GPP Quantitative data validation and cross-checks
Step-by-step guide

Introduction

The data validation process for the EU Subnational project involves three individual and parallel sub-processes:

1. One-on-one interviews with experts
2. Quantitative data validation and cross-checks
3. Qualitative data validation

The one-on-one interviews with experts consist of presenting key findings and main messages from the QRQ data to experts and evaluating their consistency through focused interviews. The quantitative data validation and cross-checks consist of using statistical techniques to evaluate the consistency of our most recent data with information gathered during previous as well as in relation to data collected by third-party sources. Finally, qualitative data validation consists in outlining the overall political context in each country to assess if the collected data is aligned with the current political background.

This document outlines the step-by-step to follow during the quantitative validation process for the General Population Polls (GPP). The document presents an overview of the process with a description of the tasks needed for its successful implementation.

Overview of the process.

In overall terms, the validation process will be implemented in two phases:

1. Country data validation.
2. Outcomes validation.

1. Country data validation.

When the polling companies first submit their data, we only count on the data from single countries or a single polling company at the best. In other words, we cannot see the whole picture yet. During this phase, we focus on validating individual countries, and we perform two different checks:

- *Representativeness checks*
- *Patterns and correspondence checks*

For the ***representativeness checks***, a set of three tests will be applied to evaluate the consistency and representativeness of the survey data:

- Checking the sample distributions by sociodemographic group. *During the pretest, the analysis is performed only at the country level, while during the full fieldwork submission we extend the analysis to the NUTS region level.*

- Evaluating the missing values per variable. *During the pretest, the analysis is performed only at the country level, while during the full fieldwork submission we extend the analysis to the NUTS region level.*

- Reviewing the difficulty index score per observation (only for face-to-face surveys).

For the ***patterns and correspondence checks***, we will perform an internal and external validation.

- The *internal validation*.
- The *external validation*.

a. Internal validation

Only applied during the full fieldwork submission. The internal validation consists of evaluating the consistency of the country data with previous GPP data for that same country. Therefore, the analysis that will be conducted is as follows:

- T-test applied to a set of key variables to identify significant changes across time. Data from previous rounds of the GPP will be used during this step.
- Create regional rankings of key variables to gain a regional perspective on the performance of each region. Regional outliers will be flagged for their review.

The result of the internal validation will be a summary table listing which questions presented issues according to the previous steps.

b. External validation

Applied during both, pretest and full fieldwork. The external validation consists of evaluating the consistency of the country data with the most recent available data from the mapped

Third-Party Sources (TPS). This is usually the stage at which the implementation becomes a burden due to its complexity.

The main advantage of working with data from the European Union is the massive volume of information available. The European Union has a huge set of survey data available: (flash, special, and standard) Eurobarometers, European Social Survey, European Value Survey, Fundamental Rights Survey, among many others. On top of that, every country in the European Union has a good set of past data for the Rule of Law Index. Therefore, the main challenge of designing a validation process for the EU Subnational project is to develop a validation process that considers the vast amount of information, but it does not become a huge burden in its implementation.

To tackle this dilemma, we propose the use of a *multi-level validation process*. Under this approach, the analysis starts with a general overview of the data in its first level. If- *and only if* - issues are found in this level, a second and more in-depth level is activated, and so on. For this project, we propose a three-level system.

i. *First level: Analysis of the aggregated scores*

At the end, the external validation process is a *flagging system,* and most of the burden in its implementation is due to the huge number of individual flags to review. We propose to start the analysis by reducing the number of flags. In other words, to reduce its dimensionality. Instead of analyzing individual questions per topic, we propose to analyze the aggregated behavior of groups of questions. In this case, sub-pillars.

We compare aggregated scores to the most recent scores from the Rule of Law Index and to aggregated scores per sub-pillar from our TPS dataset. The flagging system works as follows:

- A sub-pillar is flagged if the difference between scores is higher than 10 points.
- A pillar is flagged if at least half of the sub-pillars within that pillar are flagged.
- A polling company is flagged if at least half of the pillars are flagged.

ii. *Second level: Analysis at the question level*

Each pillar will have a set of “*quasi-comparable*” matches. In other words, we will identify which questions from the EU-GPP questionnaire have similar questions in our

TPS database. We will analyze these matches if a pillar is flagged during the first-level analysis. In this case, we review all the questions from that pillar.

iii. *Third level: Extreme outlier detection*

If a question presents an extreme deviation from their benchmark, an in-depth review of that question will be performed. By an extreme deviation, we refer to a 25-point deviation or more in their score.

2. *Outcomes validation*

The final stage in our data verification process involves validating the estimated outcomes constructed with our dataset. This process is designed to ensure the confidence and accuracy of the data we present, aligning it with the patterns under analysis. The expected result from this stage is a document containing recommendations on what should be published and what should be withheld based on both internal and external analyses.

Due to the data validation carried out during the full fieldwork process, we anticipate these advisories will be factored into the estimation phase. As a result, the internal and external data validation methodologies at this stage diverge. The analyses we will undertake in this process are as follows:

- Detection of outliers at the sub-national level (internal data validation).
- Ranking comparisons validation (external validation).

TPS data gathering.

TPS data gathering encompasses selecting, downloading, cleaning, and integrating third-party databases essential for the validation of the gathered GPP data. Each of these tasks involves multiple steps, which we will overview here to provide a clearer understanding of the process:

a. *Selection of external sources*

In selecting the external sources, the EU research team created a catalog of indicators, questions, and variables collected by third-party organizations. This entailed gathering information and metadata from external sources and structuring it into a single document.

Following this research effort, the team categorized the information based on its alignment with the conceptual framework. This allowed us to track the link of individual variables to each one of the pillars and sub-pillars within the EU theoretical framework. Below, we offer some statistics to describe the scale of this process:

- We chose 1,711 indicators and questions from 57 different external sources.
- From this universe, we pre-selected 744 indicators from 44 different sources.
- In the pre-selected group, every pillar encompasses a minimum of 60 variables.
- The only sub-pillars lacking associated external information are 1.8, 2.5, 7.5, and 8.7.
- Sub-pillars with less than three and more than one external variable include 1.9, 4.4, 7.6, 8.2, 8.4, 8.5, 8.6, and 8.7.

From this catalog, the team filtered down sources by considering factors such as comparability, source type, accessibility, and cost-efficiency. Once a final selection of sources is achieved, all indicators within this data source are included in the TPS dataset. The team will compile a data codebook that includes the following information for each selected external indicator:

- The original variable name.
- The name of the external source or database.
- The variable name in the TPS dataset, following a standardized naming convention.
- The year of data collection.
- The type of data source.
- The pillars and sub-pillars to which each variable is related to.

b. Downloading, cleaning, and harmonization of the external data

Following the selection of main variables categorized by pillar from the catalog, the downloading process offers two distinct options: acquiring the original source, which may contain multiple selected variables or retrieving data points from individual Excel files or website dashboards. Because of the above, the folder structure would be organized by source of information.

The conditions for downloading a database are as follows:

- We exclusively select the most recent year of the database.
- The most recent year must be from 2019 or later.
- The source must encompass information for all European Union countries.

Once the external data is downloaded, the data is wrangled as follows:

- Identify the indicators within the data.
- Subset and produce a new data set containing only the variables of interest.
- Re-orient the indicators according to their theoretical relation with the Rule of Law.
- Normalize the indicators to fit within a zero-to-one scale.
- Aggregate variables at the country level.

c. Data merging

Once all the necessary information has been wrangled and harmonized, we integrate the information into a single dataset. The integration process involves merging the already harmonized and cleaned databases into a master data set containing information such as country codes, standardized country names, and other project-related variables. Within this database, we will incorporate all sources sharing the same observations and employ the country code as the unique identifier. It's important to note that the integrated TPS database should not exceed 26 observations.

Outcomes Presentation

For the country data validation, a country report will be produced summarizing the flags and issues identified by country. These reports will be featured in HTML format and will have the following structure:

- Main findings
- Sample representativeness
- Patterns and consistency checks
    - Internal validation
    - External validation

For the validation of the outcomes, we will construct a comprehensive platform that systematically organizes all the information by country. This platform will feature multiple layers of information, beginning with the most high-level data and allowing users to drill down into more specific details. Users will have the capability to disaggregate information by sub-pillar, question, and even by region where feasible. The primary objective of this platform is to serve as a valuable tool for researchers and analysts, enabling them to efficiently identify anomalies in the data or potential reliability issues.