Institut für Finanz- und Aktuarwissenschaften ifa

Interpretable Ensembles

Enhancing Claim Frequency Modeling with External Socioeconomic Factors

- Ulm Actuarial Day
- Tobias Baur – joint work with Dr. Johannes Schupp
- March 2026

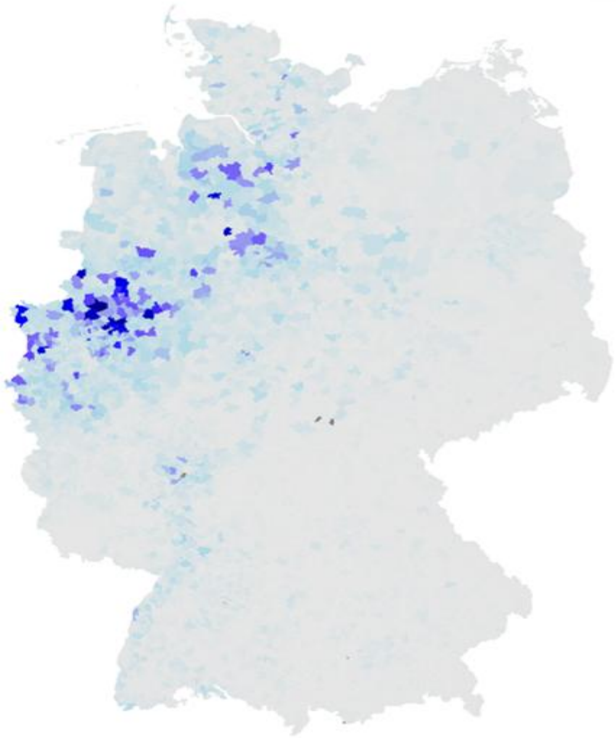
Introduction

Modeling Claim Frequencies in P&C and Data (focus on burglary and theft insurance)

- Standard Approach: Compound Poisson Model (see Klugman, Panjer, Willmot (2012))
  - separate modeling of claim frequency (focus here; rare event) and claim severity
- Real-World dataset with over 27.5 Mio observations from burglary & theft insurance (less than 1% claim frequency)
- Well established number of **14 tariff factors**

- Additional **socioeconomic factors** that help to better understand the risks
  - Socioeconomic data increasingly available, e.g. EU Data Act
  - Better predictions but increased computational cost, see NAIC (2025)
    - → **14 + 87 additional features**

Accumulated no. of claims



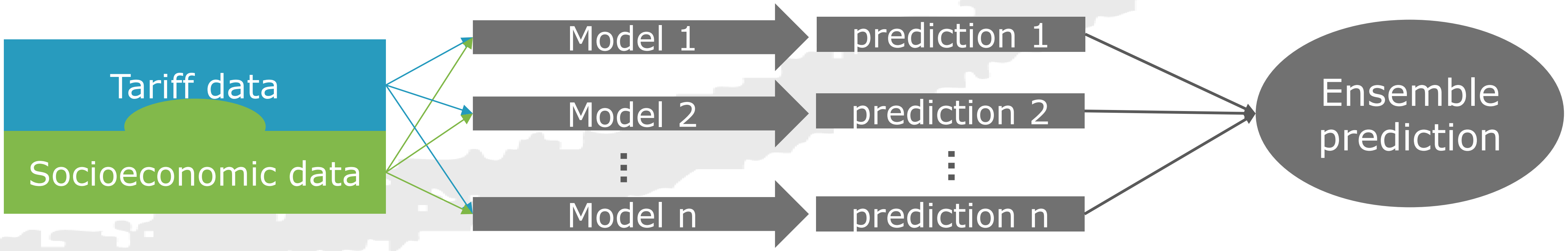
Introduction
Contribution

- **Research Contributions:**
  - What is the **effect of additional socioeconomic factors** on the forecasts? How can the relevant factors for the forecasts be selected?
  - **How much data** is needed to obtain stable forecasts? What additional information can be obtained from very large amounts of data (assessment of uncertainty)?
  - How can **procedures** be adapted so that they are suitable **for very large datasets**? How can procedures effectively extract the benefits of large amounts of data?

Introduction

Models

- Random Forests, neural nets, transformers → high accuracy but low interpretability
- (regularized) GLMs: interpretable, well-known
  - Existing algorithms have high memory requirements when dealing with data volumes of this size.
- **Classic Approach: PCA - Transformation of socioeconomic data →** Compression of (additional) data
  - New factors are not sparse!
- **Ensemble Methods:** Reduction of memory requirements for each base learner
  - Combine interpretability of GLMs with higher accuracy of ensembles



Benchmarks: (regularized) GLMs & PCA

Principal Components Analysis (PCA)

Two Approaches:

- Principal Compontent Analysis (PCA) using **all 87 socioeconomic factors**
  - Subsequent use of the first x principal components

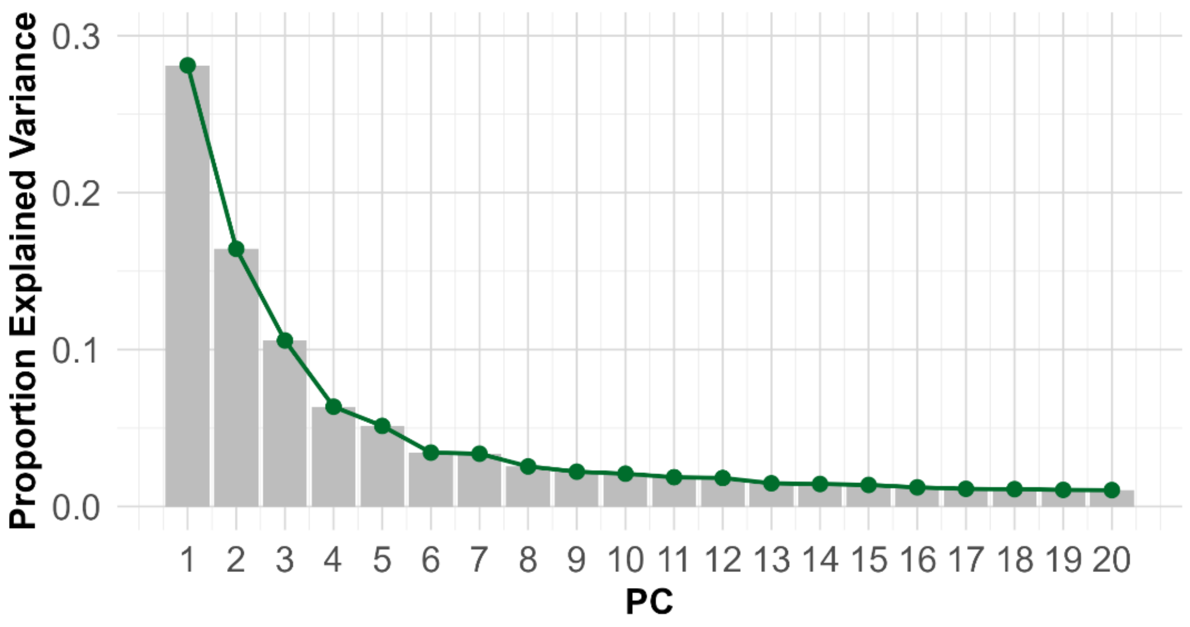
- Principal Component Analysis (PCA) on **7 groups of factors**
  - PCA performed on each group
  - Subsequent use of the respective first principal component

→ Each compared to regularized GLMs trained solely with tariff data

Benchmarks: (regularized) GLMs & PCA

Model comparison

Model	Test-Deviance	Improvement Rate*
Intercept-Only	204,620.9	* in comparison to Intercept-Only
GLM (λ = 0)	193,590.1	5.39%
Ridge (λ = 1.3)	193,589.3	5.39%
6 PCs	191,485.0	6.41%
8 PCs	191,002.0	6.65%
10 PCs	190,928.7	6.69%
12 PCs	190,742.6	6.78%
14 PCs	190,593.6	6.85%
PCA with groups	191,467.6	6.42%

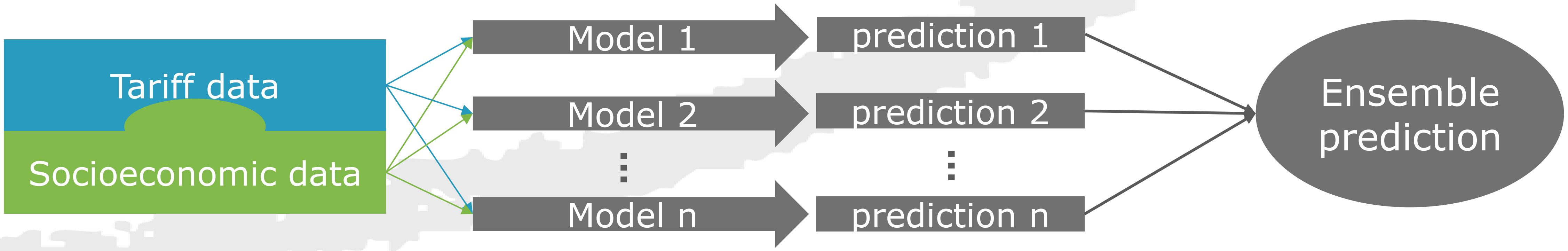


- Additional socioeconomic features improve predictive performance
- BUT:
  - PCA leads to **dense features** → sharp rise in memory usage per new component
  - additional PCs provide only marginal performance gains while substantially increasing matrix size
    → further components are not included (model matrix expands from 4.5 GB to 8.5 GB)

Innovative Methods for exploiting large data

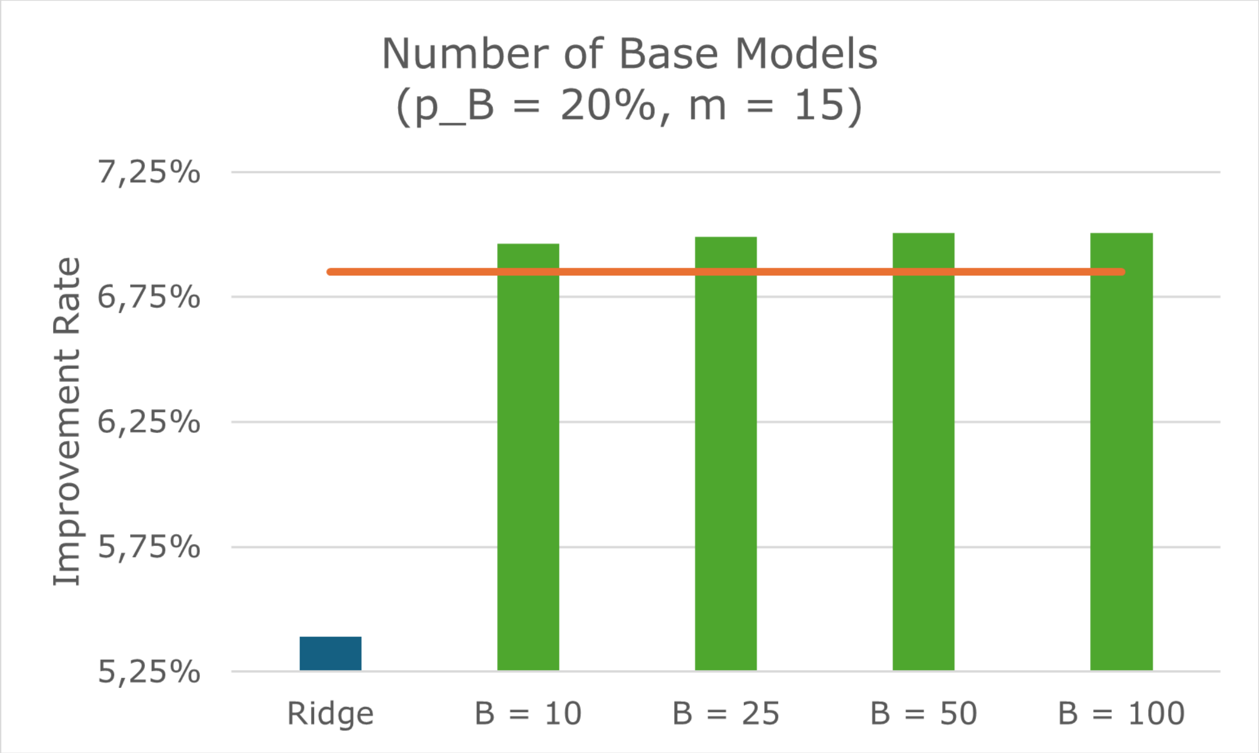
Approach

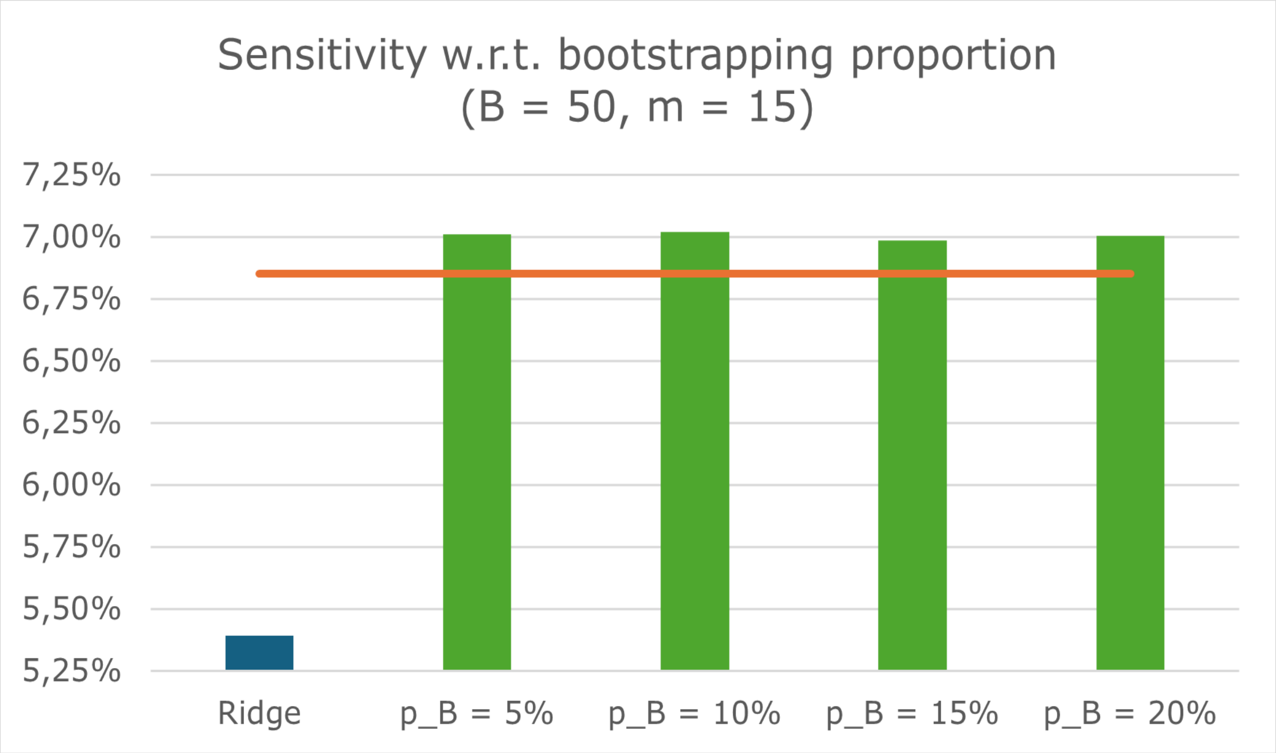
- Ensemble of GLMs: For each model ***B***
  - Bootstrapping: random proportion of training data (e.g. p_B=20%)
  - Bagging: All tariff factors + ***m*** randomly (weighted) chosen socioeconomic factors
    - Breiman: Random Forests (2001), Bagging Predictors (1996)
- Averaging predictions of individual base learners
  - Adapted approach from Ensemble-methods



Innovative Methods for exploiting large data

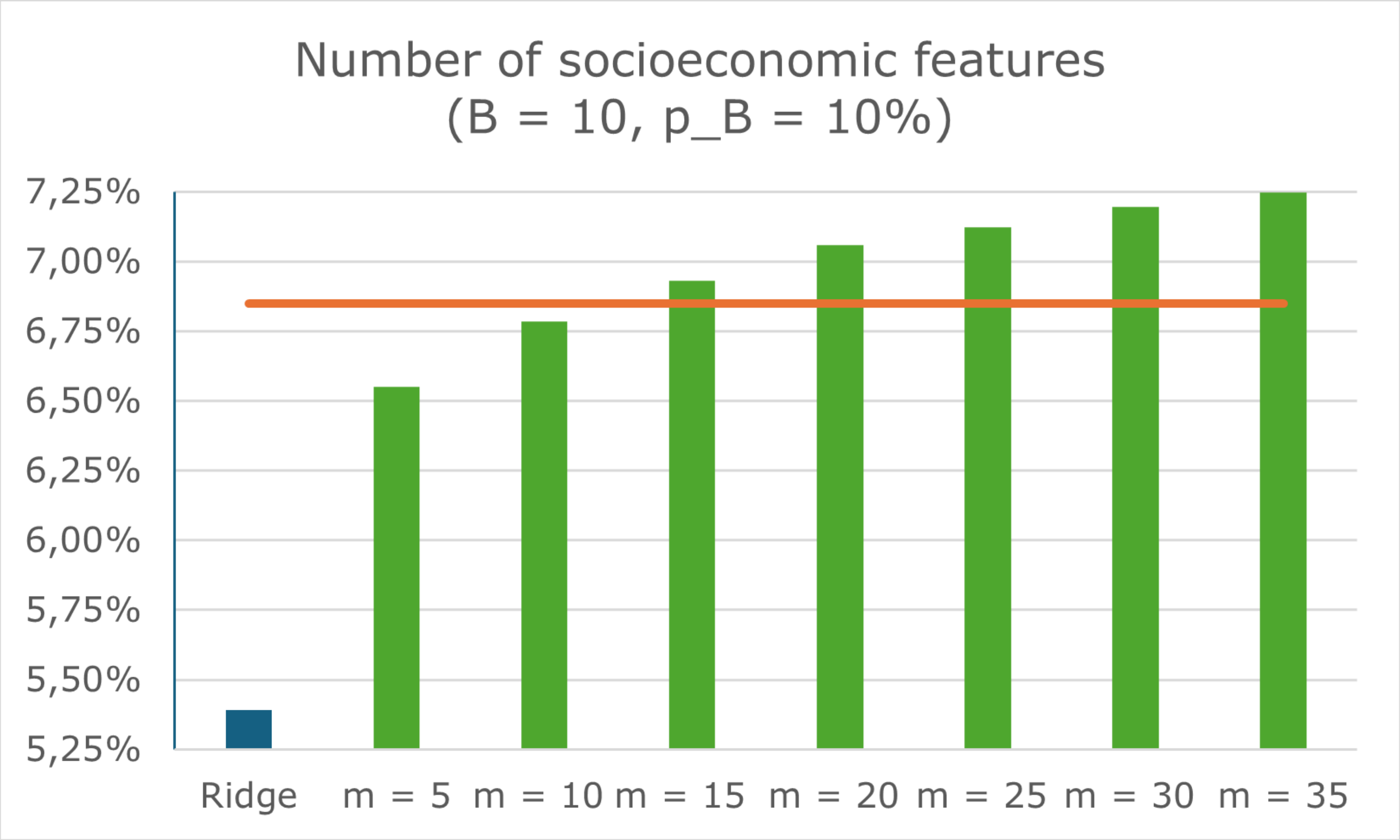
Sensitivity Analysis w.r.t. important parameters





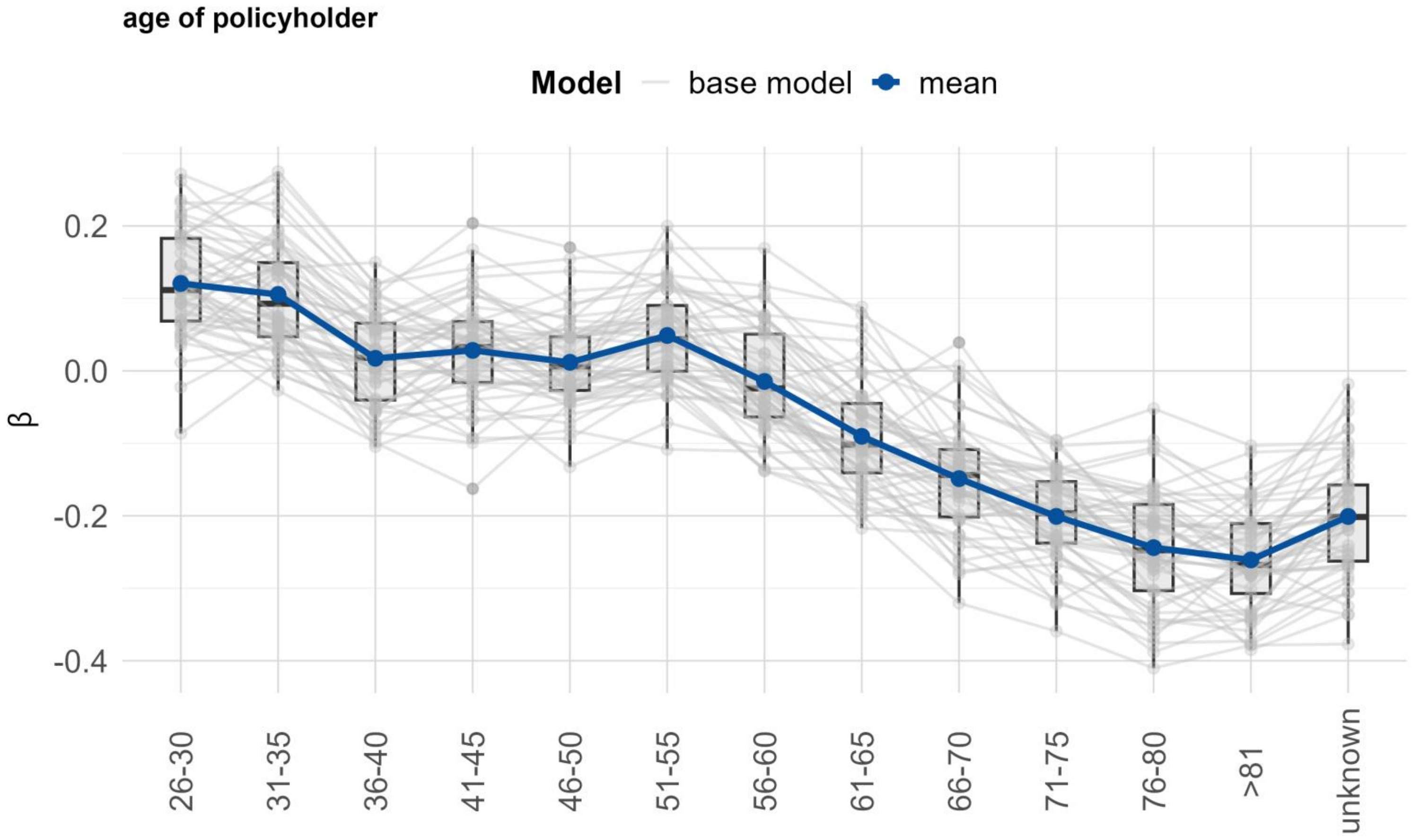
Innovative Methods for exploiting large data

Sensitivity Analysis w.r.t. important parameters



Innovative Methods for exploiting large data

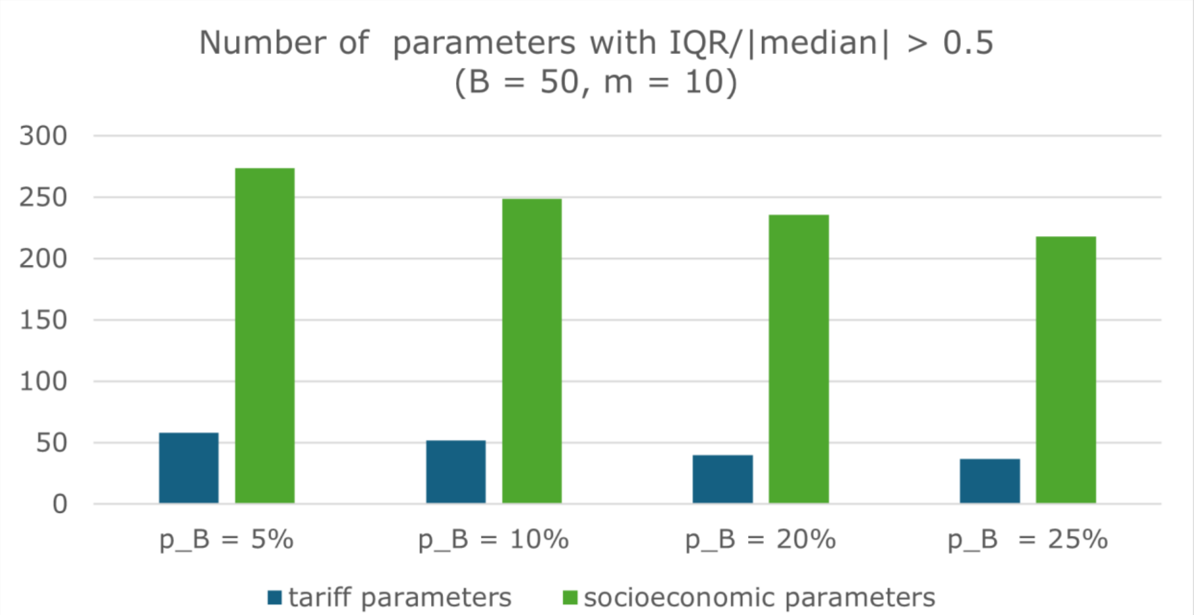
Uncertainties in the marginal effects

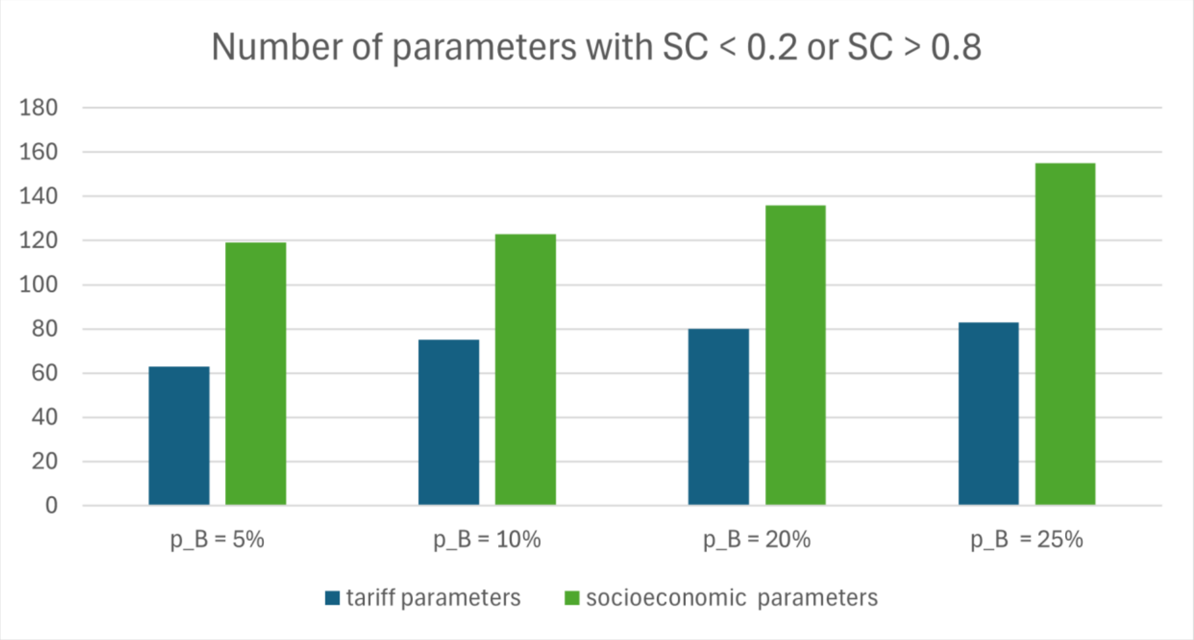


- Results stabilize with ensemble
  - $B = 50$, $p_B = 10\%$, $m = 15$
- Even for important tariff factors significant variation (Each model uses 2.5 Mio observations)
  - volume of a small/medium book
- Predictions remain fully interpretable
- Only 10% memory requirement
- Computation time linear in B

Innovative Methods for exploiting large data

Stability of parameters / consistency of sign





IQR = interquartile range, third quartile (75%) – first quartile (25%)

- SC = proportion of base learners in which a parameter is estimated with a negative (or positive) value. (Rocha, Wang, & Yu, 2018)
- SC close to 0 (or 1) indicates consistently negative (or positive) influence of parameter on target variable

Increasing the Bootstrapping Proportion per base model increases the stability of the parameters

Summary

- Incorporating socioeconomic factors via **PCA** improves predictive performance (to ~ 6.8%)
  - **But**:
    - Resulting compressed data not ‘sparse’ → numerically disadvantageous

- Key results for **Ensemble** approach:
  - Better predictive performance (~ 6.8% → 7.2%)
  - Full interpretability of individual factors
  - Reduced memory requirement per base learner → applicable to (very) large datasets
  - Visualization of variability of marginal distributions and assessment of parameter stability

References

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

EU Data Act, European Union (2023). Regulation (EU) 2023/2854 of the European Parliament and of the Council — Data Act. Official Journal of the European Union, L 2023/2854

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). Loss models: From data to decisions (4th ed.). John Wiley & Sons.

National Association of Insurance Commissioners. (2025). Auto insurance database report. NAIC.

Rocha, G., Wang, P., & Yu, B. (2018). Asymptotic distribution and sparsistency for L1-penalized parametric M-estimators, with applications to linear SVM and logistic regression. https://arxiv.org/abs/0908.1940