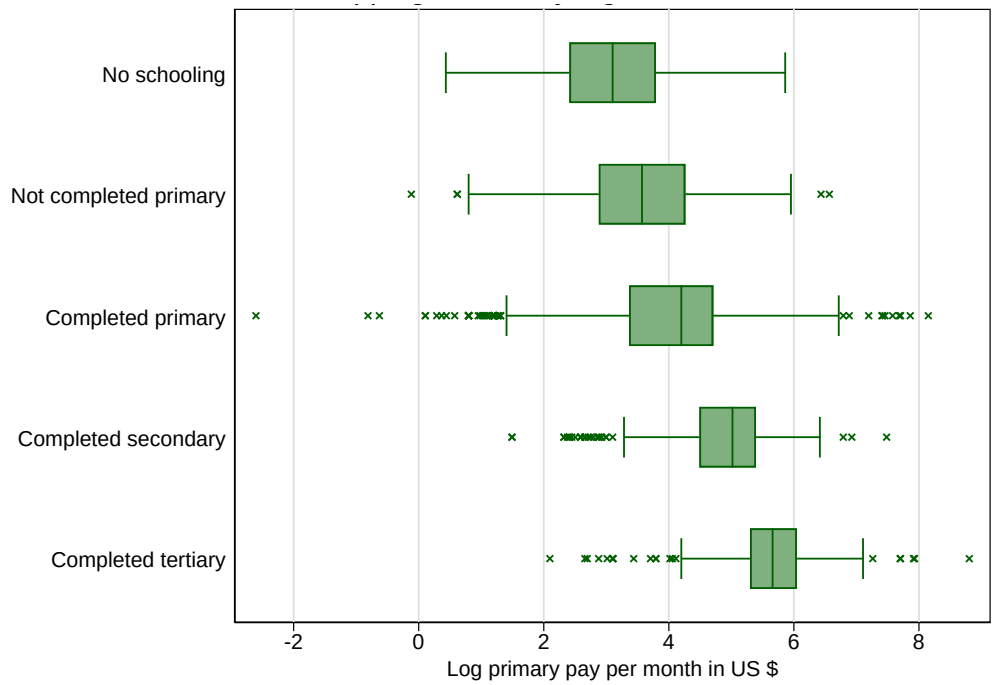
representation of the outliers in terms of log of primary pay by occupational category and highest level of education (Figures 1 and 2).

Figures 1 and 2 give a clear indication that the occupational category and highest level of education variables are likely to be good predictors of primary pay in the imputation process. In terms of highest level of education, the median values, for example, increase as the level of education increases from no schooling through to having completed tertiary education. The pattern of median values for occupational category is not quite as clear-cut as there is not a distinct hierarchy of occupation. However, when each category is taken in turn and compared to other categories, the median values appear to be plausible.

Figure 1: Primary pay outliers by occupational category: Tanzania

Source: authors' calculations using HBS 2011/12.

Figure 2: Primary pay outliers by highest level of education: Tanzania



Source: authors' calculations using HBS 2011/12.

In Stata, data points are plotted separately if they are more than 1.5 times the interquartile range distant from either the upper or lower quartile. This method of identifying outliers gives a good starting point to identify implausible values. However, as outlier calculation depends on the scale being used, the original rather than logarithmic scale was used to identify outliers. Outliers were therefore identified as values that are more than 1.5 times the interquartile range distant from either the upper or lower quartile, by either occupational category or highest level of education. In practice, there are not any outlying values using this definition at the lower end of the distribution. A value was classified as implausible if it was an outlier in terms of primary pay by occupational category or highest level of education (or both). These implausible values were recoded as 'missing, needing imputation'.

Finally, *consumption* data were explored. Monthly consumption per equivalent adult (*aec* in the original dataset) is a variable created by NBS for the purpose of producing poverty estimates.[15] Percentile values were calculated for the *aec* variable for cases aged 15 or over. It was decided that values above the 99th percentile are implausible and should be recoded as 'missing, needing imputation'. This threshold is approximately equivalent to the value at 10 times the interquartile range away from the upper quartile for all cases.[16] It is interesting to note that 40 per cent of the cases set to missing due to consumption being implausibly high are also classified as implausible in some way in terms of primary pay. It will be the task of the imputation procedure to find

[15] The household-level consumption aggregate comprises all food and non-food consumption with the exception of housing-related expenditure, actual rent, and imputed rental values for homeowners, use values for large durable items, and household-level investments. The consumption aggregate is divided by an equivalence scale that adjusts consumption for differences in household size and composition and differences in consumption needs between children and adults (NBS, 2014b: 48–52).

[16] This is for all cases together rather than by occupational category or highest level of education as previously described. Note that outliers at the lower end of the distribution were judged to be unproblematic as there are not any values which fit the standard definition of an outlier (i.e. a value greater than 1.5 times the interquartile range away from the lower quartile).

appropriate values to replace the missing values for these variables, and for all other missing values in the dataset.

The final dataset for imputation contains 4,283 cases (out of the original 4,328 cases who are working for pay and aged 15 or over) as all cases with missing information for primary pay where imputation is not needed were deleted.[17] The number of missing values requiring imputation is shown in Table 3. There are 409 cases (approximately 10 per cent) that need to have primary pay imputed. The proportion is much lower for the covariates, and as discussed above, gender, age and highest level of education do not have any missing values.

Table 3: Number of cases in HBS requiring imputation by variable, for employees only: Tanzania

Variable	Missing, needing imputation—implausible value	Missing, needing imputation—other reason	Good value
Gender	0	0	4,283
Age	0	0	4,283
Highest level of education*	0	0	4,214
Occupational category	0	35	4,248
Monthly consumption per equivalent adult	42	0	4,241
Primary pay	355	54	3,874

Note: the table includes employees aged 15 and over only and excludes 45 cases where primary pay is missing and imputation is not required as the individuals are not planning to return to their primary activity.

* There are also 70 cases missing a value but where imputation is not needed (coded –99). These are cases where the individual is recorded as currently in school but also receiving primary pay. Note that one of these cases was recoded as such on the basis of age only as information on the individual's current schooling was unavailable.

Source: authors' calculations using HBS 2011/12.

4 Imputation approaches implemented: Tanzania

Four imputation methods were implemented using the Tanzanian dataset in the first instance, and were then repeated using the Zambian dataset. The methods were: linear prediction (in Stata code '*predict, xb*'); predictive mean matching (PMM) (in Stata code, '*mi impute pmm*'); and two variants of SRMI: SRMI Regress (in Stata code '*mi impute chained (regress)*') and SRMI PMM (in Stata code '*mi impute chained (pmm)*').

The basis for each imputation technique is a regression model or models. For linear prediction and standard PMM, this is an ordinary least squares (OLS) regression model as the main variable of interest is continuous (primary pay). As regards the two SRMI models, these are predicated on sequential regression models, where a combination of OLS and logit models are used.

The multiple imputation approaches (PMM, SRMI Regress, and SRMI PMM) produce a number of complete datasets (Ragunathan et al. 2001). The user specifies the number of discrete imputations *(M)* to be produced, with each imputation $m = 1,\ldots, M$ generating a separate complete dataset. In the case of SRMI Regress and SRMI PMM, each separate imputation is generated

[17] This is a group of individuals (45 cases) not planning to return to their primary activity and therefore effectively no longer in the group working for pay.

through a process of iterative model fitting, with the user able to specify how many iterations should be performed to generate each imputation.

4.1 Linear prediction

An OLS regression model was fitted using the same covariates as in the other methods but the imputation was achieved by straightforward prediction (in Stata code '*predict, xb*'). Prior to the regression, covariates with missing values (adult-equivalent consumption and occupational status) were imputed using a hotdeck approach.[18]

4.2 Predictive mean matching

The PMM method operates by taking the predictions from the specified regression model and creating predicted values of the variable to be imputed for all cases, including missing cases. The algorithm identifies 'donors' (also referred to as 'nearest neighbours') with predictive values as close as possible to the predictive value for each case with missing data. It then substitutes the actual value of one of the respective donor cases (chosen at random) for the missing value. The number of donors is set when the model is specified. The PMM method is a multiple imputation technique, and so this process is repeated a specified number of times, and on each occasion the algorithm introduces a small amount of random error.

Regarding the number of donors, there are no clear recommendations in the literature, except that too many donors can result in increased bias, and too few donors can increase the variability of the estimates (e.g. StataCorp 2017). The number of donors was tested and an examination of the kernel density plots of the imputed values suggested that for the Tanzania dataset, five donors was optimal.

Fifty imputations were specified. When specifying multiple imputation models, it is possible to either have additional cases added to the dataset ('long'), or for additional imputed variables to be added ('wide'). Given that the objective was to produce a new input dataset for TAZMOD, the 'wide' option was used.

For PMM it is a prerequisite that covariates have no missing data. Occupational class (*loc*) was an important predictor, and so as a preliminary step the missing occupational class variable was imputed using hotdeck imputation. This generated a new occupational class variable (*loc_i*).

4.3 Two variants of SRMI: SRMI Regress and SRMI PMM

SRMI is especially useful where the dataset contains a number of variables with missing values or multiple variables of varying data type (e.g. a mixture of continuous and categorical) as the model form can be specified for each individual variable with missing data.

The SRMI approaches involve a number of iterations for each imputation. The algorithm works through each iteration in turn, starting with the variable with the smallest number of missing values and regressing this variable on the variables with complete data, using the model form specified by the user (e.g. OLS for continuous outcomes; logistic regression for binary outcomes). Next, the variable with the second smallest number of missing values is regressed (again using the model form specified by the user) on the variables with complete data plus the newly imputed variable. This continues until the variable with the largest number of missing values is regressed on the

[18] Using Stata ado file *hotdeckvar* by Matthias Schonlau.

variables with complete data plus the variables that have already been subject to the imputation (i.e. the versions of these variables that include the imputed values). The process then repeats as a second iteration, again starting with the variable that had the fewest values missing originally, but this time regressing it on the variables with complete data plus the imputed versions of the variables with originally incomplete data.

As noted above, the number of imputations, and the number of iterations per imputation, are set by the user. Rubin (1986) found that 5–20 imputations is usually sufficient to achieve all reasonable efficiency, although the need to limit the number of imputations and iterations was primarily motivated by lack of computational resources, which has become less of a problem with advances in computing power. SRMI can be implemented in a number of statistical software programs, including Stata by using the '*mi impute chained*' command (StataCorp 2017).

As emphasized in Section 2, the primary objective is to impute missing and implausible incomes, drawing upon relevant covariates in the dataset. However, some of the covariates also have missing or implausible values, and so the SRMI approach will impute missing values across all the variables that are entered into the model.

The details of the process of developing the final imputation specification are set out in the Technical Note, and the variables that were used are listed in Appendix A. The Technical Note also contains details of the diagnostic steps that were undertaken, which confirmed that the imputation specification was appropriate.

In addition to the SRMI Regress model, a variant was produced which incorporates the PMM approach: SRMI PMM. This was specified in a similar way to the SRMI Regress model, but as with the standard PMM model, covariates with missing data were first imputed using a hotdeck technique.

4.4 Imputing values for artificial missing data

As a separate exercise, and in order to assess the four methods described in this section more thoroughly, artificial missing income data were introduced to the Tanzania dataset and imputed.

The process of introducing artificial missing data was undertaken by creating a subset of the input dataset containing just those individuals with employment income (i.e. income from primary pay in the case of Tanzania). Next, missing data were introduced as follows: each observation was assigned a random number which was then used to generate decile groupings. Ten separate files were created. In each file, 10 per cent of employment incomes were set to missing based on the decile of random numbers. That is, in the first file, all cases in decile 1 had their employment income set to missing, in the second file cases in decile 2 had their employment income set to missing, and so on. The four imputation techniques (linear prediction, PMM, SRMI Regress, and SRMI PMM) were then applied to each of the 10 separate files. Having run the four imputation techniques, the observations containing imputed income from each of the 10 files were then extracted and appended so that a complete file was created where all the cases had imputed income data which could then be compared to the original (observed) income data for each of the four imputation techniques.

5 Results: Tanzania

This section presents the results for the four imputation methods that were applied to the Tanzanian income data.

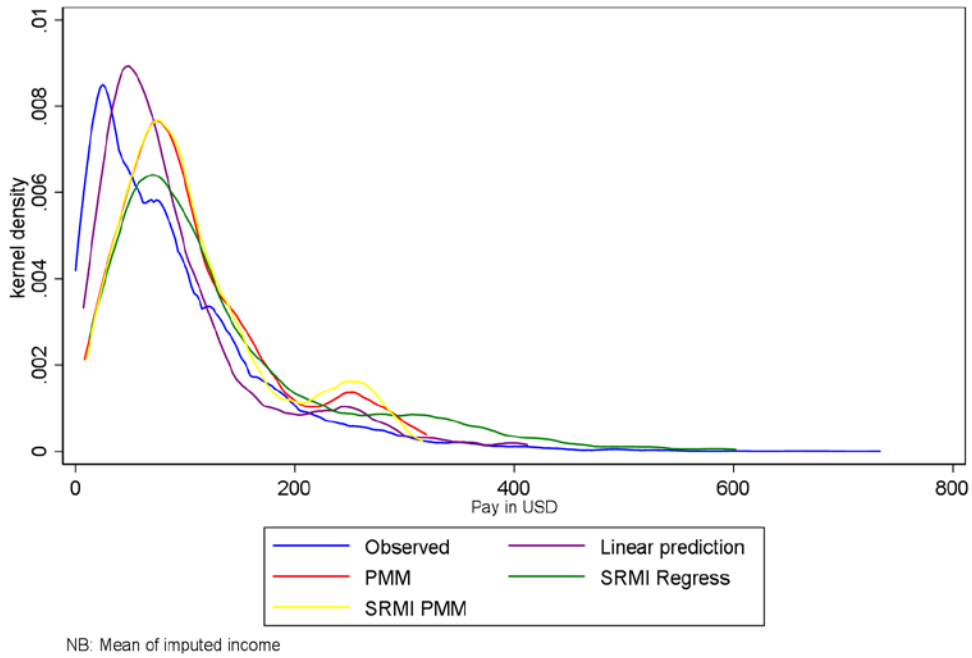
Results are presented for the four imputation techniques with reference to the imputed employment income (primary pay) data (Section 5.1); for the four imputation techniques that were applied to the artificial missing employment income datasets (Section 5.2); and with respect to the impact of the imputed data on simulated direct taxes in TAZMOD using the means of the imputations for each of the multiple imputation methods (Section 5.3), and also using each imputation separately for each of the multiple imputation methods (Section 5.4).

5.1 Impact of applying the four imputation techniques to the Tanzanian dataset

The results of the diagnostic tests were satisfactory. Examples are presented in the Technical Note and include scatterplots with LOWESS (locally weighted scatterplot smoothing) lines fitted and—for the two SRMI methods—trace plots.

Figure 3 provides an overview of the results by comparing the distribution of observed and imputed employment income for the four imputation methods that were applied to the data. For the three multiple imputation techniques the mean value of imputed income is used. For all four imputation methods, the imputed employment income has a similar distribution to the observed employment income.

Figure 3: Comparison of observed and imputed employment income distributions: Tanzania



Source: authors' calculations using HBS 2011/12.

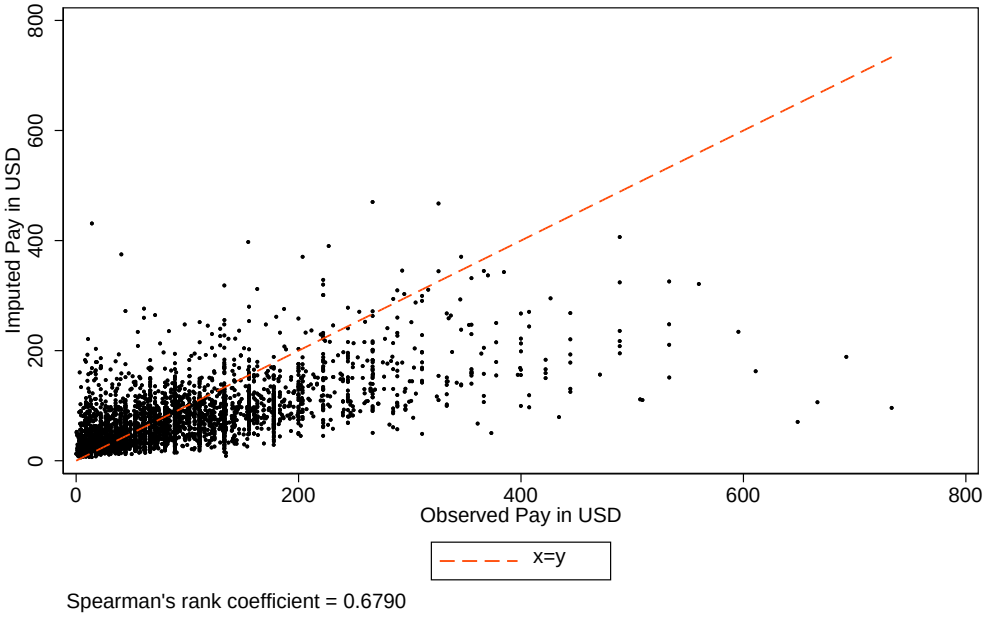
5.2 Impact of applying the imputation techniques to artificial missing income data in the Tanzanian dataset

This sub-section presents results for the more stringent tests, using artificial missing data (explained in Section 4.4). The following scatterplots show the imputed and observed income data

for the four methods: linear prediction (Figure 4); PMM (Figure 5); SRMI Regress (Figure 6); and SRMI PMM (Figure 7). The figures also provide the Spearman's rho correlations.

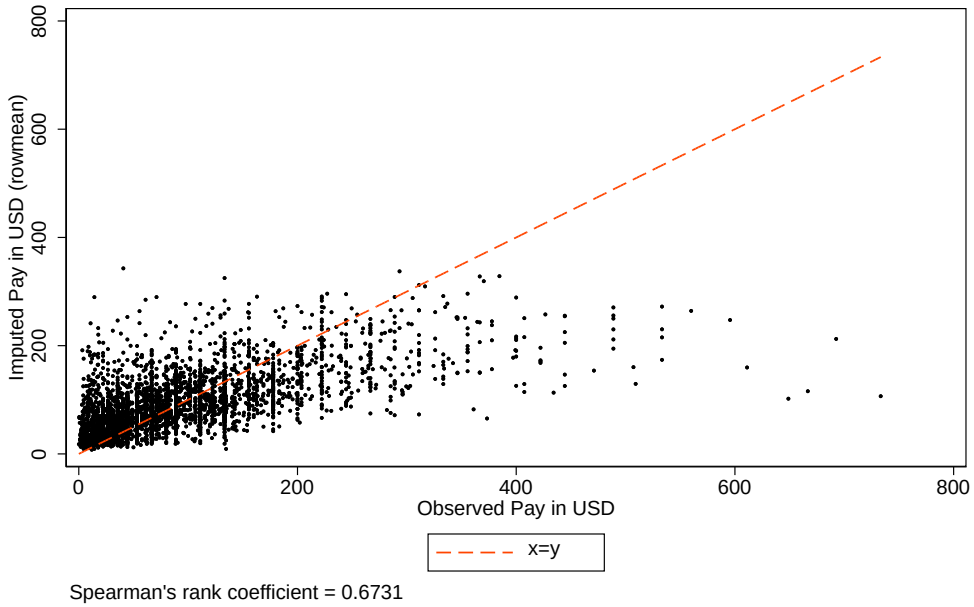
All four methods generate similar results, in that the correlations between imputed and observed are similar, and—with the exception of a small number of cases in Figure 6—the cases with the highest observed incomes have been imputed with much lower values. Of the three multiple imputation approaches, PMM and SRMI PMM produce very similar results. The SRMI Regress results correlate slightly less well, whereas the linear prediction model has a slightly higher correlation. Overall, for the Tanzanian dataset, there seems to be little difference between the four methods, even when applied to artificial missing data.

Figure 4: Scatterplot of imputed and observed employment income data using linear prediction: Tanzania



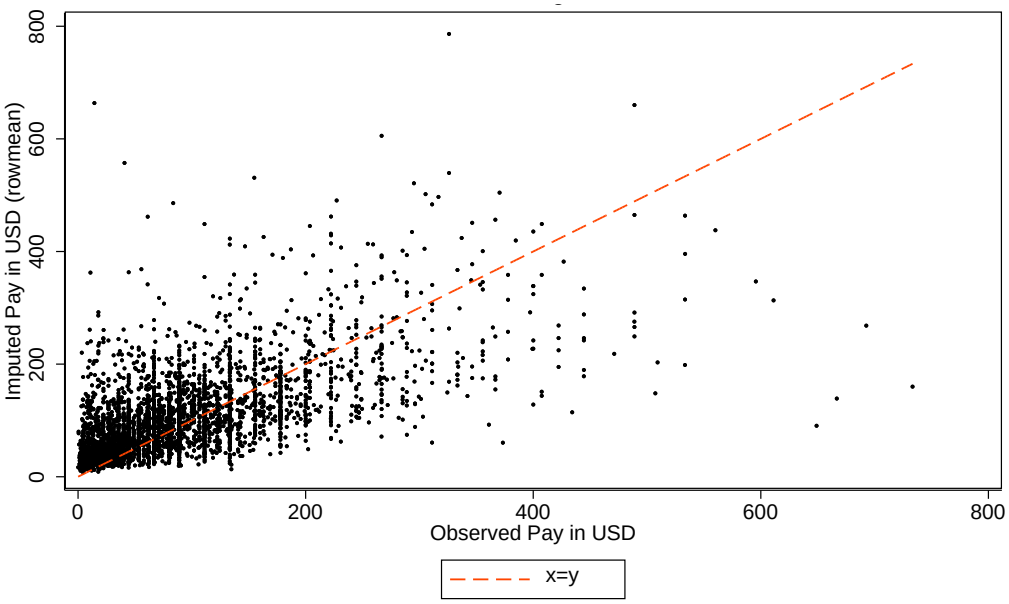
Source: authors' calculations using HBS 2011/12.

Figure 5: Scatterplot of imputed and observed employment income data using PMM: Tanzania



Source: authors' calculations using HBS 2011/12.

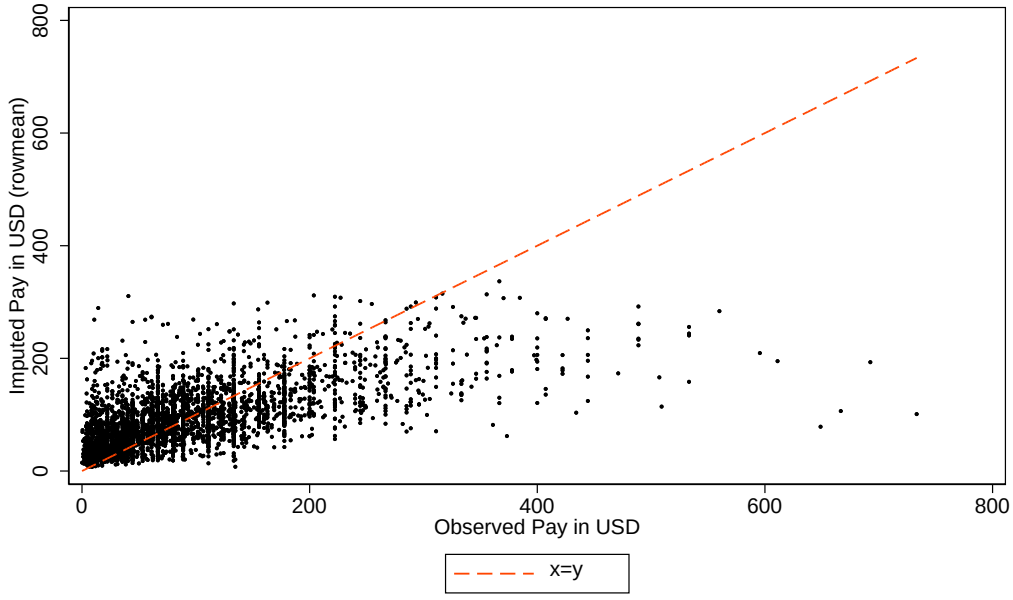
Figure 6: Scatterplot of imputed and observed employment income data using SRMI Regress: Tanzania



Spearman's rank coefficient = 0.6657

Source: authors' calculations using HBS 2011/12.

Figure 7: Scatterplot of imputed and observed employment income data using SRMI PMM: Tanzania



Spearman's rank coefficient = 0.6736

Source: authors' calculations using HBS 2011/12.

5.3 Impact of imputed income on TAZMOD's simulated direct taxes

As emphasized in the introduction, the reason for exploring the quality of income data and embarking on various imputation methods is to improve the quality of the input datasets that underpin the tax-benefit microsimulation models TAZMOD and MicroZAMOD. To illustrate the impact that the imputation processes have had on the input dataset for Tanzania, the results from each of the four imputation methods were incorporated into the underpinning dataset for TAZMOD and the tax-benefit system for 2015 was simulated.

Table 4 presents the results for direct taxes. Each of the imputation methods has resulted in the simulation of less direct tax, falling from 493 per cent of reported direct tax (before outlier adjustment), to 167 per cent (after outlier adjustment—see Section 2), to a low of 127 per cent (using the linear prediction method, or the SRMI PMM method). However, the differences between the datasets for which missing and implausible employment income data were imputed using linear prediction, SRMI PMM, and PMM are negligible.[19]

Table 4: Simulated direct taxes: Tanzania

Version of HBS dataset	1 Simulated direct taxes 2015 (TZS million)	2 Reported direct taxes 2015 (TZS million)	3 Percentage simulated (simulated/reported)
Before outlier adjustment[20]	11,751,885	2,382,952	493.2
After outlier adjustment	3,980,848	2,382,952	167.1
Imputed income—linear Prediction	3,030,183	2,382,952	127.2
Imputed income—PMM	3,040,163	2,382,952	127.6
Imputed income—SRMI Regress	3,088,225	2,382,952	129.6
Imputed income—SRMI PMM	3,035,923	2,382,952	127.4

Note: PIT restricted to those working in the formal sector. For each of the three multiple imputation approaches, the mean of the *M* imputations was used.

Source: authors, based on simulations using TAZMOD v1.8 and HBS 2011/12.

5.4 An alternative approach to calculating simulated direct taxes in Tanzania: running the microsimulation model multiple times

The results presented in Section 5.3 used new input datasets with imputed income values for those cases with missing or implausible reported values, whereby the mean of the *M* imputation results was calculated and used for each of the multiple imputation approaches. This approach has the advantage of allowing developers and users of the microsimulation models to underpin the models with a single input dataset and to simulate taxes and benefits (and calculate poverty and inequality indices) in the usual way.

However, an alternative method to utilizing the results from the three multiple imputation approaches is to run the microsimulation model separately for each of the *M* imputations, and then calculate the mean of the *M* simulated outputs, with accompanying confidence intervals.

Table 5 shows the results for Tanzania when the microsimulation model TAZMOD was run separately 50 times for each of the multiple imputation approaches. Results are presented for PIT. So, for example, the 50 imputed datasets that were generated using the PMM method were each merged back into TAZMOD's main underpinning dataset to generate 50 new input datasets for TAZMOD: each new input dataset contained imputed values for the missing and implausible employment income cases that had been generated by one imputation. The procedure for running TAZMOD multiple times efficiently is elaborated in the Technical Note.

[19] The computing power required for the different methods varied considerably for the Tanzanian data: SRMI PMM took a number of hours to run, whereas PMM took only minutes to run.

[20] This table presents total direct taxes and not just PIT as is presented in Table 1.

Table 5: An alternative approach to calculating simulated taxes in Tanzania using the results from the multiple imputation models

Imputation approach	1 Mean PIT (TZS million)	2 PIT standard error	3 PIT 95 per cent confidence interval (TZS million)		4 Direct tax (TZS million)
PMM	2,993,162	2,031	2,989,079	2,997,244	3,063,941
SRMI Regress	3,100,618	6,155	3,088,249	3,112,988	3,171,397
SRMI PMM	2,991,208	1,698	2,987,796	2,994,620	3,061,987

Note: column 1 is the mean amount of PIT simulated by TAZMOD using 50 input datasets, for each imputation approach shown. Column 4 is the sum of column 1 (PIT) plus presumptive tax (a constant) and can be compared with column 1 in Table 4.

Source: authors, based on simulations using TAZMOD v1.8 and different derivations of the HBS 2011/12.

The pattern of results presented in Table 5 is largely consistent with the pattern of results presented earlier, in that the simulated outputs from all three approaches are relatively similar in magnitude, with the two variants of PMM being very similar to one another and the outputs from SRMI Regress being slightly higher. The confidence intervals around the means are very tight for all three imputation approaches. The results from all three imputation approaches are higher than the external statistics shown in column 2 of Table 4, which is again consistent with the results presented earlier (column 1 of Table 4). The simulated results for direct tax presented in Table 4 (column 1) and Table 5 (column 4) are similar to one another, although those in Table 5 are slightly higher than those in Table 4 for all three multiple imputation approaches. This finding is due to the different ways in which the mean values are calculated for PIT: averaging the *M* imputed income values by person has the effect of dampening down the mean value for the persons due to the skewed nature of the income distribution in Tanzania, resulting in fewer higher-band taxpayers than might actually be the case. In contrast, producing *M* simulations from the *M* imputed datasets and then calculating the mean increases the likelihood of each constituent imputation containing more higher-band taxpayers, which then feeds through to the mean of the simulated taxes shown in Table 5 (column 1).

In summary, the alternative approach to simulating taxes produces similar patterns of results to the methodologies presented earlier for Tanzania, which are also presented for Zambia and South Africa below. Comparison of the two sets of results for Tanzania suggests that generating a single input dataset for microsimulation by assigning the mean of the *M* imputed values to each relevant person may result in slightly lower imputed incomes than might otherwise be expected, due to the skewed nature of the income distribution in Tanzania. However, overall, the results presented here show a reassuring degree of correspondence.

6 Summary of the Zambia findings

As highlighted in Section 2, the primary concern with the Zambia LCMS 2015 was the apparent under-enumeration of income, leading to a considerable under-simulation of direct taxes in MicroZAMOD. In this regard, the Zambian case provides a useful contrast to the Tanzanian case, where direct taxes were over-simulated due to an apparent over-enumeration of incomes reported in the Tanzania HBS 2011/12.

Similar steps for data cleaning, income validation, and identification of implausible incomes were undertaken for the LCMS dataset to those set out in Section 3 for the Tanzania HBS dataset, and again the focus was on employee income from primary employment, specifically 4,868 individuals

who reported their main economic activity status as 'in wage employment'. During the data preparation stage, it was identified that one-fifth (19.2 per cent) of the 4,868 relevant cases contained missing or implausible income data (see Appendix B for further details). Importantly, and unlike the Tanzanian case, it was identified that the Zambian employment income data had a bimodal distribution, with employees in the formal sector exhibiting a notably different distribution of income to employees in the informal sector (see Figure B1 in Appendix B).

The same four imputation methods that were applied to the Tanzanian data (described in Section 4) were also applied to the Zambian data: linear prediction, PMM, SRMI Regress, and SRMI PMM. Whereas in Tanzania it was necessary to deal with missing data on the covariates of occupational category and adult-equivalent consumption as well as employment income, in Zambia only the covariate of adult-equivalent consumption required imputing alongside employment income. For the linear prediction and PMM imputation approaches, missing values on adult-equivalent consumption were imputed using a simple hotdeck approach, as was also the case for the Tanzania dataset. For the two variants of SRMI, adult-equivalent consumption was imputed in the same way as employment income (using the Regress or PMM model form as appropriate).

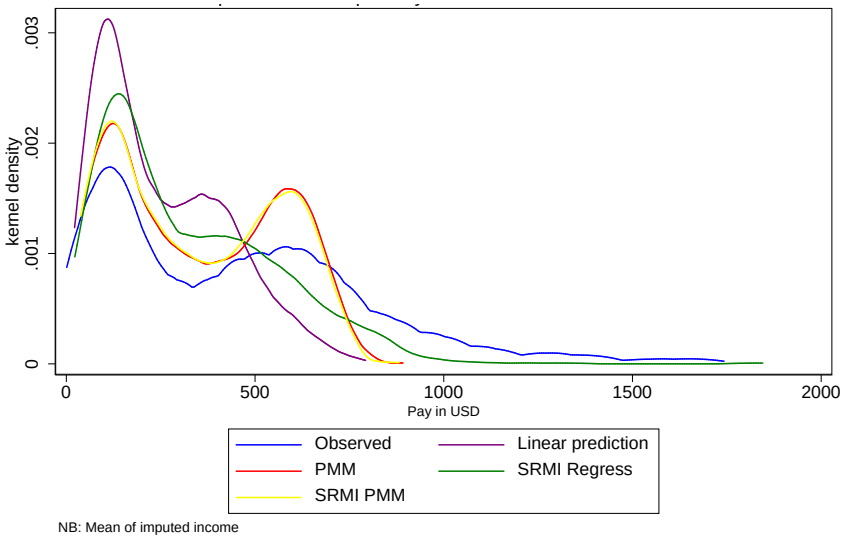
The models were specified in the same way as for the Tanzanian data. So, for the two variants of the PMM imputation approach (PMM and SRMI PMM), the models were specified to calculate the imputation value using five nearest neighbours. For the three multiple imputation approaches (PMM, SRMI Regress, and SRMI PMM), the models were specified to produce 50 imputations. For the two SRMI approaches, the models were specified to perform 100 iterations to produce each separate imputation. Upon completion of each multiple imputation method (PMM, SRMI Regress, and SRMI PMM), a final imputation result was derived for that method by calculating the mean of the 50 imputations produced by that method. This was not necessary for the linear prediction method as only a single imputation value was produced.

The initial imputation results—presented in Appendix C—were disconcerting as none of the four imputation methods produced a distribution of imputed values that corresponded closely to the distribution of observed values. Given the bimodal distribution of observed values and the important differential between the distributions of those in formal employment and those in informal employment noted in Appendix B, the imputation processes were therefore repeated separately by formal/informal employment status. The imputation results for the formal and informal groups were then appended together to reconstitute a full dataset (i.e. the 4,868 cases that reported their main economic activity status as being 'in wage employment').

Figure 8 presents the distributions of imputed employment income for each of the four methods when the respective models are specified separately for those in formal and informal employment. The resulting distributions from all four imputation approaches are clearly modified by specifying the models separately for formal/informal employment status, with the linear prediction and SRMI Regress approaches showing a closer correspondence with the distribution of observed values while the PMM and SRMI PMM approaches show an increased magnitude of bimodality.

The results presented in Figures 8 and A3.1 illustrate that, for Zambia, the choice of imputation method will have an important determining effect on the final set of imputed values. However, it is not possible to draw firm conclusions on the suitability of the imputation approaches solely by comparing the distributions of imputed values against the distribution of observed values. The distribution of unobserved employment income values may legitimately differ from the distribution of observed values and, indeed, might be somewhat expected under the assumptions of the MAR missing data mechanism.

Figure 8: Comparison of observed and imputed employment income distributions, with each imputation approach run separately for formal and informal employment status: Zambia



Source: authors' calculations using LCMS 2015.

In order to explore the four methods further, they were applied to artificially missing income data, and again the models were run separately for formal and informal employees. The results are presented in Appendix D, but in summary the scatterplots show moderately strong positive relationships between observed pay and imputed pay for all four imputation approaches.

Lastly, in order to illustrate the impact that the imputation processes have had on the input dataset for Zambia, the results from each of the four imputation methods were incorporated into the underpinning dataset for MicroZAMOD and the tax-benefit system simulated accordingly. Table 6 presents the results. Each of the imputation methods has resulted in the simulation of slightly more direct tax, rising from 31.3 per cent in the initial version of MicroZAMOD's input dataset, to a high of 33.4 per cent using the dataset that contains cleaned income data using the PMM method.

Table 6: Simulated direct taxes: Zambia

Version of LCMS dataset	1 Simulated direct taxes 2015 (ZMW million)	2 Reported direct taxes 2015 (ZMW million)	3 Percentage simulated (simulated/reported)
Initial version of input dataset	3,129	10,005	31.3
Imputed income—linear prediction	3,195	10,005	31.9
Imputed income—PMM	3,346	10,005	33.4
Imputed income—SRMI Regress	3,328	10,005	33.3
Imputed income—SRMI PMM	3,332	10,005	33.3

Note: the amounts in column 1 comprise PIT plus turnover tax. For each of the three multiple imputation approaches, the mean of the *M* imputations was used. The amounts in column 2 comprise PIT, turnover tax, and withholding tax.

Source: authors, based on (column A) simulations using MicroZAMOD V2.0 and (column B) Ministry of Finance (2016: 28, 30).

It is evident that the application of the imputation techniques to the Zambian data results in relatively small increases in the amount of simulated direct taxes. This is discussed further in Section 8.

7 An application of PMM using a South African dataset and artificial missing income data

In this penultimate section, the PMM approach that was applied to the Tanzanian and Zambian income data is tested on a third dataset: the South African National Income Dynamics Study (NIDS) Wave 4 v1.1 (Chinhema et al. 2016; SALDRU 2016). NIDS is administered by the Southern Africa Labour and Development Research Unit (SALDRU) at the University of Cape Town, and is one of two datasets that underpin the South African tax-benefit microsimulation model SAMOD. NIDS Wave 4 was conducted in 2014, and the income and expenditure data have been deflated to a June 2014 time-point. Survey data on income have been used more extensively in South Africa than in Tanzania and Zambia, and the NIDS income data have received particular attention. The NIDS income data were further cleaned by the authors when preparing the data for inclusion as one of the underpinning datasets for SAMOD, and it was found to perform well as an underpinning dataset for SAMOD when compared to external validation data including tax statistics from the South African Revenue Service (Wright et al. 2016).

For this reason, it was decided that the income data in NIDS, specifically income from paid employment, could be used as an exacting test of the income imputation methods used on the Tanzanian and Zambian income data. The PMM method was selected as this yielded results that were as persuasive as the other multiple imputation methods in the Tanzanian case but required considerably less computational time to run.

The first step was to fit a series of OLS models to ascertain suitable covariates. The selected covariates were similar to those used in Tanzania and Zambia, but with the addition of population group. Some of the covariates had very small numbers of missing cases and these were first imputed using a hotdeck technique.

In order to test the PMM method, artificial missing income data were introduced to the NIDS dataset, in the same way as described for Tanzania in Section 4.4. Ten separate files were created, each containing 10 per cent missing data for the income from employment[21] variable, based on the decile groupings of the assigned random numbers. The PMM imputation technique was then applied to each of the 10 separate files. The observations containing imputed employment income from each of the files were then extracted and appended so that a complete file was created in which all the cases had imputed employment income data. Having generated a file containing 50 imputed values for employment income, these were averaged in order to be compared to the original (observed) employment income data.

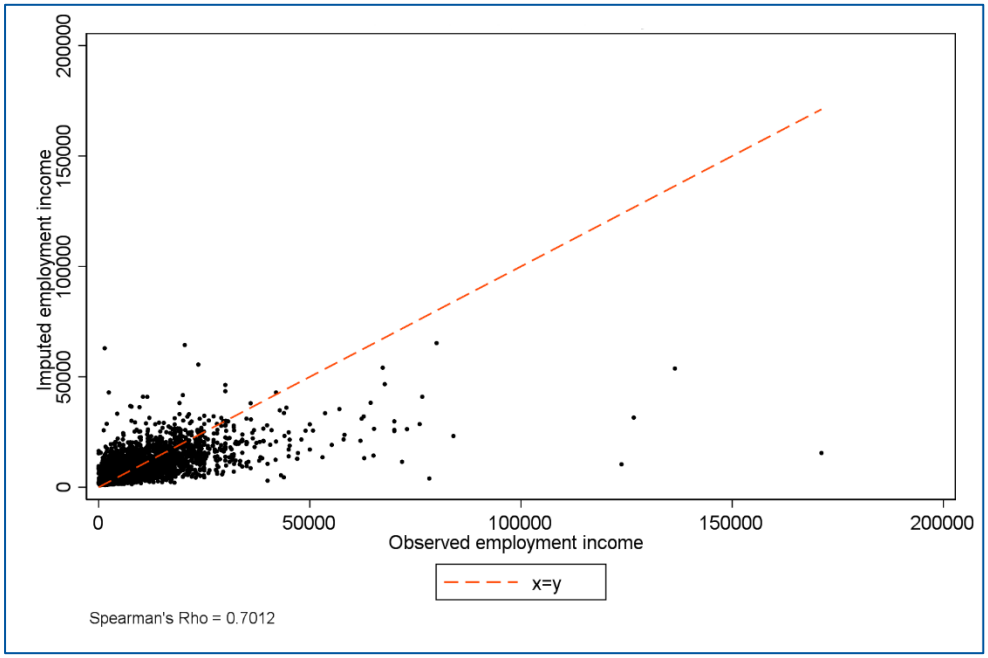
Figure 9 shows a scatter plot comparing observed employment income and imputed employment income. There is a reasonable correlation between observed and imputed employment income (0.701), but it is notable that the highest observed incomes have all been imputed with much lower values and, as will be demonstrated, this affects the amount of tax simulated. It is inherent in the method that the highest observed incomes will generate lower imputed values as the maximum

[21] In NIDS there is no distinction between primary and secondary pay and therefore gross income from employment was used.

possible imputed value for any given case is the value of the highest observed value, and so when taking a mean over 50 imputations the final imputed amount will almost inevitably be lower.

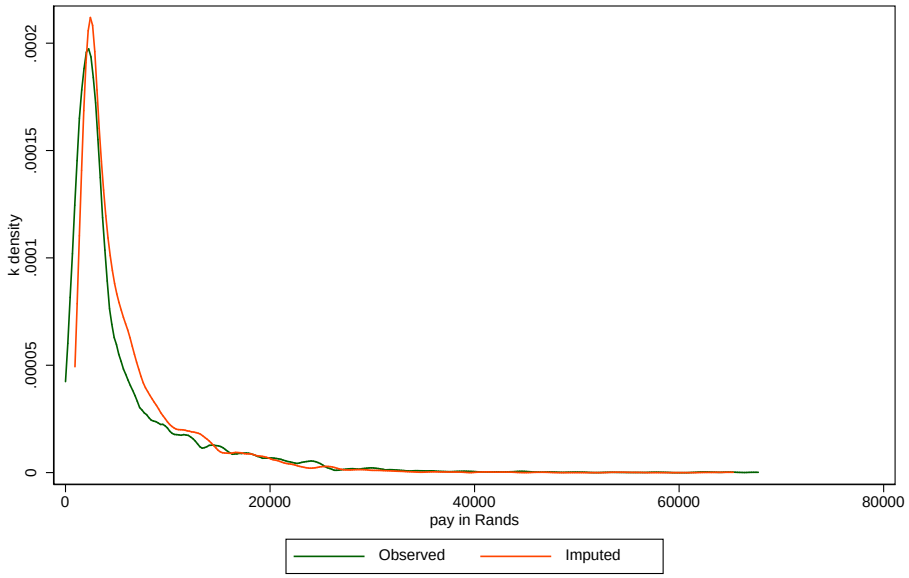
Figure 10 presents kernel density plots of the distributions of the observed and imputed employment income, truncating the observed income at R100,000 per month. As can be seen, the distribution of the imputed employment income is very similar to the distribution of the observed employment income, suggesting that the method has worked well.

Figure 9: Scatterplot of observed and imputed monthly income from employment: South Africa



Source: authors' calculations using NIDS Wave 4 v1.1.

Figure 10: Kernel density plots of observed and imputed monthly income from employment: South Africa



Source: authors' calculations using NIDS Wave 4 v1.1.

Finally, Table 7 compares simulated tax and benefit results from SAMOD v6.6 using the NIDS dataset with the observed employment income data (column 1), and using a dataset containing the imputed employment income data (column 2). Overall, the simulations using the dataset with wholly imputed employment income data yield 77.5 per cent of the amount of direct taxes when compared to simulations using the original dataset. This is a stringent test for the imputation method and the findings are encouraging. It is also notable that the simulations of social transfer expenditure—most of which are means-tested—are largely unaffected by the use of wholly imputed data on employment income.

Table 7: Simulated taxes and benefits using original and imputed employment income, 2014: South Africa

	1 Using original employment income (R million)	2 Using imputed employment income (R million)	3 Percentage change
Total annual government revenue through direct taxes and social insurance contributions, of which:	230,620	178,798	77.5
direct taxes	217,126	164,492	75.8
social insurance contributions (employee and employer)	13,475	14,306	106.1
Total annual government expenditure on social transfers, of which:	145,377	144,000	99.1
child benefits	64,974	63,717	98.1
disability benefits	20,232	20,004	98.9
pension benefits	60,170	60,279	100.2

Notes: imputed employment income obtained using PMM.

Source: authors' calculations using SAMOD v6.6 with NIDS Wave 4 v1.1.

8 Conclusions and recommendations

In this paper, findings have been presented from an exploration of the quality of the income data in the Tanzania HBS 2011/12 and the Zambia LCMS 2015 datasets. A single income variable was used as a case study—income from primary employment.[22] The quality of this variable in the underpinning datasets of the Tanzania and Zambia tax-benefit microsimulation models, TAZMOD and MicroZAMOD respectively, is likely to be one of the main contributors to the outcome of simulations of PIT. It was demonstrated in Section 2 that the two datasets present contrasting challenges, with the over-simulation of direct taxes in Tanzania and under-simulation of direct taxes in Zambia.

A series of steps were applied to each dataset as part of the process of data cleaning, income validation, and identification of implausible incomes. An account of these steps is included for Tanzania in Section 3 to demonstrate the extensive data exploration and preparation procedures

[22] Other income types were considered but on preliminary exploration no suitable covariates could be found for the imputation models.

that were necessary before being able to consider any imputation of missing data. A summary of these steps is included for the Zambia dataset in Appendix B.

Having identified missing and implausible cases in both datasets for the variable of interest, four different imputation methods were applied: linear prediction, PMM, SRMI Regress, and SRMI PMM. The imputation models were described and results for the Tanzanian dataset included in Sections 4 and 5 respectively, and summarized for Zambia in Section 6. In the case of Zambia, the models were run separately by formal/informal employment status due to a bimodal distribution that was driven by formality (Figure B1) which was not apparent in the Tanzanian data.

As a further and more stringent set of tests, the imputation models were also applied to the two country datasets having introduced artificial missing data, in order to enable a comparison of imputed and observed cases. The PMM method was also applied to artificial missing data in a South African dataset that contains income data that has been much more extensively interrogated and so can be regarded as a Southern African 'gold standard' dataset for income data. The method was found to work well in this context (Section 7).

Having run various diagnostic tests, the imputed income data was incorporated into the underpinning datasets of the two microsimulation models, TAZMOD and MicroZAMOD, and it was identified that the imputation process improved the simulated taxes for both countries. Less tax was simulated in the case of Tanzania, and more tax was simulated in the case of Zambia. Further details of the analysis are included in the Appendices and in the accompanying Technical Note.

The results and diagnostic tests presented here and in the Technical Note demonstrate that, in general, the four imputation approaches do tend to produce similar outputs to one another. However, the results are not identical across the four approaches, for either Tanzania or Zambia. Furthermore, it has been shown that there are differences between Tanzania and Zambia in the patterning of the outputs across the four imputation approaches. It is therefore not possible to state unequivocally that any one of the four imputation methods is best in all scenarios. Future applications of the income data cleaning and imputation processes to other country datasets should therefore involve a careful review of the results and diagnostic tests in order to select the most appropriate method for the country and dataset in question.

The analysis presented here only 'scratches the surface' in terms of interrogating the quality of the income data in the Tanzania and Zambia datasets, and there are many ways in which the work could be developed further. For example, the types of missing data explored in this paper only pertain to item missing data (values that are either missing or implausible): the issue of unit missing data has not been explored here. Given the fact that MicroZAMOD simulates such a low proportion of reported taxes even after imputation (though noting that the external validation data include withholding taxes which are not simulated), it seems likely that there is a problem of unit missing data in Zambia. That is, the Zambia dataset probably does not contain a sufficient number of high earners.

As mentioned at the outset, it is important to have confidence in the quality of the income data that underpin tax-benefit microsimulation models. It is hoped that the findings in this paper will assist researchers working on these and other country datasets that contain income data that have so far been largely unexplored. However, it is an unavoidable fact that each dataset will have its unique challenges, and so cleaning of income data is not a process that can simply be replicated without careful consideration of the challenges of each particular dataset. That having been said, it is encouraging to see that the imputation methods explored seem to have assisted in the process of cleaning two very different datasets that presented distinct challenges. It has also been

demonstrated that the imputation of missing and implausible data in Tanzania yielded better results than manual cleaning alone.

Lastly, the analysis presented here presumes both that the external validation data are accurate, and (much less plausibly) that there is full compliance by taxpayers. The tax-benefit microsimulation models TAZMOD and MicroZAMOD currently incorporate the assumption that the tax rules are precisely adhered to by individuals, and it is an underlying premise of the analysis presented in this paper that by cleaning the income data the simulations of PIT should become more closely aligned with the external validation data on taxes. It is only possible to conclude that the income data have been improved by imputation if these assumptions are correct. Further information on the extent of compliance in each country is needed in order to unpick whether, for example, an over-simulation of PIT reflects unclean data, or simply poor compliance. Without this information, one runs the risk of cleaning missing data, but missing the bigger picture.

References

Ardington, C., D. Lam, M. Leibbrandt, and M. Welch (2005). 'The Sensitivity of Estimates of Post-Apartheid Changes in South African Poverty and Inequality to Key Data Imputations'. CSSR Working Paper 106. Cape Town: Centre for Social Science Research, University of Cape Town.

Azur, M.J., E.A. Stuart, C. Frangakis, and P.J. Leaf (2011). 'Multiple Imputation by Chained Equations: What Is It and How Does It work'. *International Journal of Methods in Psychiatric Research,* 20(1): 40–49. https://doi.org/10.1002/mpr.329

Beegle, K., C. Carletto, B. Davis, and A. Zezza (2015). 'Households and Income in Africa'. In C. Monga and J.Y. Lin (eds), *The Oxford Handbook of Africa and Economics: Volume 1: Context and Concepts*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199687114.013.22

Chinhema, M., T. Brophy, M. Brown, M. Leibbrandt, C. Mlatsheni, and I. Woolard (eds) (2016). *National Income Dynamics Study Panel User Manual.* Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

CSO (2012). *Living Conditions Monitoring Survey Report 2006 and 2010*. Lusaka: Living Conditions Monitoring Branch, Zambia Central Statistical Office (CSO).

CSO (2016). *2015 Living Conditions Monitoring Survey (LCMS) Report*. Lusaka: Zambia Central Statistical Office (CSO).

Decoster, D., J. Pirttilii, G. Wright, and H. Sutherland (forthcoming). 'SOUTHMOD: Simulating Taxes and Social Protection for Development: An Introduction'. *International Journal of Microsimulation*.

Ferreira, F.H.G., S. Chen, A. Dabalen, Y. Dikhanov, N. Hamadeh, D. Jolliffe, A. Narayan, E.B. Prydz, A. Revenga, P. Sangraula, U. Serajuddin, and N. Yoshida (2016). 'A Global Count of the Extreme Poor in 2012: Data Issues, Methodology and Initial Results.' *Journal of Economic Inequality*, 14(2): 141–72. https://doi.org/10.1007/s10888-016-9326-6

Graham, J. W. (2009). 'Missing Data Analysis: Making it Work in the Real World'. *Annual Review of Psychology*, 60: 549–76. https://doi.org/10.1146/annurev.psych.58.110405.085530

Groves, R.M., F.J. Fowler, M.P. Couper, M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Heeringa, S.G., B.T. West, and P.A. Berglund (2017). *Applied Survey Data Analysis*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Lacerda, M., C. Ardington, and M. Leibbrandt (2007). *Sequential Regression Multiple Imputation for Incomplete Multivariate Data Using Markov Chain Monte Carlo*. DataFirst Technical Paper 5. Cape Town: DataFirst, University of Cape Town.

Leyaro, V., E. Kisanga, M. Noble, G. Wright, and D. McLennan (2017). 'UNU-WIDER SOUTHMOD Country Report: TAZMOD v1.0, 2012, 2015'. Helsinki: UNU-WIDER.

Little, R.J., and D.B. Rubin (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley & Sons. https://doi.org/10.1002/9781119013563

Ministry of Finance (2016). *2015 Annual Economic Report*. Lusaka: Ministry of Finance, Republic of Zambia.

Nakamba-Kabaso, P., S. Nalishebo, D. McLennan, M. Kangasniemi, M. Noble, and G. Wright (2017). 'UNU-WIDER SOUTHMOD Country Report: MicroZAMOD v1.0, 2015'. Helsinki: UNU-WIDER.

NBS (2014a). *Tanzania Household Budget Survey: Main Report 2011/12*. Dar es Salaam: Tanzania National Bureau of Statistics (NBS).

NBS (2014b). *Tanzania Household Budget Survey: Technical Report 2011/12*. Dar es Salaam: Tanzania National Bureau of Statistics (NBS).

Raghunathan, T., J. Lepkowski, J. van Hoewyk, and P. Solenberger (2001). 'A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models'. *Survey Methodology*, 27(1): 85–95.

Rubin, D.B. (1986). 'Basic Ideas of Multiple Imputation for Nonresponse'. *Survey Methodology*, 12(1): 37–47.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rubin, D.B. (1996). 'Multiple Imputation after 18+ Years'. *Journal of the American Statistical Association,* 91: 473–489. https://doi.org/10.1080/01621459.1996.10476908

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

StataCorp (2017). *Stata Multiple-Imputation Reference Manual Release 15*. College Station, TX: StataCorp LLC.

Sutherland, H., and F. Figari (2013). 'EUROMOD: The European Union Tax-Benefit Microsimulation Model'. *International Journal of Microsimulation*, 6(1): 4–26. https://doi.org/10.34196/ijm.00075

University of Essex (2017). EUROMOD software v2.0.5. August 2017.

Wright, G., M. Noble, H. Barnes, D. McLennan, and M. Mpike (2016). 'SAMOD, a South African Tax-Benefit Microsimulation Model: Recent Developments'. UNU-WIDER Working Paper 2016/115. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2016/159-8

Appendix A: Variables entered into the SRMI models—Tanzania

Table A1: Variable names, descriptions, and categories

Variable	Description	Categories
dgn	Gender	0 = Female 1 = Male
dag _LS	Age in years, logged and mean-centred	n/a
deh2	Highest education completed	0 = Not completed primary education 1 = Completed primary education 2 = Completed secondary education 3 = Completed tertiary education 9 = Currently in education
loc	Occupation category	1 = Senior officials and managers 2 = Professionals 3 = Technicians and associate professions 4 = Clerks 5 = Service and sales workers 6 = Skilled agricultural 7 = Craft and trades workers 8 = Plant and machine operators 9 = Elementary occupations
urban_rural	Urban/rural/Dar es Salaam code	1 = Urban 2 = Rural 3 = Dar es Salaam
ppr_LS	Persons per room, logged and mean-centred	n/a
good_floor	Binary flag to indicate decent floor material in house	0 = No 1 = Yes
good_walls	Binary flag to indicate decent wall material in house	0 = No 1 = Yes
good_roof	Binary flag to indicate decent roof material in house	0 = No 1 = Yes
good_light	Binary flag to indicate decent source of energy for lighting (electricity or solar)	0 = No 1 = Yes
good_cook	Binary flag to indicate decent source of energy for cooking (electricity or gas)	0 = No 1 = Yes
good_toilet	Binary flag to indicate decent toilet (flush, ventilated pit, compost/eco)	0 = No 1 = Yes
shared_toilet	Binary flag to indicate whether household shares a toilet with others	0 = No 1 = Yes
good_wash	Binary flag to indicate whether household has a place to wash hands with soap and water	0 = No 1 = Yes
piped_water	Binary flag to indicate access to piped water (to house, within yard, or to public pipe within 500 m of house)	0 = No 1 = Yes
aec_LS	Adult-equivalent household monthly consumption, logged and mean-centred	n/a
pay_LS	Primary pay per month, logged and mean-centred	n/a

Source: authors.

Appendix B: Summary of data cleaning, income validation, and identification of implausible incomes—Zambia

The data preparation stage for the Zambia LCMS dataset is described here, and was very similar to that documented for the Tanzania HBS dataset in Section 3.

The analysis focused on individuals aged 15 and over who reported their main current economic activity status as being 'in wage employment'.[23] There were 4,868 such cases in the LCMS 2015 dataset. These cases should, in theory, report an income value for the question 'How much is your regular gross monthly salary/wage including regular allowances and transport allowances, regular overtime, retention allowance, from the main job?'.[24] However, as noted in Section 2, sizeable proportions of this sub-group either reported zero income or had missing data for this question.

Of the 4,868 cases that met the inclusion criteria, 3,935 reported a positive employment income, 408 cases reported a zero value for employment income, and 525 cases had a missing value for employment income. As such, almost one-fifth of the cases within this subset did not report a positive monthly income from their main job, despite stating that their main economic activity status was 'in wage employment'. The 408 cases reporting a zero income were set to missing due to a zero income being regarded as implausible, resulting in a total of 933 cases requiring imputation.

A set of covariates were identified as being potential predictors of employment income in Zambia. In terms of demographic and labour market variables, the following were selected: age; sex; highest level of education attained; occupational category; a formal/informal employment indicator; and an indicator of whether the individual suffered from a chronic illness. In addition to these personal characteristics, a number of household and locational variables were also selected as covariates. The two locational variables were: rural/urban; and an indicator of whether the household had lived in the same district for 12 months or more. Nine material asset covariates were selected: ownership of bed; mattress; table; sofa; television; computer; clock; and electric iron. Seven covariates relating to the quality of housing were selected: number of persons per room; type of dwelling; material of roof; material of floor; source of energy for lighting; source of energy for cooking; and type of toilet. Finally, adult-equivalent household consumption was included as a covariate.

Each of the selected covariates was subjected to manual cleaning by assessing internal consistency between relevant variables and also assessing distributions of responses within variables. At the end of the manual cleaning phase, all covariates were fully coded, with no missing values.

The next stage of the data preparation entailed checking for outlier cases in the employment income and adult-equivalent consumption variables. The distribution of employment income was judged to be plausible and so it was concluded that no adjustments should be made for outliers (unlike in Tanzania). The distribution of adult-equivalent consumption was assessed for each of the constituent categories of occupational classification; formal/informal employment status; and highest level of education attained. Where the reported consumption value was greater than the 99th percentile value or less than the 1st percentile value, the reported value was deemed to be implausible and set to missing.

---

[23] LCMS (2015: section 5, question 1).

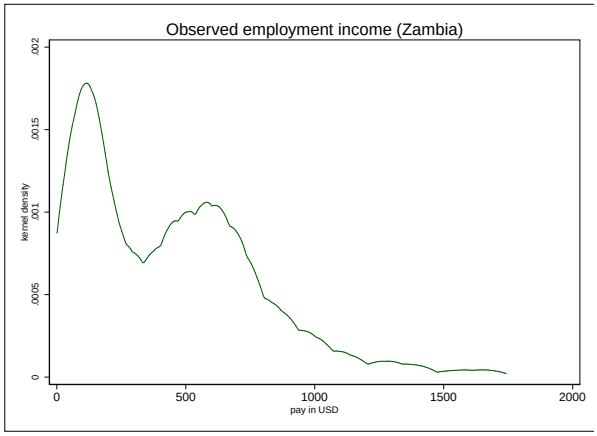
[24] LCMS (2015: section 6, question 27).

With regards to the adult-equivalent consumption value, all 4,868 cases reported a value,[25] and of these cases 82 were deemed to have an implausibly high consumption value and were therefore set to missing, while a further 82 cases were deemed to have an implausibly low consumption value and were therefore also set to missing. In total, therefore, 164 cases were set to missing on adult-equivalent consumption, representing 3.4 per cent of the total cases that stated their main economic activity status as 'in wage employment'. These 164 cases required imputation.

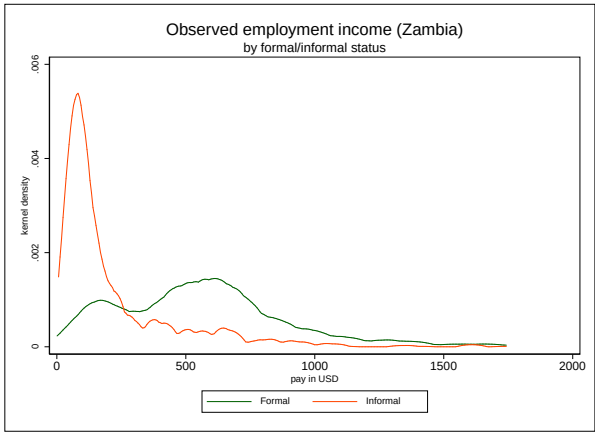
It was identified that the employment income data had a bimodal distribution and this was explored further to inform the imputation stage. Figure B1 shows two kernel density plots relating to employment income. In Panel A, the distribution of observed values is shown for employment income across the 3,935 cases that did report a positive income value in the survey and it is clearly evident that the distribution is bimodal. Further analysis revealed notable differences in the distributions of reported employment income between those classed as being employed in the formal sector and those classed as being employed in the informal sector, and this is presented in Panel B. As one might expect, those employed in the informal sector typically earn considerably less than those employed in the formal sector, although there is a degree of overlap between the distributions at the lower end of the income scale. This provided important context for the analyses of imputation methods summarized in Section 6.

Figure B1: Distribution of employment income

A: Employment income



B: Employment income by formal/informal status



Source: authors' calculations using LCMS 2015.

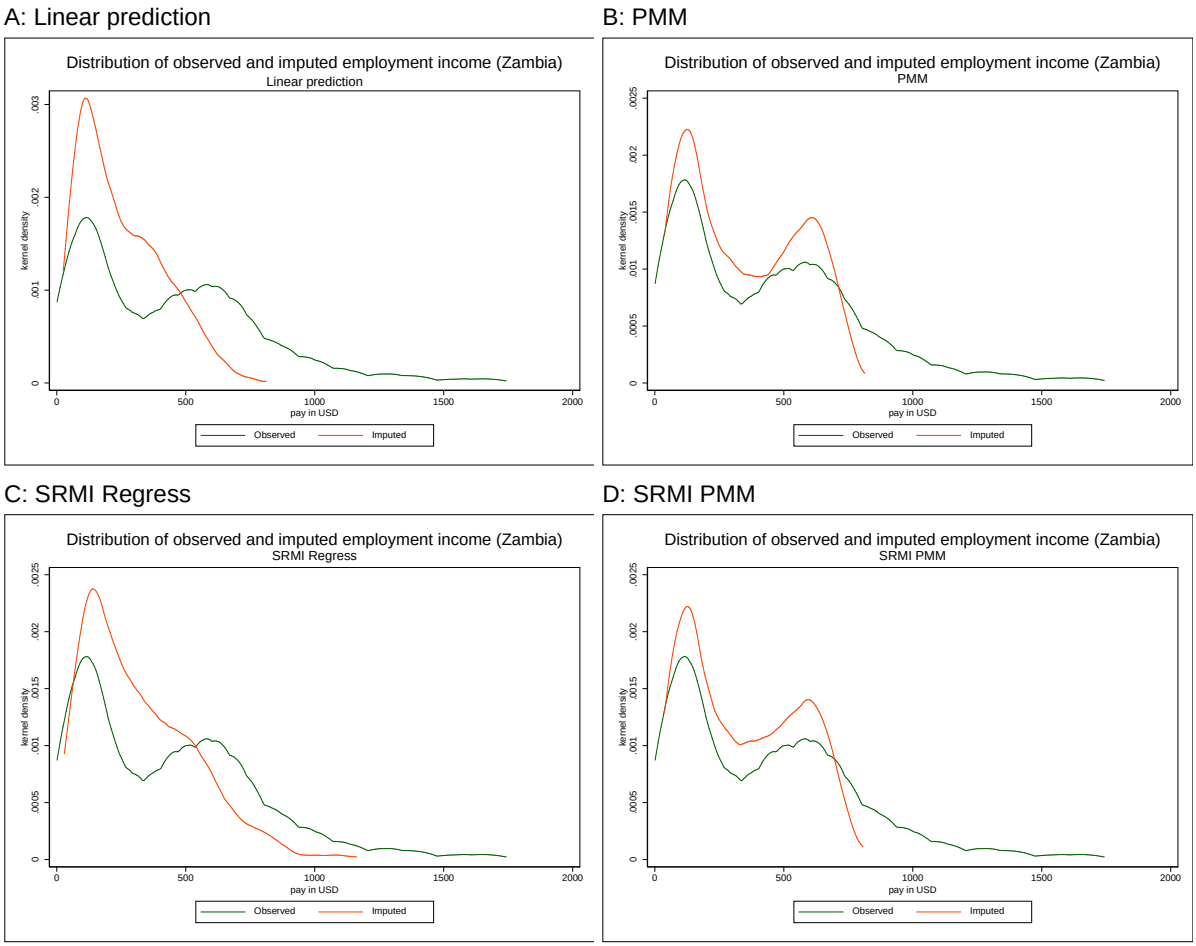
Appendix C: Initial imputation results—Zambia: before the imputation models were run separately for formal/informal employees

The initial results of the four imputation approaches—before the imputation models were run separately for formal/informal employees—are presented in Figure C1. The panels show the initial results for the linear prediction method (Panel A); the PMM method (Panel B); the SRMI Regress method (Panel C); and the SRMI PMM method (Panel D). In each of the panels, the green line shows the distribution of observed values reported by the 3,935 cases that did not require imputation (and, as such, is the same distribution as shown in Panel A of Figure B1), while the red

[25] One case reported a zero value but this was manually recoded to equal the lowest positive reported consumption value, prior to the outlier identification process being undertaken.

line shows the distribution of imputed employment income values for the 993 cases that did require imputation.

Figure C1: Comparison of observed and imputed employment income distributions



Source: authors' calculations using LCMS 2015.

It is evident that through the initial run of the models—before they were run separately for formal/informal employees—none of the four imputation methods produced a distribution of imputed values that corresponded closely to the distribution of observed values. Furthermore, while the two variants of the PMM approach produced relatively consistent distributions (as one might expect), and while the linear prediction and SRMI Regress approaches produced relatively consistent distributions (again, as one might expect), there is a notable difference between the PMM and non-PMM approaches. The two PMM approaches produced bimodal imputed values that have a degree of correspondence with the distribution of observed values, whereas the linear prediction and SRMI Regress approaches are clearly unimodal. However, the upper end of the distribution of imputed values on the two PMM approaches is seriously curtailed compared to the observed values, whereas the SRMI Regress approach generates a range of imputed values that corresponds very closely to the range of observed values.

In response to these initial findings the imputation process was repeated separately for those in formal employment and those in informal employment, and these results are presented in Section 6.

Appendix D: An application of the imputation methods to artificial missing data—Zambia

To further validate the results of the Zambia imputations, the four imputation methods were tested on artificial missing income data. The 3,935 cases with observed employment income were randomly allocated into one of 10 equally sized subsets; then the technique of artificially setting each subset's employment income to missing and re-running the imputation models was performed, as described for Tanzania in the main text. In a slight modification to the Tanzanian analysis, for Zambia the models were run separately for formal and informal employment status for each of the 10 equally sized subsets of observed cases.

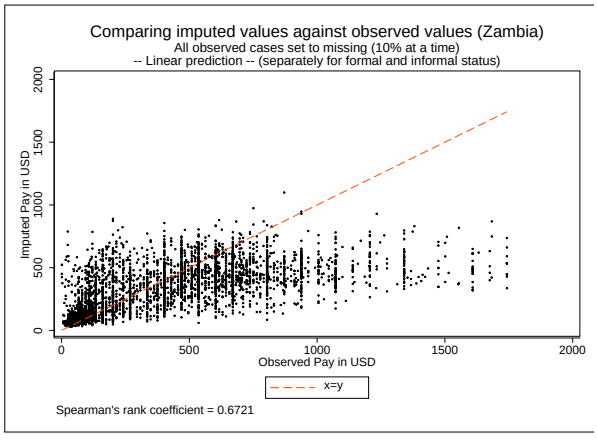
The process of setting observed cases to be artificially missing enables the resulting imputed values to be compared against the actual observed values for each of the 3,935 cases with observed values. Adopting the same approach as was presented in relation to Tanzania in Section 5.2, Figure D1 shows the scatterplots of observed versus imputed values for each of the four imputation methods applied in Zambia. These scatterplots should be read in conjunction with the kernel density plot provided in Figure D2 in which the distributions of imputed values based on artificial missing data are compared against the distribution of observed values.

The scatterplots show moderately strong positive relationships between observed pay and imputed pay for all four imputation approaches. The strong bimodal nature of the distributions generated through the two variants of the PMM approach is apparent in the scatterplots. It is also apparent that the two PMM approaches do not produce the same range of imputed values that are seen in the linear prediction and SRMI Regress approaches.

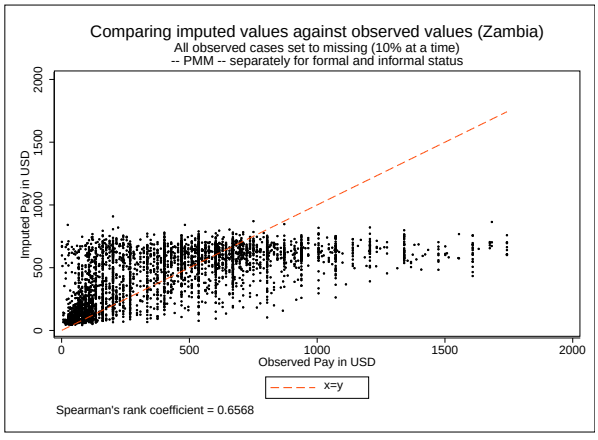
The Spearman rank coefficient provides a useful summary statistic to help qualify the relationship between imputed values and observed values. As is evident from Figure D1, the imputation method that produced the highest rho value is the linear prediction model at 0.6721. However, all four imputation methods produce similar rho values, ranging from a high of 0.6721 to a low of 0.6568.

Figure D1: Scatterplots of observed versus imputed values with observed cases artificially set to missing prior to imputation—Zambia

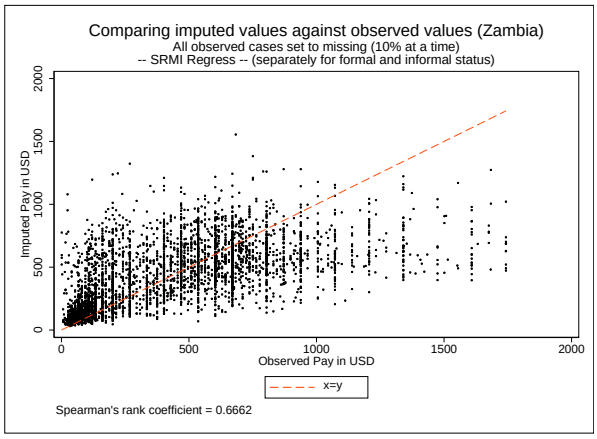
A: Linear prediction



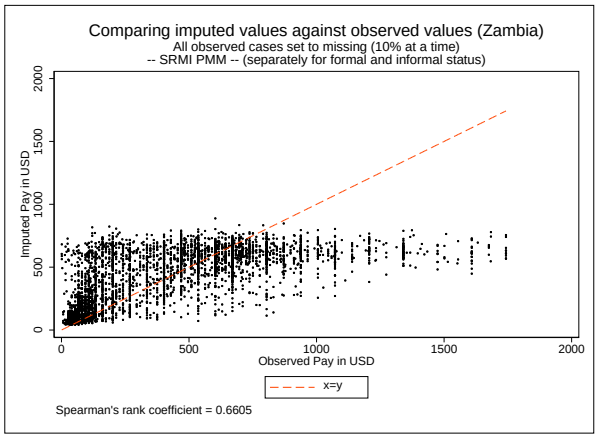
B: PMM



C: SRMI Regress



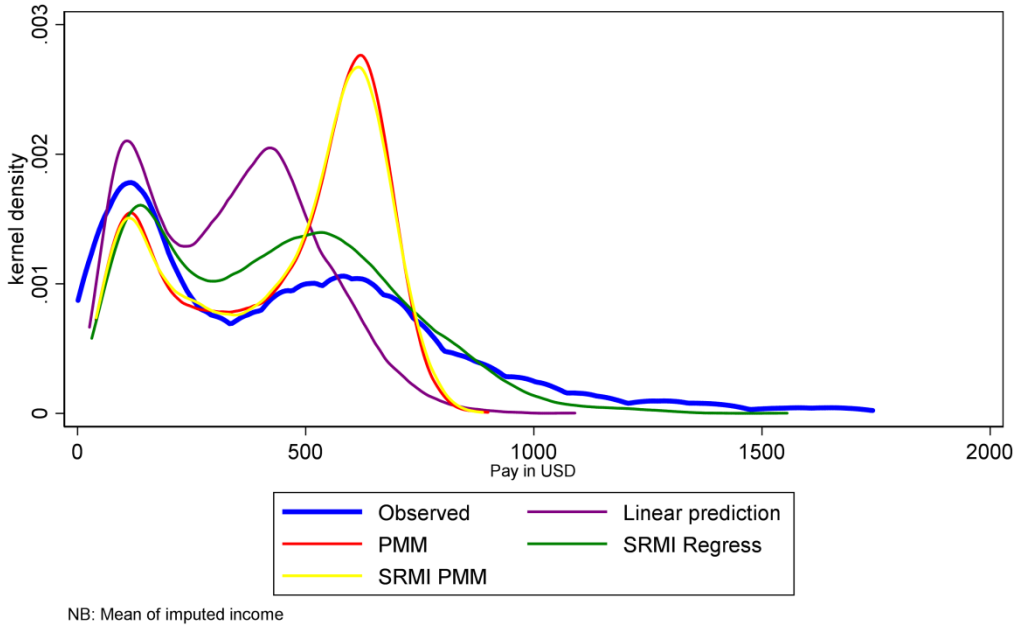
D: SRMI PMM



Source: authors' calculations using LCMS 2015.

These findings from the scatterplots are also evident in the kernel density plot presented in Figure D2, where the two PMM approaches can be seen to produce very similar results, with large concentrations of imputed values in the lower, and (to an even greater extent) upper-middle sections of the observed distribution, while the SRMI Regress and linear prediction values stretch further along the *x*-axis than the PMM variants.

Figure D2: Distributions of imputed values and observed values, having set cases artificially to missing prior to imputation—Zambia



Source: authors' calculations using LCMS 2015.

Appendix E: List of acronyms

CSO	Central Statistical Office, Republic of Zambia
EUROMOD	Tax-benefit microsimulation model for the European Union
HBS	Household Budget Survey (Tanzania)
LCMS	Living Conditions Monitoring Survey (Zambia)
MAR	missing at random
MCAR	missing completely at random
MICE	multiple imputation using chained equations
MicroZAMOD	Zambian tax-benefit microsimulation model
MNAR	missing not at random
NBS	National Bureau of Statistics, United Republic of Tanzania
NMAR	not missing at random
PIT	personal income Tax
PMM	predictive mean matching
SRMI	sequential regression multiple imputation
TASCO	Tanzania Standard Classification of Occupations
TAZMOD	Tanzanian tax-benefit microsimulation model
TRA	Tanzania Revenue Authority
TZS	Tanzanian shillings
ZMW	Zambian Kwacha