Fagereng, Andreas, Holm, Martin Blomhoff, and Torstensen, Kjersti Næss. ‘Housing Wealth in Norway, 1993–2015’. 1 Jan. 2020 : 65 – 81.
DOI: 10.3233/JEM-200471

Housing Wealth in Norway, 1993 – 2015[*]

Andreas Fagereng
BI Norwegian Business School and Statistics Norway

Martin Blomhoff Holm
University of Oslo and Statistics Norway

Kjersti Næss Torstensen
Norges Bank and Statistics Norway

We provide a new estimate of household-level housing wealth in Norway between 1993 and 2015 using an ensemble machine learning method on housing transaction data. The new housing wealth measure is an improvement over existing data sources for two reasons. First, the model outperforms previously applied regression models in out-of-sample prediction precision. Second, we extend the sample of estimated housing wealth by including cooperative units, non-id apartments, and cabins.

Keywords: Machine learning, housing wealth, house prices
JEL Classification Codes: C50, D10, D63, R30

[*] Corresponding author: Martin Blomhoff Holm, Department of Economics, University of Oslo; tel: (+47) 46410134; e-mail: m.b.holm@econ.uio.no; web: https://sites.google.com/site/martinblomhoffholm/. The article should not be reported as representing the views of Statistics Norway or Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Statistics Norway or Norges Bank.

1 Introduction

The ideal data to study household economic behavior contain panel information on consumption, income, balance sheets, and household characteristics. Administrative tax data are almost ideal, but often suffers from an important drawback: housing wealth is often reported at values that do not reflect market values. In Norway, housing wealth has been undervalued in administrative tax data as values have been related to original transactions prices dating back in time. Furthermore, once tax authorities update property values, the homeowner may have an incentive to complain when the tax-assessed value is too high is, but not when it is too low. These potential systematic errors could contaminate inferences made by researchers. For example, if housing wealth is biased, measures of wealth inequality would also be biased.

Here, we document how we construct an updated value of the housing wealth stock for the universe of Norwegian households from 1993 to 2015. We apply two methods: ensemble machine learning techniques and hedonic regressions. Since 2010, the Norwegian Tax Administration has applied hedonic regressions to measure housing wealth.[1] The *estimated* property values[2] after 2010 are therefore reliable and contain limited systematic measurement errors. However, although the regression models used by the tax authority perform well, our preferred machine learning method significantly outperforms the regression models based on out-of-sample predictive performance. For example, 95% of predicted house prices from the machine learning model are within 20% of observed transaction prices compared with only 67% for the hedonic regression models.

Beyond re-estimating housing wealth, we extend the sample of estimated housing units in three dimensions: cooperative units, cabins, and non-unique apartments. First, about 16% of housing units in Norway are cooperative units and these units are not included in the estimation of market values used by Statistics Norway (see e.g. Kostøl and Holiløkk, 2010). We estimate housing values of cooperative units going back to 1993. Second, we include cabins in the sample of estimated housing units. Third, although there is an official id system for housing units in Norway, this system does not provide all housing units with unique identifiers in some densely populated areas.[3] This non-uniqueness implies that many housing units are either missing or have mis-measured housing wealth in the tax data. We use repeated population censuses to identify

[1] These regression models are based on the results in Kostøl and Holiløkk (2010) and are updated yearly by Statistics Norway. Recently, Stubhaug (2017) has estimated housing wealth using this approach going back to 1993.

[2] We distinguish here between estimated property values and the property values observed in the tax registry. The estimated property values come directly from a hedonic regression model and is unbiased. The property values observed in the tax registry are adjusted after complaints from homeowners.

[3] For example, several apartment buildings in Oslo have the same building id, implying that individual apartments are not uniquely identified by the official unit id.

these housing units to supplement the official data.

Equipped with the novel measure of property values for all housing units, we compute housing wealth at the household level using ownership data. With the new housing wealth numbers, the share of households with negative net wealth in the administrative data drops from between 22% and 34% to around 17% in all years in our sample.[4] After 2010, a period when the tax authority has a relatively precise measure of housing wealth, the share of households with negative net wealth drop from about 22% to 17%. About half of this drop is explained by the re-estimated housing wealth numbers and the rest is explained by including a larger share of the housing stock.

This paper proceeds as follows. Section 2 first describes the data. Section 3 next presents the estimation methods before we evaluate the predictive performance of these methods in Section 4. The two next sections evaluates the new data by first comparing it to other available house price indexes in Norway in Section 5 and then constructing a measure of housing wealth and net wealth in Section 6. Section 7 concludes.

2 Data

This section presents our data sources and the properties of the final datasets. The discussion is restricted to regular housing. Appendix A provides details on cooperative units.

Data sources. The main dataset with transactions and housing unit characteristics is built on four separate sources.

1. **Land registry (L)**. The land registry[5] from the Norwegian Mapping Authority (Kartverket) is the official registry of housing in Norway and contains detailed information on all properties, including housing unit characteristics and ownership information. It is annual from 2005 to 2015.

2. **Ownership data (O)**. The ownership dataset[6] from the Norwegian Tax Administration contains ownership information and housing unit characteristics. In addition, it contains the estimated value from the tax authority which we use as a cross-check in the validation of our estimated house prices. The data is annual and available to us from 2013 to 2015.

[4] 17% is more in line with households surveys from other countries. For example, about 10% have negative net worth in 2007 in the Survey of Consumer Finances (SCF), a little more than 10% have negative net worth in the PSID in 2007 Pfeffer, Schoeni, Kennickell, and Andreski (2016), while less than 10% have negative net wealth in the Italian Survey of Household Income and Wealth (SHIW).

[5] Includes both "Grunnbok" and "Matrikkel".

[6] Skatteetatens egne Sentrale eiendomsregister (SERG) is based on the land registry, but includes also all cooperatives.

3. **Transaction data (T)**. The transaction dataset is based on information from the land registry. The dataset contains all transactions of regular housing and cabins between 1993 and 2015.[7] It contains the housing id, information about buyers and sellers, and the transaction price. In addition, the dataset contains housing unit characteristics from the land registry.

4. **Population census (C)**. We use information from the last three waves of the population census (1990, 2001, and 2011) containing information on housing unit characteristics, all individuals living in the house, and ownership status.

Data construction. We construct a dataset of characteristics over our sample period by combining information from all four sources using the unique id from the Norwegian mapping authority. Table 1 lists the included housing characteristics. Some housing characteristics exist in multiple datasets and all datasets contain missing variables for some observations. Two datasets sometime have different values for the same characteristics. In most cases, the values are very similar and simply rounding errors. Nevertheless, we impose a ranking of the datasets according to a judgment of their reliability. The ranking is: ownership data, transaction data, land registry, and population census.

[Table 1 here]

For some characteristics, the data source only exists after 2005. To construct a dataset for the full sample going back to 1993, we assume that the housing unit had the same value for the characteristics in the years prior to 2005. Some housing units do not have a unique id from the mapping authority. In this case, there is no information on the housing unit from the land registry and the ownership registry. These properties with incomplete id's are most prevalent among apartments in large cities. To obtain information on such housing units, we use the population census data. The census data does not contain the unique id from the mapping authority, but we can find unique housing units by combining ownership information from the census and the individual's id in the transaction data. Further, we impute ownership fractions from the census data by counting the number of adults living in the housing unit and being registered as owners.[8]

Detailed ownership registry is only available from 2005 onwards. However, using the transaction data we extend the ownership dataset to 1993.

[7] We only have transaction data for cooperatives from 2007 to 2015.

[8] The population census has no information about ownership fractions, but it contains all family members that live in the housing unit. We impute the ownership fraction by first removing all persons below 35 in the housing unit and then give every remaining household member an equal share of the housing unit.

We combine information from all four data sources to construct two datasets: the characteristics dataset and the ownership dataset. The characteristics dataset is unique on housing unit-year observations and used to estimate property values. The ownership dataset is unique on person-housing unit-year observations and contains only the person id, the housing unit id, and the ownership fraction, and is used to construct household-level housing wealth.

Sample, variable, and index selection. The next step is to select an estimation sample from our characteristics dataset. First, we select housing units where the whole property is transacted as a market transaction, the total price of the property is greater than the equivalent of NOK 800,000 (USD 100,000) in 2015, the size of the housing unit is greater than 20 m^2 and less than 500 m^2, and the number of rooms is less than 16. In addition, we exclude transactions in cooperatives from the estimation because these units often have a share in cooperative debt that we do not observe.[9] We further trim the top and bottom 1% on the price per m^2.

We also select the variables and indexes we want to estimate. We rely on the set of variables and specifications already used by Statistics Norway to estimate property values Kostøl and Holiløkk (2010). Table 1 presents an overview of the variables included in the dataset. We estimate the regressions separately for three house types (apartments, chained-houses, and regular houses) within 22 regions,[10] producing housing wealth measures with both geographic and type-specific variation. Since we include a measure of zones within geographical regions, which includes dummies for municipality, city district, and zip codes, there is additional geographical variation within the 22 regions. The cabin sample is separated into only three broader regions (East, West, and North) to allow for a sufficient number of market transactions within each region.

3 Estimation methods

We apply two methods to estimate property values: hedonic regressions and an ensemble machine learning method. This section describes each of these methods.

[9] The observed transaction prices for cooperative units is not the market price of the housing unit itself, but the price of the unit net of its share of cooperative debt. Since we do not observe the unit's share of cooperative debt, we cannot use the transaction values of cooperative units in the estimation setup as for regular housing.

[10] The regions are: Oslo and Bærum, Bergen, Trondheim, Stavanger, Østfold, Akershus (excl. Bærum), Hedmark, Oppland, Buskerud, Vestfold, Telemark, Aust-Agder, Vest-Agder, Rogaland (excl. Stavanger), Hordaland (excl. Bergen), Sogn og Fjordane, Møre og Romsdal, Sør-Trøndelag, Nord-Trøndelag, Nordland, Troms, and Finnmark. For cabins, we use wider regions: Østlandet (county no. 1-9), Vestlandet (county no. 10-15), and Nordlandet with Trøndelag (county no. 16-20).

Hedonic regressions. The idea behind hedonic regressions is to use the information of housing characteristics to predict property values in the transaction data. The tax authority in Norway already applies hedonic regressions to estimate housing wealth based on the results in Kostøl and Holiløkk (2010).[11] The hedonic regressions are therefore the benchmark to which we compare our machine learning approach.

There are two choices that determine the performance of an hedonic regression, the functional form and the choice of variables to include. We choose a functional form close to Kostøl and Holiløkk (2010) that has already been applied to Norwegian data

$$\begin{aligned}\log(p_{i,j,t}/m^2_{i,j,t}) = \beta_{0,j} + \beta_{1,j,t}\log(m^2_{i,j,t}) + \beta_{2,j,t}\log(m^2_{i,j,t})^2 + time_{t,j} \\ + type_{j,t} + zone_{i,j,t} + density_{i,j,t} + age_{i,j,t} + garage_{i,j,t} + psize_{i,j,t} + \epsilon_{i,j,t}\end{aligned} \tag{1}$$

where i is the housing unit, $j \in [1, 69]$ is the type-location index,[12] t is year, m^2 is square meters (bruksareal), *time* is a quarterly index, *type* is a fixed effect for house types within the broad type, *zone* is a fixed effect for zones within the geographical region (municipality in counties or neighborhoods in cities), *density* is a dummy which is one if the house is in densely populated area $\{0, 1\}$, *age* is a set of age-groups for the housing unit $\{0, 5, 15, 25, \geq 35\}$, *garage* is a dummy for the existence of a garage $\{0, 1\}$, *psize* is a set of plot sizes in squared meters $\{0, 500, 1000, 1500, \geq 2000\}$, and ϵ is an error term. All coefficients are time-dependent. Our housing price specification therefore allows for substantial time and zone heterogeneity also within geographic areas and housing types.

Machine learning. The second method is machine learning. A challenge when using regression models for predictions is overfitting. By including more variables, we get a better in-sample fit, but could potentially worsen the out-of-sample performance because the model becomes too specific to the estimation sample (overfitting). This trade-off between in-sample fit and out-of-sample predictive performance is at the heart of all machine learning methods.

A common feature of machine learning algorithms is that they are estimated partially based on pseudo out-of-sample predictive performance. In particular, the data is separated into folds in each estimation. The algorithm uses data from N-1 folds and evaluate the model by predicting the Nth fold. The final model is an averaged model selected on its pseudo out-of-sample performance. The estimation step therefore finds the best prediction model conditional on model complexity.

[11] Stubhaug (2017) use a similar approach to construct a housing wealth measure in Norway back to 1993.

[12] 3 types and 22 locations for regular housing and cooperatives, and 3 locations for cabins. The 22 locations corresponds to the same geographic areas used in the official housing price data from Statistics Norway.

In addition, machine learning methods require a tuning step. All machine learning algorithms contain a measure of model complexity, the regularizer. For example, the regularizer is the maximum depth of the trees in a regression tree model. An important step is to choose optimal complexity. If we choose a model with high complexity, we get a good in-sample fit, but could worsen the out-of-sample performance due to overfitting. We therefore tune our model by performing the estimation procedure for a wide range of regularizers and choose the combination of tuning parameters that yield the best predictive performance within our estimation sample. After choosing the optimal regularizer, we perform the model estimation by running the N-fold procedure described above with the tuned model.

There are many machine learning algorithms and the appropriate method depends on the data availability and the research question.[13] We use an ensemble approach that weights together predictions from three different algorithms: regression tree, LASSO, and random forest. The estimation is finally a two-step procedure. First, we tune each algorithm to best fit our data. Second, the weights in the final ensemble prediction are calculated based on the relative out-of-sample performance of each algorithm. Our approach is based on the codes in Mullainathan and Spiess (2017).[14]

4 Predictive performance

In this section, we describe the relative predictive performance of the two estimation procedures: the hedonic regression and the ensemble machine learning method. In addition, we compare our models to using the tax values directly.[15]

Prior to estimating the models, we first randomly select 1/8 of the sample of transactions that we keep out of all estimations and tuning steps ("the holdout sample"). We then apply all methods on the remaining 7/8 of the sample.[16] We evaluate predictive performance based on two criteria: the root mean squared error of the prediction (RMSE) and the distribution of deviations between predicted and transacted housing values in the holdout sample. The regression is specified in logs, so the root mean squared error can be interpreted as the average percentage deviation of the model from the observed transaction price.

[Table 2 here]

[13] For recent surveys of machine learning for economists, see Varian (2014) and Mullainathan and Spiess (2017).

[14] The estimation is performed in R and based on the following packages: randomForest (Liaw and Wiener, 2002), rPart (Therneau and Atkinson, 2018), and glmnet (Friedman, Hastie, and Tibshirani, 2010).

[15] We adjust the tax values in the tax reports by the inverse of the tax discount to get the assessed value by the tax authority.

[16] To be precise: this means that the machine learning method use eight (N) folds within the 7/8 sample we select.

Table 2 presents the results. The ensemble machine learning method significantly outperforms the hedonic regression, both in-sample and out-of-sample. The out-of-sample root mean squared error (RMSE) is 0.11, suggesting that the prediction from the ensemble method deviates on average by 11% from the observed transaction price, a significant improvement on the 26% obtained by the hedonic regression model. Furthermore, 95% of the predictions from the ensemble model is within +/- 20% of the observed transaction price, compared with 67% for the hedonic regression.[17] Since the tax authority has applied hedonic regressions since 2010 and other housing wealth papers on Norwegian data either rely on tax data or use hedonic regressions (Stubhaug, 2017; Aaberge and Stubhaug, 2018; Eika, Mogstad, and Vestad, 2020), our new housing wealth dataset is an improvement on already existing data on Norwegian housing wealth. In addition, the machine learning method and the hedonic regression models outperform the usage of tax data.

Table 2 also presents predictive performance for cabins. The prices of cabins are challenging to estimate because there are few transactions within relevant geographic regions. We therefore rely on three large geographic regions (East, West, and North) to estimate our cabin models. Although there will be substantial heterogeneity within regions, the machine learning methods are still an improvement on the already existing data for cabins. The out-of-sample RMSE of our ensemble method is 0.37 and 55% of cabin prices are within +/- 20% of the observed cabin transaction prices. While this is better than the hedonic regression models, there is still scope for improving the estimates of cabins using more local information and characteristics of the cabin.

[Table 3 here]

Table 3 presents the median adjustment ratios between our housing predictions and the tax data. The adjustment ratios can be interpreted as a measure of the bias in the tax data. A number above one means that the median house price in the tax data is undervalued. There are three notable observations. First, for regular housing, the bias is relatively small after 2010 since the tax authority started to use hedonic regression models. However, it is consistently above one, suggesting that the tax values still are systematically undervalued.

Second, before 2010, when the tax authority did not apply regression models, there is substantial time-variation in the bias, ranging from about 1 to 1.9. The size of this bias is related to the discrete intervals at which the tax authority decided to revise their tax numbers. For example, there is a drop in bias in the specific years (1993, 2001, 2006, 2008, and 2009) when the Norwegian Tax Administration revised their house price index.

Third, the bias is greater for cabins. Cabins are systematically undervalued in the tax registry

[17] The 67% varies by year and has become significantly better during the past couple of years. For example, in the last version of the hedonic estimation equations, the number is 74.5% (Takle and Medby, 2020).

and the errors have increased over time. In the beginning of our sample, the researcher has to double the tax values to get close to the observed market value. In the most recent years in the sample, one instead has to multiply the tax data by more than 4. The median cabin is therefore valued at around 20-25% of its market value in the tax registry after 2013.

5 House price indices

To assess the reliability of our data, we construct house price indices at the country, county, and municipality level and compare these with other available data sources.

Constructing house price indices. The first step in constructing a house price index is to transform the prediction results into individual property values. Our model predicts the log price per squared meters. The value of an individual property is therefore

$$\widehat{p_{i,t}} = exp\left(\log(\widehat{p_{i,t}/m^2_{i,t}})\right) * exp\left(\frac{\widehat{rmse}^2}{2}\right) * m^2_{i,t} \tag{2}$$

where $\log(\widehat{p_{i,t}/m^2_{i,t}})$ is the predicted log price pr squared meters from the model, $\widehat{rmse}$ is the in-sample root mean squared error of the prediction model within the location-type, and $m^2_{i,t}$ is the squared meters of the housing unit. We adjust by the root mean squared error since we estimate the model in logs and the expected price of a log-normal distribution should be adjusted by the standard deviation (≈ RMSE).

We directly observe the transaction price of some housing units. In these cases, we use the predicted price rather than the transaction price for two reasons. First, Anundsen and Larsen (2018) show that there is mean-reversion in house prices at the micro-level. They document that an excessively low or high sell price in one transaction is not repeated in the next transaction. Hence, the predicted house price is, on average, a more precise estimate of the potential market price of the house. Second, we acknowledge that there may be measurement errors in the transaction data. In our estimation models, we reduce the impact of outliers by carefully selecting the sample. By relying on the average model rather than individual observations, we limit the extent to which individual measurement errors may affect the estimated property prices.

We next compute the average house price growth within geographical units by taking the average of all predicted house prices. Our index is therefore volume-weighted and is an index that is representative for the existing housing stock.

[Figure 1 here]

Comparison with other house price indices. There are two other available sources of house price indexes in Norway: the official price index from Statistics Norway and the price index from Eiendomsverdi. Figure 1 compares annual house price growth from 1997 to 2015 from the machine learning approach, Statistics Norway, and Eiendomsverdi. Reassuringly, the house price index from the machine learning methods evolves similarly to the official series from Statistics Norway.

Both the machine learning index and the official house price index from Statistics Norway differ from Eiendomsverdi's house price index. One important difference is that the index from Eiendomsverdi is transaction-weighted while the two others are volume-weighted.[18] Our index and that of Statistics Norway therefore present the average house price growth of the existing housing stock, while the index from Eiendomsverdi presents the average house price growth of transacted units. Since some housing types tend to be transacted more often than others, the two approaches yield different aggregate house price indexes.

We also construct disaggregated house price indexes that we compare with the disaggregated house price indexes from Statistics Norway. Table 4 presents the annual growth rate for four cities in Norway: Oslo and Bærum, Bergen, Trondheim, and Stavanger. The machine learning house price index follows the same pattern as the official house price index from Statistics Norway also at more granular levels.

[Table 4 here]

6 Housing wealth and net wealth

Our goal is to construct an estimate of housing wealth for the universe of Norwegian households. In this section, we describe how we go from our predicted house prices to housing wealth and net wealth.

Housing wealth. In Section 5, we compute the price of individual properties. To obtain a measure of housing wealth, the next step is to distribute the properties to individuals according to ownership shares. In particular, if individual k own fraction f of property i, then that individual has housing wealth $\hat{w}_{k,i,t} = \hat{p}_{i,t} f_{k,i,t}$ in property i. Total housing wealth of individual k is $\hat{W}_{k,t} = \sum_i \hat{w}_{k,i,t}$. Housing wealth for a household is the sum of each household member's housing wealth.

[18] See Boug, von Brasch, and Takle (2018) for a discussion of how house price indexes vary depending on being volume or transaction weighted.

For many housing units, we do not have complete information for each relevant characteristic and ownership fraction. In this case, we impute housing wealth using estimated ratios from the sample of households with both estimated housing wealth and tax-assessed housing wealth. We construct these ratios by first estimating housing wealth at the household level using the method above. Since tax values are discounted, we discount the estimated housing wealth and calculate ratios for each housing type (regular housing, cooperatives, and cabins) for each household.[19] Using this method, we obtain the ratio between estimated housing wealth and tax-assessed housing wealth for households. We then compute the median yearly adjustment ratios at the country and county-level within each property type (owner, coop, and cabin). We presented the yearly adjustment ratios at the country-level in Table 3.[20] About half of our final housing wealth values are estimated while the remaining half are adjusted tax values using county-level adjustment ratios.

[Figure 2 here]

Net wealth. We define net wealth as the sum of all assets minus all liabilities. Assets include deposits and cash, mutual funds, stocks, private equity, bonds, vehicles, housing wealth, and outstanding receivables (mostly private loans). Liabilities include all types of debt (mortgage, consumer loans, student loans, and the share of debt in cooperatives). We aggregate all data to the household level when we compute wealth.

One concern when using administrative tax data is that housing wealth could contain systematic measurement errors. In particular, since liabilities are measured precisely in the tax data, a systematic undervaluation of housing wealth implies that many households have negative net wealth in the data, but should have positive net wealth. Figure 2 reveals that this concern is justified when using tax data by comparing the share of households with negative net wealth in three cases. First, the dashed line shows the share of household with negative net wealth when we compute net wealth using tax data directly (only adjust by discounting in the tax framework). The share of households with negative net wealth varies significantly over time and is correlated with the bias in housing wealth presented in Table 3. Second, to produce the dotted line, we replace the data from the tax administration with our new measures of housing wealth, but we do not

[19] It is important that we use the appropriate discounting. For example, if a household owns only one house and lives in it, this is the primary property and should have a discount of 75% in the tax registry. If a household instead owns multiple properties, these are discounted differentially depending on the property's tax status. For the latest tax discount rules for houses in Norway, see https://www.regjeringen.no/no/tema/okonomi-og-budsjett/skatter-og-avgifter/skattesatser-2018/id2575161/.

[20] County-level adjustment ratios are available upon request.

include the additional observations of non-unique housing units. The share of households with negative net worth is reduced by between 3 and 12 percentage points, revealing a systematic undervaluation in the tax data. Third, the solid line illustrates the share of households with negative net worth when we include our new measures of housing wealth and additional housing units. The share of households with negative net wealth drops further by 3-4 percentage points. In the last part of the sample about half of the drop is explained by the increase in housing wealth for already existing units, while the other half is explained by new observations. Compared with the two other samples, the share of households with negative net wealth with our preferred measure of housing wealth is stable across time around 17%, suggesting that our new measure of net wealth contains less systematic measurement errors.

7 Conclusion

In this paper, we construct a new estimate of household-level housing wealth in Norway between 1993 and 2015. We estimate housing wealth using a machine learning method and show that our new housing wealth measure is a significant improvement on existing data in Norway for two reasons: (i) our prediction model outperforms previously applied regression models in out-of-sample precision and (ii) we extend the sample of estimated housing wealth by including cabins, cooperatives, and non-id apartments. We find that using our new housing values, the share of households with negative net wealth drops to 17%, which is about 5 percentage points lower than using unadjusted values.

Although the machine learning methods significantly outperforms the hedonic regression models by the tax authorities, there is still room for improvement. In particular, all estimation methods are more precise with more data. The combination of geographical regions and housing types with a low number of transactions may therefore contain large measurement errors. Housing wealth is therefore precisely measured in large cities where there are many transactions per year, but less so in rural areas and for specific housing types. In particular, cabins and apartments in rural areas still contain relatively large measurement errors in our sample and should still be treated with caution.

8 Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 851891). The authors would like to thank the editor, an anonymous referee, Erlend Eide Bø, Ådne Cappelen, Morten Grindaker, Ragnar Juelsrud, and Gisle J. Natvik for comments and suggestions.

References

Aaberge, R., and M. E. Stubhaug (2018): "Formuesulikhet i Norge 1995–2016," *Statistics Norway Analyse*, 18, 2018.

Anundsen, A. K., and E. R. Larsen (2018): "Testing for micro efficiency in the housing market," *International Economic Review*, 59(4), 2133–2162.

Boug, P., T. von Brasch, and M. Takle (2018): "Hvorfor spriker boligprisindeksene til Eiendom Norge og SSB?," 7, 2018.

Eika, L., M. Mogstad, and O. L. Vestad (2020): "What can we learn about household consumption expenditure from data on income and assets?," *Journal of Public Economics*.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1), 1–22.

Kostøl, A., and S. E. Holiløkk (2010): "Reestimering av modell for beregning av boligformue," *Statistics Norway Notater*, 39, 2010.

Liaw, A., and M. Wiener (2002): "Classification and Regression by randomForest," *R News*, 2(3), 18–22.

Mullainathan, S., and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31(2), 87–106.

Pfeffer, F. T., R. F. Schoeni, A. Kennickell, and P. Andreski (2016): "Measuring wealth and wealth inequality: Comparing two US surveys," *Journal of Economic and Social Measurement*, 41(2), 103–120.

Stubhaug, M. E. (2017): "Housing and wealth inequality," *Working Paper*.

Takle, M., and P. Medby (2020): "Modell for beregning av boligformue," *Statistics Norway Notater*, 9, 2020.

Therneau, T., and B. Atkinson (2018): *rpart: Recursive Partitioning and Regression Trees*R package version 4.1-13.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28(2), 3–28.

A Data on cooperatives

Cooperative housing constitutes about 16 % of all housing in Norway. Owners of a cooperative unit also owns a part of the cooperative. On average, cooperatives contain smaller units such as apartments or chained houses with a mean square meter of 63. The typical owner of a cooperative unit is younger and the prevalence of single households is higher than in the rest of the population.

The cooperative housing law (borettslagsloven) regulates the use and ownership of cooperative housing. The law limits letting of the unit to maximum 3 years, under the condition that the owner (or close relative) lived in the unit for at least one of the two previous years. Furthermore, the main rule in cooperative units is that an owner(s) can only own one unit in the cooperative and the owner(s) has to be an individual. For our purpose, this is useful because it implies that the address registry is a good indicator of ownership in the years where we do not directly observe ownership.

In the Norwegian data, there is a separate register for cooperative units. We construct the ownership register by combining information from the ownership register, tax register, population census, and address register. In this section, we explain how we construct the dataset of housing unit characteristics and ownership information for cooperatives.

The data sources. We use five data sources to construct the dataset of ownership in cooperative housing.

1. **Ownership data.** The ownership dataset[21] from the Norwegian Tax Administration contains ownership information and housing unit characteristics. The data is annual and available to us from 2013 to 2015.

2. **Transaction data.** The transaction data contain all transaction of cooperative units between 2006 and 2015. It contains the housing id, information about buyers and sellers, housing unit characteristics, and the transaction price.

3. **Population census**. Statistics Norway conducts the population census approximately every ten years (1990, 2001, and 2011). The population census contains information on housing unit characteristics, all persons living in the house, and information about ownership status.

4. **Tax records.** The tax records contain the tax values for each individual in Norway. We use positive values in the tax registers for housing cooperatives as an indicator that the person owns cooperative housing in a given year.

[21] Skatteetatens egne Sentrale eiendomsregister (SERG) is based on the land registry, but includes also all cooperatives.

5. **Address data.** The Norwegian address data contain information on the inhabitants of each cooperative unit in Norway between 1993 and 2015. Since the law surrounding cooperative units restricts letting, we use the address data to indicate ownership of a cooperative unit.

Data construction. The main challenge in creating the data on cooperative units is to impute ownership. After 2006, we have transaction data and can track ownership, but prior to 2006, we have no information from official registries about ownership. We therefore combine information from all four other data sources to track ownership between 1993 and 2015.

Prior to 2006, the population census provides information on ownership in 1990 and 2001. In addition, we use the tax data and address data to plausibly select owners of individual properties. We denote a person as an owner if both of the following conditions are true:

1. They are listed as owners of a cooperative unit in at least one year in our sample (either from the ownership data, the transaction data, the population census).

2. They still live in the property (from address) or are still listed as owners in the tax registry.

Using this method, we recover information about approximately 200,000 cooperative units with 320,000 unique owners in 1993, to approximately 600,000 cooperative units with 1,100,000 unique owners in 2015.

In addition, we combine information from all sources to construct a data set of housing characteristics similar to that for regular housing. Similar to regular housing, we impose an ordering of the reliability of the sources when multiple sources predict different values. We use the same ordering as in Table 1 except that we do not have any observations from the land registry. In addition, we assume that all cooperative houses are located in areas with high population density.

B Tables

Name	Description	Values	Source Priority			
			1	2	3	4
log(*price pr* m^2)	log (p/m^2)	continuous	T			
log(m^2)	log m^2	continuous	O	T	L	C
log(m^2)^2	(log m^2)^2	continuous	O	T	L	C
log(m^2)^3	(log m^2)^3	continuous	O	T	L	C
time	date in year-quarter	indicator				
type	housing type	indicator	O	T	L	C
zone	municipality, city district, and zip code	indicator	O	T	L	C
density	in dense area	0/1	O	T	L	C
age	age-groups	0/5/15/25/≥35	O	T	L	C
garage	have garage	0/1	L			
psize	size of plot in m^2	0/500/1000/1500/≥2000	L			
rooms	number of rooms	1/2/3/4/5/≥6	O	T	L	C
floor	floor number	discrete	O	T	L	C

Table 1: Variables included in the estimation dataset

Notes: The left part of the table presents the variables included, their description, and the truncated values in the estimation. The right part shows the priority of data sources. T = transaction data, O = ownership data, L = land registry, and C = population census.

		RMSE		Out-of-sample predictive performance						
				percent deviation from transaction value						
	Method	In-sample	Out-of-sample	(-∞, -20)	[-20,-10)	[-10,-5)	[-5,5]	(5,10]	(10,20]	(20,∞)
Housing	Ensemble ML	0.089	0.110	0.026	0.051	0.098	0.657	0.093	0.049	0.026
	Hedonic regression	0.241	0.264	0.170	0.144	0.096	0.207	0.092	0.133	0.158
	Tax Values		0.670	0.184	0.067	0.045	0.118	0.054	0.101	0.431
Cabins	Ensemble ML	0.172	0.368	0.237	0.127	0.082	0.169	0.066	0.106	0.213
	Hedonic regression	0.350	0.378	0.283	0.132	0.068	0.123	0.060	0.094	0.240
	Tax Values		0.918	0.439	0.056	0.029	0.086	0.031	0.085	0.273

Table 2: Predictive performance of our estimation methods, 1994-2015.

Notes: In-sample refers to the estimation sample (7/8th of the full sample), while out-of-sample refers to the holdout sample (1/8). RMSE is the root mean squared error = $\sqrt{\frac{1}{N}\sum_i \left(\log \frac{\hat{p}_i}{m_i^2} - \log \frac{p_i}{m_i^2}\right)^2}$. Out-of-sample predictive performance shows the distribution of transaction prices minus predicted prices in the holdout sample. *Ensemble ML* is our preferred machine learning approach, *hedonic regression* refers to estimation using equation (2), and *tax values* is using the tax values directly.

Year	Regular Housing	Cabins
1993	0.955	2.260
1994	1.037	2.405
1995	1.000	1.775
1996	1.086	2.000
1997	1.218	2.136
1998	1.277	2.261
1999	1.404	2.619
2000	1.451	2.664
2001	1.348	2.532
2002	1.430	2.698
2003	1.551	2.896
2004	1.704	3.199
2005	1.886	3.508
2006	1.681	2.992
2007	1.723	3.073
2008	1.590	2.872
2009	1.456	2.673
2010	1.179	3.490
2011	1.178	3.718
2012	1.161	3.406
2013	1.143	4.347
2014	1.137	4.668
2015	1.106	5.257

Table 3: Median yearly country-level adjustment ratios for tax values.

Notes: We construct these ratios by first computing the ratios of the predicted tax value divided by the value from the tax registry at the household level. The table reports the median ratio within housing types and years.

	Oslo & Bærum		**Bergen**		**Trondheim**		**Stavanger**	
Year	ML	SSB	ML	SSB	ML	SSB	ML	SSB
2006	12.98	15.27	19.50	17.52	10.79	11.38	22.65	25.35
2007	17.21	11.26	7.28	8.70	7.92	7.08	22.10	19.90
2008	0.57	-4.46	-4.72	-8.14	-0.00	-3.23	2.54	0.80
2009	1.35	3.27	2.21	0.78	1.17	1.58	1.79	1.05
2010	8.57	8.60	9.24	11.57	12.61	11.98	14.51	14.47
2011	9.96	9.86	11.91	9.82	9.07	10.97	9.44	9.57
2012	9.71	8.09	7.44	6.42	8.93	8.39	6.65	7.59
2013	3.83	3.16	5.83	5.44	6.70	5.89	3.28	2.42
2014	4.30	2.83	7.44	4.38	1.38	3.16	0.37	-1.89
2015	7.83	10.25	4.36	7.53	6.62	5.71	-1.12	-3.85

Table 4: Annual house price growth for Oslo & Bærum, Bergen, Trondheim, and Stavanger. The machine learning approach (ML, this paper) and the house price index from Statistics Norway (SSB).

C Figures

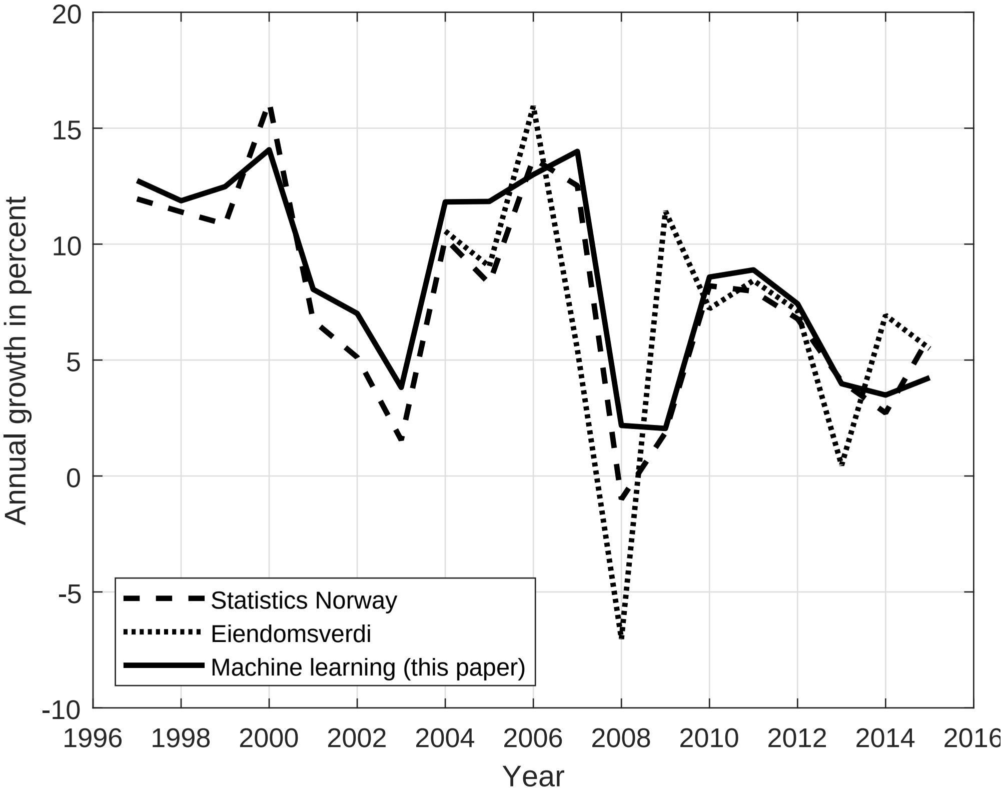

Figure 1: Annual House Price Growth, 1997 - 2015

Notes: The graph shows the annual nominal house price growth during a given year from three different sources: Statistics Norway, Eiendomsverdi, and this paper. Our data is fourth quarter so we compute annual house price growth as the four-quarter growth rate.

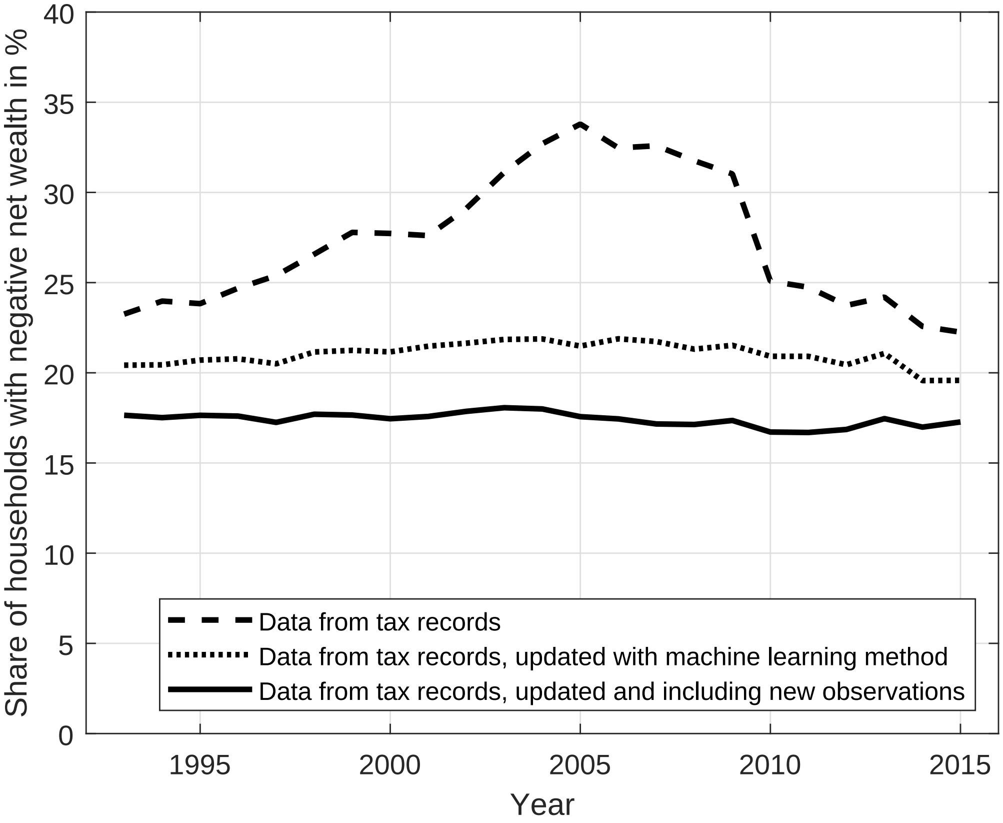

Figure 2: Share of individuals with negative net wealth, 1993 - 2015.

Notes: The figure shows the share of households with negative net wealth. The dashed line displays the share in the original data where we have only adjusted for the discounting by the tax administration. The dotted line shows the share where we take the observed housing wealth in the original tax data, but replace these observations with re-estimated housing wealth. The solid line displays the share where after we re-estimate housing wealth and include the housing unit observations of cooperative units, cabins, and non-id apartments.