Evaluation of Machine Learning Models for Non-Standard Data Structures

Roman Hornung[1,2], Malte Nalenz[3], Lennart Schneider[3,2], Andreas Bender[3,2], Ludwig Bothmann[3,2], Florian Dumpert[4], Bernd Bischl[3,2], Thomas Augustin[3], Anne-Laure Boulesteix[1,2]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany

[2]Munich Center for Machine Learning (MCML), Munich, Germany

[3]Department of Statistics, LMU Munich, Munich, Germany

[4]Federal Statistical Office of Germany, Wiesbaden, Germany

June, 26th, 2025

Background

This presentation is based on the following paper:

Hornung, R., Nalenz, N., Schneider, L., Bender, A., Bothmann, L., Dumpert, F., Bischl, B., Augustin, A., Boulesteix, A.-L., 2023.
Evaluating machine learning models in non-standard settings: An overview and new findings.
arXiv:2310.15108. Statistical Science (to appear).

Background

- Machine learning (**ML**) applications in **official statistics** frequently deal with **complex data structures** most of which violate the standard i.i.d. assumption.

- Such data structures include spatial and clustered data, as well as data under concept drift.

- The presence of these complex structures **can introduce bias** in generalization error (**GE**) **estimates** derived from **ordinary resampling** methods like cross-validation.

- **Tailored resampling** techniques are **necessary** for each data structure to obtain (largely) unbiased GE estimates.

Background

- A **shared aspect** of these specialized procedures is that they **ensure** that the **test data** relate to the **training data** in the **same way** that the **future data** relate to the data used for constructing the ML model.

- I will give an **overview** of such methods for the following **selection of complex data structures**: clustered data, spatial data, unequal sampling probabilities, concept drift, and hierarchically structured outcomes.

- This overview **synthesizes** insights from the **existing literature** and **our simulation studies**.

Clustered data

- In official statistics clustered data are frequently encountered, where **observations are grouped** based on structural relationships (e.g., individuals within a household).
- **When** the observations are **randomly assigned** to the training and test datasets, members of the **same cluster** appear **in both**.
- As a result, the **ML models within** the **resampling** tend to work **better on** the **test data** sets **than** on observations from **independent clusters**.
- **If** the **aim** of the prediction is to **predict** the outcome of observations from **new clusters**, this can lead to **underestimation** of the GE.

Clustered data

- **Empirical studies** have consistently shown **severe** GE **underestimation when** clusters involve **repeated measurements** of the **same entities**.
- **Our simulations** suggest that this **underestimation** tends to be **modest for** clusters constituted of **distinct entities**. **However**, the presence of **cluster-constant covariates** can exacerbate this effect.
- The **recommended** solution is **cluster-level resampling**, ensuring that training and test datasets comprise entirely **separate clusters**, thus avoiding any overlap.

Spatial data

- Spatial data, common in official statistics, inherently exhibit **spatial correlation**—observations **near** each other tend to be **more similar** than those farther apart.

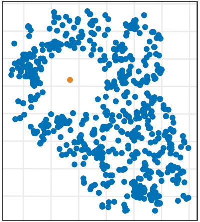
- It is **well known** from the literature that spatial correlation **must be taken into account** when **estimating the GE** to avoid underestimation.

- **Spatial cross-validation** methods are used here. These methods provide **adjustable parameters** to **fine-tune** the **separation** between training and test data.

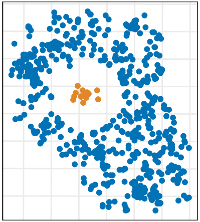
- The **selection** and **customization** of these methods **depends strongly on the application**, especially the **proximity** of the **new data** relative to the existing data.

Spatial data: Popular variants of spatial CV

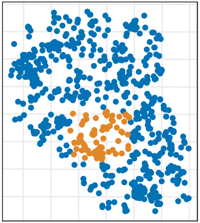
Leave−one−out CV
(w/ buffer)



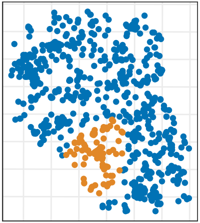
Leave−one−disc−out CV
(w/ buffer)



Leave−one−block−out CV with
geometric blocks (w/o buffer)



Leave−one−block−out CV with
clustered blocks (w/o buffer)



Unequal sampling probabilities

- In official statistics, the observations often have **unequal sampling probabilities**, e.g., to represent minorities adequately.
- The **resulting samples** are **not representative** of the population to which the ML model is to be applied.
- **Neglecting** this leads to **biased GE estimates**.
- This bias can be **corrected using** the **Horvitz-Thompson** theorem, **inversely weighting errors** by sampling probabilities.
- Our simulations indicate that the **bias** of conventional GE estimation **varies by ML method**. Yet, **Horvitz-Thompson**-based correction **consistently** ensured **unbiased** GE estimates.

Concept drift

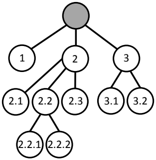
- Concept drift refers to **changes in the distribution** of data **over time**, **affecting predictive performance** in official statistics, where models are applied over long periods and concepts evolve.

- Accurately accounting for concept drift is **critical** in **GE estimation** to avoid **bias**.

- Our simulation results indicate that **largely unbiased** GE estimates can be obtained by using the **most recent observations as test data** ("out-of-sample validation").

- They also underscore the importance of **frequent model updating** and ensuring the **test dataset** is **not very small**.

Hierarchically structured outcomes



- Many official statistics **classification systems** are hierarchically structured, notably **tree-structured**.
- Despite the availability of **various evaluation metrics** for hierarchical predictions, **resampling** method **recommendations** are **lacking**.
- **Hypothesizing** that **stratified** cross-validation has **lower bias** and **variance** for hierarchical outcomes, we compared it with ordinary cross-validation in a simulation study.
- **Ordinary** cross-validation **slightly underestimated performance**, whereas **stratified** cross-validation showed **no bias** except for very small datasets; the **variances** of the estimates **did not differ** between the two approaches.

Figure from Hornung et al. (2023)

Summary of GE estimation for complex data structures

- **Clustered data:** Implement **cluster-level resampling** to prevent cluster overlap between training and test data.
- **Spatial data:** Apply **spatial cross-validation** techniques, ensuring the prediction goal drives the selection and customization of these methods.
- **Unequal sampling probabilities:** **Adjust** GE estimates **using** the **Horvitz-Thompson** theorem for bias correction.
- **Concept drift:** Apply out-of-sample validation by using the **most recent data** for **testing**; avoid very small test datasets.
- **Hierarchically structured outcomes:** Use **stratified cross-validation** to estimate performance, which avoids the slight bias associated with ordinary cross-validation.

References

Hornung, R., Nalenz, N., Schneider, L., Bender, A., Bothmann, L., Dumpert, F., Bischl, B., Augustin, A., Boulesteix, A.-L., 2023.
Evaluating machine learning models in non-standard settings: An overview and new findings.
arXiv:2310.15108. Statistical Science (to appear).

Brenning, A., Lausen, B., 2008.
Estimating error rates in the classification of paired organs.
Stat. Med. 27(22), 4515–4531.

Schratz, P., Becker, M., Lang, M., Brenning, A., 2021.
Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R.
arXiv:2110.12674.

Holbrook, A., Lumley, T., Gillen, D., 2020.
Estimating prediction error for complex samples.
Can. J. Stat. 48(2), 204–221.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., Petitjean, F., 2016.
Characterizing concept drift.
Data Min. Knowl. Disc. 30(4), 964–994.

Kohavi, R., 1995.
A study of cross-validation and bootstrap for accuracy estimation and model selection.
In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137–1143.