ML Poverty: Using Machine Learning to estimate poverty rates in Switzerland at the canton level

Prepared by:

Yara Abu Awad, Christopher Sulkowski, Matthias Widmer, Lancelot Marti, Raphael de Fondeville, Stephan Haeni

Agenda

- SILC, Absolute Poverty & Register Data
- Accounting for Complex Survey Design
- Infrastructure & Model Selection
- Next steps

Aims of the ML Poverty Project

Goal: Estimate poverty indicators for all cantons

Current practice: Phone questionnaire with “only” 18’000 people.

- Robust only for ‘Large Regions’.
- Already quite expensive.

→ Federal Council requests cantonal level. Uncertainty estimation also needed.

New solution: Machine Learning algorithms to improve estimation**.**

Source: BFS

Poverty Indices

The different poverty concepts applied by the FSO are:

- Absolute Poverty - more on this later
- Relative Poverty - 60% of median equivalised disposable income
- Material Deprivation

Absolute Poverty

- A needs-based definition of a social subsistence level that not only guarantees physical survival but also a minimum level of participation in social life
- People are considered as poor if they do not have the means to buy goods and services that are necessary for a socially integrated life
- Disposable income is compared to cost of: Basic needs (food, clothing, personal care, transport, entertainment, education), Housing, Other necessary expenditure such as liability insurance

Statistics on Income and Living Conditions (SILC)

- Annual survey in over 30 European countries
- Switzerland: start in 2007
- Sample size: 8’500 households / 18’000 persons (4 year rotating panel)
- 99.7% CATI-interviews conducted by a private survey institute in 3 languages (German, French, Italian)
- Extensive use of administrative data for salaries and benefits from social security, other incomes collected in the interview
- Average interview duration: 51 minutes (for grid, household and individual questionnaire together)

SILC: Sampling

- The sampling plan uses a proportional, stratified design which is structured around the seven major geographical regions (level NUTS2)
- Use of weights to adjust for non-response, loss to follow-up and for calibration to the reference population
- Change of the sampling framework from directories of landline telephone numbers to population registers in 2014

Data Linkage

- Article 14*a* of the Federal Statistics Act (FStatA; SR 431.01) provides the legal basis for the linkage of data for statistical purposes
- Linkage is carried out using a unique identifier assigned to all Swiss residents
- Unique identification number is encrypted and certain other identifiers are removed
- Access to the data is limited

Features used

- A total of 124 features were created from Register data
- Personal Information: Age, Sex, Country of Birth, Marital Status, Location of residence
- Income information: Income, Social assistance, Unemployment benefits
- Dwelling information: number of residents, number of rooms, building characteristics
- Geographic information: number of visible mountain peaks, noise exposure from rail, travel time to the next agglomeration center

Using weights to train models & calculate metrics

”Failing to account for sampling weights in gradient boosting models may limit generalizability for data from complex surveys, dependent on sample size and other analytic properties. In the absence of software for configuring weighted algorithms, post-hoc re-calculations of unweighted model performance using weighted observed outcomes may more accurately reflect model prediction in target populations than ignoring weights entirely.”

MacNell, Nathaniel, et al. "Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting." *Plos one* 18.1 (2023): e0280387.

Creating folds for model evaluation

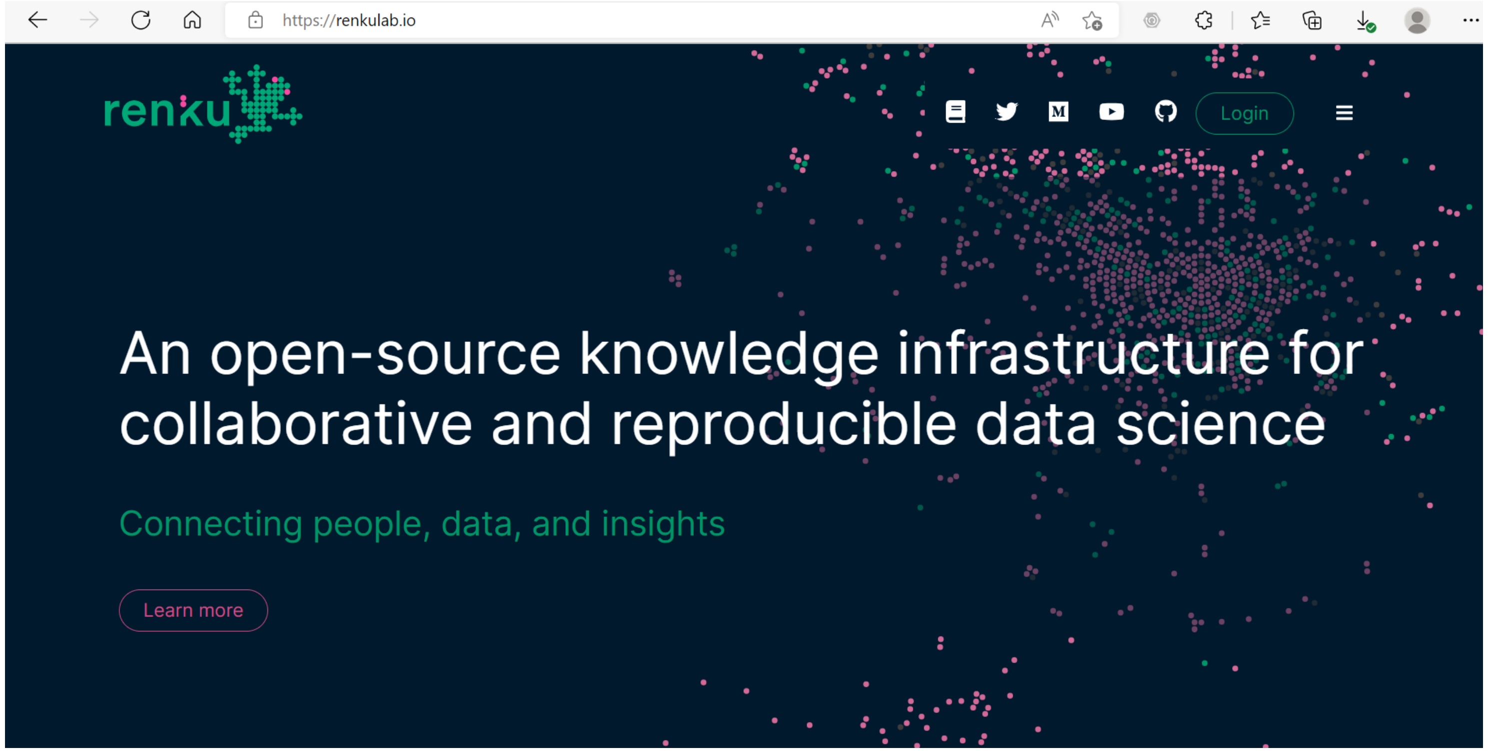
Creating CV folds:

" For cluster sampling, partition the data at the level of the Primary Sampling Unit (PSU). All elements from a given PSU should be placed in the same fold, so that the folds are a random partition of PSUs rather than of elementary sampling units. (Note: with multistage sampling, after a first-stage cluster sample there is further subsampling in the selected clusters. Since there is no straightforward way for CV to mimic this subsampling, we simply form folds at the PSU level even in multistage samples.)

For stratified sampling, make each fold a stratified sample of units from each stratum. Create SRS CV folds separately within each stratum, then combine them across strata."

Wieczorek, J., Guerin, C., & McMahon, T. (2022). K-fold cross-validation for complex sample surveys. *Stat*, *11*(1), e454.

Infrastructure



GitLab

docker

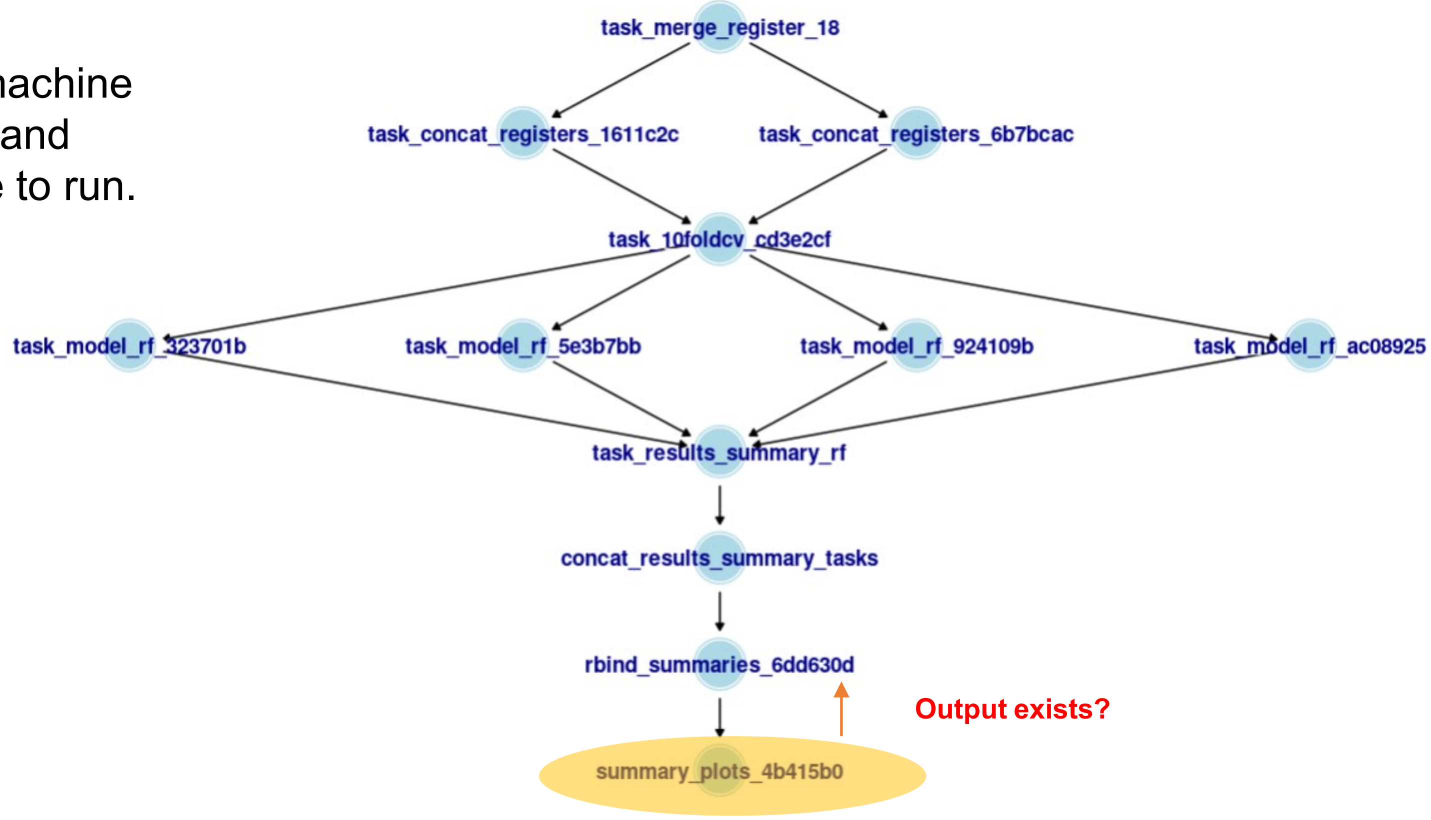
Coding and Libraries used

- All code is written in R (as requested by collaborators)
- Use h2o package for: Gradient Boosting, Random Forest, Logistic Regression (for comparison), Elasticnet (for feature selection)
- Use tensorflow package for neural network fitting
- Both packages are industry standard

Code Structure: Example

In this project, we have a DAG for each machine learning model, *gbm_grid_search_dag.R* and *rf_grid_search_mawi.R* that we would like to run.

Check if summary_plots exists.
⇒No. Check if rbind_summaries exists.



Modeling strategies

- GIS **vs** no GIS
- 2018-19-20 **vs** 2018 / 2019 / 2020
- Random forest vs Gradient Boosting Model

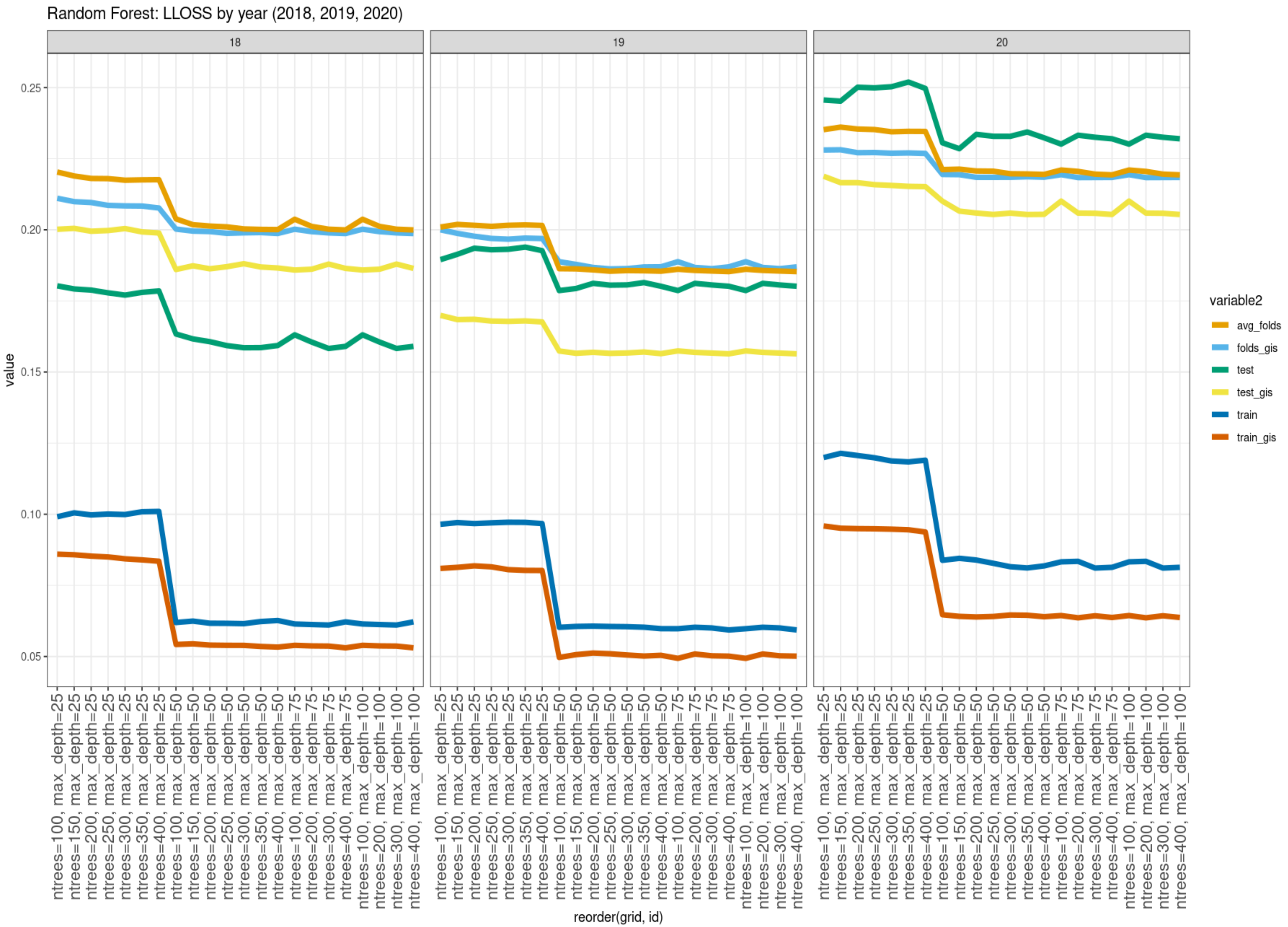
Metrics

1- Log loss

$$L_{\log}(y,p) = -(y\log(p) + (1-y)\log(1-p))$$

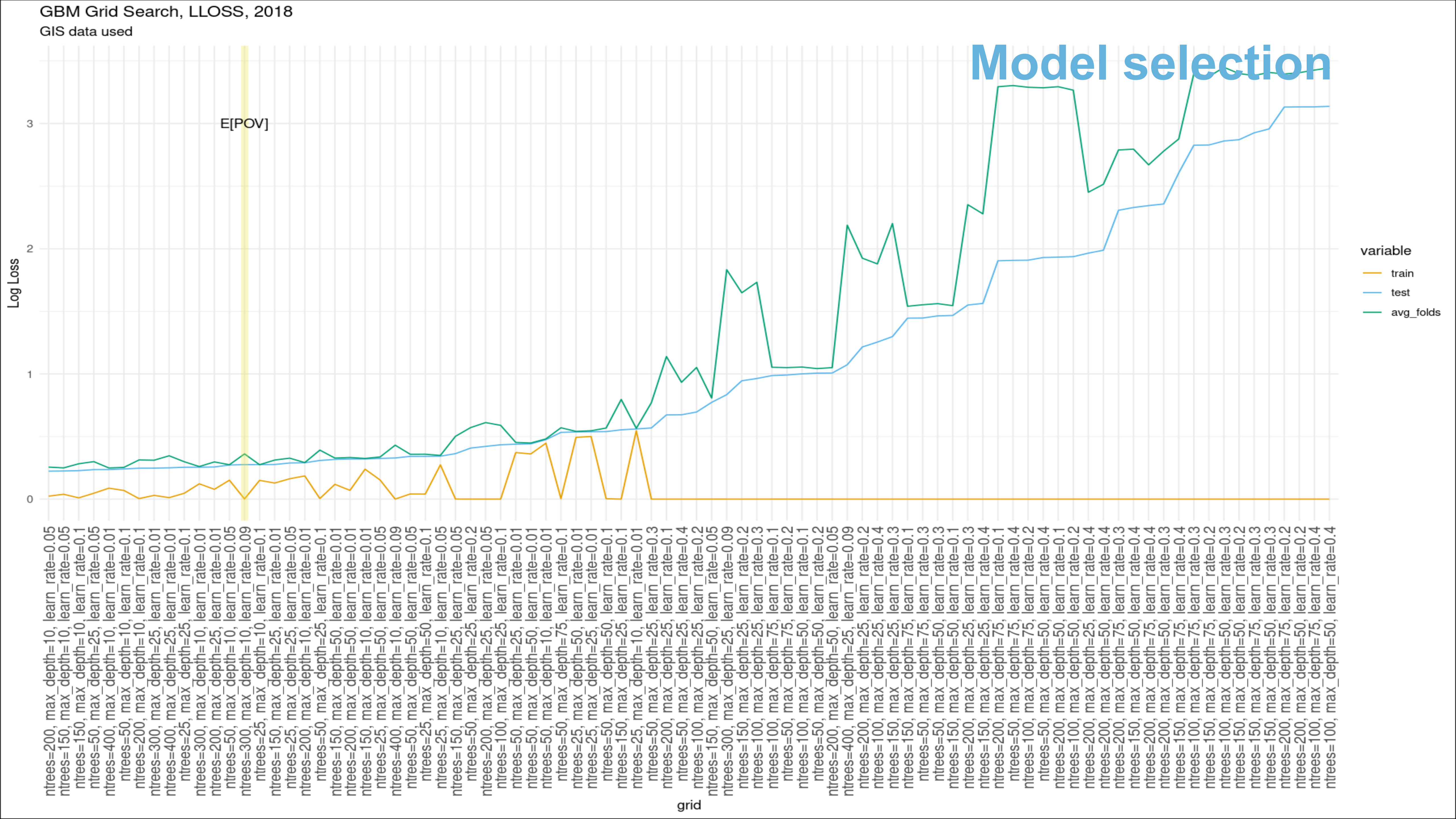
p : predicted probability of y = 1
y : true value

2- Difference between observed poverty rate and predicted poverty rate (weighted)



Fitting data to all 3 years with and without GIS

Model selection



Next Steps

- Generating predictions at the population level: post-hoc sensitivity analyses
- Estimating the uncertainty of the predictions: first using bootstrapping to estimate model uncertainty and then propagating this uncertainty into the Bernoulli approach illustrated by the Austrian team while accounting for clustering by household
- Use of privacy preserving techniques for dissemination of information: due to the potential of disclosure of these results at a smaller resolution (i.e. smaller than canton), we would like to use differential privacy when doing so and are looking into what this means for geographic data in collaboration with the OpenDP project.