saspri

Southern African Social Policy Research Insights

Enhancing the Quality of Income Data in Surveys for Microsimulation Models in Africa

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

WIDER Development Conference, Helsinki, 13th September 2018

University of Essex

UNITED NATIONS UNIVERSITY
UNU-WIDER

The challenge 1/2

- Good quality income data is required for tax- benefit microsimulation models

- However, although income data is collected in many sub Saharan African surveys it is rarely used (*c.f.* Consumption data) and there is concern about its quality.

- Initial analysis of the income data for Tanzania and Zambia revealed several issues:
  - Missing income values
  - Implausibly high and/or low income values

The challenge 2/2

- Early versions of TAZMOD *apparently* simulated far too much direct tax whilst MicroZAMOD simulated far too little, compared to external administrative tax data
- Why?
  - validation data e.g. accrual versus cash-flow basis
  - compliance e.g. informal sector
  - quality of income data in surveys

The income variable of interest

- Employment income (*yem*) was selected to test imputation methods for two reasons
  - Major contributor to over simulation of direct taxes in Tanzania. 72% of initial direct taxes simulated attributable to income from employment.
  - More practically, we couldn't find suitable covariates to be able to model self employed income (*yse*), income from agriculture (*yag*) or other taxable income (*yot*)
- Prior to imputation the process of constructing *yem* was revisited to identify missing/implausible values and set these to missing

Identifying implausible incomes

- Identifying missing incomes is relatively straightforward but does need to take into account e.g. periodicity i.e. income may be present but periodicity absent or unquantifiable.
- Manual checks revealed outliers at top end of distribution that were implausible e.g. paid 'hourly'.
- Using the raw (untransformed) primary pay values, outliers were identified as values that were 1.5 times the interquartile range away from either the upper or lower quartile.
- Outlier identification was performed by occupation category and by highest level of education.
- Approximately 10% of Tanzanian employment income cases required imputation.

Further cleaning of covariates

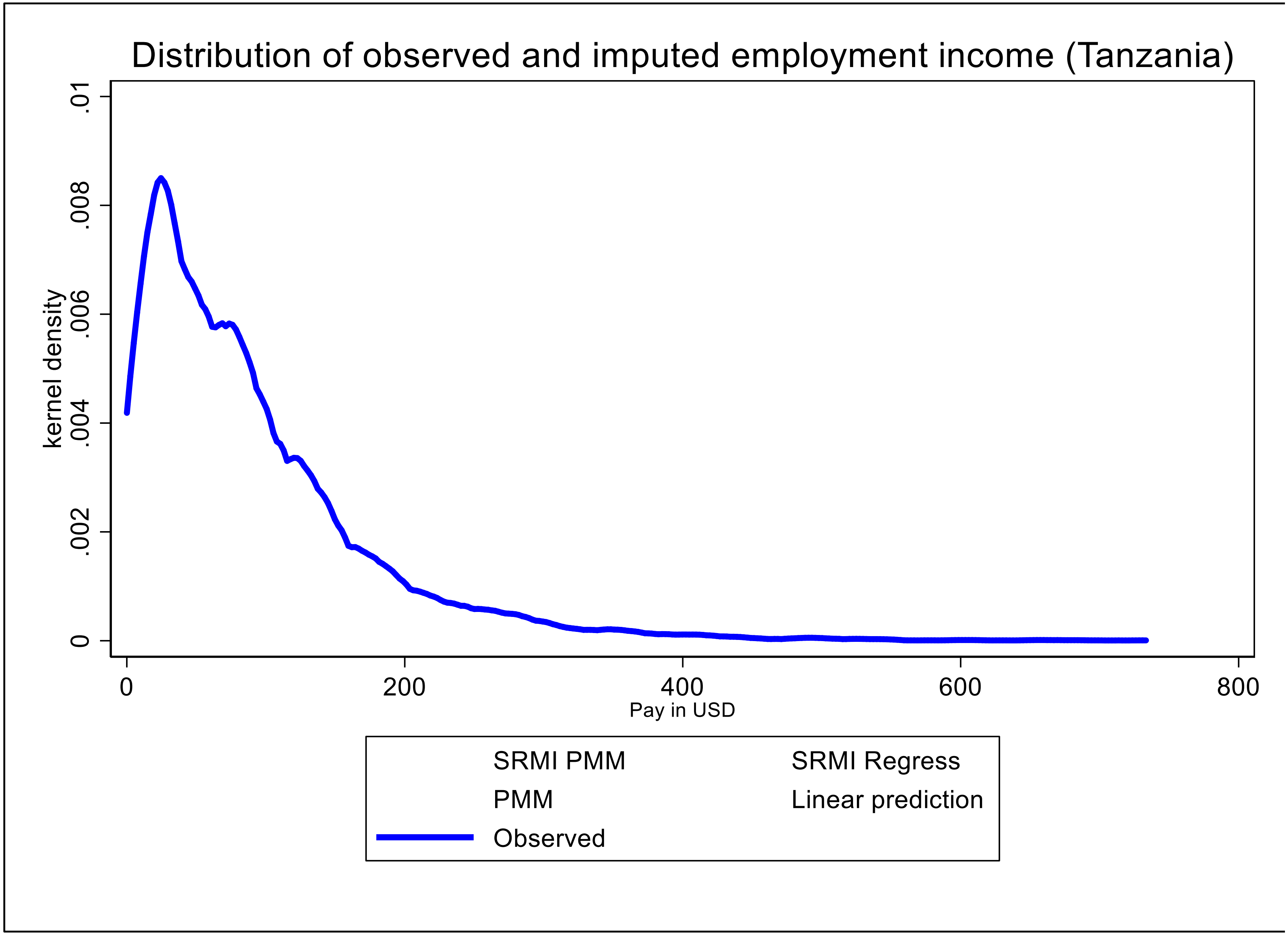
- All the imputation models require the identification of predictor variables
- The extent of missing and implausible values was explored for a set of variables: gender (*dgn*), age (*dag*), level of education (*deh*), labour market status (*loc*) as well as a range of 'living environment' variables
- Gender, age and level of education had very few missing data and could be cleaned more readily e.g. *deh* using a combination of age and current education grade

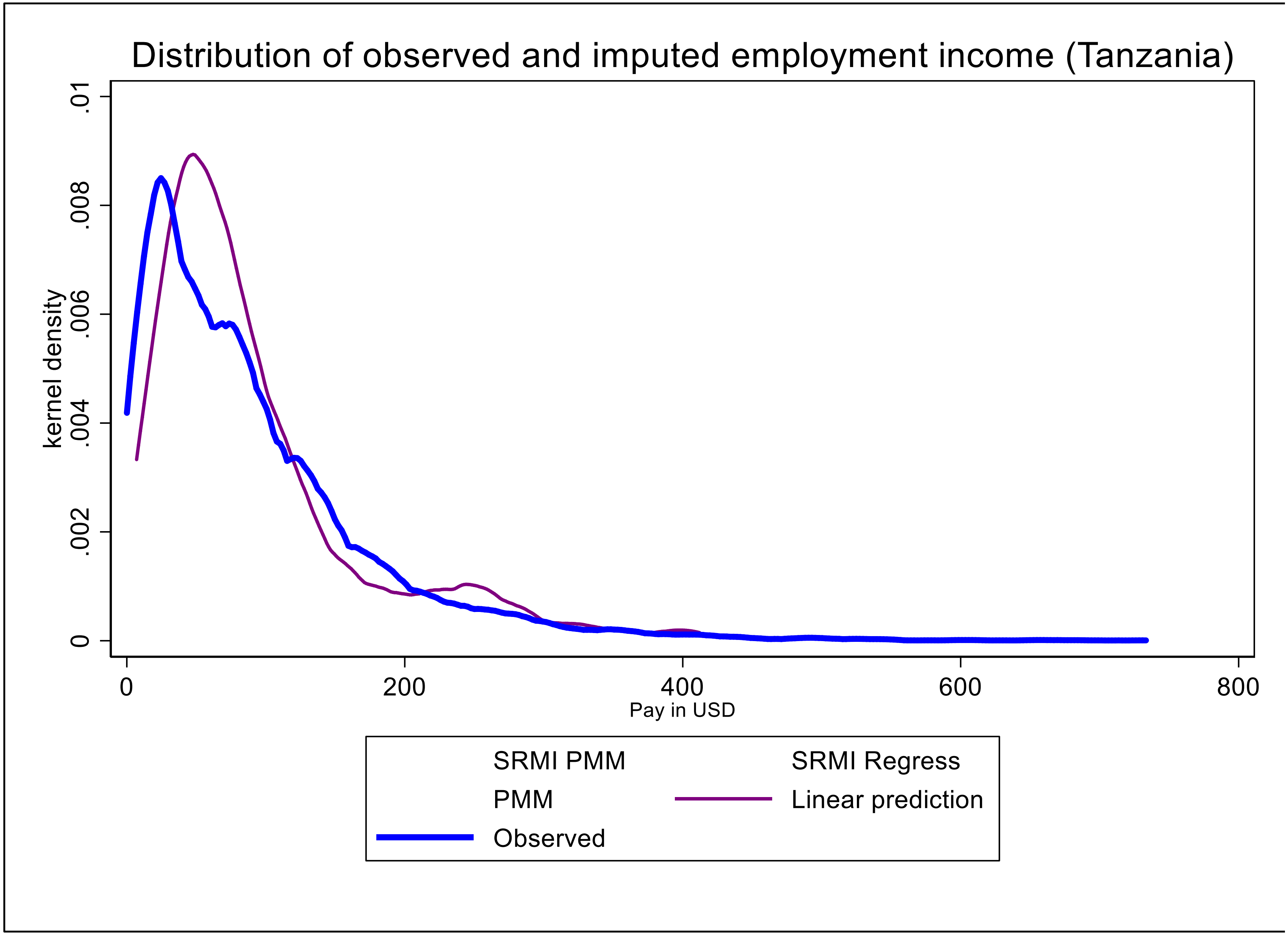
Four Imputation Methods tested

- Single imputation method
  - Simple linear prediction
- Three multiple imputation methods
  - Predictive Mean Matching (PMM)
  - Two variants of Sequential Regression Multiple Imputation - SRMI (aka Multiple Imputation using Chained Equations – MICE)
    - SRMI Regress
    - SRMI PMM

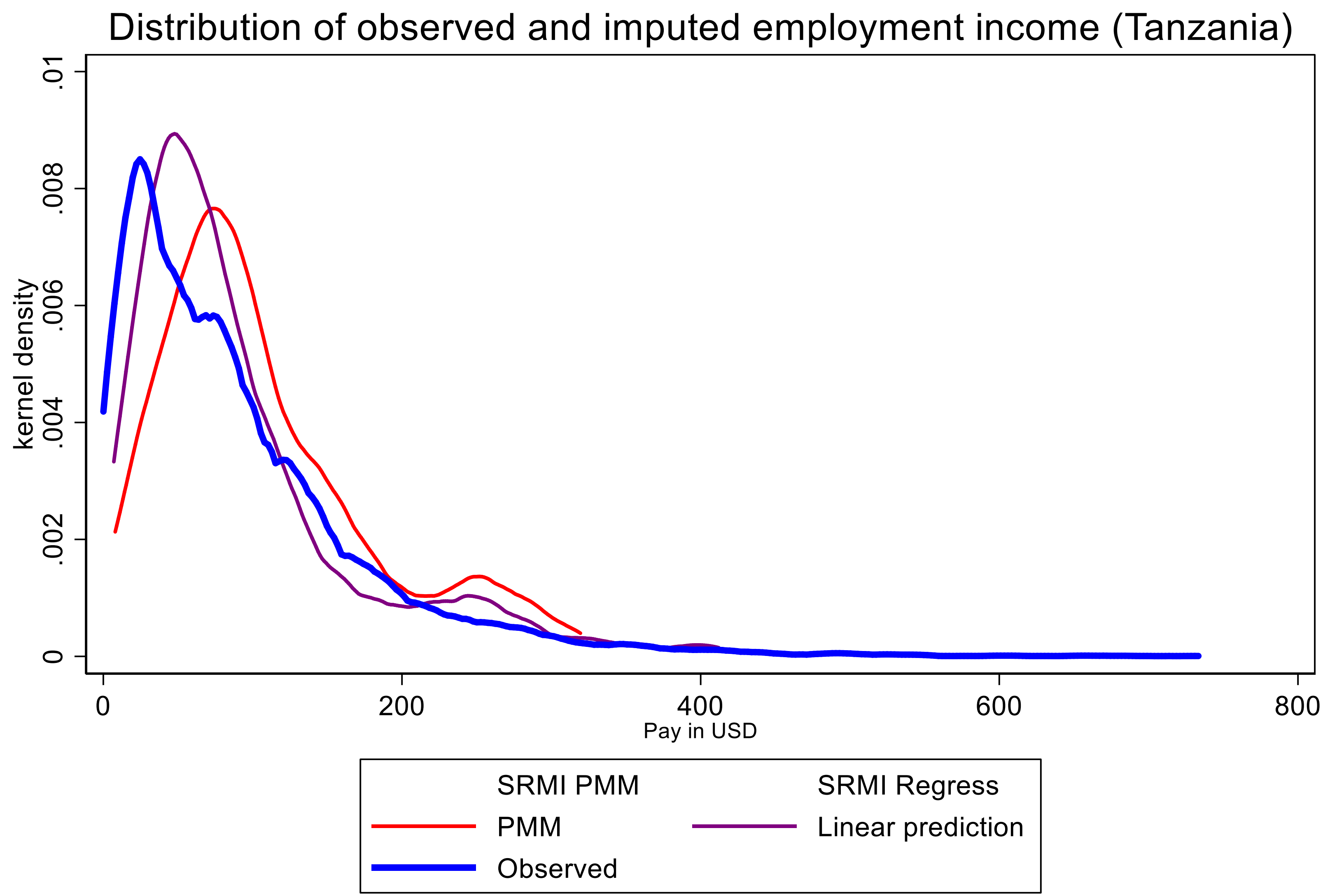
General principles of Imputation methods

- The basis for each imputation technique is a regression model or models.
  - For linear prediction and standard PMM, this is an OLS regression model as the main variable of interest is continuous (primary pay).
  - For the two SRMI models, these are predicated on sequential regression models, with a combination of OLS and multinomial logit models.
- 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*=50*)* and - for the SRMI approaches - the number of iterations per imputation (100).


Distribution of observed and imputed employment income (Tanzania)
kernel density
0
.002
.004
.006
.008
.01
0
200
400
600
800
Pay in USD
SRMI PMM
SRMI Regress
PMM
Linear prediction
Observed

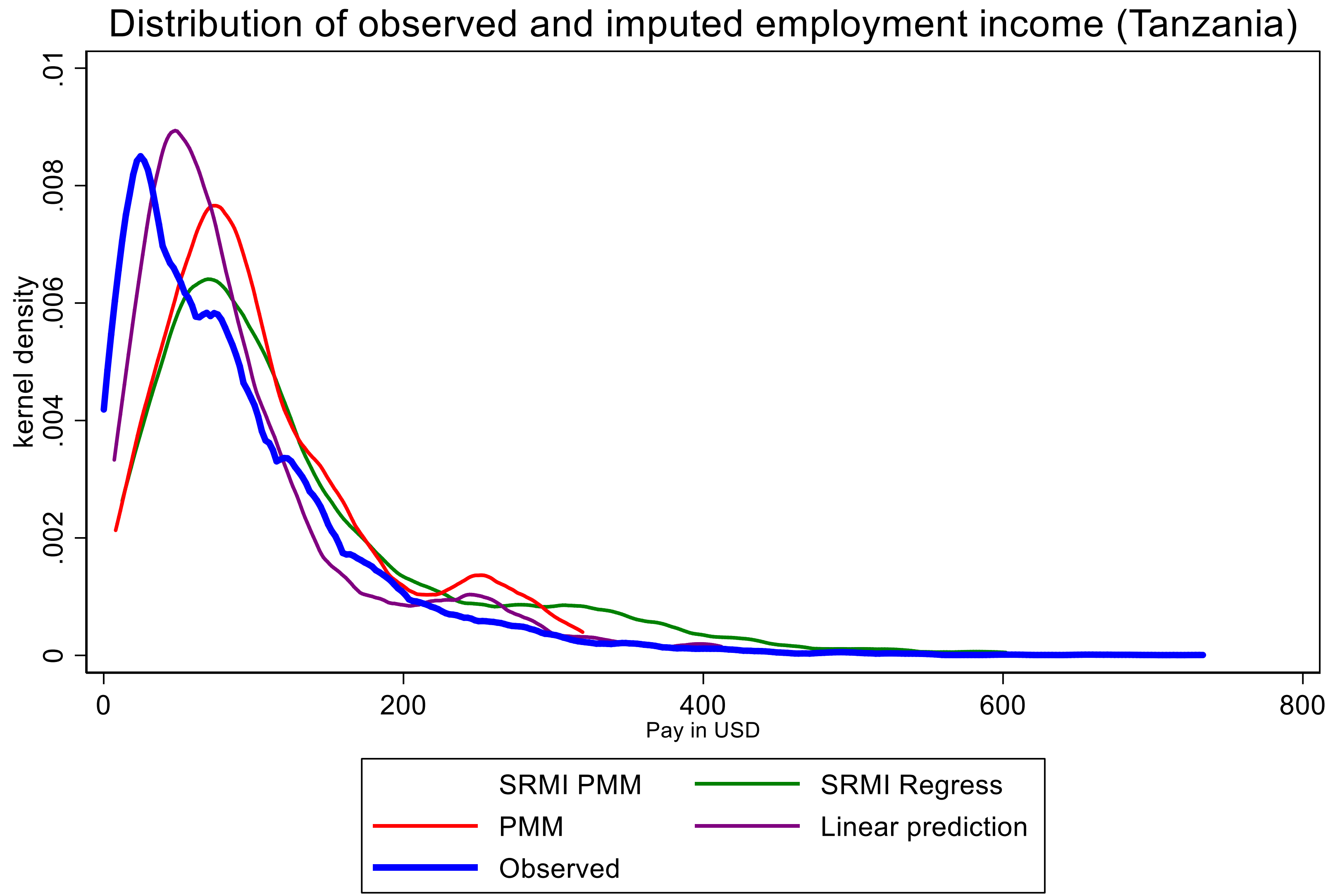


Distribution of observed and imputed employment income (Tanzania)
kernel density
0
.002
.004
.006
.008
.01
0
200
400
600
800
Pay in USD
SRMI PMM
SRMI Regress
PMM
Linear prediction
Observed

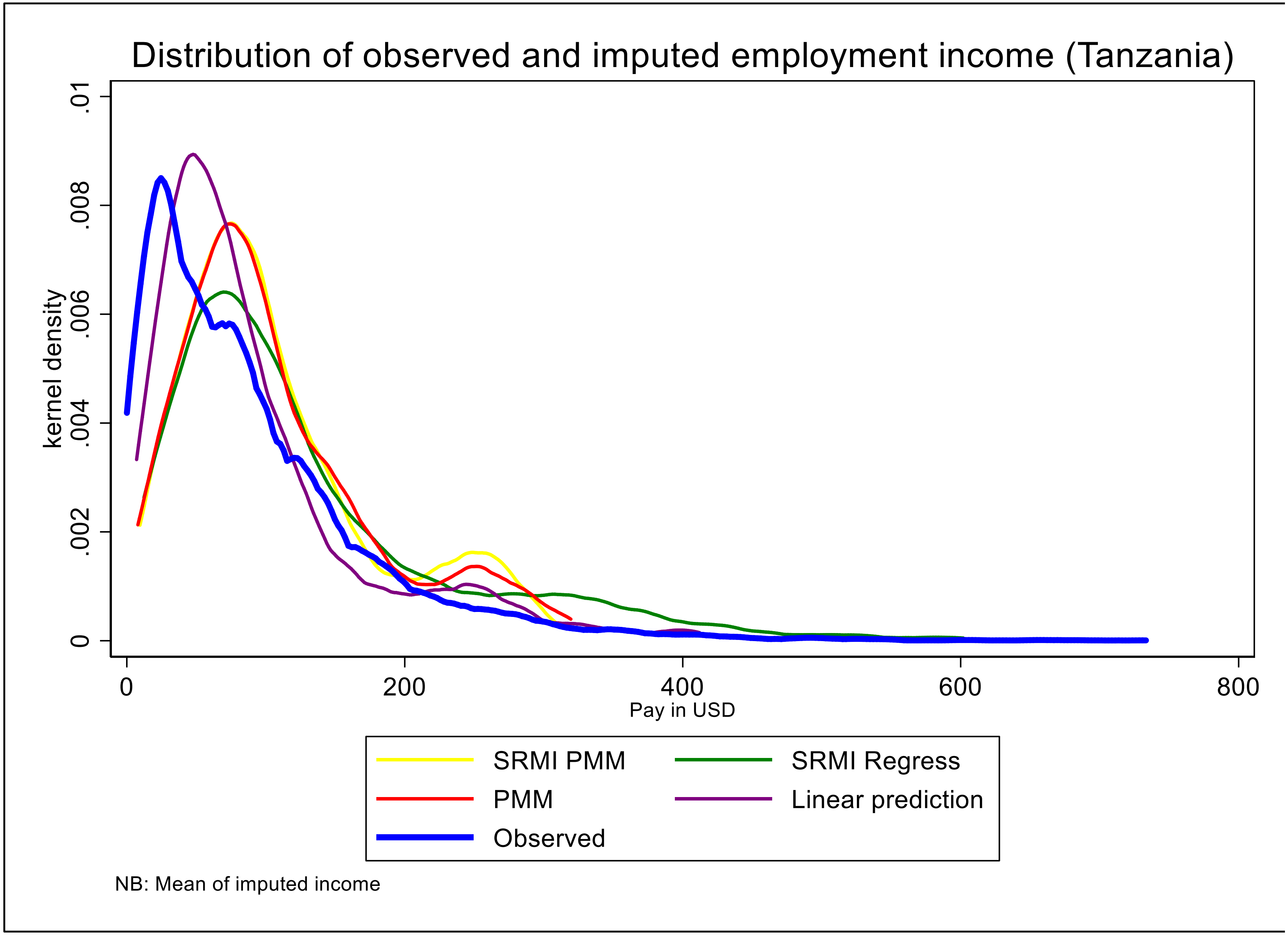


NB: Mean of imputed income

saspri University of Essex UNITED NATIONS UNIVERSITY UNU-WIDER



NB: Mean of imputed income



Distribution of observed and imputed employment income (Tanzania)
kernel density
0
.002
.004
.006
.008
.01
0
200
400
600
800
Pay in USD
SRMI PMM
SRMI Regress
PMM
Linear prediction
Observed
NB: Mean of imputed income

Results - Tanzania

Version of HBS Dataset	A Simulated Direct Taxes 2015 (TZS Million)	B Reported Direct Taxes 2015 (TZS Million)	C % simulated (Simulated/ Reported)
Before adjustments to income*	11,751,885	2,382,952	493.2
After constraining outliers to 99th pctile	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

Source: Simulations using TAZMOD Version 1.6 and HBS 2011/12.
* But after limiting simulations of PIT paid by employees to the formal sector

Testing the PMM approach in South Africa

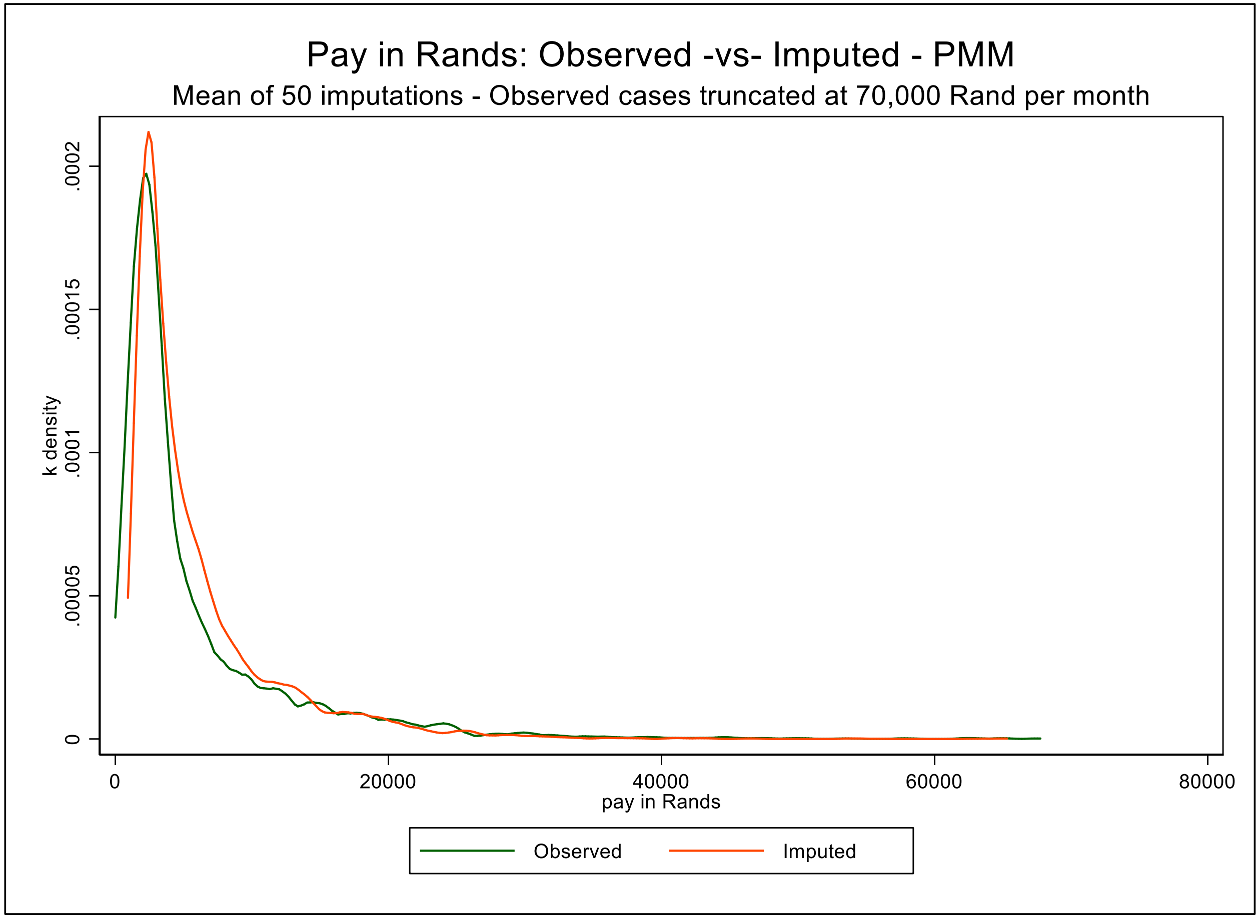
- The South African National Income Dynamics Study (NIDS) Wave 4 version 1.1 is one of two datasets underpinning SAMOD
- Survey data on income has been used more extensively in South Africa than in Tanzania and Zambia, and the NIDS income data has received particular attention.
- It was found to perform 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).
- It was therefore an ideal data set on which to test one of the methods (PMM) using artificial missing data

Validation using artificial missing data

Artificial missing data was introduced as follows

- Each observation was assigned a random number which was then used to generate decile groupings.
- Ten separate files were created, with each file containing income data set to missing for ten percent of the cases based on these decile groupings.
- The imputation technique(s) were then applied to each of the ten separate files.
- Having run the imputation technique(s), the observations containing imputed income from each of the 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.

South Africa - PMM on artificial missing data



University of Essex

UNITED NATIONS UNIVERSITY
UNU-WIDER

Results – South Africa using PMM

National Income Dynamics Study Wave 4 (2014)	A Using original employment income 2014 (R Million)	B Using imputed employment income 2014 (R Million)	C % change
Total Annual Revenue (direct taxes and Social Insurance of which:	287,029	233,125	81.2
- direct taxes	273,554	218,877	80.0
- Social insurance contributions (employer/ee)	13,475	14,247	105.7
Total expenditure on social transfers of which:			
- child benefits	145,443	144,485	99.3
- Disability benefits	65,017	64,083	98.6
- Pension benefits	20,232	20,034	99.0

Source: Authors calculations using SAMOD Version 6.5 with NIDS Wave 4 Version 1.1.
Notes: Imputed employment income obtained using PMM.

Dealing with multiple imputations

- One option may be to calculate a simple mean or median of the *M* imputed income values for each separate person in the dataset and assign this as the final imputed income value for the relevant person. (Used here)

- Another option is to retain all *M* imputations and instead run the microsimulation model *M* times (using an automation command in Stata) to generate *M* sets of simulated outputs, which can then be combined in some way.
  - Advantage: Could allow estimation of standard error and thus ci around result
  - Disadvantage: not particularly 'user friendly'.

Conclusions

- Meticulous data preparation (prior to any imputation) is essential and will vary by dataset.
- Manual adjustments to income outliers may be useful e.g. capping at 99th percentile (Tanzania) but not always (Zambia)
- Choice of imputation technique seems to make little difference to simulated results
- All imputations seem to improve the input datasets

Thank you

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. The results and their interpretation presented here are solely the authors' responsibility.

saspri

University of Essex

UNITED NATIONS UNIVERSITY UNU-WIDER

TAZMOD and MicroZAMOD references

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

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

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

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

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

UNU-WIDER (2018). https://www.wider.unu.edu/project/southmod-simulating-tax-and-benefit-policies-development

SOUTHMOD available for:

Ecuador

Ethiopia

Ghana

Mozambique

Tanzania

Viet Nam

Zambia

HEKIMA NI UHURU

saspri

University of Essex

iSER

EUROMOD

UNITED NATIONS UNIVERSITY

UNU-WIDER