25 November 2019

Economic Commission for Europe

Conference of European Statisticians

Expert meeting on measuring poverty and inequality: SDGs 1 and 10

Geneva, Switzerland
5–6 December 2019
Item F of the provisional agenda
Emerging issues: Machine learning

LEARN4SDGis–A Machine Learning based Poverty Mapping Exercise in Austria

Note by Statistics Austria*

This contribution informs on preliminary results obtained by Statistics Austria during a EUROSTAT grant on merging geospatial information and statistics. The project LEARN4SDGis aims to extend the use of geographic information systems to sample data for which detailed regional estimates are normally not available. More specifically, dissemination of important indicators for the global Sustainable Development Goals (SDGs) shall be supported. This involves the production of maps with an enhanced geographic resolution, compared to direct estimates from relatively small survey samples in the social statistics domain. The project explores new data sources on spatial distributions, registers and their integration with sample data by machine learning algorithms. A first application is presented by high resolution mapping of EU-SILC based poverty estimates for which a wealth of yet unused relevant auxiliary information exists in Austria.

Subnational estimates from survey data are in high demand but scarcely available due to sample size restrictions. Activities in the field of small area consistently prove that precision enhancement is possible with several methods as long as reliable auxiliary information is available. However, conventional estimation methods rarely include geospatial information. This is partly because the research community appears much more concerned with estimation methods than expanding substantially the range of auxiliary variables. This is also a consequence of traditional modelling approaches which quickly reach a limit as regards the number of parameters to be estimated.

The contribution explains how several machine learning algorithms were trained upon EU-SILC sample data to predict the poverty status of each individual in the population. Algorithms can rely on extensive register information on work and transfer incomes which make up 80% of total income in EU-SILC. While major income components of individuals appear well covered in register data, household

* Prepared by Matthias Till (matthias.till@statistik.gv.at). Contributors from Statistics Austria: Thomas Glaser, Johannes Gussenbauer, Ingrid Kaminger, Alexander Kowarik, Sibylle Saul, Alexandra Wegscheider-Pichler.

composition represented in such data is not fully coherent with the actual living circumstances of respondents which implies that regional poverty rates need to be approximated. In this process, information on the spatial distribution of potentially predictive variables (including accessibility) can be considered.

Several runs of cross validation are considered in the evaluation of the results, to avoid overfitting and give an indication of the variance of different approaches. Preliminary results suggest that predictions on unit level work remarkably well but certain geographical patterns deserve further investigation. For instance, urban regions as well as frontier regions (possibly rated to labour mobility across borders) still appear to exhibit implausibly high poverty rates. These are likely to be artefacts and may require strengthening the spatial component in the estimation (e.g. by shrinkage).

This contribution will present the results available by the time of the conference, including poverty maps (or heat maps) on the level of districts, municipalities and even enumeration districts. The discussion shall include challenges with regard to privacy and the political sensitivity in disseminating such "experimental" statistics.