

Matthias Till
Directorate Social Statistics

EFGS - Manchester
October, 2019

LEARN4SDGis-PovertyMaps

Machine Learning based on
Registers & EU-SILC Sample Data

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?



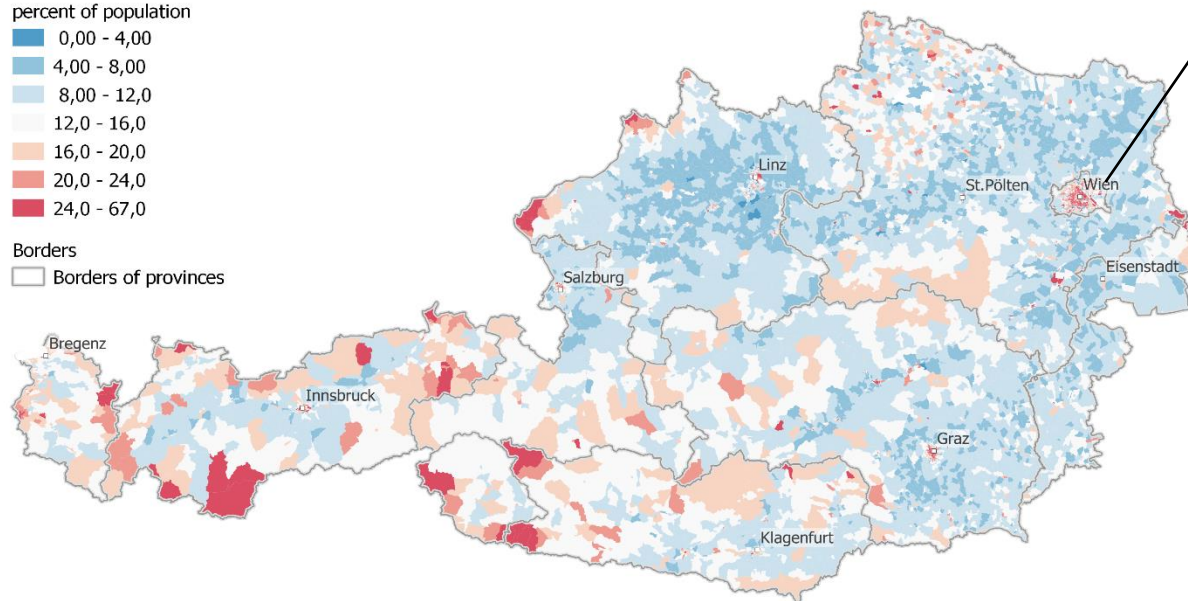
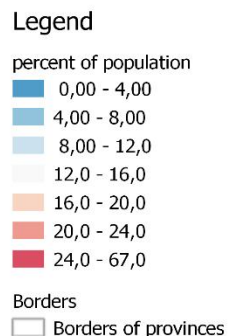
Experimental Results: Enumeration Districts

At-risk-of-poverty-rate in AT: ~ 14%

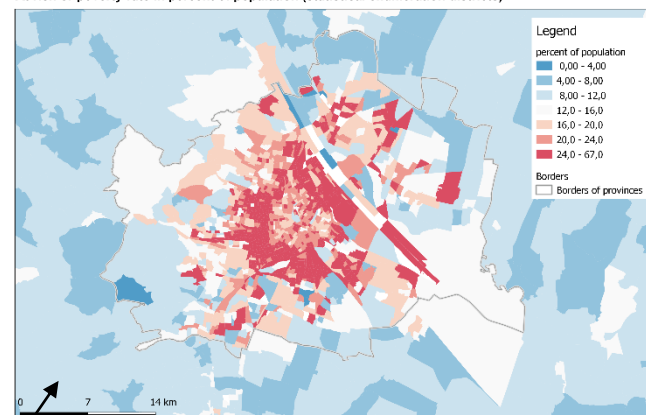
Concentrations: city/remote/border regions

> 24% near Vienna outer ring

favourable places (< 8%) are recognisable



At-risk-of-poverty-rate in percent of population (statistical enumeration districts)



Caveats for publications

- Ensure consistency (e.g. EU-SILC 3yr average for NUTS2)
- Highlight experimental
- Apply knn-smoothing
- Suppress small n
- Plausibility rules (eliminate high poverty in areas with high property prices)

Conclusions from Validations & Outlook

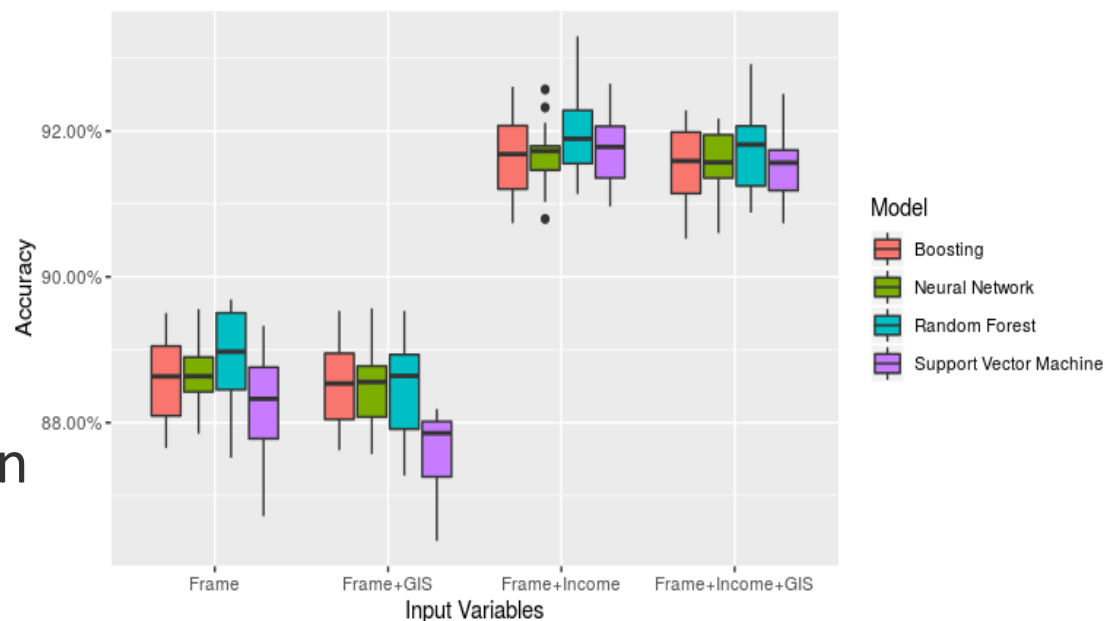
Plausible Machine Learning predictions are feasible for Poverty Maps

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

⇒ RF preferred as a simple & pragmatic choice



To be confirmed when health & education indicators will be estimated (possibly without similarly strong predictors)

comments & feedback are welcome: matthias.till@statistik.gv.at