Global Trade Analysis Project
https://www.gtap.agecon.purdue.edu/

This paper is from the
GTAP Annual Conference on Global Economic Analysis
https://www.gtap.agecon.purdue.edu/events/conferences/default.asp

Gender Disaggregated Labor Database: Harmonizing industry and occupation variables to international classifications systems[1]

Cristian I. Jara Nercasseau[2], Maryla Maliszewska, Claudio E. Montenegro[3], Raimundo Smith Mayer [4], Israel Osorio Rodarte[5], Javiera Petersen Muga[6] & Huanjun Zhang[7]

This paper describes the construction of the gender disaggregated labor database (GDLD). The GDLD is a microeconomic-based global database that provides detailed accounts on employment levels, wages, and skill qualification of labor at disaggregated economic activity, occupation category, and gender. The data was constructed by harmonizing industry and occupation variables to international classification systems. This database can be used in global trade models interested in assessing the links between gender, employment, and poverty.

JEL classification:

Keywords: gender, wage gap, poverty, equilibrium models, labor force

1 Introduction

The Gender Disaggregated Labor Database (GDLD) is …

The GDLD relies on previous harmonization efforts that work on top of on nationally representative household surveys. The database could be used to describe, at a finer level of disaggregation, internationally comparable statistics employment and remunerations. Until now, only data in broad economic categories is available for cross-country comparisons. This database fills an important information gap in global gender statistics by providing detailed accounts on employment levels and relative wages at a finer economic activity level and occupation category than is usually available.

---

[1] This paper is a product of the staff of the International Bank for Reconstruction and Development/The World Bank. The findings, interpretations, and conclusions expressed in this paper do not necessarily reflect the views of the World Bank, the Executive Directors of the World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work. This material should not be reproduced or distributed without the World Bank's prior consent. The authors remain solely responsible for the views expressed, interpretations, conclusions and any errors.

[2] Economics Department, University of Chile

[3] Economics Department, University of Chile; The World Bank; and the German Development Institute (DIE)

[4] Economics Department, University of Chile

[5] The World Bank

[6] Economics Department, University of Chile

[7] The World Bank

The database has a direct application in computable general equilibrium (CGE) modeling, particularly on the missing link between gender and international trade. Our database contributes to this by fulfilling the shortage of data, that is relevant for the CGE modeling. However, other types of modeling frameworks, including macroeconomic modeling, could also benefit from this database. The thirst for new data in the economic world is not new. This is specially the case when dealing with disaggregated level than extend beyond what is usually available in macroeconomic statistics. For example, breakdowns of labor by activity, occupation, education, or geographical area are not usually available across countries. In the same sense figures of poverty by any covariate are scarce.

Gender is integral to the process of development. Gender development outcomes are both strong determinants and co-dependent on a variety of development goals. At the household level, for instance, gender gaps in educational attainment are a strong determinant of households' allocation of labor and the roles played by family members in providing care; it is well-established that mothers' educational attainment is a better predictor of the educational achievement of household's offspring; largely explained by the fact that women, having different spending patterns, tend to allocate a larger share of expenditure on health, education, and well-being of their children. On aggregate too, gender is strongly interlinked with economic performance. Labor participation of women can play an important role in raising the rate of potential output, especially in countries where women's' labor force participation has been historically low. Similarly, economic activities that rely on women work force, such as manufacturing of wearing apparel and textiles, can shape the comparative advantage across nations, especially when opened and integrated with global value chains.

As a result, the gender dimension is crucial to the formulation of sound economic policy. It is well-established that a careful examination of the impact of economic policy should consider the gender dimension to address the multitude of impacts across different segments of the population– and that this examination should extend beyond the microeconomic approach. Macroeconomics, including trade policy, is increasingly considering gender a fundamental aspect for the design of economic policy. Not only macroeconomic policy can have sizable and long-lasting effects on gender-related outcomes; but also, existing gender inequities can influence the effectiveness of macro-economic policy. Typically, macroeconomists are interested in the effects on labor force participation, financial inclusion, trade diversification, firm performance, intra-household choices, and public investment.

As of now, important data gaps on gender statistics on labor conditions still remain. Particularly, comparable data across countries on detailed sector of employment is scarce. This database contributes to fill this gap. Traditionally, only employment at an aggregate level are published at the broader economic activity (agriculture, industry, services)[8]. The World Economic Forum's Global Gender Gap

[8] This data was produced by the International Labor Organization (ILO). It is available at the World Bank Gender Data Portal.

Report[9], for instance, present aggregated indicators for earning gaps only for skilled workers and lacks sufficient detail on women's sector of employment, labor volumes, and earning by detailed economic activity. The International Labor Organization produce detail statistics on employment and earnings, constructed using tabulations from national statistical offices. Broad aggregations in global gender statistics are insufficient for a careful examination of the links between international economic policy, gender and poverty.

This paper is structured as follows. Section 2 describes sources of data sources used in this paper. Section 3 provides details about the harmonization process of the microeconomic databases, including the construction of concordance tables and the creation of decision rules for assigning codes, particularly when there isn't a 1-to-1 match between codes in classification systems. Section 4 presents an application of the database: based on the microeconomic harmonized data, global statistics on employment are disaggregated by activity, the skill-level of labor and gender. This process involves the reconciliation of microeconomic based statistics and macroeconomic aggregates. The resulting statistics can be used to disaggregate labor accounts in the GTAP v10 database. Section 5 concludes by suggesting good practices in terms of data gathering, compilation and discusses areas for future development.

2 Sources of data

Data sources include individual-level record data from household surveys, documentation, meta-data, and internationally comparable employment statistics. This information is scattered in a variety of sources. Primarily, microdata was accessed from the harmonized collections from the World Bank and the Luxembourg Income Study. While access to record-level microdata has historically faced cumbersome restrictions, it is becoming more common that statistical offices provide public access to microdata (and code). Internationally comparable employment statistics were obtained from the World Bank (World Development Indicators), the International Labor Organization (ILOSTAT), and the Global Trade Analysis Project Database.

A variety of data sources have been used to collect documentation and meta-data information on national classification systems. This information has been systematically organized and is available for public use. The meta-data contains, at the household survey level, information regarding activity and occupation classification systems gathered from a variety of sources, including the U.N. Statistical Division website, national statistical agencies, household survey technical documents and questionnaires, among other sources. The systematic availability of meta-data is a serious bottleneck for developing research that relies on comparable international labor statistics. This is a labor-intensive process that can well be regarded as a global public good. The information gathered for the construction of our database, while

[9] http://reports.weforum.org/global-gender-gap-report-2016/

limited in scope, contributes to this effort and can be accessed at the GDLD World Bank data portal[10].

Microdata is accessed from the World Bank collection of household surveys and from the Luxemburg Income Study. Micro-data for developing countries was mainly accessed from the World Bank collection of household surveys. The World Bank harmonization efforts are organized by geographical region[11]. On top of regional databanks, there exists additional harmonization efforts that contain a broader coverage, with a limited number of variables. The World Bank collections tend to focus on nationally representative surveys used for the official measurement of poverty, rather than the use of surveys used to measure labor outcomes, such as labor force surveys. The Luxembourg Income Study (LIS) permits remote processing of household surveys information. LIS contains some information about classification systems.

2.1 United Nations' national classifications database

The United Nations has compiled global information about national and international classification systems. The U.N. Statistics Division has published updated lists of national classification systems in use (as of April 2018)[12]. These lists include the names of the classification systems in use, by country. For activities, it contains the classification system's names for 121 countries and for occupation classifications, it documents 76 countries. In a previous version released in 2012, a more in-depth survey about each classification system was compiled. While this information is considered outdated, it contains valuable information for performing historical harmonization and it can be still accessed through web-repositories[13]. A next round of this survey was planned for 2018, but as of today, it has not been implemented.

2.2 World Bank collection of harmonized household surveys

The main source of the GDLD are the World Bank household survey harmonization collections. Given our objective of creating disaggregated data by industry and occupations, the natural point of departure for this exercise was the household surveys with individual or household level observations. Within the World Bank, there are two global collections of frequent use: the International Income Distribution Database (I2D2), and the Global Micro Database (GMD). The I2D2 evolved from the need of presenting a global picture on key development aspects of

---

[10] The data is in the World Bank Intranet website at http://datatopicsqa.worldbank.org/gess/

[11] Since countries in the same region tend to share common language, classification practices and standards. In an initial phase for this project a subset of countries was selected to address the feasibility of the project. These countries were Colombia, Egypt, Indonesia, Cambodia, Lao's PDR, Sri-Lanka, Lesotho, Malaysia, Thailand, and Vietnam.

[12] https://unstats.un.org/unsd/classifications/Nationalclassifications

[13] For instance, see: https://web.archive.org/web/20160302223812/http://unstats.un.org/unsd/cr/ctryreg/ctrylist2.asp

global trends on inequality, education, labor, and employment; mostly to inform World Development Reports and other World Bank flagships. At present, I2D2 has more than 2,000 household surveys that cover 162 countries. The GMD is a similar harmonization effort, but its main objective (and comparative edge) is to provide comparable statistics on poverty. GMD is the official database for the monitoring the World Bank Twin Goals of Poverty and Shared Prosperity. Given that the most relevant variables for this project are the original industry and occupation classification systems, which are more commonly recorded in I2D2 than in GMD, this project rely more on I2D2's infrastructure. Nevertheless, there is considerable overlap between the two databases and the harmonization effort can be extended to GMD or other type of harmonization.

On top of industry and occupation variables, this project relies on the existence of a larger set of pre-harmonized set of variables. The harmonized household surveys follow common harmonization procedures and data structures. Across the board, we have used some already pre-existing variables to cross-tabulate industry and occupation. These variables include age, gender (self-reported), employment status, education (in categorical levels and/or years of schooling), labor force status, wages (self-reported), time unit of payment, welfare aggregate); depending on the harmonization and household survey additional variables can be retrieved.

Inconsistencies in harmonization efforts were reported to corresponding teams. One common inconsistency is the improper handling of the original industry/occupation variables, and the lack of proper referencing or documentation with regards to them. Typically, statistical offices create national classification systems by adapting the international family of classifications to their own needs; mostly to incorporate a set of items (activities, occupations, products) that result most relevant to them. Those differences appear at higher levels of disaggregation and typically occur when countries increase the existing nomenclatures. As it will be discussed later, many of these problems required manual adjustments.

2.3 Luxembourg Income Study

The Luxembourg Income Study (LIS) is a cross-national data center which servers a global community of researchers, educators, and policymakers. LIS acquires datasets with income, wealth, employment, and demographic data from high- and middle-income countries, harmonizes them to enable cross-national comparisons, and makes them publicly available in two datasets, the Luxembourg Income Study Database (LIS) and the Luxembourg Wealth Study Database (LWS). Their meta data is stored in the METIS search tool which provides immediate access to a comprehensive set of documentation about the LIS database.

3 The harmonization procedure

To the best of our knowledge for published macro wage and employment data, only employment or wage at an aggregate level is harmonized at the broader economic activity or ISIC 1-digit levels (21 sections in ISIC Rev 4 and 17 sections in ISIC Rev 3.1). Thus,

extracted information from micro household surveys to get wage bill splitting by GTAP 65 sectors, gender and skill is necessary. GDLD develops this objective to provide a global good in the form of a documented database that complements statistics on employment and labor incomes disaggregated by gender and detailed economic activity.

This raw household survey data of GDLD is derived from more than 2000 pre-harmonized household surveys for 162 countries on the International Income Distribution Database (I2D2) in The World Bank and Luxembourg Income Study (LIS) for Europe countries. Given the objective of creating disaggregated data by very detail industry levels, GDLD harmonized the key variables necessary for this study, on the top of six modules including (i) Database, (ii) Demographic, (iii) Dwelling, (iv) Education, (v) Labor, and (vi) Welfare in individual levels (i.e., the observation is every person.), that

1. information of individual and household;
2. demographic information: age, gender, education in category levels and years of schooling
3. labor force and employment status
4. original industry and occupation
5. self- reported wages in local currency and payment unit

GDLD mainly works in the workflow focusing on mining the survey meta data, harmonizing local industry classification to International Standard Industrial Classification Revision 4 (ISIC Rev 4) and occupation to International Standard Classification of Occupations (ISCO 08) as detail as possible, and calculating the monthly wage (in local currency and US Dollars).

The first step in constructing the database is to examine the quality of country's household survey and the identification of national classification of industry and occupation. It was decided that to be used the latest survey with disaggregated level of original industry and occupation variable.

One of main approaches and tasks of GDLD is identification of national classification system used by the country's household survey, in order to mapping local classification of industry to international standards. To do this point, we searched in survey documentation, national statistical office, United Nation, and other published articles. In most cases, the countries have adjusted the international classifications based on their own necessities. Thus, national customize mapping and meta data are created per country. The process of looking for a reliable information source to identify the industrial and occupation classifications used for the country to construct the specific survey we want to process was certainly one of the big challenges of this database.

For the creation of each country GDLD harmonized data, we use several correspondence tables depending on the previously identified classifications. In addition, since the final purpose is building the data consistent to GTAP10 65 sectors, which was built on ISIC. The GDLD firstly create the mapping from local classification to ISIC. Then concordant

the ISIC to GTAP using the information on the correspondence table between ISIC Revision 4 and GTAP 10.

Typically, statistical offices create national classification systems by adapting the international family of classifications to their own needs; mostly to incorporate a set of items (activities, occupations, products) that result most relevant to them. Those differences appear at higher levels of disaggregation and typically occur when countries increase the existing nomenclatures. As it will be discussed later, many of these problems required manual adjustments.

Since the simulation require wage per year, GDLD also identified the wage and payment unit of wage (weekly, monthly, etc.). Then calculate the monthly wage in local currency and US Dollars.

Our approach, at first, was to process those countries with the highest disaggregation of industry and occupation, and that were cover a larger share of global GDP or population coverage.

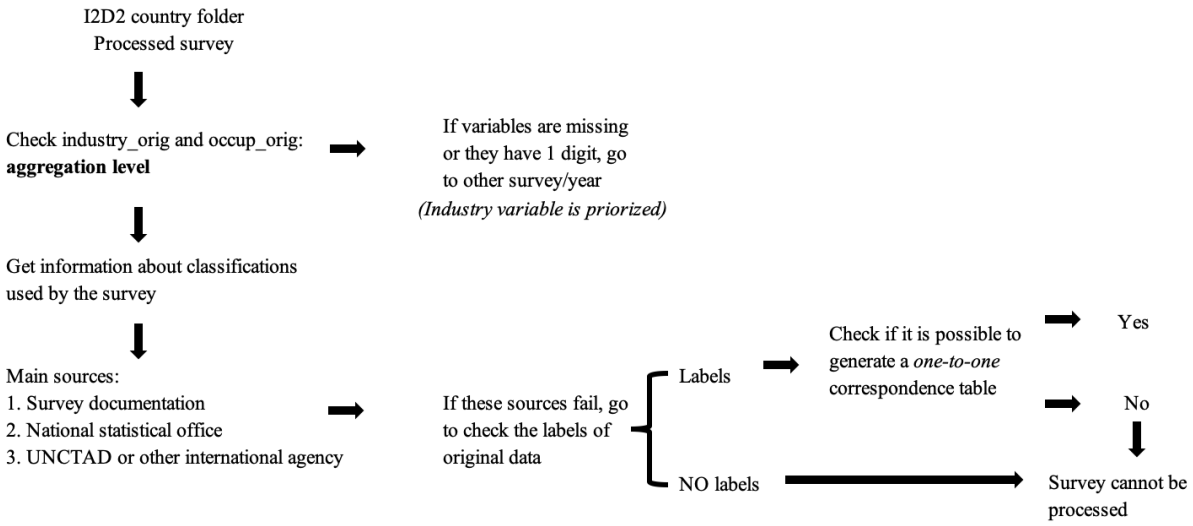
The I2D2 database has (in general) several years of data for one country and so one could, in principle, choose which year to use for a particular country.

It was decided that to be used have to be near 2014 (because of data availability in I2D2), and also have to have industry and occupation at a relative high level of disaggregation. This implies that if country X has data for 2014, but industry and occupation were highly aggregated, then it could be better use a different year than 2014 (earlier or later, but the closer to 2014) with higher disaggregation of industry and occupation.

3.1 Identifying national classification systems

The first step in constructing the database is to examine the quality of country's household survey and the identification of national classification of industry and occupation. Figure XX1 shows how this process is done for each country. As it was pointed out earlier, this project relied in data and harmonization variables from the I2D2. The newest country' survey processed by I2D2 is the first and best option to be selected by GDLD. However, in the cases that economic activities or occupation variable were missing or in very aggregated lelel, the previous I2D2 household surveys are selected for each country. In order to meet the timely request, GDLD did not select the surveys before 2000.

Figure XX1. Identification classification systems flow



One of the most important purpose of GDLD is the contribution of collecting "metadata" for every survey, in order to identify national classification system used by the country's household survey. To do this, we relied on three main information sources:

1) **Survey documentation:** this is the most reliable source of information regarding the classification used to construct the database of the survey. In some cases, national statistical office shows the current official classification of the country, but that classification does not fit with the survey we want to process. This is either because the country uses more than one classification system, or because we use a survey from previous years.

2) **National statistical office:** if the survey documentation is missing, then the second best information source is the official statistical department of the country. The information required is the name of the classifications used to construct the survey we want to process. If that information is missing, then we rely on the name of classification systems used by the country in the same year of the survey.

3) **UNCTAD or other international agency:** when both survey documentation and national statistical office information are missing, we search on information from international agencies. The information needed is the same that when we look at national statistical office. ILO, for instance, has information about occupation classification but not always this information is complete. In that case, we use cross reference to determine which classification was used to construct the data of de country's household survey.

In the case the three information sources failed, we used the labels of the original variable to do a one-to-one correspondence table with the International Standard for Industry

Classification (ISIC Revision 4). However, labels of original variables were not always in the data. This method was used, for instance, to process the Russian Longitudinal Monitoring Survey of 2016.

The case of Colombia is a good example of how the three information sources explained above can give different information. UNCTAD's tables of national classifications indicates that Colombia uses a local adaptation of ISIC Revision 4 (*CIIU Rev. 4 A.C*) as the current national classification. It also notes that this classification was adopted on February 1st of 2012. Colombia statistical department shows the same information. However, the official documentation of GEIH 2017 (*Gran Encuesta Integrada de Hogares*), the survey selected to be processed as GDLD, points out that a local adaptation of ISIC Revision 3 (*CIIU Rev. 3 A.C*) was used to construct the dataset.

The process of looking for a reliable information source to identify the industrial and occupation classifications used for the country to construct the specific survey we want to process was certainly one of the big challenges of this project. Thirty nine countries could not be processed because we did not find proper information regarding national classification systems. As the correct identification of national classifications is a necessary condition for processing surveys, improving the access to official national information would be a major step forward in this project.

3.2 Creating correspondence tables

For the harmonization process in GDLD, we use several correspondence tables depending on the previously identified classifications. This process is valid for both, industry and occupation's classifications, and is the main part of the harmonization.

3.2.1 Industry

To generate industry variables in the GDLD data, we rely on the information of the GTAP version 10 codes. Version's 10 nomenclature is built from the Central Product Classification (CPC v.2.1) and the International Standard for Industry Classification (ISIC Revision 4). Compared to last version (GTAP version 9), GTAP version 10 considers 65 sectors, which includes agriculture, food, resource extraction, manufacturing, and service activities to describe all economic activities in each country. It has more manufacturing and services sectors than previous versions. There are three new sectors in manufacturing, namely: Chemicals, Pharmaceuticals, and Rubber products, that were previously aggregated as a single "Chemical, Rubber, and Plastics (crp)" sector. It also distinguish the Electrical Equipment sector separately from other machinery. In the service sector, GTAP 10 now represents Accommodations and Food Services, Warehousing, Real Estate Activities, Education and Health Services, which were previously included in aggregated Trade, Other transport, Other business and Other government services sectors, respectively.[x] Table XXX of the annex summarizes all the 65 sectors of GTAP 10.

As previously mentioned, GTAP 10 sectors were built based on ISIC Revision 4. This implies that to generate GDLD variables, we must find the correspondence between the local industry classification and ISIC Revision 4 in the first place. This process is described in Figure XX2.

Figure XX2: Industry correspondence tables flow

Most of the industry classifications in our sample are identical or are based in international standard classifications. This means that in most of the cases, official correspondence tables with ISIC Revision 4 already exists, and we only have to deal with the level of aggregation of the local classification. Specifically, we merge the raw data with the ISIC Revision 4 correspondence table according to the digit of aggregation, and then assign the local codes to GTAP 10 using the information on the correspondence table between ISIC Revision 4 and GTAP 10. Most common classifications used in our sample of surveys (or in which local classifications were based on) are:

1. ISIC Revision 2
2. ISIC Revision 3
3. ISIC Revision 3.1
4. ISIC Revision 4
5. NACE Revision 2
6. NAICS

The process described above, represents the "ideal" case in which all the information is available, and the local classification is identical to other international standard classification. Nevertheless, there are two departures from this ideal scenario: 1) There is no direct official correspondence table with ISIC Revision 4; 2) official documentation states that local classification is based on an international classification, but some of the codes do not match the official ones.

The way in which we deal we the first departure, is to include an additional correspondence table in the process. That is, we use a correspondence table between the local industry classification and some other international classification different from ISIC Revision 4 in

the first place. Then, with this information, we assign codes using a correspondence table between the "auxiliary" classification and ISIC Revision 4. We usually have to include this additional step in the process when the classification used in the survey is not "updated". For instance, industry codes from the Household Integrated Survey (HIS) 2013 from Georgia, were based on NACE Revision 1.1 according to the official documentation of UN.[xx] An official correspondence table between NACE Revision 1.1 and ISIC Revision 4 does not exist, but there is one between NACE Revision 1.1 and NACE Revision 2. We use this table and merge it with the correspondence table between NACE Revision 2 and ISIC Revision 4.

The second departure is a little more difficult to deal with. Particularly, in many cases, surveys documentation declared to use a classification X (or to be based on it), but when looking at the codes in detail, there are many departures from the official codes. For instance, a country X declared to use ISIC Revision 3 to classify their industries, but when looking at the codes in detail, some of them do not appear in the official correspondence table. This means, that official tables are not the unique solution in all cases, and official documentation or the information that appears in raw data (e.g: industry variable labels) are crucial, as pointed out in the workflow of Figure XXX. More specifically, we use official tables only as a first step when we face situations like this, and the rest of the codes that do not appear on these tables are assigned manually using the information of the survey.

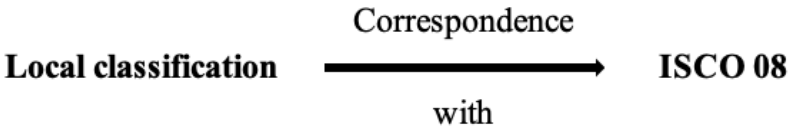
3.2.2 Occupation

The process of creating corresponding tables to assign occupation codes is essentially the same to the industry process. The only difference is that GDLD occupation variables are based on ISCO-08 codes. This means that our challenge is to find/create a correspondence table between the local classification and ISCO-08. The most common classifications we find in our sample are:

1. ISCO-88
2. ISCO-08
3. SOC 2010

Moreover, both processes faces the same potential departures from the ideal case in which local classification is identical to the international standard classification, and an official correspondence table already exists. And the way to deal with them is exactly the same. This process is described in Figure XX3.

Figure XX3: Occupation correspondence tables flow



3.3 Assignment rules

This section summarizes the methods by which industrial and occupational classification codes are assigned to individuals. The assignation process depends on the classification utilized within the survey, information obtained according to what it is explained in section 3.1. Our baseline classifications are ISIC Revision 4 for industries, and ISCO 08 for occupations. If a survey already utilizes one of these categories, the industrial or occupational code is taken as given. If a survey doesn't utilize one of these categories, we merge a concordance table from original industrial or occupational codes to ISIC Revision 4 codes and ISCO 08 codes and apply an algorithm to assign only one of them per local code. In the case of industrial classifications, we add an extra step, applying the assignation algorithm to match one GTAP Version 10 code per ISIC Revision 4 code.

The assignation algorithm of industrial codes is based on the trade weight of each sector within the countries exports. This implies that larger sectors have a higher likelihood of being assigned if one local industry code can be assigned to multiple ISIC Revision 4 codes, or if one GTAP Version 10 code can be assigned to multiple ISIC Revision 4 codes.

The assignation algorithm of ISCO 08 occupational codes, if necessary, is completely random.

3.3.1 Trade Weights Data

The raw data source comes from the UN's Comtrade Dataset[x1]. This dataset contains country-product level bilateral values and quantities of exports between the years 2003 and 2015, products are classified by their 6 Digit HS Codes according to revisions HS 2002, HS 2007 and HS 2012. Country Codes are classified using their official UN codes.

This dataset is transformed into a country-year-HS code level dataset of exports, for which we merge a concordance table between HS Codes and ISIC Revision 4 industrial codes at the 4th digit. For this dataset we compute total exports at the country-year-ISIC Revision 4 Code. We then calculate total exports by country-year for ISIC Revision 4 Codes at 3, 2 and 1 digit of aggregation.

Finally, we merge the concordance table between ISIC Revision 4 and GTAP Version 10 to the previous datasets and calculate total exports at the country-year-GTAP Version 10 Code.

3.3.2 From Local Industrial Classification to GTAP Version 10

To reduce the dataset to one ISIC Revision 4 Code per person we use the following algorithm:

a) Apply the concordance table between the local industrial classification to ISIC Revision 4. This implies that the relation between industrial codes is potentially a many to many concordance, which means that multiple local codes can be assigned to multiple ISIC Revision 4 codes.

b) Draw a uniform at the person level.

c) Call the trade weights dataset corresponding to the country-year pair of the survey we are analyzing and at the digits of aggregation of the concordance between local industrial codes and ISIC Revision 4. If trade data is not available for the survey year, we take the average exports by sector between the two closest previous and posterior years to the survey. If there are no posterior years, we take the closest previous year available. If the country is not available, we don't consider trade data for assignation. The result is a dataset at the person-year-local industry code-ISIC Revision 4 industry code level, with total export values of the ISIC Revision 4 industry code.

d) Calculate the share of trade of each ISIC Revision 4 industry code within local industry codes at the person level, by computing total trade within the local person-industry code pair. This implies that each ISIC Revision 4 industry code will be assigned a value between 0 and 1, and that the sum of these values within each individual will be one.

e) Compute accumulated trade weights at the person level, then if the uniform drawn in step a) lies in the range between two accumulated trade weights, the local industry code is assigned to the ISIC Revision 4 code represented by that accumulated weight.

f) If trade data is missing each ISIC Revision 4 sector within local industry at the individual level has the same likelihood of being assigned.

The following example will illustrate the assignation algorithm:

The original dataset contains a person identifier, with a local industry classification as an attribute.

Person Id	local industry
Id1	local sector 1
Id2	local sector 2

Step a) Crosswalk between local industries and ISIC Revision 4 Industries, in this case, local sector 1 is assigned isic sector 1, and local sector 2 is assigned both isic sector 2 and isic sector 3. This expands the data to a person-isic industry level industry dataset.

Person Id	local industry	isic industry
Id1	local sector 1	isic sector 1
Id2	local sector 2	isic sector 2
Id2	local sector 2	isic sector 3

Step b) Draw a uniform at the individual-level.

Person Id	local industry	isic industry	uniform
Id1	local sector 1	isic sector 1	0.3
Id2	local sector 2	isic sector 2	0.7
Id2	local sector 2	isic sector 3	0.7

Step c) and d) Call Trade Weights and Calculate Accumulated Export Shares. In this example, ISIC sector 1 exports 1000, ISIC sector 2 exports 500 and ISIC sector 3 exports 2000. As local sector 1 is matched to only ISIC sector 1, ISIC sector 1 accumulated export

share is 1. As local sector 2 is assigned to both ISIC sector 2 and ISIC sector 3, different export values are assigned to each ISIC sector within local sector 2, then we calculate their respective export shares and accumulated export shares. If trade data were missing, the exports shares would be identical among ISIC industries within local industries.

Person Id	**local industry**	**isic industry**	**uniform**	**isic exports**	**export share**	**accumulated export share**
Id1	local sector 1	isic sector 1	0.3	1000	1	1
Id2	local sector 2	isic sector 2	0.7	500	0.2	0.2
Id2	local sector 2	isic sector 3	0.7	2000	0.8	1

Step e) Assign Industry. For ID1, as the uniform draw of 0.3 lies between 0 and 1, then isic sector 1 is assigned. For ID2, as the uniform draw of 0.7 lies between 0.2 and 1, then isic sector 3 is assigned.

Person Id	**local industry**	**isic industry**	**uniform**	**isic exports**	**export share**	**accumulated export share**	**final isic industry**
Id1	local sector1	isic sector 1	0.3	1000	1	1	isic sector 1
Id2	local sector2	isic sector 2	0.7	500	0.2	0.2	isic sector 3

Id2	local sector2	isic sector 3	0.7	2000	0.8	1	isic sector 3

The final dataset will now have one ISIC Revision 4 code per local industry code and will be at the person level.

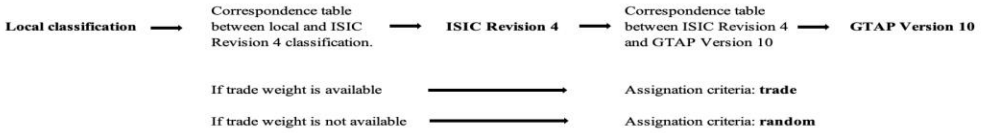
Person Id	**local industry**	**final isic industry**
Id1	local sector1	isic sector 1
Id2	local sector2	isic sector 3
Id2	local sector2	isic sector 3

This method assures that the assignation of industrial codes is random (that is representative of the economy of the country), and that the crosswalk between local industry codes and ISIC Revision 4 codes will be stable.

3.3.3 [add a subtitle]

An identical algorithm used to assign one ISIC Revision 4 code per local industry code is used to assign one GTAP Version 10 code per ISIC Revision 4 codes. The difference is that export values are now calculated within GTAP version 10 sectors rather than ISIC Revision 4 sectors. The example above applies by changing local industry to ISIC Revision 4 industry, and ISIC Revision 4 Industry to GTAP Version 10 sector.

Figure XX4. From Local Industrial Classification to GTAP Version 10 flow



3.3.4 From Local Occupational Classification to ISCO 08

The following algorithm is used to assign ISCO 08 occupation classifications to local occupation classifications.

a) Apply the concordance table between the local occupational classification to ISCO 08, this implies that the relation between occupation codes is potentially a many to many concordance, which means that multiple local codes can be assigned to multiple ISCO 08 occupational codes.

b) Draw a uniform.

c) Sort dataset according to the uniform's realization.

d) Keep the first observation for each individual.

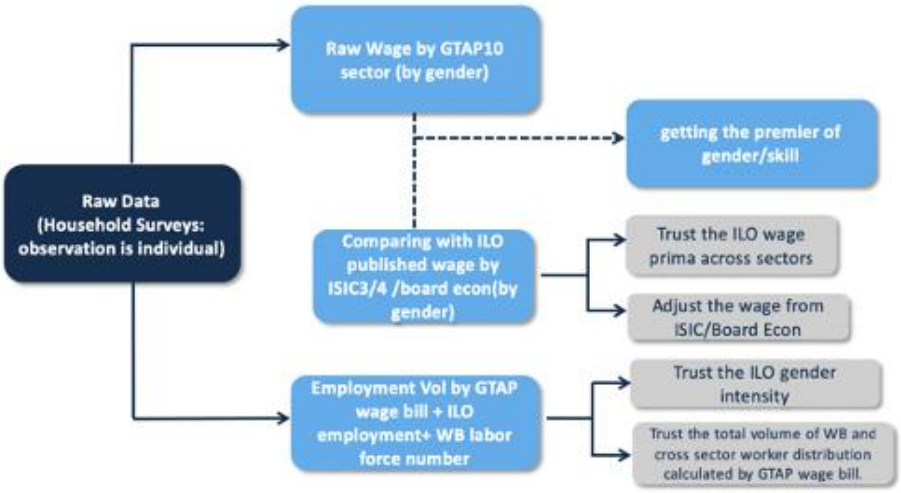
This assignation algorithm is completely random and does not take into account the productive structure of the country.

4 An application: Link with GTAP v10 database

Regarding the labor value-added, CGE model requires wages and total employment volumes for different types of workers (gender, skill) in each disaggregated industry for every country. This section covers key demand and technical aspects behind the construction of the GDLD and provides an overall perspective on the dataset's underlying advantages and caveats.

The data for wage and employment by type of workers (female, male, skilled and unskilled) are generated mainly based on the household surveys harmonized by The Gender Disaggregated Labor Database (GDLD) by 65 GTAP 10 sectors, supplemented with the harmonized earnings and employment distribution by International Labor Organization (ILO) and employment data from World Bank Open Data. Figure 1 shows the data processing.

Figure D.1. Procedure for Wage and Employment Volume



In micro dataset construction, our approach, at first, is going through as many resources as possible to check if household survey data and documentation available, especially for the local classification. Based on the data and documentation availability, GDLD harmonized 78 countries (survey datasets) with the identified disaggregation of industry to ISIC and GTAP classification and occupation, as well as 15 countries with LIS tabulation data sets, covering more than 70% GDP and almost 80% of population, respectively.

However, since household surveys are sampling surveys with national representative, it is more important that the responses given are accurate and thorough. The industry level in this paper is very disaggregated. Most sampling surveys are national representative, which will match the total population or other national wide feature. But when paying attention to a specific sector details, it cannot guarantee all disaggregated sectors are selected and all types for workers are interviewed (e.g. enough skilled female workers working in big farms).

Also, in recent years, decreasing response rates and data errors have challenged the usefulness of some surveys and resulted in lower quality data. For instance, some respondents give inaccurate information about their personal finances (esp. in wage) (Meyer, 2015).

The final goal of this session is preparing employment volume and average wages by 65 GTAP sectors of activities, by worker types (female skilled[1], male skilled, female unskilled and male unskilled) for 141 GTAP regions, matching the value-added macro data in GTAP database.

However, per the caveats of GDLD above, the household data is limited by

1) the coverage (not covering all 65 sectors, nor all four types of workers.) when calculating total number of workers;

2) and accurate in self-reported wage when calculation the average wage of every type of worker.

In order to solve both problems, on top of GDLD household survey data, we refer ILO employment and monthly earning data, World Bank Open Data and GTAP10 Value-added data as validation, to calculate the wage and total employment volume in every GTAP sectors and country, for four groups of workers: females skilled workers, males skilled workers, female unskilled workers and male unskilled workers.

The GDLD provides a global micro level dataset to make the tabulation of wage. However, this initial wage 1) comes from sampling survey and 2) is self-reported by the respondents. Including bias because of outliers and sampling bias as well as the reporting error. It's a

common issue of sampling survey methodology. One way to reduce errors in survey data is by linking this information to existing administrative or third-party data sets (Meyer, 2015). This would allow for an external validation of survey responses.

In this paper, we link the initial wage for GDLD household survey data to the labor data published by ILO. The underlying strategy is using the relative wage and skilled worker wage premia (the ratio of wage for skilled workers to that for unskilled workers) from household surveys for four worker groups cross 65 disaggregated sectors, and adjust the wage, to make sure same wage in the aggregated sector (21 sectors in ISIC Rev 4 1-digit level) are close to ILO's.

The ILO database systems compile the largest set of labor-specific statistics with global coverage. ILO published three gender, wage and employment related harmonized macro tables, including "Mean nominal monthly earnings of employees by sex and economic activity", "Employees by sex and economic activity (Thousands)" and "Employment distribution by economic activity (by sex)". Ideally, these tables are gender-ISIC specific tabulation with cleaned and reasonable data for every year (and wage in local currency and USD). However, some regions or years are not available for the full data or only harmonized to board economic activities.

149 countries are included by both "Mean nominal monthly earnings of employees by sex and economic activity[2]" table by ILO and GTAP 10 database, 134 out of which are following ISIC 1digit level industry sectors by gender. (Even the ILO was not aggregated to more disaggregated levels like 65 GTAP sector, nor by skill levels.).

There are 3 steps to calculate the average wage:

1. using GDLD surveys as GTAP regions, to calculate the wage per month in USD for four worker groups by GTAP 65 sectors in the country. Prior to the tabulation, we exclude the highest 0.15% and lowest 0.15% wage for every board economic activity in the country, i.e. keep the observations in three standard deviation.
2. then adjusting the wage per month in USD in step 1, to match the wage of ILO in the aggregated sector. For example, GTAP sector "coal", "oil" and "gas" could be aggregated to ISIC Rev 4 1-digit section "B: mining and quarrying". Suppose the females' mean wages for "coal", "oil" and "gas" are w1, w2, w3, respectively and mean of "mining and quarrying" is w from household survey, while the wage of ILO for "mining and quarrying" is w_ILO. The gap from GDLD household survey wage and ILO is w_ILO/w. Then we multiply the gap to w1, w2, w3, in order to make sure same average wage in the aggregated sector (mining and quarry) with ILO's.
3. substituting the missing by the ILO wages of aggregated sector. Using the example above, if the female wage of "coal" is missing, we use the ILO female wage for "mining and quarrying" to substituting the female workers' mean wage of "coal", and use the skill workers' wage premia of other non-missing sectors, i.e. "oil" and "gas".

In this section, for every country, every GTAP sector, we will get wages in 2014 US dollars for four worker groups, and skill worker wage premia for female and males.

The wage bill per sector per region in GTAP database indicates the value-added by labor, which is the total wage, i.e. mean wage times employment volumes in this sector of this region. Given the wage per sector above, we can easily calculate the total volume, simply dividing the wage bill of GTAP database by the mean wage of this sector. Then using the share of four work groups from GDLD household surveys, to make sure the total volume of aggregated sector is the same with ILO employment data[3], and total volume of the country is close to the labor force number in 2014 published by World Bank Open Data.

In this section, we used the household data in GDLD, ILO employment data and World Bank data in different dimensions that

1) within the sector, the share of four worker groups in GDLD household surveys per sector,

2) across sector, the relative volume as GTAP wage bill distribution (consider the wage above),

3) for the region, the total volume as the labor force of World Bank Open Data.

Even with the above action, there are still a lot of missing values in wage or share of four worker groups, if both household survey and ILO data of this country are not available. We can infer that, similar wage bill and capital value-added for the same sector in the close geographic area will have the similar volume share and wage premia. Thus, we use the regression of wage bill, capital value-added and geographic area to predict the approximately wage, wage premia and volume shares for four worker groups.

We have worked on making wage, wage premia and labor shares consistent. The data contains all 141 GTAP 10 regions. The GTAP 10 value added of labor has been split into volumes and wages for females, males, skilled and unskilled workers. The sum of volumes matches each country total labor force, as published by the World Bank.

[1] One difference between our database and GTAP is that skill levels are defined by years of schooling contrasting with a skill level defined by broad occupation category. In some applications with a CGE model (Bussolo et al., 2010) the workers' level of qualification were defined as skilled and unskilled, using an ad-hoc threshold of 9 years-of-schooling. In this paper, we use 9+ year schooling to define skilled worker in low and middle lower income countries and that of 13+ for high- and upper middle income countries.

[2]ilo.org/ilostat/faces/oracle/webcenter/portalapp/pagehierarchy/Page32.jspx;ILOSTATCOOKIE=Ax1Gk807yqspCaA-ceQguA65pOQOUnKISOuDmQxN-pCnXTKHdj7w!-1162039553?indicator=EAR_4MTH_SEX_ECO_CUR_NB&subject=EAR&locale=EN&datasetCode=A&collectionCode=YI&_adf.ctrl-state=u9cdm4wub_257&_afrLoop=105225482411893&_afrWindowMode=0&_afrWindowId=null#!%40%40%3Findicator%3DEAR_4MTH_SE

X_ECO_CUR_NB%26_afrWindowId%3Dnull%26subject%3DEAR%26locale%3DEN%26_afrLoop%3D1052254824118 93%26datasetCode%3DA%26collectionCode%3DYI%26_afrWindowMode%3D0%26_adf.ctrl-state%3Dm1xla4kvp_4

[3] . https://www.ilo.org/ilostat/faces/oracle/webcenter/portalapp/pagehierarchy/Page32.jspx?locale=EN&subject=EAR&indicator=EAR_4MTH_SEX_ECO_CUR_NB&datasetCode=A&collectionCode=YI&_afrLoop=85258474741968&_afrWindowMode= state%3Du9cdm4wub_257

5 Conclusion

The main contribution of the GDLD was indicated by its global coverage, which can be distinguished by more disaggregated levels of industry and workers' level of education.

In micro dataset construction, our approach, at first, is going through as many resources as possible to check if household survey data and documentation available, especially for the local classification. Based on the data and documentation availability, GDLD harmonized 78 countries (survey datasets) with the identified disaggregation of industry to ISIC and GTAP classification and occupation, as well as 15 countries with LIS tabulation data sets, covering more than 70% GDP and almost 80% of population, respectively.

However, since household surveys are sampling surveys with national representative, it is more important that the responses given are accurate and thorough. The industry level in this paper is very disaggregated. Most sampling surveys are national representative, which will match the total population or other national wide feature. But when paying attention to a specific sector details, it cannot guarantee all disaggregated sectors are selected and all types for workers are interviewed (e.g. enough skilled female workers working in big farms).

Also, in recent years, decreasing response rates and data errors have challenged the usefulness of some surveys and resulted in lower quality data. For instance, some respondents give inaccurate information about their personal finances (esp. in wage) (Meyer, 2015).

6 References

Aguiar, Angel, Maksym Chepeliev, Erwin Corong, Robert McDougall, & Dominique van der Mensbrugghe. "The GTAP Data Base: Version 10." *Journal of Global Economic Analysis* (forthcoming).

Annexes

Table XXX: The 65 sectors in GTAP 10P3

No	Codes	Description
1	pdr	Paddy Rice
2	wht	Wheat
3	gro	Cereal grains nec
4	v_f	Vegetables, fruit, nuts
5	osd	Oil seeds
6	c_b	Sugar cane, sugar beet
7	pfb	Plant-based fibers
8	ocr	Crops nec
9	ctl	Bovine cattle, sheep and goats, horses
10	oap	Animal products nec
11	rmk	Raw milk
12	wol	Wool, silk-worm cocoons
13	frs	Forestry
14	fsh	Fishing
15	col	Coal
16	oil	Oil
17	gas	Gas
18	omn	Minerals nec
19	cmt	Bovine meat products
20	omt	Meat products nec
21	vol	Vegetable oils and fats
22	mil	Dairy products
23	pcr	Processed rice
24	sgr	Sugar
25	ofd	Food products nec
26	b_t	Beverages and tobacco products
27	tex	Textiles
28	wap	Wearing apparel
29	lea	Leather products
30	lum	Wood products

31	ppp	Paper products, publishing
32	p_c	Petroleum, coal products
33	chm	Chemical and chemical products
34	bph	Pharmaceutical and medical products
35	rpp	Rubber and plastic products
36	nmm	Mineral products nec
37	i_s	Ferrous metals
38	nfm	Metals nec
39	fmp	Metal products
40	ele	Electronic equipment
41	eeq	Electrical equipment
42	ome	Machinery and equipment nec
43	mvh	Motor vehicles and parts
44	otn	Transport equipment nec
45	omf	Manufactures nec
46	ely	Electricity
47	gdt	Gas manufacture, distribution
48	wtr	Water
49	cns	Construction
50	trd	Trade
51	afs	Accomodation, food and beverage services
52	otp	Transport nec
53	wtp	Water transport
54	atp	Air transport
55	whs	Warehousing
56	cmn	Communication
57	ofi	Financial services nec
58	isr	Insurance
59	rsa	Real estate
60	obs	Business services nec
61	ros	Recreational and other services
62	osg	Public administration and defence, compulsory social security
63	edu	Education
64	hht	Human health and social activities
65	dwe	Dwellings

[1] Official documentation of GTAP v.10 database can be found at https://www.gtap.agecon.purdue.edu/databases/v10/index.aspx

[2] For example, Eurostat RAMON (Reference and Management of Nomenclatures), Index of Correspondance Tables, https://ec.europa.eu/eurostat/ramon/relations/index.cfm?TargetUrl=LST_REL&StrLanguageCode=EN&IntCurrentPage=9

[x] Aguiar, et al. (2019) provides additional information on GTAP 10.

[xx] For more information on Georgia's industry classification go the link: http://web.archive.org/web/20151023044140/http://unstats.un.org/unsd/cr/ctryreg/ctrydetail.asp?id=1120

[x1] This dataset is publicly available in this webpage https://comtrade.un.org/Data/.

Table 1 Household Survey List in The Gender Disaggregated Labor Database

Country Code	Country Name	World Bank Region	Survey Name	Survey year	GDP[a] share	Population[b] share
AUS	Australia	East Asia & Pacific	HILDA	2015	1.67	0.33
CHN	China	East Asia & Pacific	CGSS	2013	15.86	18.36
FJI	Fiji	East Asia & Pacific	HIES	2008	0.01	0.01
IDN	Indonesia	East Asia & Pacific	SAKERNAS	2009	1.21	3.50
KHM	Cambodia	East Asia & Pacific	CLFCLS	2012	0.03	0.21
MNG	Mongolia	East Asia & Pacific	LFS	2014	0.02	0.04
PHL	Philippines	East Asia & Pacific	LFS	2013	0.39	1.39
SLB	Solomon Islands	East Asia & Pacific	HIES	2005	0.00	0.01
THA	Thailand	East Asia & Pacific	HSES	2011	0.59	0.91
TLS	Timor-Leste	East Asia & Pacific	LFS	2010	0.00	0.02
VNM	Vietnam	East Asia & Pacific	LFS	2010	0.29	1.27
AZE	Azerbaijan	Europe & Central Asia	AMSSW	2015	0.05	0.13
BLR	Belarus	Europe & Central Asia	LFS	2016	0.07	0.13
GEO	Georgia	Europe & Central Asia	HIS	2013	0.02	0.05
HUN	Hungary	Europe & Central Asia	HBS	2008	0.18	0.13
MDA	Moldova	Europe & Central Asia	LFS	2015	0.01	0.05
MNE	Montenegro	Europe & Central Asia	LFS	2011	0.01	0.01
POL	Poland	Europe & Central Asia	HBS	2011	0.68	0.50
RUS	Russian Federation	Europe & Central Asia	RMLS	2016	1.93	1.91
SVN	Slovenia	Europe & Central Asia	HBS	2004	0.06	0.03
TJK	Tajikistan	Europe & Central Asia	JMSC	2013	0.01	0.12
TUR	Turkey	Europe & Central Asia	HLFS	2015	0.89	1.07
XKX	Kosovo	Europe & Central Asia	LFS	2014	0.01	0.02
ARG	Argentina	Latin America & Caribbean	EPHC_2	2014	0.60	0.59
BOL	Bolivia	Latin America & Caribbean	EH	2015	0.05	0.15
BRA	Brazil	Latin America & Caribbean	PNAD	2015	2.18	2.77

CHL	Chile	Latin America & Caribbean	CASEN	2015	0.35	0.24
COL	Colombia	Latin America & Caribbean	GEIH	2014	0.38	0.65
CRI	Costa Rica	Latin America & Caribbean	ENAHO	2012	0.07	0.06
DOM	Dominican Republic	Latin America & Caribbean	ENFT	2015	0.09	0.14
ECU	Ecuador	Latin America & Caribbean	ENEMDU	2015	0.13	0.22
HND	Honduras	Latin America & Caribbean	EPHPM	2014	0.03	0.12
HTI	Haiti	Latin America & Caribbean	EEEI	2007	0.01	0.15
MEX	Mexico	Latin America & Caribbean	ENIGH	2010	1.43	1.71
NIC	Nicaragua	Latin America & Caribbean	EMNV	2014	0.02	0.08
PER	Peru	Latin America & Caribbean	ENAHO	2015	0.26	0.43
SLV	El Salvador	Latin America & Caribbean	EHPM	2014	0.03	0.08
URY	Uruguay	Latin America & Caribbean	ECH	2015	0.07	0.05
DJI	Djibouti	Middle East & North Africa	EDESIC	2015	0.00	0.01
EGY	Egypt	Middle East & North Africa	ELMPS	2005	0.29	1.29
IRQ	Iraq	Middle East & North Africa	HSES	2012	0.26	0.51
JOR	Jordan	Middle East & North Africa	LFS	2016	0.05	0.13
LBN	Lebanon	Middle East & North Africa	LBN	2011	0.07	0.08
MAR	Morocco	Middle East & North Africa	ENSLE	2009	0.14	0.47
TUN	Tunisia	Middle East & North Africa	HBS	2010	0.05	0.15
USA	United States	North America	CPS	2018	23.89	4.31
AFG	Afghanistan	South Asia	ALCS	2013	0.02	0.47
BGD	Bangladesh	South Asia	HIES	2010	0.32	2.18
BTN	Bhutan	South Asia	BLSS	2017	0.00	0.01
IND	India	South Asia	NSS_SCH10	2011	3.18	17.74
LKA	Sri Lanka	South Asia	HIES	2016	0.10	0.28
MDV	Maldives	South Asia	HIES	2009	0.01	0.01
NPL	Nepal	South Asia	LSS	2010	0.03	0.39
PAK	Pakistan	South Asia	LFS	2014	0.36	2.61
AGO	Angola	Sub-Saharan Africa	CENSUS	2014	0.12	0.39
BWA	Botswana	Sub-Saharan Africa	BCWIS	2009	0.02	0.03

ETH	Ethiopia	Sub-Saharan Africa	UEUS	2016	0.10	1.39
GMB	Gambia	Sub-Saharan Africa	IHS	2015	0.00	0.03
KEN	Kenya	Sub-Saharan Africa	IHBS	2005	0.10	0.66
LSO	Lesotho	Sub-Saharan Africa	HBS	2010	0.00	0.03
MLI	Mali	Sub-Saharan Africa	EPAM	2010	0.02	0.25
MOZ	Mozambique	Sub-Saharan Africa	IOF	2014	0.02	0.39
MUS	Mauritius	Sub-Saharan Africa	HBS	2012	0.02	0.02
MWI	Malawi	Sub-Saharan Africa	LFS	2013	0.01	0.25
NAM	Namibia	Sub-Saharan Africa	LFS	2014	0.02	0.03
NER	Niger	Sub-Saharan Africa	ECVMA	2014	0.01	0.28
RWA	Rwanda	Sub-Saharan Africa	EICV	2013	0.01	0.16
SDN	Sudan	Sub-Saharan Africa	NBHS	2009	0.05	0.54
SLE	Sierra Leone	Sub-Saharan Africa	LFS	2014	0.00	0.10
SOM	Somalia	Sub-Saharan Africa	HFS	2016	0.01	0.20
SSD	South Sudan	Sub-Saharan Africa	SSD	2009	0.00	0.17
SYC	Seychelles	Sub-Saharan Africa	HBS	2006	0.00	0.00
UGA	Uganda	Sub-Saharan Africa	UNHS	2016	0.03	0.57
ZAF	South Africa	Sub-Saharan Africa	QLFS_Q1	2017	0.43	0.75
ZMB	Zambia	Sub-Saharan Africa	LCMS	2015	0.03	0.23
ZWE	Zimbabwe	Sub-Saharan Africa	LFS	2011	0.04	0.22
AUT	Austria	Europe & Central Asia	LIS		0.53	0.12
CHE	Switzerland	Europe & Central Asia	LIS		0.82	0.11
CZE	Czech Republic	Europe & Central Asia	LIS		0.28	0.14
DEU	Germany	Europe & Central Asia	LIS		4.66	1.10
DNK	Denmark	Europe & Central Asia	LIS		0.41	0.08
EST	Estonia	Europe & Central Asia	LIS		0.04	0.02
FIN	Finland	Europe & Central Asia	LIS		0.32	0.07
GBR	United Kingdom	Europe & Central Asia	LIS		3.29	0.88
GRC	Greece	Europe & Central Asia	LIS		0.25	0.14

LTU	Lithuania	Europe & Central Asia	LIS	0.06	0.04
LUX	Luxembourg	Europe & Central Asia	LIS	0.08	0.01
SVK	Slovak Republic	Europe & Central Asia	LIS	0.12	0.07
GTM	Guatemala	Latin America & Caribbean	LIS	0.09	0.22
PRY	Paraguay	Latin America & Caribbean	LIS	0.05	0.09

Source: [a]GDP (Current US$), World Bank Open Data, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?

[b] Population, total, World Bank Open Data, https://data.worldbank.org/indicator/SP.POP.TOTL