NSDS Status Table

Contents

Progress report as of January 2023

The following table presents the status of National Strategies for the Development of Statistics (NSDS) in International Development Association (IDA) borrower countries, Least Developed Countries, Low and Lower-Middle Income Countries, and some Upper-Middle Income Countries (in order to report on the whole of the African continent), as of January 2023. Non–IDA countries are identified in the table with an asterisk. The information is drawn from three principal sources:

- Direct information provided by countries to PARIS21;
- Websites of key development partners (in particular the World Bank's Bulletin Board on Statistical Capacity); and
- Websites of countries' national statistical offices.

The table is sorted by geographical region and provides information on existing strategy and next. It should be noted that while existing strategies may not necessarily follow the NSDS guidelines, most strategies currently being designed do follow them. This table is updated annually, and is available on the PARIS21 website at http://www.paris21.org.

1 PARIS21's overall support to data science in partner countries

PARIS21 produces the Partner Report on Support to Statistics (PRESS) annually to report on trends in support to statistics. The methodology is applied retrospectively for all previous years to ensure comparability over time. This document presents the methodology.

2 Area1 – Process analysis boosts efficiency

2.1 Example: Improving census operations using log data generated by survey devices – Collaboration with IBGE Brazil

2.2 How your country can benefit from this type of activity

3 Area2 – Complementing administrative data using comprehensive data sources

3.1 Example: Improve sampling efficiency using water consumption data – Collaboration with INE Chile

Knowing occupancy status of dwellings is crucial a for national statistical offices, as it forms the basis for conducting efficient surveys and censuses. However, collecting this information manually can be extremely time-consuming and expensive so naturally new ways for estimating the occupancy status using novel Data Science techniques have been explored recently. One approach consists in using water consumption data in combination with other administrative data to estimate this occupancy status.

As one of only two South American OECD member countries, Chile had the opportunity to obtain this water consumption data. Being conscious of the eminent potential of making use of Data Science on non-traditional data sources, Chile's National Institute of Statistics INE is looking to become a forerunner in applying recent advances in Data Science techniques to census and survey operations. For this purpose, INE participated in the second round of the Data Science Accelerator Programme, a twelve-week program for the use of novel Data Science methods in the development of official statistics initiated by the Office of National Statistics of the United Kingdom. PARIS21 was embracing the collaboration with INE on their far-reaching endeavor: estimating the occupancy status from water providers data. It aimed at informing survey operations on the occupancy status of household by estimating it from water consumption data. The underlying assumption consists of the fact that water consumption significantly differs from occupied to non-occupied or only seasonally occupied households. This in turn could assist during the sample design for surveys or even eventually the census, making these operations more efficient by reducing costs due to a reduced number of visits.

In the first step, water providers data from 48 municipalities was matched against up-to-date civil registry data. This made sure that trusted information on the occupancy status for each unit was available that was present in the water providers data. This database was then enriched with neighborhood-level information about the age and income group in the hope that this data could be exploited by the models later on to make predictions more precise. On top of this, the database was inspected for potential biases that could arise from using only units for which water consumption data was available, but none were detected.

In the second step, the models were selected starting from simpler techniques going to more sophisticated approaches. As a baseline model, a multinominal logit model was chosen, which predicts a probability for each unit being occupied, non-occupied and only seasonally occupied. The more sophisticated approaches comprised decision trees, random forests and the gradient boosting framework XGBoost (in order of their flexibility). After adapting the models for our purposes one thing became quickly: As good as the results

were for occupied units (up to 95% of accuracy), it was a higher accuracy on the non-occupied that gives the most bang for the buck. Initially we were able to reach an accuracy of 50% on non-occupied units which is already far above what one would get guessing at random. After tuning and tweaking the models towards optimal performance, finally a top performance of 80% on non-occupied units was reached. As every percent of increase in accuracy can result in more omitted households during a survey operation and therefore reduced costs, this methodology has the potential to be applied not only in future surveys throughout Chile but also in many other national statistical offices around the world.

3.2 How your country can benefit from this type of activity

4 Above and beyond - Further possible applications