JoIn Methodololgy Note

Jobs Indicators Database (JoIn)

The Global Jobs Indicators Database (JoIn) presents 61 labor supply indicators for 150 countries and was compiled from more than 1200 surveys. It includes data at the national level, for the overall population as well as disaggregated by age (youth/adults), gender (male/female) and geography (urban/rural).

In order to test the quality of the data and detect potential errors in the indicators, we designed a system of quality checks, following the "Q-check" approach of the Poverty team. The system consists on six independent basic quality checks that detect anomalies in the data. Each check produces a positive or negative result at the smallest possible level of dissaggregation of the data, that is, for a particular country/year/survey/indicator. Intuitively, this represents a "cell" in the dataset. A positive result (a flag) suggests that there might be an issue wih the data, a negative result (no flag) means that no problem has been detected and a null result (NA) is returned when the check is not applicable for that particular data-point.

The results of the checks are then combined to classify data issues in three groups: potential weights problems, potential sampling problems and potential derivations problems.

In this document we present each one of the checks and explain how to interpret the combinations of checks.

Basic Quality Checks

Check 1. Ranges

The first check, named "Ranges", flags data points-that fall outside of their expected range. It is applied to all (61) indicators, although the expected range varies by indicator.

Appendix A contains the complete of the indicators in the dataset, along with their defined range.

Check 2. Derivations

Some indicators can be calculated as a function of other indicators included in the dataset. We call these indicators "derived indicators", and they are calculated directly from the microdata. However, we could calculate them from the indicators that definde them, and the result should be the same.

For instance, the dependency ratio is defined as the ratio of dependent population (children and elderly) over working age population (youth and adults). Since our database includes the population shares, we can manually calculate the dependency ratio from them. This should equal the reported dependency ratio. The "Derivations Check" detects cases where this does not happen, because of accumulated rounding error or otherwise.

It does so by flagging indicators that fulfill the following two conditions:

1. Differ from the manually computed value by 10% or more

2. Differ from the manually computed value by 0.1 units or more.

We take into account both the relative and the absolute distances to ensure that we do not flag indicators with a very small value that differ from their manually computed number by a large amount in terms of percentages but much less if we look at the absolute difference (for instance, if a dependeny rate were reported to be 0.18 and its manually computed value were 0.10, the percentage difference would be 80% but it would only represent a distance of 0.08 units).

The five indicators checked in Check 2 are listed in Appendix B, along with their definitions in terms of other indicators.

Check 3. Sums

Check 3 verifies that groups of indicators that are computed as fractions of a total are indeed equal to one. For instance, the economic activity shares (percent employment in agriculture, industry and services) should add up to one.

The check flags all indicators taking part in a sum, when the sum differs from one by 10% or more.

Appendix C presents a list of the indicators to which check 3 is applied.

Check 4. External data comparison

Check 4 compares the indicators in the database to equivalent indicators from other sources, and flags them if they differ from them by 10% or more.

The main origin of the external data is the World Development Indicators (WDI) database, accessed through the Stata package `wbopendata`. For indicators not available in WDI, we use ILOSTAT, accessed through the SMDX connector `sdmxHelp`.

Appendix D contains a list of all indicators and the code of the external variable used for each external comparison.

Check 5. Outliers

Check 5 detects exceptional rates of growth of an indicator between consecutive years, and flags them.

This check is limited to series that are at least two data-points long and to indicators that are *not* measured in shares, that is, all counts and ratios.

The rate of growth is calculated following the usual formula and adjusting for years between consecutive data-points $\frac{x_t - x_{t-1}}{x_{t-1}} \cdot (y_t - y_{t-1})$, where x_t is the value of an indicator at time t and y_t is year t. A data-point will be considered an outlier and flagged in the following cases:

1. If the adjusted absolute rate of growth between consecutive data-points is equal or greater than 1, which corresponds with an increase or decrease of 100%, *and* it is not followed by another jump of 100% growth or more in the opposite direction. For long enough series, this corresponds to a break in the series.
2. If the adjusted absolute rate of growth between consecutive data-points is greater than 100% and is followed by another adjusted absolute rate of growth greater than 100% in the opposite direction. This corresponds to a "pure" outlier, which is at a large enough distance of the trend.

A list of the indicators covered by Check 5 can be found in Appendix E.

Check 6. Tails

The last Tails check detects extreme values in the entire distribution of each indicator, regardless of what country or year they belong to. It aims to complement the Outlier test (Check 5) by checking all data-points, including indicators not covered by Check 5 and data points that are left out of it because they are not part of a long enough series.

To detect extreme observations, we first sort each indicator by its value, from lowest to highest, disregarding the survey where the data come from. Then, we compute the rate of growth between consecutive data-points. In general, we should expect very low rates of growth between observations. However, extreme observations are characterized by being having a significantly higher or lower value than the rest. If we think in terms of the shape of the distribution, extreme observations will be at the tails of the distribution (and hence the name of this check). This implies that the absolute rate of growth between an extreme observation and its preceding one will be very high. We follow this principle to flag observations that have a rate of growth higher than 100 times the average of that indicator.

This check is applied to all (61) indicators in the database.

Combinations of checks: buckets

We use the results of the checks and the analysis behind them to categorize potential problems in the data. This section describes how we classify surveys into three categories.

Weights problems

Survey weights are used to obtain estimates of population parameters for a given survey. When the weights are incorrect, the population parameters can take unreasonable values. We follow this idea to classify as having weights problems surveys where the population estimates are very far away from the external data.

We differenciate between “pure” weights problems, where the population shares are correct but not the aggregate value, and “mixed” weights problems, when the survey presents incorrect aggregate population value as well as incorrect population shares.

“Pure” weights problems satisfy the follwing conditions:

1. The population variable is flagged in Check 4
2. No population share is flagged in Check 4

We mark a survey as having “mixed” weights problems if:

1. It is not categorized as having potential "pure" weights porblems
2. The population variable is differs by more than 50

Sampling problems

The sampling of a survey is the selection of a subset of the population that will be used to estimate characteristics of the whole population. A sampling that is not representative of the entire population might bias certain indicators.

We indentify surveys with potential sampling problems if any of the population shares satisfies the following conditions:

1. Differs by 50% or more from the external data and
2. Differs by 0.1 units or more from the external data

Just like we did for the “Derivations” check, we do not rely exclusively on relative (percent) distances in order to avoid flagging surveys where the absolute distance is tiny.

Derivation problems

The derivation problems are defined as a direct application of Check 4. We say a survey presents potential derivations issues when any of the “derived” indicators (Appendix B) are flagged.