UiO : Department of Mathematics
University of Oslo

Interpretable House Price Prediction Using a Collection of Local Machine Learning Models

WEAI 97th Annual Conference
Portland (Oregon)

Anders Hjort, Johan Pensar, Ida Scheel, Dag Einar Sommervoll

Table of contents

Introduction

- An Automated Valuation Model (AVM) estimates the price of a dwelling at the current time
- Many approaches: Hedonic model, repeat sales model, nearest neighbor regression etc.

Introduction

- An Automated Valuation Model (AVM) estimates the price of a dwelling at the current time
- Many approaches: Hedonic model, repeat sales model, nearest neighbor regression etc.
- Norwegian banks use the dwelling as mortgage collateral $\Longrightarrow$ banks need AVMs to monitor the value of their collaterals

Introduction

- An Automated Valuation Model (AVM) estimates the price of a dwelling at the current time
- Many approaches: Hedonic model, repeat sales model, nearest neighbor regression etc.
- Norwegian banks use the dwelling as mortgage collateral $\Longrightarrow$ banks need AVMs to monitor the value of their collaterals
- Using statistical methods for mass appraisal is not a novel idea: Bailey et al. (1963), Rosen (1974)
- The state of the art for AVMs are machine learning models like gradient boosted trees or random forest; Baldominos et al. (2018), Sing et al. (2021), Kim et al. (2021), Hjort et al. (2022)

Introduction

- Many of the most popular machine learning models are black box models
- The field of Explainable AI aims to look inside the black box through methods like SHAP (Lundberg et al. (2017)), LIME (Ribeiro et al. (2016))
- These are post-hoc explanations
- Some, most notably Rudin (2019), argue that one should use inherently interpretable models instead; *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*

Introduction

We seek to answer the following questions:

1. Can Generalized Additive Models (GAM) match the performance of the state-of-the-art machine learning methods for house price prediction?

Introduction

We seek to answer the following questions:

1. Can Generalized Additive Models (GAM) match the performance of the state-of-the-art machine learning methods for house price prediction?
2. Can the GAMs be improved by constructing an ensemble of local models rather than a single global model?

We answer this by studying $N = 29\,931$ transactions from the Oslo (Norway) in 2018-2019

Table of contents

A map of Oslo

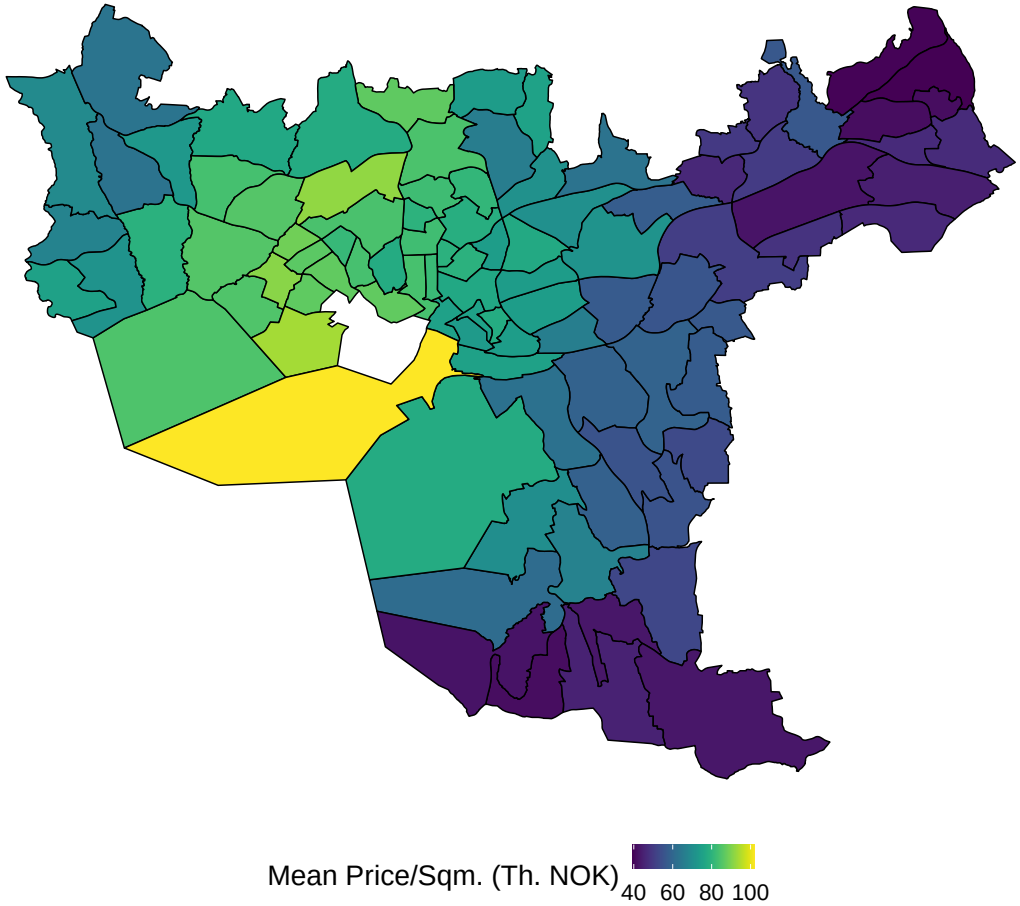


Figure: The $K = 96$ city districts in the data set with mean sale price per m^2 (measured in thousand NOK). 1 NOK $\approx$ 0.1 USD.

The Oslo (2018-2019) data set

Variable	Unit	Mean	St. Dev.	Min	Max	Type
Sale Price	NOK (mill.)	4.69	2.14	1.26	67.5	Numerical
City District	-	-	-	-	-	Categorical
Sale Date	months	9.57	6.00	1.00	24.00	Numerical
Altitude	m	90.27	61.68	0	480	Numerical
Size	m^2	65.63	24.24	15.00	370.00	Numerical
Floor	-	3.02	1.89	-4	14	Numerical
Bedrooms	-	1.79	0.76	0	9	Categorical
Dwelling Age	years	61.27	37.4	0	218.00	Numerical
Balcony	-	0.75	0.43	0	1	Binary
Elevator	-	0.37	0.48	0	1	Binary
Units On Address	-	20.54	27.49	0.00	274.00	Numerical
Coast Distance	*m*	3,160	2,395	5	12,201	Numerical
Lake Distance	*m*	966.60	497.37	31	3,183	Numerical
Nearby Homes	-	2,815.72	1,589,6	100	6,746	Numerical
Nearby Buildings	-	166.66	144.38	6	1,323	Numerical

Table of contents

A decision tree

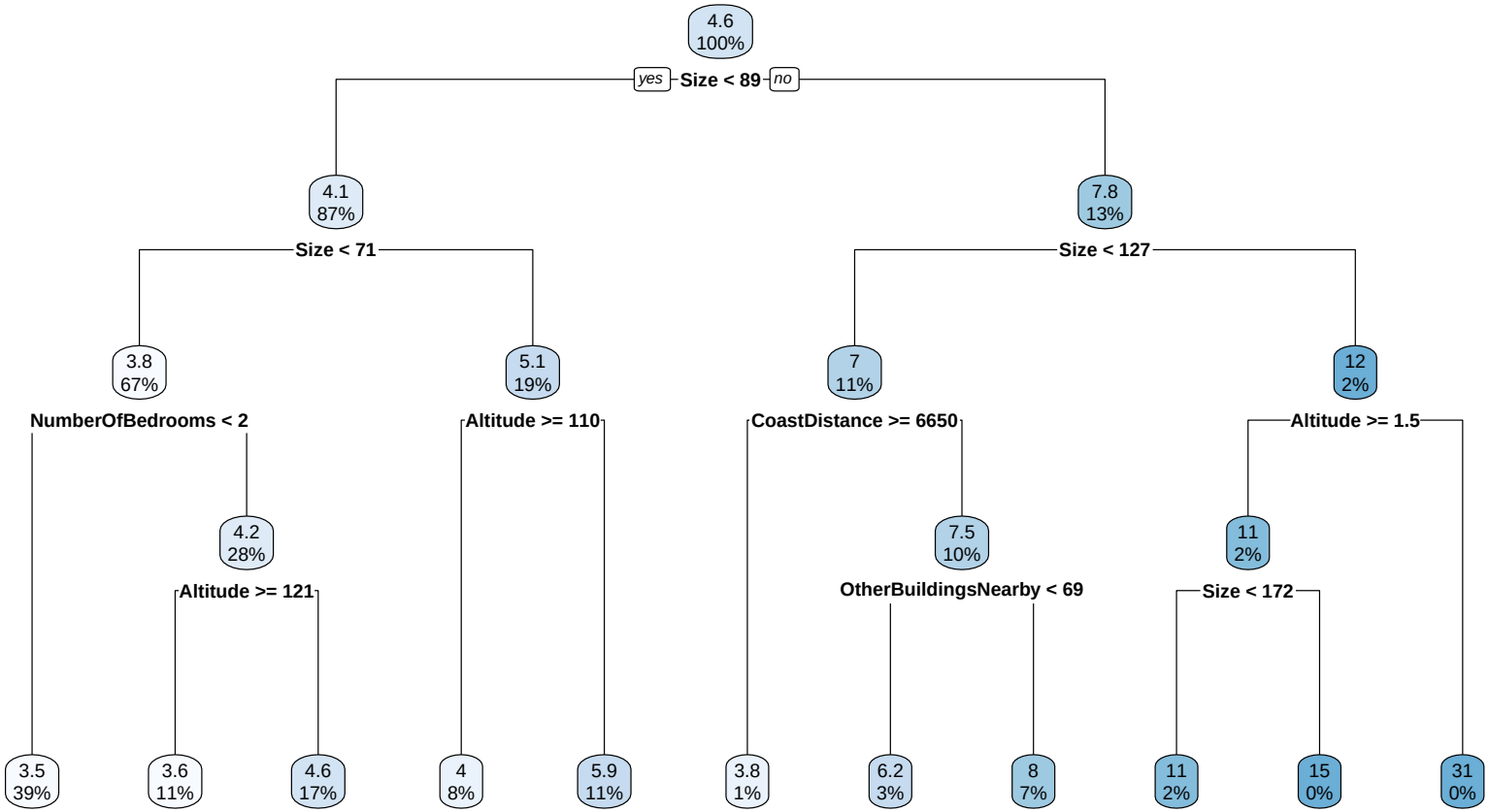


Figure: A decision tree with max depth of 4, dividing feature space into 11 distinct regions with a constant prediction in each region.

Gradient boosted trees

A gradient boosted tree ensemble trains a sequence of (thousands of) trees, each tree trained on the residuals from the previous tree. Introduced by Freund (1995), Freund and Schapire (1996), popularized and made accessible by the XGBoost implementation (Chen et al. (2016)).

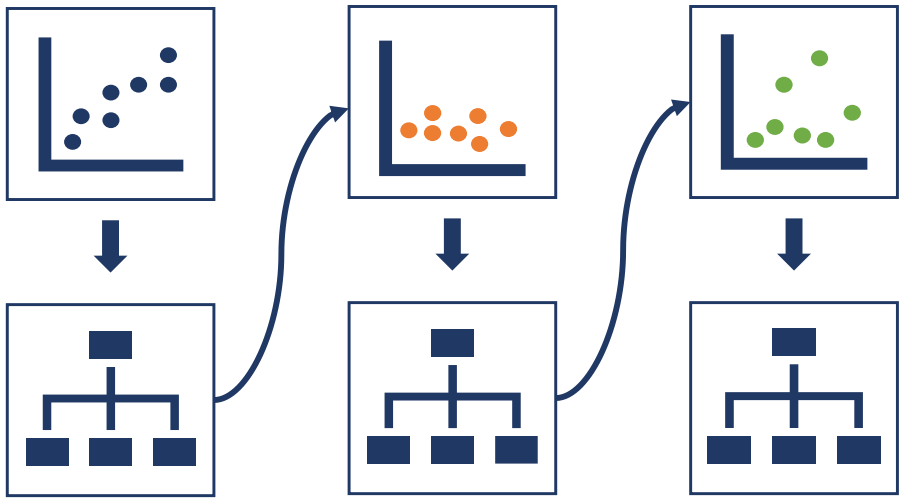


Figure: A visual explanation of how a sequence of boosted trees work.

Many tree stumps in a sequence form a GAM

Decision trees with a single split are called tree stumps. Lou et al. (2013) demonstrate how gradient boosted trees with tree stumps can be written as a Generalized Additive Model (GAM):

$$\hat{y} = f_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p).$$

No interactions; yields a highly interpretable model.

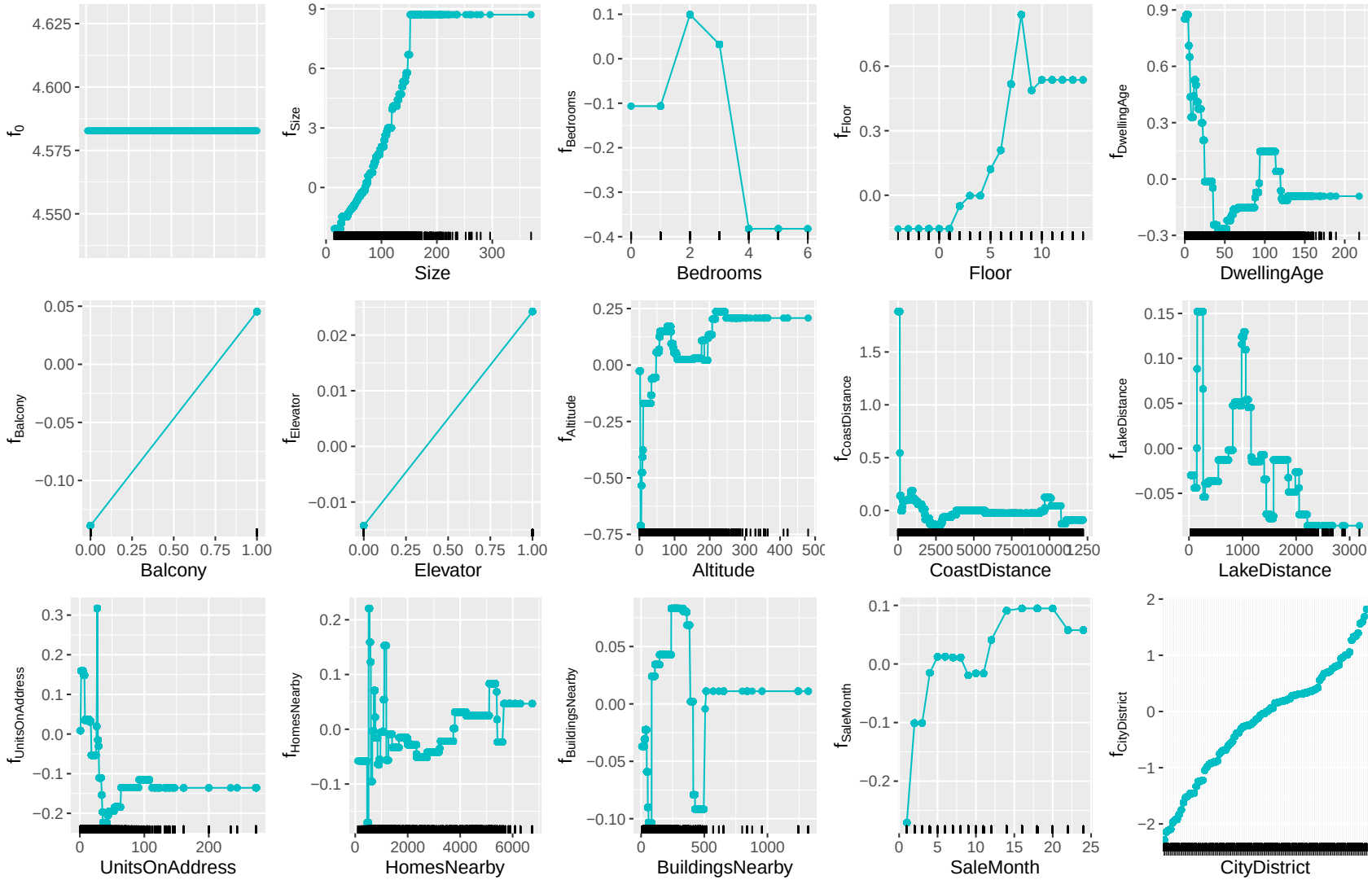
Table of contents

Gradient boosted trees outperform GAM

Model	RMSE (%)	MdAE (%)	R^2 (%)
GAM (tree)	15.5	8.4	82.3
Gradient boosted trees	**10.1**	**5.4**	**89.1**

Table: Predictive performance on $N_{test} = 14\,965$ observations. All methods have used a learning rate of $\eta = 0.05$ and $M = 10\,000$ iterations. The gradient boosted trees is trained with XGBoost and with trees of depth $D = 8$.

The shape functions



Explanations are available a priori

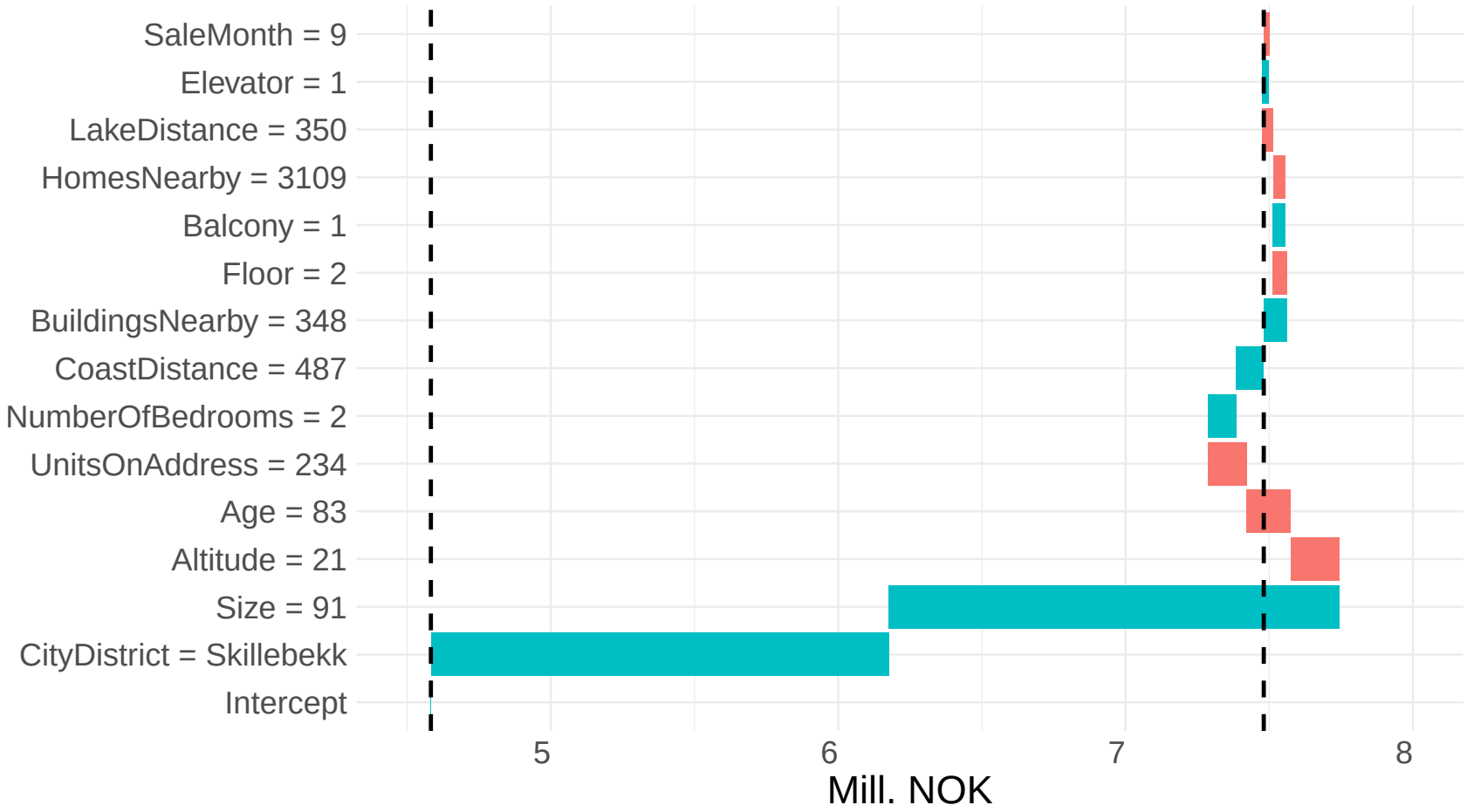


Figure: An explanation of the prediction $\hat{y} = 7.48$ from the GAM model for one specific dwelling. The true sale price was $y = 7.30$.

Table of contents

Local models?

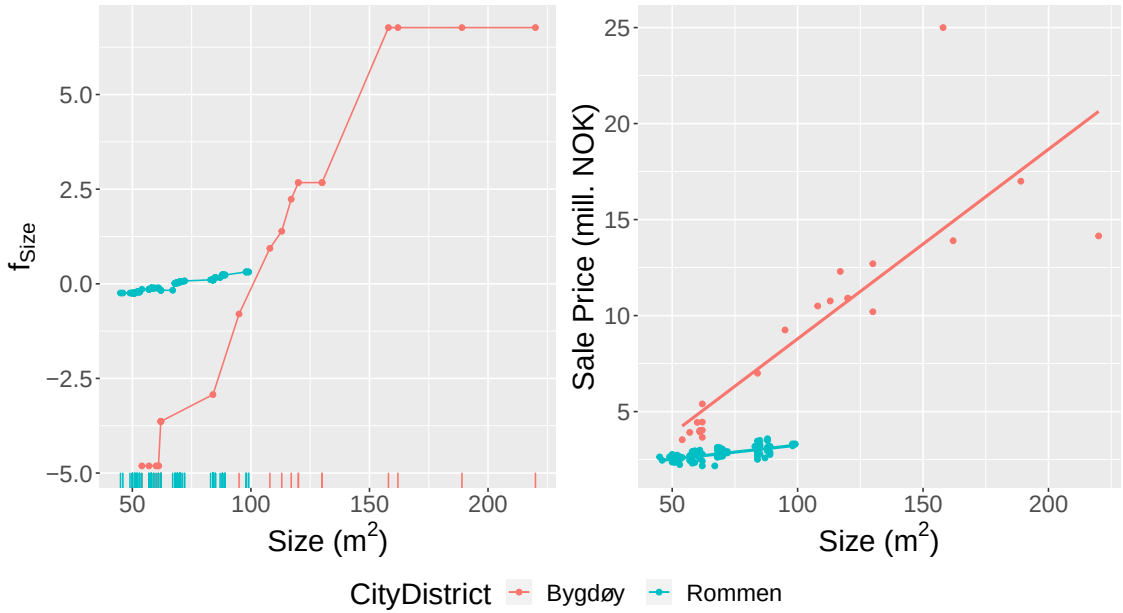
Challenge:

A GAM without interactions assumes/forces equal shape of $f_j(x_j)$ in each city district.
What about trying the polar opposite, i.e., a separate GAM model for each city district?

Local models?

Challenge:

A GAM without interactions assumes/forces equal shape of $f_j(x_j)$ in each city district.
What about trying the polar opposite, i.e., a separate GAM model for each city district?



A method for clustering of city districts

Start with $K = 96$ local models. Cluster by out-of-sample performance: If model trained on City District A gives a good performance on City District B, cluster the two.

A method for clustering of city districts

Start with $K = 96$ local models. Cluster by out-of-sample performance: If model trained on City District A gives a good performance on City District B, cluster the two.

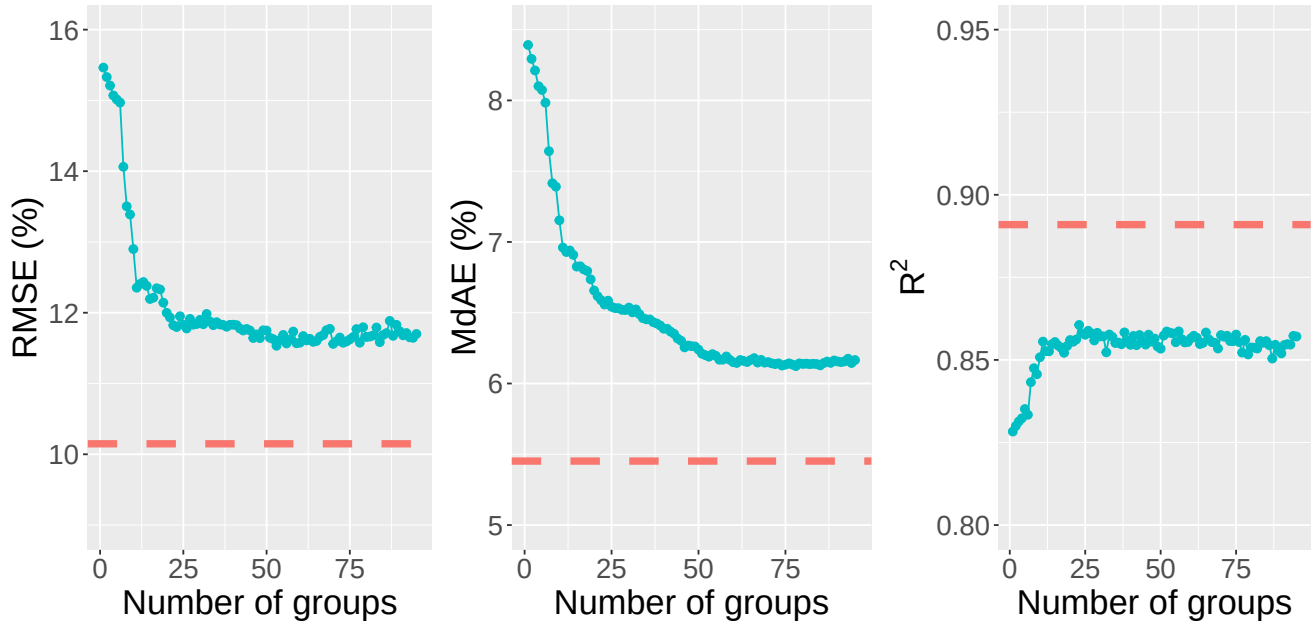


Figure: Simulation results with a *greedy* algorithm as the workhorse for clustering.

Clustering results

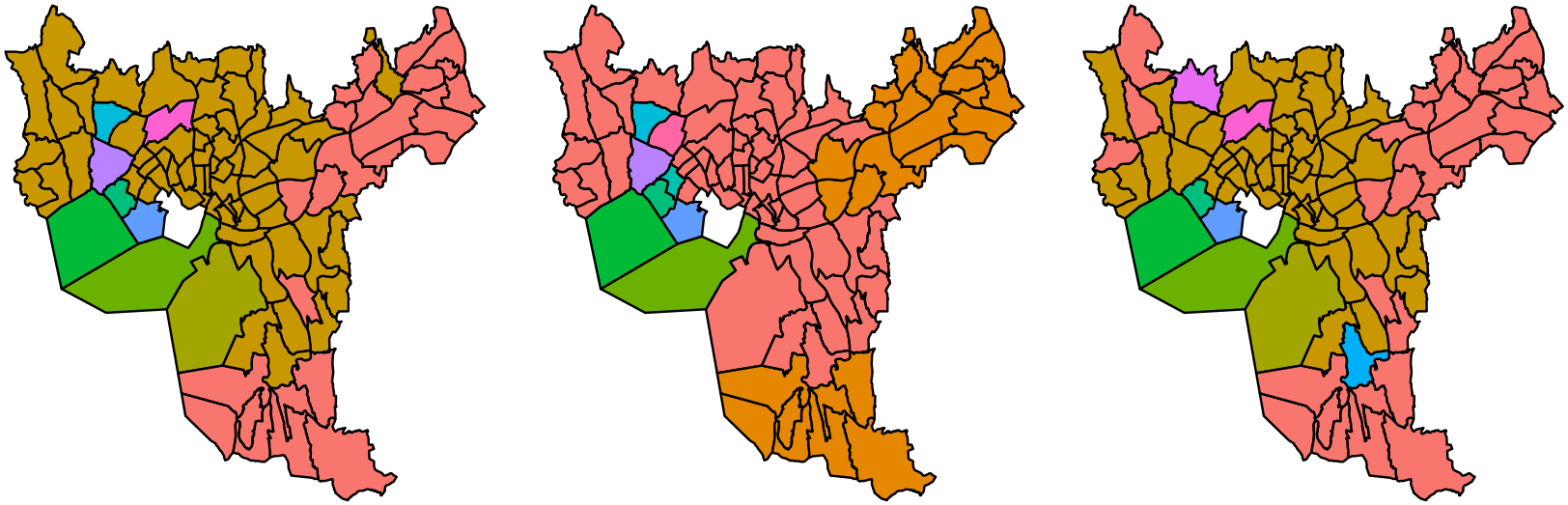


Figure: Three examples of how the $K = 96$ city districts are clustered into 10 groups.

Results

Model	RMSE (%)	MdAE (%)	R^2(%)
GAM (1 group)	15.5	8.4	82.3
Gradient boosted trees	**10.1**	**5.4**	**89.1**
GAM (greedy, 96 groups)	11.7	6.2	85.7
GAM (greedy, 53 groups)	11.5	6.2	85.8
GAM (greedy, 25 groups)	11.8	6.5	85.8
GAM (greedy, 20 groups)	12.0	6.7	85.6
GAM (greedy, 10 groups)	12.9	7.2	85.1

Conclusion

- Generalized Additive Models (GAMs) provide house price prediction models that are fully interpretable as opposed to black box models like gradient boosted trees.
- GAMs have significantly worse performance than gradient boosted trees; 15.5% RMSE vs. 10.1%
- We introduce a greedy clustering algorithm and cluster together city districts based on similarity. This ensemble of local GAMs improve performance to 11.5% RMSE, while still remaining fully interpretable

Further research

- Larger data set? The dwellings in Oslo are probably highly homogeneous. Interesting to try this on all of all of Norway rather than just Oslo
- Mixed effects models are good at handling different effects for different city districts – but do they work with tree models?
- Make better use of the spatial dimension when clustering; city districts that are close should probably be clustered together!

UiO : Department of Mathematics
University of Oslo

Anders Hjort, Johan Pensar, Ida Scheel, Dag Einar Sommervoll

Interpretable House Price Prediction Using a Collection of Local Machine Learning Models

WEAI 97th Annual Conference
Portland (Oregon)