UiO : University of Oslo

House Price Prediction with Confidence: Empirical Results from the Norwegian Market

Anders Hjort [1]

[1]Department of Mathematics, University of Oslo & Eiendomsverdi AS

anderdh@math.uio.no

Introduction

Automated Valuation Models (AVMs) are models used for risk regulating purposes by banks and other financial institutions to get an assessment of the estimated value of a portfolio of dwellings. Tree-based machine learning models like random forest and gradient boosted trees are often the preferred choice for this task due to their high predictive accuracy (Hjort et al. 2022), but uncertainty quantification has historically been a major challenge for these models. In this study we try to tackle this challenge by utilizing a set of techniques from the conformal prediction (CP) literature. We compare three different methods uncertainty quantification on a data set consisting of $N = 29\,933$ transactions from Oslo (Norway) from 2018-2019, and analyze the methods in terms of empirical coverage and the size of the produced confidence regions.

Methodology

We use a random forest model (500 trees, max depth of 10) to create a point prediction $\hat{y}_{n+1} \in \mathbb{R}$ given the known features $x_{n+1} \in \mathbb{R}^d$. We aim to construct a confidence region $C(x_{n+1})$ s.t. $P(y_{n+1} \in C(x_{n+1})) \geq 1 - \alpha$ for the true value y_{n+1} and some confidence level α. We use the following three conformal prediction methods:

1. **Normalized split CP**. We calibrate the confidence intervals on the weighted residuals $R_i = \frac{|y_i - \hat{y}_i|}{\sigma(x_i)}$, where $\sigma(x_i) = \exp\{\gamma \cdot \hat{\mu}(x_i)\}$ and $\hat{\mu}(x_i)$ is a GAM model fitted on $\log|y_i - \hat{y}_i|$ from the training set, as described by Bellotti et al. 2021, and γ is a hyper parameter set to be 0.8.
2. **Conformalized quantile regression (CQR).** We use quantile regression (via the `quantregForest` package) and conformalize the confidence regions via the methods described by Romano et al. 2019.
3. **Mondrian CQR.** We follow the same procedure as in CQR, but in a Mondrian fashion. We construct the confidence regions separately for each of the 15 different city districts in Oslo. The motivation behind this is that house prices vary significantly in different parts of a city. This is also seen in the data set; the mean price per m^2 varies from 42 500 in the cheapest city district (Stovner) to 88 200 NOK in the most expensive city district (Frogner).

For all the methods we divide the full data set into a training set, calibration set and test set of equal size. We construct confidence regions at $\alpha = 0.1$.

The data set

The data set consists of $N = 29\,933$ transactions of apartments from the open housing market in Oslo (Norway) between 2018 and 2019. 1 NOK $\approx$ 0.1 USD per August 2022.

Table 1: The variables in the data set with summary statistics for the numerical variables.

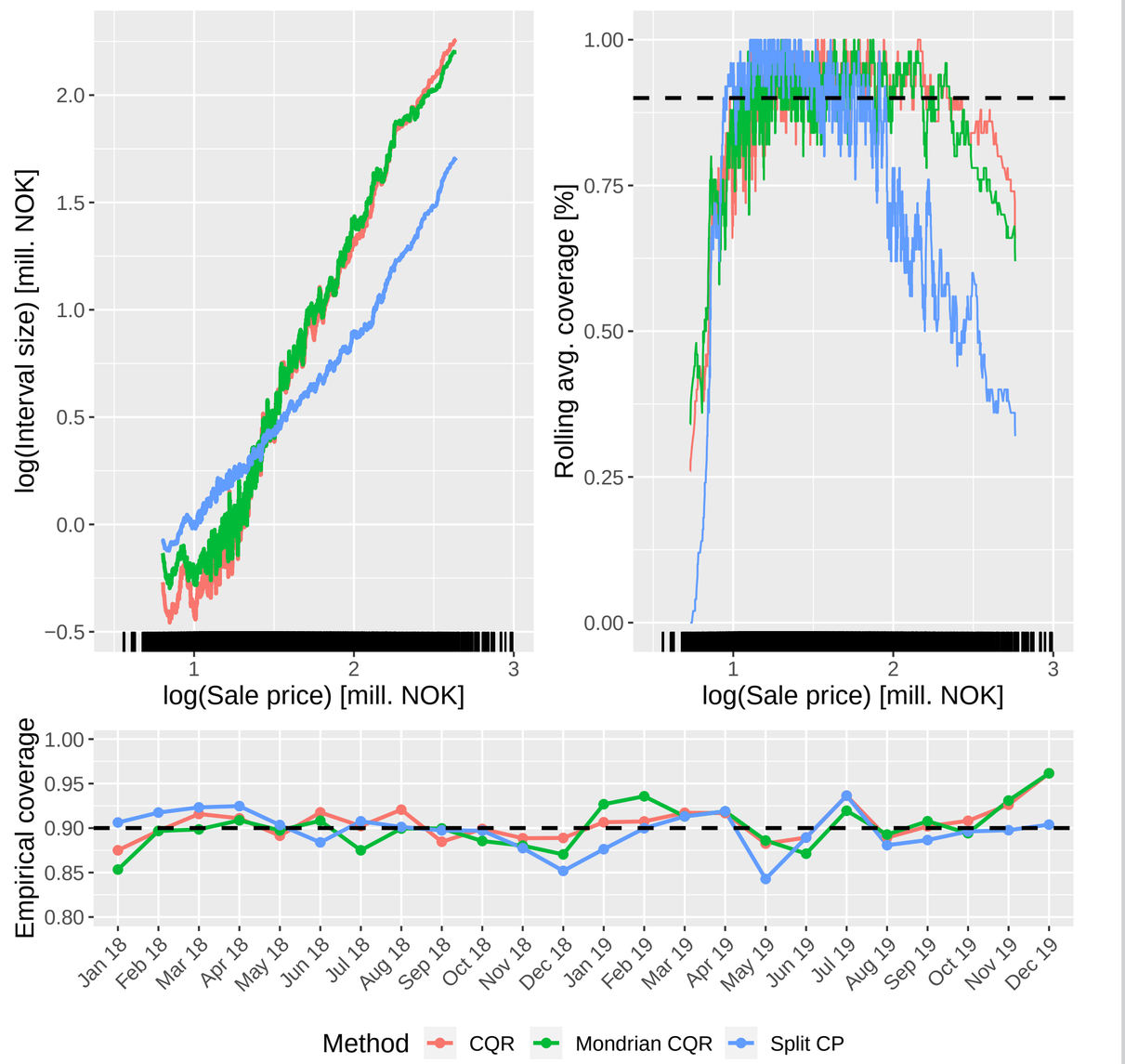
Variable	Unit	Mean	Median	St. Dev.	Min	Max	Type
Sale Price	NOK (mill.)	4.69	3.93	2.14	1.26	67.5	Numerical
City District	-	-	-	-	-	Categorical	
Sale Date	months	12.54	13.00	6.78	1.00	24.00	Numerical
Altitude	m	90.27	76.00	61.68	0	480	Numerical
Size	m^2	65.63	63.00	24.24	15.00	370.00	Numerical
Floor	-	3.02	3.00	1.89	-4	14	Numerical
Bedrooms	-	1.79	2.00	0.76	0	9	Categorical
Dwelling Age	years	61.27	60.00	37.4	0	218.00	Numerical
Balcony	-	0.75	1.00	0.43	0	1	Binary
Elevator	-	0.37	0.00	0.48	0	1	Binary
Units On Address	-	20.54	12.00	27.49	0.00	274.00	Numerical
Coast Distance	m	3,160	2,483	2,395	5	12,201	Numerical
Lake Distance	m	966.60	911.00	497.37	31	3,183	Numerical
Nearby Homes	-	2,816	2,585	1,590	100	6,746	Numerical
Nearby Buildings	-	166.66	131.00	144.38	6	1,323	Numerical

Results

Table 2: Results from the Oslo data set at confidence level $\alpha = 0.1$. Interval sizes are given in million NOK, where 1 NOK $\approx$ 0.1 USD per August 2022.

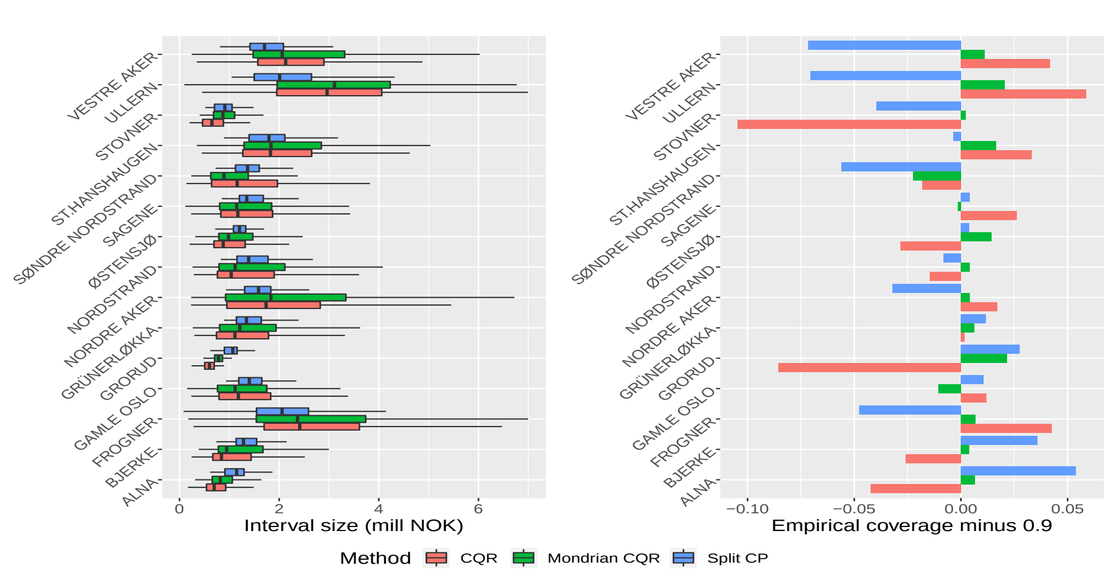
Method	Coverage (%)	Mean interval size	Median interval size
Split CP	89.54	1.85	1.61
CQR	90.25	1.79	1.23
Mondrian CQR	90.40	1.85	1.25

Figure 1: Results for the Oslo data set at confidence level $\alpha = 0.1$. **Left:** Rolling average interval sizes vs. actual sale price. **Middle:** Rolling average empirical coverage vs. actual sale price. **Bottom:** Empirical coverage over time.



Results per city district

Figure 2: Box plots of results for each of the 15 city districts. **Left:** Mean interval size per city district. **Right:** Empirical coverage per city district.



References

- Bellotti, Anthony and Zhe Lim (2021). "Normalized nonconformity measures for automated valuation models". In: *Expert Systems with Applications* 180, pp. 115–165.
- Hjort, Anders, Johan Pensar, Ida Scheel, and Dag Einar Sommervoll (2022). "House price prediction with gradient boosted trees under different loss functions". In: *Journal of Property Research* 0.0, pp. 1–27.
- Romano, Yaniv, Evan Patterson, and Emmanuel Candes (2019). "Conformalized Quantile Regression". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Discussions and further work

While all the methods have empirical coverage close to 90%, the CQR methods seem to results in more well calibrated confidence regions for higher sale prices. As expected, the Mondrian CQR method yields more consistent empirical coverage across city districts. **Further research** plans include a more thorough investigation of methods that account for covariate shift, both spatially and temporally. Figure 1 shows no clear trend over time, but it will be interesting to study this over a larger time period that includes booms and busts in the housing market.