STOCKTAKING NOTE FOR DATA HARMONIZATION

Sub-Saharan Africa Team for Statistical Development
FY 2019

July 2019

BACKGROUND AND CONTEXT

Household data are of great importance for poverty and share prosperity measurements. In addition to provide rich information on poverty and living standards, such data covers substantial information on social and economic issues namely demographic characteristics, education, labor, social protection, access to amenities, and migration, at the household or individual level. However, methodological approaches in terms of variable definition, data collection, and codes definition vary from one country to another. As a result, data are underutilized, in large part because the complexity and diversity of surveys require significant time to prepare the data for analytical work. To address this shortfall, it is critical to harmonize different household's data for comparability purposes.

Comparable data sets across countries and over time are crucial for multiples reasons. To name just a few, comparable data i) improves the quality of the World Bank's analytical work by encouraging more international comparisons, ii) enriches flagship regional and global reports, and iii) facilitates the monitoring of specific indicators over time for a given country, region or even the entire world.

Over the past decade, data harmonization has aimed to produce harmonized household surveys data all over SSA countries. A lot of effort has been exerted so far, but unlike in the past where the task was concentrated on building comparable data on welfare aggregate and poverty estimation, the effort now includes other modules, including Household characteristics, Individual characteristics and Labor module. Moreover, a substantial effort was made in Fall 2018 adopting International Income Distribution Database (I2D2)[1] labor variables as the new Labor module for SSAPOV.

From the period of October 2018 to May 2019, the SSATSD harmonized and improved the quality of 30 surveys. For 13 of these surveys, the team obtained data from the country teams and fully harmonized and checked all four modules from scratch. For 13 other surveys, the team upgraded the surveys to the revised dictionary, which is described below. Finally, the remaining four surveys were labor force surveys for which team harmonized the labor module. Of the 30 surveys that were harmonized, 6 were included for the first time in Povcalnet in the Spring 2019 poverty update. This note presents a list of surveys harmonized during this period, and discusses challenges faced, factors of success, as well as areas of improvement for future.

[1]The I2D2 is a global harmonized household survey database primarily focused on Labor force survey

1. UPDATE OF THE DICTIONARY

As mentioned above, the team significantly upgraded the dictionary guidelines used for harmonization. These upgrades improved the quality of the data and expanded the database to include other pertinent variables for analytical work. The last update in fall 2019 improved the dictionary in the following ways:

- Added about 8 temporal and/or spatial deflators to P module by differentiating spatial and temporal deflators. Prior versions of the database included only the deflator used by the country, without identifying whether it deflated across time, across space, or both; and whether the deflator was used for national and/or international poverty calculations.
- Integrated more variables into the Labor module. The old version of the dictionary contained only a few variables, all of which used a 7-day recall period. The short labor module was created due to lack of resources to produce a full set of labor characteristics. In the new dictionary, variables with a 12 month recall period were added along with information regarding earning (wage, benefit), health insurance, contract etc. In total 36 employment related variables were added to the previous dictionary.
- Removed certain variables which are well documented in other surveys or which are rarely used in practice. [2],[3]

2. SURVEYS AND MODULES ADDED TO SSAPOV

Twenty-five household surveys from 24 different countries were harmonized according to the new dictionary, with priority given to the most recent one, and moving backward if necessary. As a consequence, the distribution of surveys year is skewed to the right as shown in graph 1. Over the 24 surveys, the average year is 2014, with a standard deviation of 3.5. Following are a list of surveys harmonized during the period of October 2018 to May 2019:

Tab 1: List of harmonized surveys

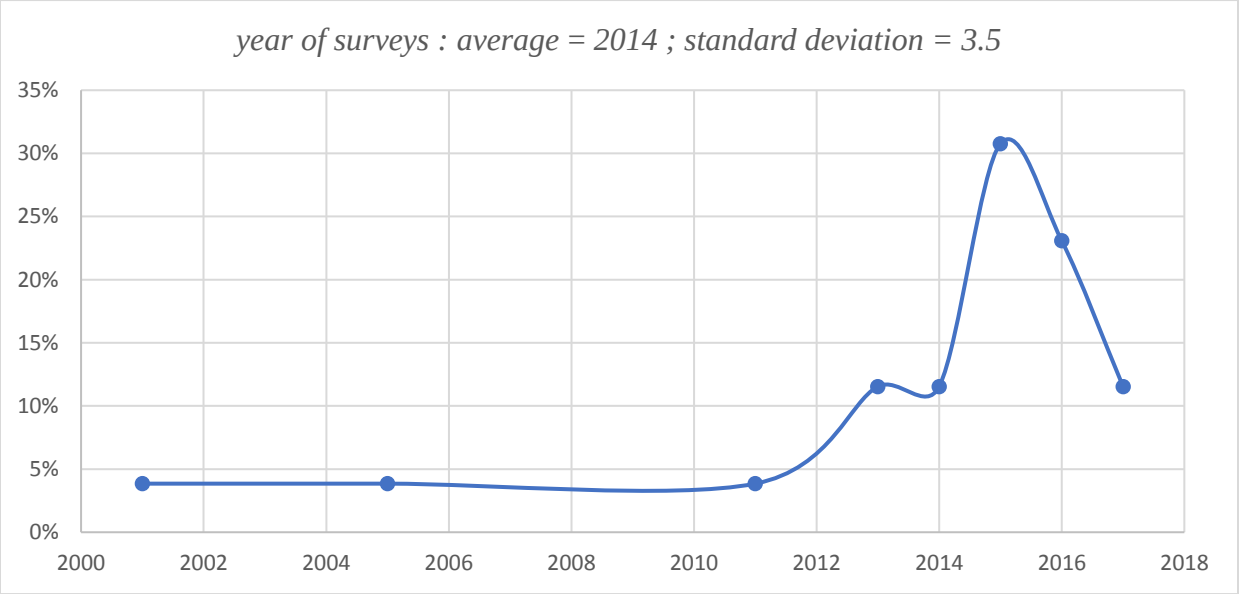
Country	Code	Year of survey	Survey name
Gabon	GAB	2017	EGEP-II
Mauritius	MUS	2017	HBS
Lesotho	LSO	2017	HBS
Uganda	UGA	2016	UNHS
Ghana	GHA	2016	GLSS-VII
Liberia	LBR	2016	HIES
Malawi	MWI	2016	IHS-IV
Rwanda	RWA	2016	EICV_V
Eswatini	SWZ	2016	HIES
Benin	BEN	2015	EMICOV
Botswana	BWA	2015	BMTHS
Ethiopia	ETH	2015	HICES

[2] Health related variables for example are documented comprehensively in Demographic Health Survey (DHS) so they were removed.
[3] Variable related to household enterprise, use of land/livestock, as an example.

Gambia	GMB	2015	IHS
South Sudan	SSD	2015	HFS
Kenya	KEN	2015	IHBS
Namibia	NAM	2015	NHIES
South Africa	ZAF	2014	LCS
Sudan	SDN	2014	NHBPS
Burkina Faso	BFA	2014	EMC
Burundi	BDI	2013	ECVM
Rwanda	RWA	2013	EICV-IV
Comoros	COM	2013	EDMC
Chad	TCD	2011	ECOSIT-III
Madagascar	MDG	2005	EPM
Cameroon	CMR	2001	ECAM-II

As previously mentioned, priority was given to recent surveys. Below is the distribution of surveys' year. Two thirds of the surveys were conducted after 2015 and less than 10 percent before 2010, meaning that the harmonized data reflect recent living standards and social household's characteristics.

Graph 1: Distribution of number of harmonized surveys across year of survey



3. CHALLENGES FACED

Several challenges emerged during the harmonization process. The most common one is the absence or inaccuracy of variables across many surveys. The team followed the general principle that variables should be set to missing when there is insufficient information in the dataset to generate them, to ensure comparability across surveys. While some variables are created as missing because they don't exist in the raw data, a non-negligible number of variables are set to missing because of inconsistency with the standardized definition. Example of such inconsistencies include using different recall periods, difficulty matching categories between raw

variable and harmonized variables, and others. This section presents challenges encountered in the harmonization process of each module.

Poverty module:

Covering about 59 variables, this module mainly includes consumption variables (food, non-food, and total consumption), spatial price indices, as well as poverty status. As far as challenges are concerned, deriving geographic variables at the subnational level is a recurrent difficulty, particularly at the third administrative level. Out of 25 surveys harmonized, only 3 have non-missing values for this variable. [4] Another major challenge is the absence of spatial and temporal deflators in several cases. Even if some surveys provide deflators, they almost always don't distinguish between whether the deflator is spatial or temporal. The team was therefore forced to field a survey to TTLs in the unit to better understand which countries used spatial and/or temporal deflation in the welfare aggregates provided.

Household module:

This module records household-level information and includes household characteristics such as housing characteristics and utilities, access to various amenities measured in terms of distances/time, and ownership of durable goods among others, making a total of 84 variables. This module also suffers from the existence of multiples missing values for some variables. The rate of missing values is particularly high for information on remittances (amount, source, relationship to the sender) and access to amenities (distance/time to water source, to school). Graph 2 below presents the percentile of surveys with missing values as a function of cumulative numbers of variables, which partly reflects the magnitude of missing values in H module.[5]

Tab 2: Distribution of share of variables missing in H module

Variables missing	*Percentage of surveys with*
0-20 percent of variables missing	*88*
20-40 percent of variables missing	*51*
40-60 percent of variables missing	*21*
60-80 percent of variables missing	*4*

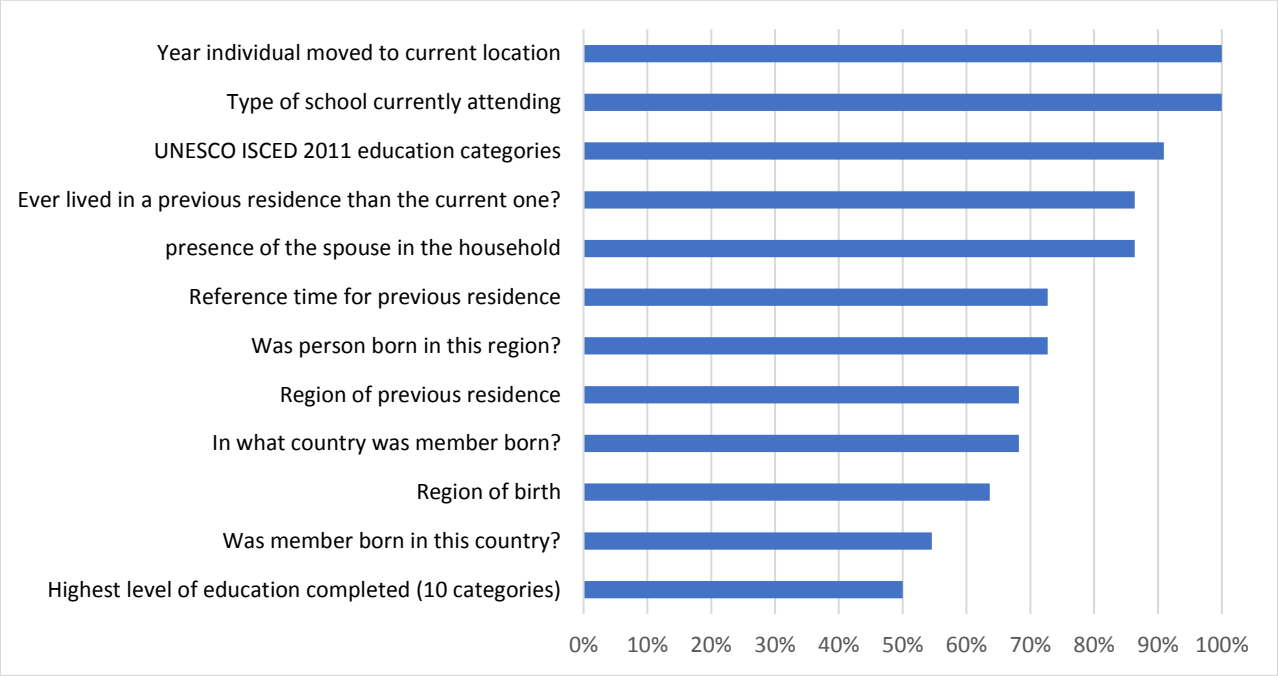
Individual characteristics module:

Covering 37 harmonized variables, this module essentially provides individuals' information on basic household identification, demographic characteristics, education and migration. When revising the dictionary, the decision was made to remove information on health and child

[4] The expression missing value for a variable or missing variable are used interchangeably.
[5] The interpretation of this graph is quite straightforward. As an example, one might read 13 variables are missing in 100 percent of surveys covered or 30 variables contain missing values in at least 70 percent of survey.

anthropometry. The rationale was that these variables are rarely present in the budget surveys used for poverty and most analysis of these topics uses the far more comprehensive Demographic and Health Surveys. Of the remaining variables, ones related to migration and disaggregated educational attainment are often missing. The *UNESCO ISCED education variable* is also difficult to harmonize, not because of a lack of data, but rather because of the difficulty of matching the country-specific categories elaborated by UNESCO to education variable in surveys.
Graph 3 below presents the proportion of variables created as missing for selected variables in the I module. [6]

Graph 2: Proportion of surveys having variable as missing value



L module:

This module covers labor market outcomes. These include: household chores; screening questions that determine primary activity, primary and secondary employment using 7-days or 12 months recall period. This module also contains several missing variables. Out of 54 variables covered, 11 variables are created as missing in all surveys. In addition, 23 variables were created as missing in 90 percent of surveys and 5 variables were created as missing in 80 percent of surveys. In 74% of cases, these variables with missing values are related to household chores, secondary employment, or employment using a 12 month recall period. The results as presented are not that surprising, considering that LSMS and other household budget survey data typically do not ask detailed question about labor market outcomes. Furthermore, in some cases the data ask labor-related

[6] The list of variables is limited to those which are missing in 50% of covered survey or more

questions but use a recall period that doesn't match the standard of 7-day or 12-month recall periods.

Box 1: *List of variables created as missing in all surveys.*

Cleaning (household chore)
Childcare (Household chore)
Elder care (Household chore)
Cooking (Household chore)
Unemployment duration
Social security for employee
Firm size in the secondary employment
Firm size in the 12- month job
Health insurance
Number of jobs
Hours of work in typical week in the 12-month job

4. CHECKING FOR QUALITY CONTROL

Throughout the process of harmonizing the data, the team attached great importance to data quality. Obtaining good quality data requires a thorough check of the accuracy and reliability of the data, to ensure that the harmonization was carried out properly. The team therefore adapted an automatic program that performs quality checks, called Q-check. The team added additional checks based on logic, consistency, and the experience working on household surveys in the SSA setting. These checks consist of non-responses rates or merging issues across modules, eliminating duplicates, making sure all coded variables should have their values within a certain range (or contain valid answers), ensuring that there is one household head per household, and many other internal cross-checking variables.[7]

Below are examples of basic checks performed by Q-check:

- The identification number for each member should be unique;
- There must be one and only one head of household;
- Persons less than 15 years should not be heads or parents or grandparents but there can be exceptions. In such a case, the proportion should be negligible, unless the poverty economist confirms the numbers. Moreover, the age difference between head and child could not be less than 12. But a verification on a case by case basis is almost always done afterwards;
- Marital status for head and spouse should be identical. If number of spouses is greater than 1, implies that head of household and spouses are polygamous;

[7] For example, it is hard to believe that a household head is 10 old year or younger or a 6 six-year-old person is married, etc.

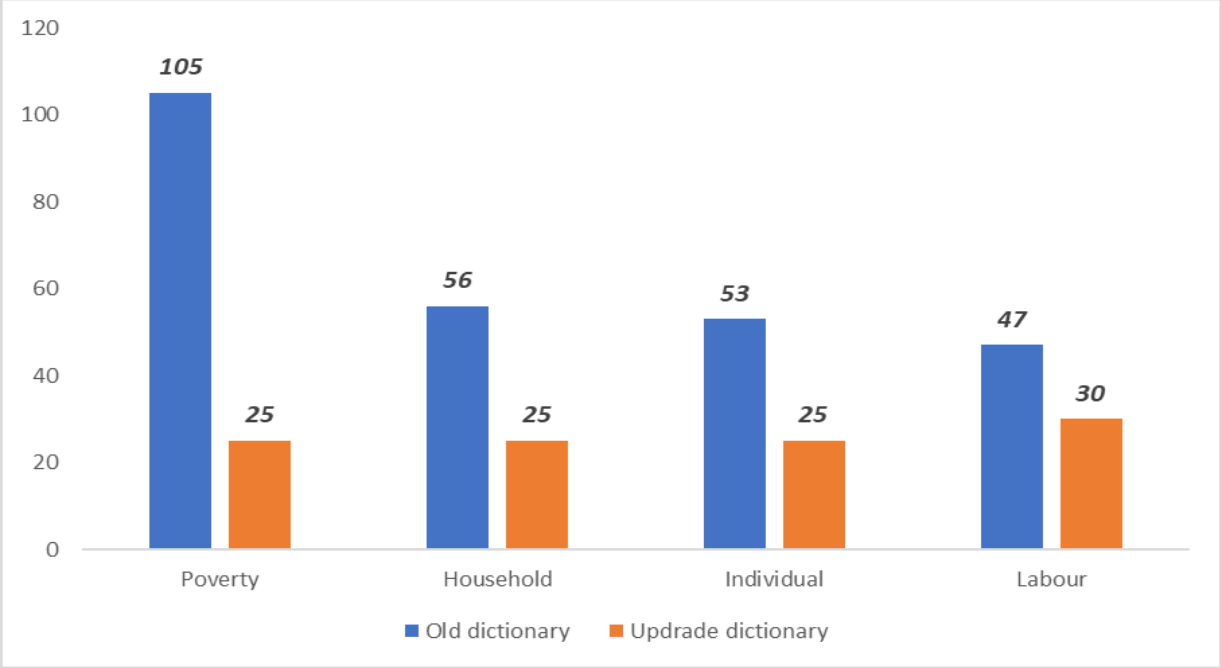
- Check age versus marital status of household members. For example, children under 12 should be single. However, depending on culture and in exceptionally cases; one may find 10-12-year-olds that are married;
- School attendance should be crosschecked with education level attained.

The Q-check program helps identify problems and allows the team to revise the code until these problems were fixed, which helps produce more consistent and reliable harmonized data.

5. DATALIBWEB

Datalibweb is one of tools used to maintain and organize The World Bank's micro data. The team has made considerable efforts to publicly disseminate this tool among staff and consultants at the Bank in several ways, including a learning session held on February 26, 2019. The current datalibweb library for Sub-Saharan Africa contains 130 P modules, 81 H modules, 78 I modules and 77 L modules. Most of surveys are harmonized according to the old dictionary. The SSATSD expects to convert all harmonized files to the updated version of dictionary, which covers labor and deflator variables more extensively than the previous version. The graph below presents the distribution of number of harmonized surveys in old and updated dictionary.

Graph 3: Distribution of numbers of harmonized surveys stored in datalibweb



6. AREAS OF IMPROVEMENT AND AGENDA FOR FISCAL YEAR 2020

Building on the lessons learned in FY 2019, the agenda for the next fiscal year is to concentrate efforts on harmonizing more surveys, starting with most recent surveys for each year. The team will focus on the following tasks:

- Adjusting the dictionary to match the new Global Monitoring Database (GMD) 2.0 expansion, which will contain additional information on geography, assets, and access to water and sanitation. The GMD is the Bank's primary harmonized data source for global poverty surveys and is derived from SSAPOV as well as the other regionally harmonized data.
- Harmonizing the latest survey for each country to GMD 2.0 specifications for the January 2020 deadline for the Spring 2020 update.
- Continuing to clean up the data. This includes: Ensuring that the database contains all harmonized surveys, that the .do files run properly and are well-structured, and that the database is properly organized.
- Finalizing and incorporating .do files from the I2D2 into the SSAPOV L module in surveys that match;
- Distinguishing between spatial and temporal deflators;
- Complete existing notes on the pace of the structural transformation and urban and rural poverty, complete a new note on MPM and continue to produce other analytical work based on harmonized data if time permits;
- Clean up the dictionary by excluding variables that are missing in all surveys or design proxies if possible;
- Advocate for the use of datalibweb among staff and consultants through trainings sessions;
- Continue providing capacity building opportunities to African students with a background in statistics or economics who are enthusiastic about pursuing graduate studies afterwards.

7. CONCLUSION

In summary, by allowing comparisons across countries and over time, these data are an excellent source to carry out analytical work or project at regional or global scale. The team has made significant progress in four areas. The first is the addition of a substantial number of surveys to the harmonization, such that the most recent survey from each country is covered. Secondly, the team expanded the SSAPOV dictionary to cover a greater number of variables in the poverty and labor modules. Third, the team implemented Q-check, an automatic tool to check for quality control, which helped to avoid inconsistencies in data. Finally, the team has been producing short notes such as: the note on Internet Access in Sub-Saharan Africa, and a polished draft of the structural transformation note, which showcase how the data can be used to address important policy issues.