Global Repository of Income Dynamics (GRID)

Motivation, Goals and Expansion Plan of the Project

1. Introduction

The goal of the **Global Repository of Income Dynamics** (henceforth GRID) project is to build a "live" open-access database that will allow researchers to retrieve, free of charge, a large number of harmonized statistics about the distribution of individual income levels and individual income changes for many countries, many years, and many sub-populations.

Over the last two decades, the interest in inequality has grown enormously in academia, policy circles, and popular media. This interest is motivated by many factors: the large increase in income concentration at the top of the distribution, its global nature (affecting countries as diverse in their institutions as the US, Sweden, or China), the observation that higher inequality has led, in some countries, to lower economic mobility and, finally, the concern that extreme inequality may distort the political process and undermine the proper functioning of democracies. Documenting patterns of inequality and mobility over time, between groups, as well as differences across countries requires researchers to access reliable big data on the income distribution and on its dynamics.

The GRID facility contains individual earnings statistics based on longitudinal data drawn from administrative (tax or social security) records for a large set of countries. The database will be regularly updated (to include new years of data and new statistics for existing countries), and quickly expanded (to include data for additional countries). So far, we have assembled 13 country teams comprising about 50 first-rate economists representing the US, the UK, Canada, France, Germany, Sweden, Norway, Denmark, Italy, Spain, Mexico, Brazil, and Argentina.

Country teams have produced research papers that summarize findings on inequality, volatility and mobility for their country based on these harmonized statistics, as well as country-specific original data analyses. A special issue of *Quantitative Economics* (a journal of the *Econometric Society*) hosting articles for each of these 13 countries and an introductory article presenting the project and the database is slated for publication in August 2022.

Further, we have received expressions of interest from research teams able to provide data that satisfy the high standards and harmonization criteria we set for Australia, Taiwan, Israel, Netherlands, Portugal, Greece, Japan, Costa Rica, and Chile among others. We plan to reach out systematically to other country teams. We have created a master code, available on our GitHub repository, that produces a long list of micro statistics in a

harmonized fashion. This resource makes it very easy for other country teams to enter the project.

Expansion plans for the database, as we explain in detail below, also include the construction of household-level statistics and statistics for earnings after taxes and transfers (both already feasible for a subset of countries). These expansion plans are important as they help keeping the pulse of many economies as they go through business cycles, policy changes, and structural reforms, and to facilitate comparative analyses across a larger set of economies.

In order to build the database and a user-friendly website front-end to access the data, we have joined forces with the University of Minnesota's Institute for Social Research and Data Innovation (ISRDI). ISRDI has built and maintains some of the largest Social Science databases in the United States, including IPUMS and the Minnesota Population Center (MPC), so they have proved extensive expertise in running projects like ours.

As we elaborate below, a key distinguishing feature of our proposed archive is its longitudinal (panel) dimension. Using large administrative panel data to study the dynamics of the income distribution allows researchers to go beyond commonly analyzed cross-sectional inequality statistics based on current income. In particular, the panel dimension allows to document properties of income volatility (both aggregate and idiosyncratic), understand the nature of income risk (e.g., the size, persistence and origin of individual income shocks), and estimate intra- and inter-generational mobility across the income distribution. It also allows comparing inequality trends in current income, the most commonly used indicator, to trends in measures of permanent income, which require availability of longitudinal data to be constructed.

The *administrative nature of the data* reduces concerns about measurement error. It therefore allows the database to produce a wide range of well-measured non-parametric (quantile-based) statistics that are informative about the shape of the distribution of earnings changes (whether they are left or right skewed, fat-tailed, leptokurtic, etc.) and analyze the nature of tail shocks, which is all but infeasible to study with small survey-based datasets. In this sense, the database builds on the latest developments in the income dynamics literature that emphasizes important nonlinearities and non-normality in earnings dynamics. Furthermore, the *granular nature of the underlying microdata* (millions of observations per year) allows to analyze dynamics by finely selected sub-groups of the population stratified by age, education, gender, race, income group, and geographical location. The ability to stratify by observable characteristics allows studying the heterogeneity in inequality and volatility measures across sub-populations, e.g., assess whether inequality and economic trends have evolved differently for more or less educated workers, for poor or rich individuals, for men or women, for large demographic groups or minorities. It helps shedding light on whether economic volatility in earnings,

asymmetry of income changes (i.e., wage losses or wage hikes), life-cycle income growth, or wage rigidity are more pronounced for certain groups than others. These are all key inputs into public policy decisions that, beyond economic efficiency, take into consideration redistribution and social insurance. The ability to compare *harmonized statistics across countries* can allow researchers to study the importance of institutions vs. the role of market forces, especially during economic downturns, and the effects of structural reforms or policy changes across the distribution.

What makes this project feasible (and hence, in our view, timely) is the recent rapid rise in the availability of large panel data sets from administrative sources covering millions of workers over several decades and providing rich and accurate data on earnings and other aspects of economic interest. While twenty years ago such data existed only for a handful of countries, there is now a much larger global coverage. No cross-country harmonized database of the kind we envision for this project currently exists. There are, however, a few that share similarities, and we explain below how our database differs from existing ones. In short, three features will set our project apart from what is already available: the focus on the *dynamics* of the earnings distribution, the *administrative* nature of the data (as well as the effort to harmonize data across countries), and the focus on *heterogeneity*.

Users accessing our database are able to download micro statistics organized by country and by finely selected subpopulation groups, such as by age, gender, year, income percentile, and combinations thereof. All the statistics present in our database (mean, variances, percentiles, mobility measures, auto-covariance matrices, etc.) are directly obtained from the raw micro (i.e., individual) data. For some countries, additional heterogeneity domains (such as race in the US, education in most Scandinavian countries, and occupation in France and Italy) will be added.

Below, we further articulate our motivation and goals and our expansion plans.

2. Motivation for our project

There already exist a number of publicly available databases that collect data on income distribution: the most prominent example is the World Income Database (WID). Additional examples include the WIID2 database at the UNU; the Luxembourg Income Study; data sets hosted at the World Bank; and various other OECD databases.

Given the similarities, it is useful to clarify the differences and explain the void that GRID intends to fill. The existing databases have three defining characteristics. They document cross-sectional, rather than dynamic aspects of the distribution; they are (mostly but not always) based on survey data (WID is, in fact, a hybrid of administrative and survey data, which may make cross-country comparisons complex); and they mostly provide aggregate (i.e., time series) statistics, such as, say, the top 1% share of income for various calendar

years, or the Gini coefficient. In contrast, our database is informed by the goal of wanting to expand/focus on three relevant aspects: (a) the dynamics, or longitudinal dimension of the micro data, (b) the administrative nature of the data, and (c) the heterogeneity within the population. We elaborate on these aspects in turn.

2.1. Dynamics

The longitudinal nature of the datasets in GRID is the key dimension of originality relative to the World Inequality Database (WID), which is entirely cross-sectional. The WID has been a source of inspiration and is serving the research community extremely well. At the same time, cross-sectional inequality measures tell only part of the story because they do not contain information on the dynamics of individual earnings and on the degree of mobility of individuals across the distribution, which requires tracking individuals over time. This distinction is key for any welfare analysis and for modulating social insurance programs to the persistence of income shocks faced by individuals in the labor market. A classic example that illustrates the pitfall of cross-sectional snapshots is the inability to draw welfare inferences from a constant poverty rate. A constant 10% poverty rate across two subsequent years is compatible with a 10% of the population being permanently poor or with the entire population facing every year a 10% chance of falling into poverty (as well as the more realistic spectrum of intermediate cases). With cross-sectional data the two cases are indistinguishable; with panel data, one can follow the fortunes of people over time and the distinction becomes immediate.

Our project is based on administrative data sources (available for all countries) which all have an explicit longitudinal component, expanding the number of dimensions available for analysis. Since units are followed over time, this will allow for the study of additional aspects of economic welfare including earnings volatility, earnings mobility across the income distribution, and tail behavior. These three sets of statistics may reveal important features of the distribution of earnings. Earnings volatility (for example, the variance of changes in earnings from one year to the next) is often interpreted as a metric for the extent of risk and uncertainty workers face in the labor market, especially if changes are involuntary (due to, e.g., job losses, wage cuts, temporary unemployment, health shocks, etc.). Earnings mobility is important for understanding whether movements across the earnings distributions persist or quickly fade away. It is also essential to distinguish inequality in current income from inequality in permanent income, and compare these two indicators across countries. Finally, tail behavior (the probability of large positive or negative income changes due to wage hikes vs wage losses; the asymmetry of the two type of changes; etc.) and more generally departure from normality allow for a more sophisticated notion and interpretation of economic risk. Our data archive would complement the WID by providing the crucial longitudinal dimension in the analysis of the evolution of world income inequality that allows researchers to dig into all these aforementioned features.

2.2. Administrative data

All the statistics that we make available on the website come from administrative archives (tax data or social security data). The pre-condition for "entering the database" is that such individual level administrative data are readily available to the contributing team, they can be made available in some more aggregated form to outside researchers at no cost, and they can be regularly updated and uploaded to the website once new waves of data come along.

Studying features of the income distribution with administrative data offer several advantages over survey data. By their nature survey data suffer from sample attrition, measurement error, and lack of representativeness of the tails (especially at the top).

Moreover, because of their small size, survey data run into statistical power problems when trying to produce non-parametric analyses. Most of these issues are not present with administrative data. Administrative data collect info on the universe of workers, so attrition is not an issue (besides death or migration); there is minimal measurement error if misreporting is sanctioned/punished by the tax authority or social security administration; the availability of millions and millions of data allows granular analyses of data for subpopulations and estimation of higher moments of the data (such as the evolution of the top 0.01% share or kurtosis statistics, say) have statistical meaning and are reliable.

Obviously administrative data have their own limitations. Relative to survey data, administrative data may suffer from under-representation at the bottom in countries where the informal sector is significant. We have instructed teams from such countries to combine administrative and survey data to assess the extent of this bias on the statistics of interest.

Another disadvantage of administrative data is that they tend to collect extremely detailed, measurement error-free information on a single domain (such as income). This disadvantage is attenuated by the ability of linking multiple sources of administrative information through identifiers (like SSN's). This flexibility gives researchers the power to combine multiple administrative data sources as a tool for real-time policy analyses and response, as demonstrated in many countries during the Covid-19 crisis.

Administrative data are also now becoming widespread. An increasing number of empirical papers published in high-quality journals use administrative data in some form or another. The growing availability of administrative data across countries allow comparative analyses that use somewhat more homogenous populations than possible if using survey data. Joint use of administrative statistics and survey data statistics for many countries can also be employed to verify whether discrepancies between data sources on

some relevant aspects of the distribution of earnings are a common or a country-specific phenomenon.

A final consideration is that access to administrative micro-data (especially for countries different from the country of birth/residence) may be prohibitive for most researchers. Our vast menu of income statistics for many sub-populations, comparable across countries, would make access to the underlying to micro data no longer necessary for many users.

2.3. Heterogeneity

The heterogeneity dimension allows us to dig deeper beyond full population averages to explore employment rates (by part-time and full-time when available), income inequality, volatility and mobility within specific year of birth cohorts as well as by key observables such as age, gender, race, education (when available), geographical location and permanent income (a proxy for skill levels). Many studies document that income inequality, risk and mobility vary substantially across groups. Including the ability to separate the data by cohort, gender, education, permanent income, geographical location, etc. allows us to study the dynamics of these specific groups.

As mentioned above, existing databases provide mostly aggregate statistics for the entire population each year. The WID database, for example, contains no stratification by observable characteristics. Crucially, our inclusion of separate statistics by demographic group will allow researchers to separate ex-ante heterogeneity in types from ex-post uncertainty in outcomes, and to understand how ex-ante heterogeneity and ex-post labor market risk vary over the business cycle and over the long-run. Each of these added dimensions is a crucial input into the design of stabilization, redistributive and social insurance policies.

2.4. Harmonization

An important goal of the GRID project is to produce statistics that are as comparable as possible across countries. Harmonization is an inherently challenging task given the differences in variable definitions and data collection methods in different countries. We have spent a great deal of effort to harmonize the statistics taking as given the underlying datasets.

The initial planning of the project lasted about 3 months and was entirely devoted to harmonization efforts and to compiling the list of statistics that all country teams would be able to produce. These steps involved several rounds of iteration between the PI's and the country teams. We started by collecting key pieces of information about each country's dataset, including the time span, the cross-sectional sample size in each year, the different measures of earnings available (and each subcomponent included in each measure), the statistical unit of analysis (i.e., whether data were reported at the individual

or household level), presence of top or bottom coding or any changes in variable definitions or sample design over time, and whether additional variables were available (such as data on capital income, self-employment income, education, occupation, hourly wage, hours or weeks worked, employer information, and geographic location, among others).

Using this information, we developed a framework to allow maximum harmonization while providing as much useful information as possible. To give an example, for the heterogeneity statistics by group where the data are pooled across years, each country uses a 20-year panel that ends in the last year of data available. For 11 of the 13 countries this date is between 2014 and 2017, giving 17 years of overlap between these countries samples. At the same time, many of these datasets go back more than 20 years, so for time-series statistics, we ask each country to use the longest available sample (subject to stable variable definitions and other details) since this does not affect comparability.

The definition of earnings also requires careful consideration for harmonization. Many country datasets do not have data on self-employment income, while most of them have a fairly comprehensive wage earnings measure that includes bonuses, overtime pay, and other irregular pay (like tips in the US and paid sick leave in several European countries). Unemployment and welfare benefits are not available in most countries, so we excluded them from the baseline earnings measure for the main database.

A second important feature of the GRID project that allows further harmonization is that all statistics for all countries are produced by one unique master code that runs in Stata. This master code generates the list of statistics we have compiled (about 1.5M data points per country) with feedback from and discussion with country teams and outside researchers. The master code has been written by two team members who have extensive expertise in this area, Serdar Ozkan and Sergio Salgado. The code is adjusted minimally and when really needed (e.g., if a country's dataset has a peculiarity not considered in the code) and such adjustments happen in consultation with us. This master code infrastructure allows further harmonization by eliminating spurious differences in statistics across countries that often result from different teams or statistical agencies writing their own code. Furthermore, it ensures that a long list of small but potentially critical steps are carried out the same way in each country (such as the thresholds used to trim the sample at the bottom, etc.).

3. Expansion plans

The first phase of the project, which is now concluded, entailed designing the website and incorporating the data for the initial set of countries. In subsequent phases we plan to extend the database in three directions: (a) additional earnings statistics, (b) additional

income concepts, (c) additional countries, and (d) additional visualization functionality to the website.

(a) Additional earnings statistics will come partly from modulating demand and supply. The website includes a "suggestion box" designed to improve the content, format and overall users' experience on the website. In addition, the website includes a "customized data request box" where we allow researchers to formulate *ad-hoc* requests about specific samples, moments, or countries. Our project manager will forward the request to our contact person in the relevant statistical agencies and together they will assess whether these additional *ad-hoc* statistics can be provided to the researcher and, eventually, incorporated in the database for other countries as well. The availability of a master code that runs exactly in the same way for all countries makes integrating new statistics in the existing website relatively straightforward and inexpensive.

(b) We also plan to look into the possibility of offering statistics for income concepts broader than earnings. For example, many of our administrative data sources already include information on taxes and government transfers, which would allow us to document the joint dynamics of gross labor income and disposable income. For other countries, it may be possible to build the equivalent of a TAXSIM program. Furthermore, many administrative data already allow to link individuals belonging to the same household and construct household-level income statistics. Once the project is under way, we plan to perform these links and document, as much as possible, similar stylized facts at the household level. More ambitious, but not entirely out of reach, is the goal of adding information on the distribution and dynamics of consumption expenditures, at least for a subset of countries in the database.

(c) Given the interest our project has already attracted, we also expect to roughly double the number of countries within a year from the launch of the initial database.

(d) Finally, the next development phase of the website will be focused on visualization. Users will be able to plot data of interest interactively, before downloading the underlying data files.

Fatih Guvenen

Luigi Pistaferri

Gianluca Violante