Protecting High-Resolution Poverty Statistics against Disclosure using Differential Privacy

Dr. Raphaël de Fondeville

Senior Data Scientist – Group Lead

Federal Statistical Office - Switzerland

Data Science
Competence Center DSCC

CNAI

Competence Network for Artificial Intelligence
Kompetenznetzwerk für künstliche Intelligenz
Réseau de compétences en intelligence artificielle
Rete di competenze per l'intelligenza artificiale

Data Science under the Rule of Law

Data Science Competence Center's Code of Conduct

Privacy	Transparency	Ethics
• Data Governance • **Data Confidentiality** • Data Security	• Explainability • Open source • Reproducibility	• Non-discrimination • Objectivity • Representativeness

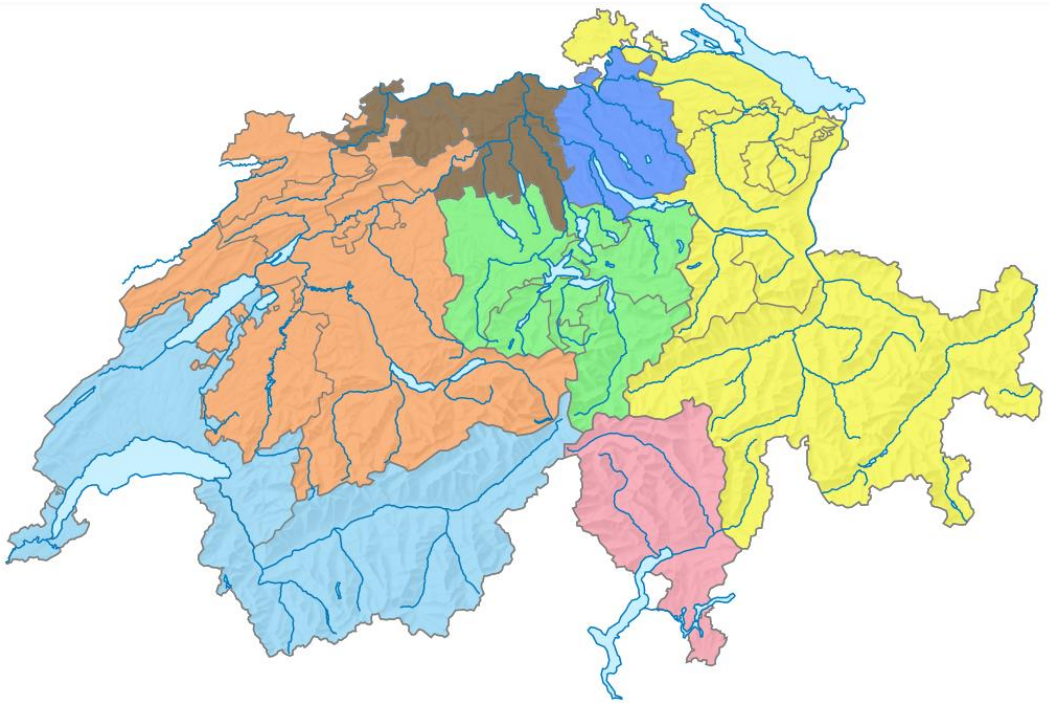
Rule of Law & Public Trust

Poverty Index Estimation

Goal: Estimating a poverty indicator in Switzerland.

Current practice: Phone questionnaire with “only” 17’000 people.

- Yearly survey.
- Representative of the Swiss population.
- Robust for “large regions” only.
- Quite expensive.



Source: FSO

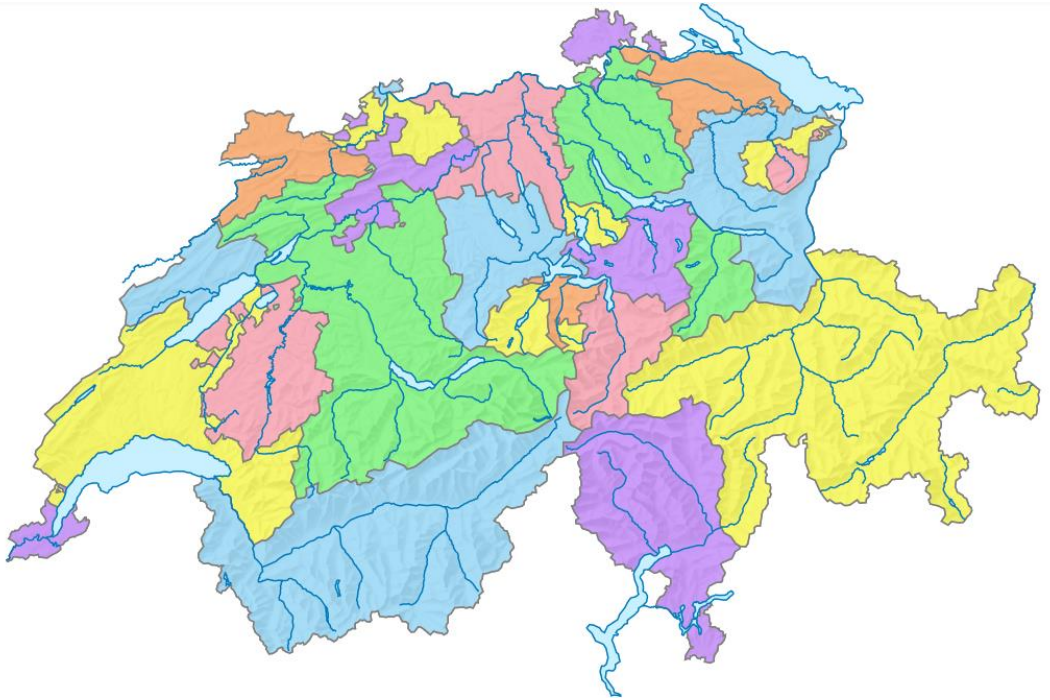
Poverty Index Estimation

Goal: Estimating a poverty indicator in Switzerland.

Current practice: Phone questionnaire with “only” 17’000 people.

- Yearly survey.
- Representative of the Swiss population.
- Robust for “large regions” only.
- Quite expensive.

➡ Federal Council requests cantonal level.



Source: FSO

Data Imputation

- Auxiliary source of data Z_i is available at **population** level: register data such as location of residence, familly status, …

- *Data Imputation*: The poverty status of each individual is predicted from auxiliary data.

$$P = \frac{1}{N}\sum_{i=1}^{N} \hat{X}_i$$

- $\hat{X}_i = X\_i$, if individual i participated to the survey.

- $\hat{X}_i = \boldsymbol{f}(Z_i)$, if individual i did not participated to the survey.

Data Imputation using Machine Learning

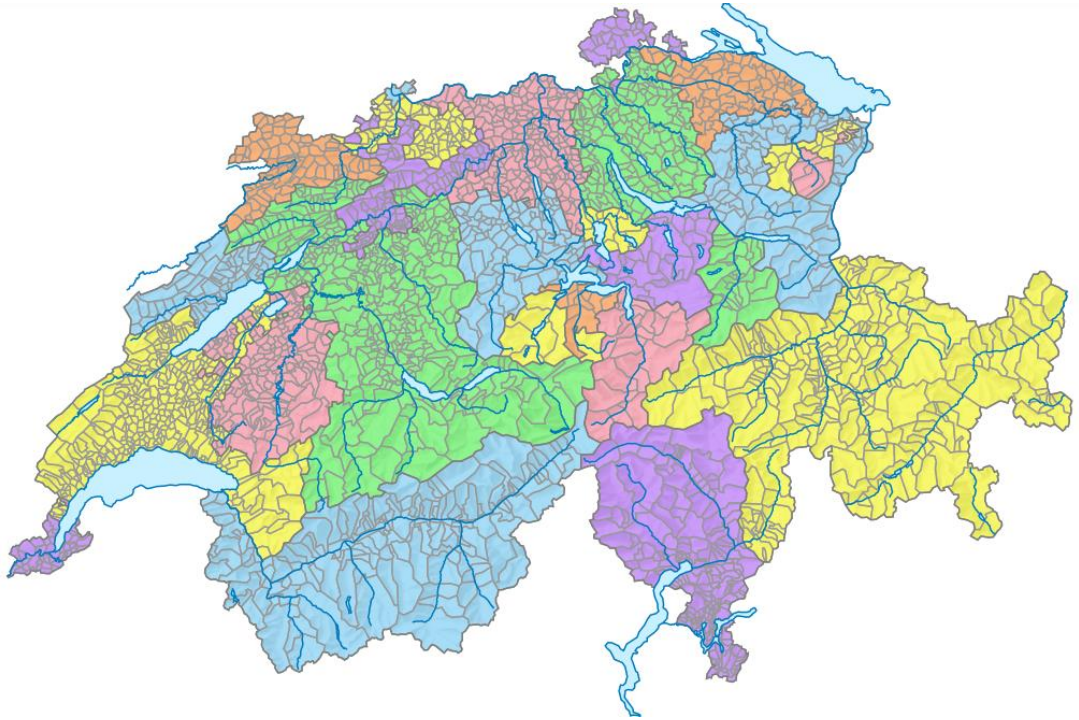
The function f linking Z to X is unknown!

→ BUT it can be learned from the data using Machine Learning (ML).

We explored multiple algorithms and quantifies their uncertainty:

- Logistic regression,
- Random forest,
- Gradient boosting,
- Neural networks.

Privacy Challenge: Disclosure Control



Source: OFS

Major privacy risk to disclose fine resolution data …

Introduction to Differential Privacy

Source Data

Population

Sex	Zip		Poverty Status
M	1005		Y
F	2012		N
F	8038		N
M	1503		N
F	1220		Y
M	3807		Y

National Statistical Office (Input and Compute Party)

Differential Privacy

Results Parties (Adversaries)

Statistical Analysis

Introduction to Differential Privacy

Source Data

Population

Sex	Zip		Poverty Status
M	1005		Y
F	2012		N
F	8038		N
M	1503		N
F	1220		Y
M	3807		Y

National Statistical Office (Input and Compute Party)

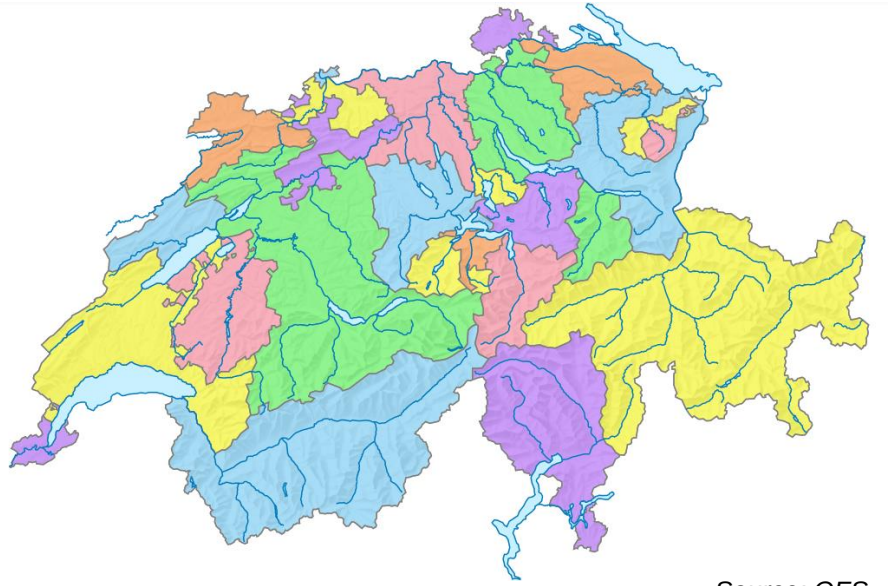
Differential Privacy Mechanism

Results Parties (Adversaries)

Statistical Analysis

Disclosure Control for Poverty Statistics

Applying DP in production is challenging [1] and requests secure and robust end-to-end data pipelines.



Source: OFS

Harvard John A. Paulson School of Engineering and Applied Sciences

OpenDP

New components were implemented in OpenDP to leverage the **data structure**.

[1] Garfinkel, S., Abowd, J., and Powazeck, S. (2018). Issues Encountered Deploying Differential Privacy. *WPES'18: Proceedings of the 2018 Workshop on Privacy in Electronic Society*, 133–137.

Data Pipelines in OpenDP

OpenDP + polars

→ End-to-end secure and efficient pipelines for data ingestion, representation and manipulation with the OpenDP library.

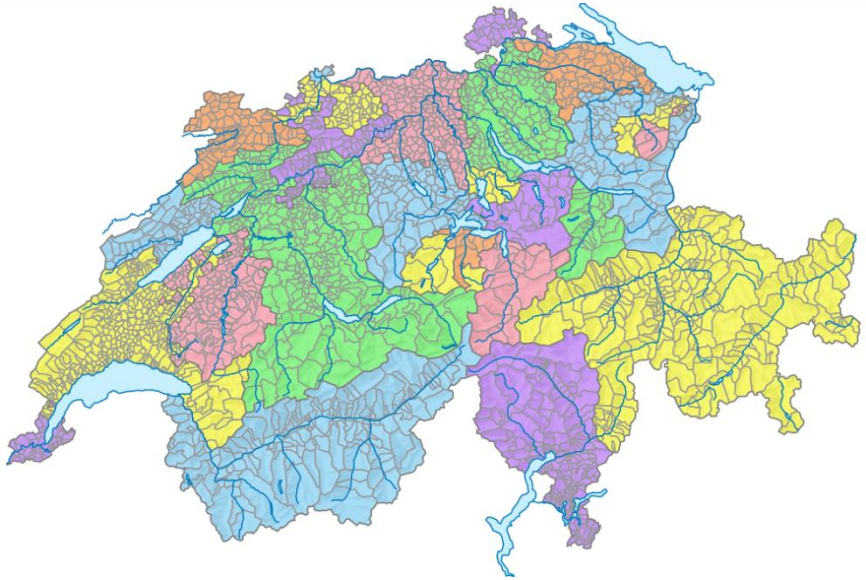
Michael Shoemate

Pauline Maury-Laribière

Anna Banaszak

Disclosure Control for Poverty Statistics

Differential Privacy does not preclude any potential discrimination or harm stemming from geographical information.



Source: OFS

Harvard John A. Paulson School of Engineering and Applied Sciences

OpenDP

Ongoing new methodological developments.