LEARN4SDGis - PovertyMaps

Machine Learning based on Registers & EU-SILC Sample Data

Matthias Till
Directorate Social Statistics
Expert Group Poverty
Geneva
December, 2019
Context & Objectives

17 UN Sustainable Development Goals (SDGs)

=> Variety of statistical domains,
=> based on sample surveys (e.g. poverty, health, education)
=> sufficient sample size for “leaving no one behind”?
=> Especially: spatial disaggregation, GIS applications?

Borrowing Strength from Auxiliary Information

=> registers, geospatial information
=> Experts for each domain required?

Machine Learning as a Generalised Approach to Enhance Spatial Resolution of Sample Estimates?
Feasibility for 5 indicators

- **Poverty**
  - Below 60% of Median Household Income
  - Europe 2020 target group (income, deprivation, work intensity)
  - Multiple poverty

- **Education**
  - Persons with educational activity in last 12 months

- **Health**
  - Persons with subjectively bad or very bad health
## Methods

**Matching auxiliary information (sample & population)**
- Register data (income)
- Geodata

**Processing**
- Training of algorithms
- Repeated split sample estimates

**Mapping synthetic data**
Algorithms tested (using R)

- Random Forest (easy, fast and reliable)
- Boosting
- Support Vector Machines
- Neural Networks (engineering necessary, not stable)

⇒ Repeated split sample 80% Training -> 20% Test
⇒ Out-of-sample prediction for all individuals in frame
⇒ Aggregation to any level
  (e.g. Raster, EA, LAU, District, NUTS...)
In-sample validation criteria

EU-SILC Split sample results:

- **Accuracy** (% correct classification)
- **Specificity** (True Positive Rate)
- **AUC** (Area under the curve – TPR/FPR)
- **MAE** (Mean absolute error)

Different model specifications:

- Framevariables
- Framevariables + Geoinfo
- Framevariables + Registers
- Framevariables + Registers + Geoinfo
Conclusion from Validation

Plausible Machine Learning predictions are feasible for Poverty Maps

Data preparation is decisive

Accuracy depends more on data (e.g. income registers!) than on the ML algorithm

Random Forest (RF) yielded highest accuracy in validation

$\Rightarrow$ RF preferred as a simple & pragmatic choice
Validation for out of sample prediction

- Multiannual Averages (Nuts3)
- Socio-economic profiles (n = 42)
  - Age X sex
  - Size of municipality
  - Type of household
  - Citizenship
  - Sex of main earner (m/f)
  - Education
- External Data
  - Tax
  - Social Assistance
  - Index developed by Economic Research Institute (WIFO)
Coherence with 3y averages (NUTS3)

Vienna
(already adjusted)

Sample Size
- 50
- 250
- 500
- 1000

Poverty Rate 3 Year Mean

Poverty Rate Machine Learning adjusted to 3 Year Mean
Coherence with 3y averages (Groups)

The LEARN4SDGis project is funded by a grant from eurostat.
Synthetic data – cautionary measures

1. Coherence: EU-SILC NUTS2 averages over 3 years
2. Smoothing: 4 nearest EA
   - X & Y coordinates of EA
   - Number of persons in EA
3. Grouped results (very low, low medium, high, very high)
4. Filter rules for EA:
   - < 50 persons
   - Missing register data > 33%
   - Property price > 200% & poverty > 20%
   (Reference value: NUTS2 x DEGURBA)
Dissemination as Experimental Statistics

The LEARN4SDGis project is funded by a grant from eurostat

www.statistik.at