Automated Valuation Services: A case study for Aberdeen in Scotland

Rainer Schulz and Martin Wersing[*]

December 7, 2020

[*] University of Aberdeen Business School, Edward Wright Building, Dunbar Street, Aberdeen AB24 3QY, United Kingdom; r.schulz@abdn.ac.uk and martin.wersing@abdn.ac.uk.

Abstract

Automated valuation services (AVSs) offered by listings platforms predict market values based on property characteristics supplied by users. We investigate the implementation of such a service for the City of Aberdeen. We fit different market value models with machine learning methods and assess them in a rolling windows procedure that mimics an AVS setting. We also investigate the ease and robustness with which the models can be implemented. We discuss how prediction uncertainty can be measured and reported to users. If implemented in the future, such a service has the potential to improve the transparency of the local housing market.

Keywords: housing market, machine learning

JEL Classification: C14, R31

1 Introduction

It is common for professionals in the residential real estate industry to use automated valuations. Banks and rating agencies use such valuations to reassess the collateral value of loan portfolios, assessors use them to estimate the taxable value of properties, and valuers use them as an input when they derive the market value of a particular property (RICS 2017).

For the general public, however, it was not common to use automated valuations. This has changed since listings platforms started to offer automated valuation services (AVSs).[1] The user, such as a house owner, submits information on the property's characteristics online and receives instantly a prediction of its market value. The service might also provide information about how certain the prediction is. Equipped with this information, the owner can then decide, perhaps after contacting an agent and a valuer, whether to list the property on the platform.

AVSs have the potential to improve the transparency of residential property markets. This would require, however, that the methods used are sound, the predictions accurate, and the information provided comprehensible for users. Platform providers are not very forthcoming regarding the models, methods, and data they are using. Providers are also reluctant to give information on the accuracy of the predictions.[2] Finally, it is also not clear how—if at all—prediction uncertainty is measured and translated into the information reported to users.

[1] Platforms such as ImmoScout24 and Immowelt-Immonet in Germany, Hometrack and Zoopla in the UK, and Redfin and Zillow in the US, offer such automated valuation services.

[2] Keeping details about the implementation private prevents that competitors can copy it. Matysiak (2017) finds that US providers are more open than European providers with respect to information on accuracy.

Academic case studies can help to clarify some of the opacity. First, case studies can assess which statistical models, if any, are accurate enough for an AVS.[3] Second, case studies can assess how prediction uncertainty should be measured and reported. Third, case studies can assess whether a particular statistical model is robust enough so that it can be implemented in an AVS. For instance, predictions must be provided promptly and lengthy computations or numerical instabilities would impede this. As the general public is neither an expert in real estate nor statistics, a model can only be based on property characteristics that users are able to provide. An AVS should also render only such statistical results that users are able to understand.

In this paper, we conduct a case study of an AVS for the housing market of the City of Aberdeen. Such a service might be implemented in the future and should be useful to all those who participate in the housing market of Aberdeen. We focus on the statistical modelling and discuss aspects relevant to the implementation of an AVS. Regarding the statistical modelling, we use machine learning methods that fit the models for out-of-sample use. We assess the performance of these models with a rolling windows approach that mimics the updating process of an AVS. We examine whether combining two or more predictions could lead to improved performance. We assess the statistical models with respect to ease and robustness of implementation. Finally, we examine how estimation uncertainty can be computed—examining the recent suggestions by Bellotti (2017) and Krause et al. (2020)—and discuss how estimation uncertainty can be reported most appropriately.

[3] As residential markets differ, it seems likely that models that work well for one market might not do so for another market. The *no free lunch theorem* from machine learning applies here, see Murphy (2012, pp. 24): no single statistical model will be best for each and every application.

There are many papers that fit different statistical models to house price data.[4] Relatively few papers have used methods from machine learning to prevent that a model overfits to the data in the training sample and then performs poorly out-of-sample. Kagie and Van Wezel (2007) use the boosting machine—a machine learning method—and compare its predictive performance with those of linear models estimated with ordinary least squares. The boosting machine performs better out-of-sample, even though Kagie and Van Wezel do not search for optimal tuning parameters. Antipov and Pokryshevskaya (2012) compare the performance of several models estimated with machine learning methods, but the results are limited, as they use listings and not transaction data. It is also not clear how they choose tuning parameters. Mullainathan and Spiess (2017) propose a useful classification of model and estimator pairings for machine learning methods and give an example that uses house prices (see in particular the online appendix of their paper). Schulz et al. (2014) use a rolling window procedure that mimics an AVS setting. They specify a flexible parametric model using cross-validation in training samples and use actual out-of-sample prediction errors to assess the estimated models. Bellotti (2017) and Mayer et al. (2019) use a similar rolling window procedure. Mayer et al. (2019) compare several flexible models estimated with machine learning methods, but do not discuss whether the different methods could be implemented easily and robustly.

The main findings of our case study are as follows. First, flexible models,

[4] Examples include: Anglin and Gençay (1996), Parmeter et al. (2007), and Haupt et al. (2010), which examine the same data set but use different models (semi-, non-, or fully parametric). Martins-Filho and Bin (2005) is another example of a semiparametric model. Bourassa et al. (2010) fit different spatial models for a given data set and compare predictions. McCluskey et al. (2013) fit linear, neural net, and geospatial models to a given data set and compare predictions.

such as penalised splines or boosting machines, outperform the linear and spatial regression models in terms of their predictive performance. For the most accurate models, about 52% (82%) of the predictions deviate by no more than 10% (20%) from the sale price. This is in line with the results in Kagie and Van Wezel (2007) and Mayer et al. (2019) and demonstrates that flexible models estimated with machine learning methods lead to accurate price predictions. We find also evidence that combining predictions from different models leads to even better predictions. Second, our preferred models generate only few very large prediction errors. This indicates that these models are robust and do not overfit to observations in training samples that have—undetected—aberrant prices. The fairly small number of characteristics that we observe will also play a role for this outcome.[5] Third, prediction intervals give the likely range for the sale price and are as such a useful measure to report prediction uncertainty in an AVS. However, standard methods to compute such intervals can be costly to implement and will perform poorly if the model is misspecified (Lei et al. 2018). Conformal prediction intervals are an attractive alternative that are valid even if the model is misspecified.

The rest of the paper is as follows. Section 2 describes the housing market of the City of Aberdeen and presents the transaction data that we use in our case study. Section 3 presents and motivates the statistical models that we consider and explains the estimators that we use. Section 4 presents the results of our analysis. Section 5 concludes. The web-appendix gives further technical details and results from additional robustness analyses.

[5] A small number of characteristics will also make it easy to check whether AVS users request plausible characteristic combinations. Naturally, a larger number of characteristics describe properties better and should result in better predictions.

2 Data

2.1 Market and raw data

Aberdeen is the third largest city in Scotland and had 228,800 residents in 2017, the year in the middle of our sample period from 2015Q3 to 2019Q4. The housing stock in 2017 consisted of 116,452 properties, mostly flats (55%), followed by non-detached (34%) and detached houses (11%). 89,635 properties were in the private housing sector, the majority in owner occupation (74%), the rest was rented out. 3,991 residential property transactions were registered for Aberdeen City in 2017, about 4% of the private stock.[6]

The raw transaction data comes from the *Aberdeen Solicitors Property Centre Ltd* (ASPC), which runs a web-based listing platform through which member firms advertise properties for sale.[7] The raw data contains the characteristics of the transacted properties, such as the number of bedrooms or the type of property. It also contains location coordinates and transaction prices obtained from member firms that prepare and witness the conclusion of sales contracts. In 2017, the raw data has 2,649 observations, which is less than the number of transactions registered officially for this year. However, as new properties are marketed directly by developers and not through the ASPC and as 1,176 new properties entered the market in 2017 (Aberdeenshire Council 2019), the ASPC raw data cover about 95% of all resale transactions.

Our AVS case study is therefore for an active mid-sized housing market where information on the majority of resale transactions becomes available in

[6] The remaining 26,817 properties were social housing and thus in the public sector, which is not relevant here. The above numbers are collated from Aberdeen City Council (2018), National Records of Scotland (2018), and Registers of Scotland (2018).

[7] The ASPC distributes also a printed register of the listings in the Aberdeen area.

a very timely manner.

2.2 Data cleaning

The raw data has 12,032 observations. We remove 59 observations that have missing or erroneous values for some characteristics.[8] We also remove observations of *unusual* properties by imposing value bounds, which corresponds roughly to trimming at the 1% level, see the web-appendix. It is unlikely that many market participants are interested in such properties and as their market value could be predicted only with substantial uncertainty, it seems reasonable not to consider them for an AVS.

After the data cleaning, the data set contains 11,908 observations. Table 1 presents summary statistics for the price and property characteristics.

[Table 1 about here.]

Three characteristics are continuous: floor area and location coordinates (statistics for coordinates are not reported). The remaining characteristics are either categorical, such as the energy certificate rating (EPC), or binary, such as central heating or garage. The characteristic room in Table 1 is the sum of the number of bed and living rooms, as it might be arbitrary in the individual case how rooms are classified and advertised.[9] The relative frequencies of sales by property types is similar to the ones reported for the housing stock in Section 2.1, which indicates a similar transaction propensity for the different property types.

[8] Erroneous values are obvious mistakes such as a property with zero rooms or location coordinates that are outside Aberdeen City.

[9] We conducted the analysis also with separate room categories. The results of this analysis are qualitatively identical to the ones reported here, see the web-appendix for details.

3 Methodology

3.1 Market value models

3.1.1 The prediction problem

The transaction price of a property can always be decomposed into

$$p = \mathrm{E}[p|\mathbf{x}] + \epsilon \tag{1}$$

where $\mathbf{x}$ is a vector of property characteristics, such as the number of rooms, and ϵ is transaction noise with $\mathrm{E}[\epsilon|\mathbf{x}] = 0$ and variance σ_ϵ^2. We call $\mathrm{E}[p|\mathbf{x}]$ the *market value function* and note that

$$\mathrm{E}[p|\mathbf{x}] = \min_{g \in \mathcal{F}} \mathrm{E}\left[(p - g(\mathbf{x}))^2\right] \tag{2}$$

This means that the market value function has—of all possible functions collected in the set $\mathcal{G}$—the minimum mean squared error for p given $\mathbf{x}$ (Goldberger 1991, Section 5.4).[10] If we knew the conditional distribution of p given $\mathbf{x}$, we could solve Eq. 2. However, we do not know this distribution and use instead candidate models $m(\mathbf{x})$ for the market value. Each candidate model depends on a vector $\boldsymbol{\theta}$ of coefficients, which we estimate with a training sample. The resulting estimator is denoted $\hat{m}(\mathbf{x})$ and is the better the smaller the *Mean Squared Prediction Error* (MSPE)

$$\begin{aligned} \mathrm{E}[(p - \hat{m}(\mathbf{x}))^2] &= (\mathrm{E}[p|\mathbf{x}] - \mathrm{E}[\hat{m}(\mathbf{x})])^2 + \mathrm{E}[(\hat{m}(\mathbf{x}) - \mathrm{E}[\hat{m}(\mathbf{x})])^2] + \sigma_\epsilon^2 \\ &= \mathrm{Bias}[\hat{m}(\mathbf{x})]^2 + \mathrm{Var}[\hat{m}(\mathbf{x})] + \sigma_\epsilon^2 \end{aligned} \tag{3}$$

where $(p, \mathbf{x})$ are observations that have not been used for the estimation of $m(\mathbf{x})$. Three remarks are in order with respect to Eq. 3. First, in the empirical

[10] The mean squared error on the right-hand side of Eq. 2 becomes smaller the more (price determining) characteristics $\mathbf{x}$ contains.

analysis, we estimate a version of the MSPE to discriminate between our candidate models $m(\mathbf{x})$. Second, the three terms of the MSPE decomposition on the right-hand side of Eq. 3 are all positive. The last term is *irreducible* variance σ_ϵ^2 of the noise and corresponds to the MSPE of the unknown market value function $\mathrm{E}[p|\mathbf{x}]$, see Eq. 1. As we have to choose and estimate $m(\mathbf{x})$, the MSPE of our candidate models will always be larger than σ_ϵ^2. Third, the remaining two terms point to a trade-off between the (squared) bias and the variance. A highly flexible $m(\mathbf{x})$ will have a small bias and, as many coefficients must be estimated, a large variance. Estimated with a training sample, $\hat{m}(\mathbf{x})$ will suffer from *overfitting* as it follows too closely the data in this sample. The opposite holds for a highly inflexible $m(\mathbf{x})$. In both cases, poor out-of-sample predictions will result. As an AVS provides, by definition, such predictions, the model with the right degree of flexibility must be found. We use established methods from machine learning to deal with this task.

Many different models $m(\mathbf{x})$ could be considered for an AVS. In our case study, we use models and estimators that lead eventually to an additive representation

$$m(\mathbf{x}) = \sum_{m=1}^{M} \theta_m b_m(\mathbf{x}, \boldsymbol{\gamma}_m) \tag{4}$$

with basis functions $b_m(\cdot)$ that transform property characteristics and might depend on further (tuning) parameters $\boldsymbol{\gamma}_m$. Once a model is estimated and implemented as an AVS, users provide vectors of characteristics $\mathbf{x}$ (which must fall within the bounds imposed on the data during estimation) and market values will be predicted instantly. We examine five different models regarding their predictive performance and ease of implementation. The first four models cover the classes given in Mullainathan and Spiess (2017, Table 2) and are: polynomial model, spline model, random forest, boosting machine. The polynomial model nests the classical linear model and we estimate it with three different estimators. The fifth is a geo-spatial model, a type of model that is frequently used in the real estate literature. The estimators for the different models require that we choose tuning parameters to regularise the complexity of $m(\mathbf{x})$, the web-appendix provides details.

3.1.2 Polynomial model

The model has the general additive structure

$$m(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{z}\boldsymbol{\theta}_0 + f_1(FA, \boldsymbol{\theta}_1) + f_2(LAT, LON, \boldsymbol{\theta}_2) \tag{5}$$

where the left-hand side makes now explicit that the function depends on the vector $\boldsymbol{\theta}$ of unknown coefficients. The vector $\mathbf{z}$ contains the constant term, indicators for categorical characteristics in $\mathbf{x}$, and quarterly time dummies; $\boldsymbol{\theta}_0$ is the vector of coefficients for these variables. The function f_1 produces a polynomial of degree d_1 in the floor area and f_2 produces a joint polynomial of degree d_2 in the latitude and longitude. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the vectors of coefficients for the terms of the two polynomials. d_1 and d_2 are both tuning parameters.[11]

We estimate the coefficient vector by minimising the penalised sum of least squares

$$S(\boldsymbol{\theta}) = \sum_{n=1}^{N} (p_n - m(\mathbf{x}_n, \boldsymbol{\theta}))^2 + \lambda \mathbf{v}_1' \mathbf{D} \mathbf{v}_2 \tag{6}$$

For the polynomial model, $\mathbf{D}$ is the identity matrix with its first entry set to zero, so that the coefficient for the constant term is never penalised. If we set $\lambda = 0$, the Ols estimator for $\boldsymbol{\theta}$ results; if we set $\mathbf{v}_1 = \mathbf{v}_2 = \boldsymbol{\theta}$, the Ridge estimator results; if we set $\mathbf{v}_1 = \mathbf{i}$ (a vector of ones) and $\mathbf{v}_2 = |\boldsymbol{\theta}|$, the Lasso estimator results. For Ridge and Lasso, λ is a further tuning parameter and the penalty term in Eq. 6 punishes variability of coefficient estimates. The Lasso

[11] For instance, with a cubic polynomial basis, $d = 3$, we obtain $f(x, \boldsymbol{\theta}) = [x, x^2, x^3]\boldsymbol{\theta}$.

estimator can go as far as shrinking estimates to zero, thereby deselecting characteristics from the model. The *regularisation* achieved by penalised least squares will increase the bias of $\hat{m}(\mathbf{x})$, but should also reduce its variance. We fit the polynomial model of Eq. 5 with three different estimators: Ols with best subset selection, Ridge, and Lasso (Hastie et al. 2009, Chap. 3.4). The Lasso seems to have a comparative advantage, as best subset selection is computational expensive and Ridge regression cannot select characteristics.

3.1.3 Penalised splines model

The model has the same structure as Eq. 5, but the functions f_1 and f_2 are cubic and thin plate splines, respectively, with coefficients collected in the vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The coefficients are estimated by minimising the penalised regression sum of squares in Eq. 6. For the spline model, we set $\mathbf{v}_1 = \mathbf{v}_2 = \boldsymbol{\theta}$ and $\mathbf{D}$ is a zero matrix, except for the diagonal elements that correspond to coefficients in $\boldsymbol{\theta}$ that multiply with a truncated term.[12] These diagonal elements are all equal to one. It follows that $\text{tr}(\mathbf{D}) = K_1 + K_2$, which is the total number of knots used in the splines. The penalty term in Eq. 6 becomes large if f_1 and f_2 are very wiggly and small if the functions are fairly smooth. K_1, K_2, and λ are tuning parameters.

[12] The cubic splines basis is $f(x, \boldsymbol{\theta}) = [x, x^2, x^3, |x - \kappa_1|^3, \ldots, |x - \kappa_K|^3]\boldsymbol{\theta}$. It extends a polynomial with truncated terms $|x - \kappa_k|$, which join at knots, $\kappa_K > \ldots > \kappa_1$. The larger the number of knots K, the more flexible will $f(x, \boldsymbol{\theta})$ be. $\boldsymbol{\theta}$ is constrained further to ensure that the function is linear beyond the boundary knots. The thin plate splines basis is a two-dimensional extension of the cubic splines basis (Wood 2017, Ch. 5).

3.1.4 Random forest

Random forests are based on regression trees (Hastie et al. 2009, Chap. 15). The leaves of trees are sets $\mathcal{S}_s$ that divide the characteristic space into non-overlapping regions. Properties that fall into the same set have the same market value. This implies

$$m(\mathbf{x}) = \sum_{s=1}^{S} \mathbf{1}(\mathbf{x} \in \mathcal{S}_s)\theta_s \tag{7}$$

where $\mathbf{1}(\cdot)$ is an indicator function that becomes ones if the characteristics $\mathbf{x}$ of a property fall into the set $\mathcal{S}_s$ and zero otherwise. θ_s is the market value for such properties. Both the sets and the corresponding market values have to be estimated with the training data. The tree estimator starts with all possible pairwise splits of the values of each characteristic in the data set.[13] It looks then for the characteristic and split that explains—of all such combinations—the price variation best. This gives the first two branches of the tree. The two branches should be split into sub-branches if this explains the price variation even better. The procedure stops once the sets belonging to the branches contain only few observations. In this instance, the leaf $\mathcal{S}_s$ has been determined. For each leaf, $\hat{\theta}_s$ is the average price of those observations that fall into $\mathcal{S}_s$.

The tree estimator has the tendency to overfit, however. Random forests improve on this by averaging a large number of individual regression trees, each tree based on a random sample from the training data set.[14] In addition, random forests randomly reduce the variables that are available for splitting each node of the tree. This ensures that the individual trees of the random

[13] If properties have up to three bathrooms, split sets are $\{1,(2,3)\}$, $\{2,(1,3)\}$ $\{3,(1,2)\}$.

[14] The random forest is also robust against outliers, as *unusual* observations will appear in only few re-sampled training samples. We deal with unusual observations by setting bounds, but additional robustness can be useful.

forest differ from each other. The fraction of variables used to determine each branch of a tree and the minimum size of the final branches are tuning parameters.

3.1.5 Boosting machine

The boosting machine builds on trees, often only on those with two branches (so-called stumps). The boosting machine is a sequential procedure in which a new tree is fitted to residuals that remain unexplained by the previous step of the sequence (Hastie et al. 2009, Chap. 10.9). The sequence starts with the average price for all observations in the training sample. The residuals from this function become the target to which trees are fitted. The tree that explains the residuals best becomes part of the estimated market value function. To prevent overfitting, the tree is not considered in full, but only in proportion λ, the so-called *learning rate*. These steps of the sequential procedure are repeated S-times. λ, S and the number of branches in each tree are tuning parameters.

3.1.6 Spatial autoregressive model

The spatial autoregressive (SAR) model is

$$m(\mathbf{x}, \boldsymbol{\theta}, \mathbf{p}) = \mathbf{z}\boldsymbol{\theta}_0 + f(FA, \boldsymbol{\theta}_1) + \theta_2 \mathbf{w}\mathbf{p} \tag{8}$$

where the function f is a polynomial of the floor area with degree d. The row vector $\mathbf{w}$ contains mostly zeros, but has entry $1/k$ for observations that are the k-nearest neighbours of the subject property. The vector $\mathbf{p}$ collects the prices in the training sample. The SAR model is similar to Eq. 5, but replaces the function f_2 with a multiple of the average price of nearby properties. Whereas the polynomial and the spline models use location coordinates to model the spatial component of market values, the SAR model uses *prices* of recently transacted nearby properties. We estimate Eq. 8 with best subset selection and the Maximum likelihood estimator. d and k are tuning parameters.

3.2 Assessment of market value models

We estimate and evaluate the performance of the different market value models with a rolling window procedure that mimics how an actual AVS would work. Our first training sample covers the period 2015Q3-2016Q2. We fit each model to the full training sample and to property type sub-samples. The split ensures that each sub-sample contains similar properties, but comes at the cost that each sub-sample has—compared to the respective full sample—a smaller number of observations that can be used for training. The first test sample is for the quarter 2016Q3. We then roll the training sample forward by one quarter and refit the models to estimate market values for subject properties in the test sample of the next quarter.[15] The procedure ends when the last test sample, 2019Q4, is reached. We implement the procedure separately with the price and the log price as dependent variable. Using log prices can help to alleviate efficiency losses due to heteroscedasticity during estimation. As users of an AVS are only interested in predictions of market values, not in predictions of log market values, we transform the log predictions back to the natural scale with the smearing estimator of Duan (1983). The web-appendix

[15] We also examined the predictive performance when the windows are rolled forward by one month. The results are comparable to those reported here, see the web-appendix. We did not examine the effect the length of the training sample. The samples covering four quarters (twelve months) produce good in-sample fits for all five models and we believe that improvements, if any at all, would be only marginal.

provides details.

We measure the performance of the candidate models based on the relative prediction errors

$$e_n = \frac{p_n - \hat{m}(\mathbf{x}_n)}{\hat{m}(\mathbf{x}_n)} \tag{9}$$

where n indicates an observation in the test sample and $\hat{m}(\cdot)$ is estimated with the observations in the training sample. Negative (positive) errors imply that the prediction of the market value is larger (smaller) than the actual transaction price. An error of zero implies that the estimated market value predicts the price perfectly. While this can happen occasionally, the irreducible noise prevents that it will be the norm. Table 2 gives the performance measures that we use to compare the different market value models.

[Table 2 about here.]

It is possible that the different measures lead to different rankings of the different models. Any assessment thus requires a—to some degree subjective—judgement about which aspects of the relative error distribution are most relevant for the application at hand. In our case study, we focus on the relative error rate RER(b) and the mean squared relative error (MSRE). Both take bias *and* dispersion of the relative errors into account. The RER(b) gives the fraction of relative errors e_n that fall within the interval $[-b, b]$. Setting b to 10% (20%) implies that RER(b) gives the fraction of observations in test samples where the predicted market value does not deviate by more than ±10% (±20%) from the transaction price. Such information is frequently used by practitioners when assessing valuation accuracy. Moreover, as RER(b) is a function of b, it allows for visual comparison of error distributions. The MSRE averages the squared relative errors, and places hence more (less) weight on absolutely large (small) errors. In addition to the RER(b) and the MSRE, we use several other performance measure to scrutinise specific aspects of the predictions from the different market value models, such as their unbiasedness.

4 Empirical results

4.1 Performance of market value models

The four implementations of the rolling window procedure are: models fit to log prices using full samples (LQ1) or property type sub-samples (LQ2) for training; models fit to prices in natural scale using full samples (NQ1) or property type sub-samples (NQ2) for training. The results for the implementations are given in Figures 1 and 2, which plot RER(b), and in Tables 3 to 6, which give detailed information on the performance measures. The tables report also the performance of a simple benchmark, which predicts the market value of properties in a test sample with the average price of all properties that have been transacted in the last quarter of the corresponding training sample. We call these benchmark predictions *unconditional*, as they ignore any further information that is available on individual properties.

[Figure 1 about here.]

[Figure 2 about here.]

Figures 1 and 2 show that the RER(b) for the penalised splines and the boosting machine model always lie—irrespective of the implementation—to the left of the RER(b) for the other models. For both models, the fraction of predictions that fall within ±10% of the price is, averaged over the three property types, always about three times larger than the fraction for the unconditional model. For the two next best models, random forest and SAR, the fraction is always about 2.8 times larger. The polynomial model comes last, irrespectively of the estimator used, but performs still much better than the benchmark. This shows clearly that property characteristics explain a substantial part of the cross-sectional variation of property prices.

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

The penalised spline model performs best under implementation LQ2, where 52.0% (82.0%) of all predictions lie within ±10% (±20%) of the prices of the properties in the test samples. This is slightly better than the performance of the boosting machine, which performs best under implementation LQ1 and produces predictions that are in 51.3% (81.1%) of the cases within ±10% (±20%) of the transaction price.[16]

[Table 7 about here.]

The penalised splines model is also the best performer over all models and implementations when the MSRE is the performance measure, as Table 7 shows.

[16] The fractions can be obtained by computing the observation-weighted averages of RER(0.1) and RER(0.2) from Table 3 (boosting machine) and Table 4 (penalised spline).

However, this time the penalised splines model performs best if implemented under LQ1. The performance of the boosting machine is again close, in particular if implemented under NQ1. As both models show similar performance, it is not obvious which one to choose for an AVS. If sufficient resources are available, such a choice might not be necessary, as predictions could be combined. We assessed this for an equally-weighted average of the predictions from the penalised splines and the boosting machine model. The RER(0.1) increases by 4.2% and the MSRE decreases by 6.7% relative to the performance of the penalised splines model on its own, see the web-appendix for details. The combination of predictions has not received much attention in the AVS literature.

There are several other interesting aspects. First, all models generate predictions that are biased downwards by about 1.5% on average, see the entries for the MRE in Tables 3 to 6. The bias comes from the fact that we predict the market value of a property *as if* it was sold at the end of the training rather than during the test sample.[17] Second, as discussed above, there might be a trade-off between using full or property type sub-samples for training. Sub-sample observations are similar, which should make model training easier, but the number of observations is small, which should make it more difficult. Table 7 gathers the MSRE for the different models and implementations. Prediction errors of the penalised splines model for detached houses, the smallest segment in the Aberdeen housing market, are less dispersed when the models are trained with full samples (MSRE of 2.2% for LQ1) than when trained with

[17] Over the period of our case study, the quality-controlled house price index for Aberdeen declined, on average, by 1.5% per quarter. A statistical model could be used to forecast changes of the price trend and these could then be used to adjust market value predictions. However, this task is separate from finding the best market value model. We do not approach it here.

sub-samples (MSRE of 2.9% for LQ2). In this case, the sample size effect dominates. For non-detached houses, the effect is marginal and goes in the opposite direction: the MSRE is 1.8% for LQ1 compared to 1.7% for LQ2. For flats, there is no difference. The boosting machine performs always better when trained with full samples, which should be the result of its adaptive estimation. Third, Tables 4 and 6 show that errors of predictions conditioned only on the property type are—measured with REV, MSRE, and MARE—the least dispersed for detached houses, which seems counter-intuitive. Once all observed characteristics are considered, the dispersion of errors for detached houses is—as expected—above those of non-detached houses. The dispersion of errors for flats is now the highest, something we would not expect, as flats are less heterogeneous than houses. However, we do not observe the location of a flat within a building —such as basement or top floor—, which is certainly a characteristic that is highly relevant for its market value.

4.2 Implementation and quality assurance

The setting up of an AVS requires that the underlying statistical model can be implemented robustly *and* maintained easily in order to provide market value estimates at low cost on a continuous basis. The implementation involves the stages of data preparation, model training and testing, fitting of the final model $\hat{m}(\mathbf{x})$, and—eventually—its integration in a web service.[18] Less robust models $\hat{m}(\mathbf{x})$ require additional attention every time the models are fitted and may produce unreliable estimates. In our experience, the maximum likelihood estimator of the SAR model can suffer from convergence issues, which require

[18] Since we fit the models to location coordinates, a web service also needs to integrate digital maps to match a street address—as provided by the user—to coordinates.

additional attention during the training stage. Models that overfit to the data in training samples can produce unreliable market value estimates. This can be seen from Table 4, where the MSRE of the polynomial (fitted with Ols) and the SAR models are several orders larger than the MSRE of the other models.[19] The other models (or other estimators for the polynomial model) balance the bias-variance trade-off through regularisation and cope much better with the problem of overfitting.

An AVS in operation needs to be backtested to monitor performance *and* identify potential improvements. Plots of cross-validated relative errors (CVRE) and realised MSRE are useful tools to monitor the quality of an AVS over time. The CVRE indicates how a newly trained model is expected to perform on future market value requests. The MSRE summarises the performance given the transacted properties in a given period. Figure 3 gives an example and shows the quarterly CVRE and MSRE of the penalised splines and boosting machine models.

[Figure 3 about here.]

The CVRE and MSRE are closely related to each other, indicating that the models perform as expected. However, deviations between them can point to problems that need to be investigated further. For instance, the unusually large MSRE for detached houses in 2019Q3 is due to a single property which receives a market value estimate substantially above its actual transaction price. This divergence cannot be explained with the characteristics of the property, so transaction noise seems the cause of it. Figure 3 shows that the overall performance of both models gets worse for flats over the period of

[19] The large MSRE is caused by a single property in the test sample that has a floor area just outside the range observed in the training sample.

our case study. During the past decade, many new flats were constructed in Aberdeen and these flats have started to enter the resale market. We do not observe a property's age and cannot control for possible vintage effects this brings. Spatial plots of the errors and tests of spatial correlation are other useful tools for backtesting. In our case, neither reveal any structure in the errors that remains unexplained (not reported).

4.3 Uncertainty of market value estimates

Users of an AVS are not only interested in the market value estimate itself, but also the uncertainty associated with this estimate. Automated valuations targeted at professional users often report interval estimates which provide a price range that is likely to cover the sale price that the property will fetch if it is sold (Krause et al. 2020). Such *prediction intervals* provide a sound conceptual framework for assessing uncertainty. Prediction intervals can be computed in different ways which may lead to significant differences in the intervals' reliability (realised coverage) and efficiency (interval length).

To demonstrate how uncertainty estimates can be computed, we examine the performance of two main approaches—standard and conformal prediction intervals—for market value estimates from the penalised splines model. Standard prediction intervals rely on asymptotic theory and may not be reliable in finite samples. Conformal prediction intervals, in contrast, require less demanding assumptions and are reliable even if the model is misspecified (Lei et al. 2018).[20] Table 8 reports the realised coverage of the standard and con-

[20] The bootstrap is a third approach for the construction of prediction intervals, which often produces better finite sample results than standard method. We refer to Krause et al. (2020) for a comparison of the bootstrap and standard approaches. For an application of

formal prediction intervals for our data. The prediction intervals are computed at the 80% and 90% confidence levels. Details are given in the web-appendix.

[Table 8 about here.]

The standard prediction intervals over-cover the nominal confidence levels. This is likely due to deviations from the normal approximation underlying their construction, which results in much wider intervals than those from the conformal method. The conformal prediction intervals have reasonable coverage rates that are closer to their nominal levels. It must be noted, however, that the length of the conformal prediction intervals depend only on the data in the training sample and not the particular properties for which a market values should be estimated. Thus, unlike standard prediction intervals, the conformal intervals will be too wide for some and too short for other properties.

Though appealing from a statistical point of view, prediction intervals are often misunderstood. A common misperception is that a $(1-\alpha)$ confidence level implies that there is $(1-\alpha)\%$ chance that the sale price will fall within the reported prediction interval. A prediction interval, however, is an *estimate* of the possible price range which either contains the price or not. Hence, a correct interpretation of the confidence level is based on the notion of calculating prediction intervals from many samples. While the interval bounds will vary from sample to sample, $(1-\alpha)\%$ of these hypothetical prediction intervals will include the sale price. To avoid such confusion, an AVS targeted at the general public could report an uncertainty score that, for instance, categorizes uncertainty according to the implied confidence level of prediction intervals that are within $\pm 10\%, \pm 20\%, \ldots$ of the market value estimates. An alternative

conformal prediction intervals see Bellotti (2017).

to such a qualitative measure are Bayesian credible intervals, which would allow probabilistic statements about the sale price itself to be made.

5 Conclusion

In this paper, we conducted a case study for an AVS for the housing market of the City of Aberdeen. We considered five statistical models and used machine learning methods to fit the models for the explicit purpose of predicting market values out-of-sample. While all of the models allow for a flexible relationship between property characteristics and price, they differ by how this flexibility is regularised in order to avoid that the models fit too closely to data in training samples. The case study produces three important insights.

First, the penalised splines and the boosting machine outperform the polynomial and the spatial autoregressive models. It does not matter much which estimator is used for the estimation of the polynomial model. The random forest shows a good performance overall. The boosting machine has been applied to house price data by Kagie and Van Wezel (2007) and Mayer et al. (2019). In both cases, however, penalised splines were not part of the set of candidate models. In our preferred implementation, the penalised spline model is effectively a decomposition of the log price into shift terms for categorical characteristics, a undetermined function for the floor space and an undetermined function for the location value. While additive models have been used to model house prices before (see Fn. 4), they have not been considered for applications that focus on out-of-sample performance. Given that the penalised splines model performs slightly better than the boosting machine and given its more intuitive form, we find it a particular attractive model. Perhaps a combination of the market value estimates from both models would be even better. The question of whether and how to combine estimates from different candidate models is a question for future research.

Second, an AVS should not only consider the out-of-sample performance of market value estimates, but must also ensure that the models can be fitted easily and that estimates can be calculated quickly. We experienced that some models are able to deal with this better than others. The SAR model, for instance, suffered at times from slow convergence, which could pose a serious problem in an AVS.

Third, an AVS must provide some indication to users how certain the market value estimates are. This requires not only the correct measure to assess such uncertainty, but requires that the information is reported in a clear manner. We have shown that prediction intervals based on conformal methods produce reliable uncertainty measures for market value estimates. Bayesian credible intervals have a more intuitive interpretation than the prediction intervals discussed here. Their implementation, however, requires to state prior assumption on the sale price distribution. This is an area for future research.

Acknowledgements

We are grateful to the three anonymous referees and the editor of the iournal for suggestions and comments that helped to improve the paper. We thank seminar participants at the Technische Universität Berlin, Fiona Stoddard, and Verity Watson for helpful comments. The usual disclaimer applies.

References

Aberdeen City Council: 2018, Local housing strategy 2018-2023.

Aberdeenshire Council: 2019, Housebuilding in Aberdeen City and Aberdeenshire.

Anglin, P. M. and Gençay, R.: 1996, Semiparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **11**, 633–648.

Antipov, E. A. and Pokryshevskaya, E. B.: 2012, Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics, *Expert Systems with Applications* **39**, 1772–1778.

Bellotti, A.: 2017, Reliable region predictions for automated valuation models, *Annals of Mathematics and Artificial Intelligence* **81**, 71–84.

Bourassa, S., Cantoni, E. and Hoesli, M.: 2010, Predicting house prices with spatial dependence: A comparison of alternative methods, *Journal of Real Estate Research* **32**, 139–160.

Duan, N.: 1983, Smearing estimate: A nonparametric retransformation method, *Journal of the American Statistical Association* **78**, 605–610.

Goldberger, A. S.: 1991, *A Course in Econometrics*, Harvard University Press, Cambridge, MA.

Hastie, T., Tibshirani, R. and Friedman, J.: 2009, *The Elements of Statistical Learning*, Springer Series in Statistics, second edn, Springer Verlag, New York.

Haupt, H., Schurbus, J. and Tschernig, R.: 2010, On nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **25**, 894–901.

Kagie, M. and Van Wezel, M.: 2007, Hedonic price models and indices based on boosting applied to the Dutch housing market, *Intelligent Systems in Accounting, Finance and Management* **15**, 85–106.

Krause, A., Martin, A. and Fix, M.: 2020, Uncertainty in automated valuation models: Error- vs model-based approaches. *Journal of Property Research*, forthcoming.

Lei, J., Sell, M. G., Rinaldo, A., Tibshirani, R. J. and Wasserman, L.: 2018, Distribution-free predictive inference for regression, *Journal of the American Statistical Association* **113**, 1094–1111.

Martins-Filho, C. and Bin, O.: 2005, Estimation of hedonic price functions via additive nonparametric regression, *Empirical Economics* **30**, 93–114.

Matysiak, G. A.: 2017, The accuracy of automated valuation models (AVMs), *Report*, TEGoVA.

Mayer, M., Bourassa, S. C., Hoesli, M. and Scognamiglio, D.: 2019, Estimation and updating methods for hedonic valuation, *Journal of European Real Estate Research* **12**, 134–150.

McCluskey, W. J., McCord, M., Davis, P., Haran, M. and Mcllhatton, D.: 2013, Prediction accuracy in mass appraisal: A comparison of modern approaches, *Journal of Property Research* **30**, 239–265.

Mullainathan, S. M. and Spiess, J.: 2017, Machine learning: An applied econometric approach, *Journal of Economic Perspectives* **31**, 87–106.

Murphy, K. P.: 2012, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge MA, London.

National Records of Scotland: 2018, Estimates of households and dwellings in Scotland, 2017, Edinburgh.

Parmeter, C. F., Henderson, D. J. and Kumbhakar, S. C.: 2007, Nonparameteric estimation of a hedonic price function, *Journal of Applied Econometrics* **22**, 695–699.

Registers of Scotland: 2018, Calendar year market review 2017: A statistical review of the Scottish residential property market, Edinburgh.

RICS: 2017, The future of valuations, *Insight paper*, Royal Institution of Chartered Surveyors, London.

Schulz, R., Wersing, M. and Werwatz, A.: 2014, Automated valuation modelling: A specification exercise, *Journal of Property Research* **31**, 131–153.

Wood, S. N.: 2017, *Generalized Additive Models. An Introduction with R*, Texts in Statistical Science, second edn, CRC Press, Boca Raton.

Table 1: Summary statistics for the transacted properties. Reports summary statistics for properties transacted during 2016Q3-2019Q4. Price is in GBP '000. Floor area is in sqm. Rooms is the sum of the number of bed and living rooms. Number of observations is 11,908.

	Mean	Std. Dev.	Min.	Max
Price	192.15	110.27	25.00	155.00
Floor area	83.73	41.12	19.00	458.00
Rooms	3.71	1.47	1.00	10.00
Bathrooms	1.25	0.49	1.00	4.00
EPC rating	3.74	0.87	2.00	7.00
Property type				
Detached	0.11			
Non-detached	0.39			
Flats	0.50			
Has				
Central heating	0.93			
Garden	0.73			
Garage	0.23			
Double garage	0.04			
Other parking	0.48			

Table 2: Performance measures. Gives description of and formulas for the performance measures based on the relative prediction errors. N_T is the number of observations in the test samples that provide the relative prediction errors e_n. $\mathbf{1}(\cdot)$ is the indicator function that takes the value one if the argument is correct and value zero otherwise.

Performance measure	Description	Formula	
Mean relative error	Arithmetic average over all errors	MRE	$= N_T^{-1} \sum_{n=1}^{N_T} e_n$
Median relative error	Middle of the error distribution	MDRE	$= \text{med}\,(e_n)$
Relative error variance	Variation of errors around their mean	REV	$= N_T^{-1} \sum_{n=1}^{N_T} (e_n - \text{MRE})^2$
Mean squared relative error	Arithmetic average over squared errors	MSRE	$= N_T^{-1} \sum_{n=1}^{N_T} (e_n)^2$
Mean absolute relative error	Arithmetic average over absolute errors	MARE	$= N_T^{-1} \sum_{n=1}^{N_T} \lvert e_n \rvert$
Relative error rate	Fraction of errors that fall within $-b \leqslant e_n \leqslant b$	RER(b)	$= N_T^{-1} \sum_{n=1}^{N_T} \mathbf{1}(\lvert e_n \rvert \leqslant b)$

Table 3: Performance measures 2016Q3-2019Q4, implementation LQ1. Shows performance measures for market value models fitted to log prices with quarterly rolling window, trained with full samples. Log market value predictions are re-transformed to natural scale using the smearing estimator of Duan (1983).

	MRE	MDRE	REV	MSRE	MARE	RER(0.1)	RER(0.2)
Detached ($N_T = 1,019$)							
Unconditional	0.8740	0.7150	0.5536	1.3174	0.8866	0.0451	0.0952
Polynomial (Ols)	-0.0107	-0.0197	0.0366	0.0367	0.1416	0.4651	0.7630
Polynomial (Ridge)	0.0190	0.0022	0.0923	0.0927	0.1559	0.4287	0.7306
Polynomial (Lasso)	-0.0076	-0.0147	0.0347	0.0348	0.1440	0.4228	0.7581
Penalised splines	-0.0008	-0.0079	0.0224	0.0224	0.1100	0.5792	0.8545
Random forest	0.0061	-0.0068	0.0262	0.0263	0.1166	0.5536	0.8378
Boosting machine	0.0090	0.0028	0.0246	0.0246	0.1137	0.5674	0.8446
Spatial autoregressive	0.0039	-0.0086	0.0267	0.0268	0.1271	0.4779	0.8073
Non-detached ($N_T = 3,632$)							
Unconditional	0.1057	-0.0601	0.2604	0.2716	0.3406	0.1892	0.4193
Polynomial (Ols)	-0.0050	-0.0167	0.0339	0.0340	0.1429	0.4309	0.7470
Polynomial (Ridge)	-0.0114	-0.0282	0.0353	0.0354	0.1476	0.4072	0.7299
Polynomial (Lasso)	-0.0060	-0.0208	0.0364	0.0364	0.1508	0.3937	0.7197
Penalised splines	-0.0036	-0.0069	0.0177	0.0177	0.1022	0.5837	0.8739
Random forest	-0.0219	-0.0260	0.0223	0.0228	0.1156	0.5361	0.8348
Boosting machine	-0.0056	-0.0073	0.0179	0.0180	0.1032	0.5757	0.8780
Spatial autoregressive	0.0018	-0.0103	0.0235	0.0235	0.1144	0.5507	0.8411
Flats ($N_T = 4,344$)							
Unconditional	-0.2910	-0.3552	0.1015	0.1862	0.38	0.0967	0.2247
Polynomial (Ols)	-0.0230	-0.0164	0.0554	0.0559	0.1865	0.3363	0.6137
Polynomial (Ridge)	-0.0271	-0.0261	0.0651	0.0658	0.2044	0.3006	0.5654
Polynomial (Lasso)	-0.0229	-0.0174	0.0602	0.0607	0.1960	0.3145	0.5847
Penalised splines	-0.0251	-0.0205	0.0349	0.0355	0.1453	0.4275	0.7426
Random forest	-0.0619	-0.0640	0.0388	0.0426	0.1607	0.3971	0.6878
Boosting machine	-0.0130	-0.0177	0.0368	0.0370	0.1450	0.4471	0.7463
Spatial autoregressive	-0.0365	-0.0382	0.0429	0.0442	0.1647	0.3831	0.6683

Table 4: Performance measures 2016Q3-2019Q4, implementation LQ2. Shows performance measures for market value models fitted to log prices with quarterly rolling window, trained with property type sub-samples. Log market value predictions are re-transformed to natural scale using the smearing estimator of Duan (1983).

	MRE	MDRE	REV	MSRE	MARE	RER(0.1)	RER(0.2)
Detached ($N_T = 1,019$)							
Unconditional	0.0061	-0.0890	0.1607	0.1606	0.2830	0.2237	0.4308
Polynomial (Ols)	-0.0056	-0.0139	0.0341	0.0341	0.1284	0.5172	0.8210
Polynomial (Ridge)	-0.0071	-0.0175	0.0312	0.0312	0.1327	0.4897	0.7856
Polynomial (Lasso)	-0.0078	-0.0130	0.0318	0.0319	0.1327	0.4769	0.8004
Penalised splines	0.0004	-0.0053	0.0288	0.0288	0.1180	0.5634	0.8535
Random forest	-0.0252	-0.0381	0.0263	0.0269	0.1207	0.5270	0.8220
Boosting machine	-0.0074	-0.0135	0.0277	0.0277	0.1182	0.5506	0.8220
Spatial autoregressive	0.6244	-0.0084	4e+02	4e+02	0.7526	0.5556	0.8348
Non-detached ($N_T = 3,632$)							
Unconditional	-0.0022	-0.1545	0.2116	0.2116	0.3272	0.1589	0.3455
Polynomial (Ols)	-0.0099	-0.0257	0.0327	0.0328	0.1361	0.4617	0.7715
Polynomial (Ridge)	-0.0105	-0.0299	0.0343	0.0344	0.1454	0.4292	0.7354
Polynomial (Lasso)	-0.0087	-0.0294	0.0340	0.0341	0.1433	0.4375	0.7472
Penalised splines	-0.0097	-0.0098	0.0164	0.0165	0.0978	0.6093	0.8910
Random forest	-0.0351	-0.0377	0.0205	0.0217	0.1126	0.5430	0.8411
Boosting machine	-0.0087	-0.0091	0.0184	0.0184	0.1033	0.5859	0.8736
Spatial autoregressive	-0.0116	-0.0129	0.0190	0.0191	0.1027	0.5931	0.8794
Flats ($N_T = 4,344$)							
Unconditional	-0.0101	-0.1076	0.1990	0.1991	0.3372	0.1724	0.3734
Polynomial (Ols)	2e+08	-0.0186	2e+22	2e+22	2e+08	0.3464	0.6225
Polynomial (Ridge)	-0.0085	-0.0244	1.1632	1.1633	0.2123	0.3187	0.5861
Polynomial (Lasso)	-0.0210	-0.0206	0.0579	0.0583	0.1914	0.3319	0.6018
Penalised splines	-0.0174	-0.0171	0.0353	0.0356	0.1450	0.4348	0.7444
Random forest	-0.0574	-0.0582	0.0387	0.0420	0.1588	0.4000	0.6988
Boosting machine	-0.0112	-0.0154	0.0389	0.0390	0.1474	0.4433	0.7386
Spatial autoregressive	3e+12	-0.0255	5e+25	5e+25	3e+12	0.3957	0.6937

Table 5: Performance measures 2016Q3-2019Q4, implementation NQ1. Shows performance measures for market value models fitted to prices with quarterly rolling window, trained with full samples.

	MRE	MDRE	REV	MSRE	MARE	RER(0.1)	RER(0.2)
Detached ($N_T = 1,019$)							
Unconditional	0.8625	0.7059	0.5465	1.2904	0.8755	0.0481	0.1011
Polynomial (Ols)	-0.0074	-0.0081	0.0299	0.0300	0.1308	0.4907	0.7886
Polynomial (Ridge)	0.0116	0.0020	0.0334	0.0335	0.1401	0.4356	0.7660
Polynomial (Lasso)	0.0038	0.0013	0.0324	0.0324	0.1396	0.4346	0.7679
Penalised splines	-0.0101	-0.0133	0.0220	0.0221	0.1103	0.5860	0.8476
Random forest	-0.0058	-0.0104	0.0241	0.0241	0.1139	0.5595	0.8397
Boosting machine	0.0000	-0.0021	0.0246	0.0246	0.1145	0.5556	0.8279
Spatial autoregressive	0.0033	-0.0087	0.0324	0.0324	0.1221	0.5113	0.8348
Non-detached ($N_T = 3,632$)							
Unconditional	0.0997	-0.0668	0.2576	0.2675	0.3393	0.1894	0.4091
Polynomial (Ols)	0.0005	-0.0104	0.0372	0.0372	0.1497	0.4017	0.7260
Polynomial (Ridge)	-0.0109	-0.0276	0.0361	0.0362	0.1516	0.3822	0.7161
Polynomial (Lasso)	-0.0044	-0.0208	0.0382	0.0383	0.1556	0.3811	0.7032
Penalised splines	-0.0034	-0.0058	0.0187	0.0187	0.1056	0.5661	0.8684
Random forest	-0.0255	-0.0278	0.0218	0.0224	0.1150	0.5300	0.8373
Boosting machine	-0.0083	-0.0094	0.0181	0.0182	0.1038	0.5735	0.8733
Spatial autoregressive	0.0012	-0.0062	0.0223	0.0223	0.1125	0.5471	0.8494
Flats ($N_T = 4,344$)							
Unconditional	-0.2951	-0.3578	0.1002	0.1873	0.3818	0.0965	0.2201
Polynomial (Ols)	-0.0202	-0.0185	0.0607	0.0611	0.1892	0.3416	0.6112
Polynomial (Ridge)	-0.0416	-0.0343	0.0575	0.0592	0.1951	0.3168	0.5861
Polynomial (Lasso)	-0.0350	-0.0260	0.0572	0.0584	0.1931	0.3248	0.5868
Penalised splines	0.0065	-0.0191	1.4519	1.4519	0.1810	0.4109	0.7093
Random forest	-0.0749	-0.0721	0.0375	0.0431	0.1620	0.3969	0.6842
Boosting machine	-0.0308	-0.0289	0.0347	0.0357	0.1442	0.4388	0.7422
Spatial autoregressive	-0.0344	-0.0388	0.0435	0.0447	0.1647	0.3874	0.6713

Table 6: Performance measures 2016Q3-2019Q4, implementation NQ2. Shows performance measures for market value models fitted to prices with quarterly rolling window, trained with property type sub-samples.

	MRE	MDRE	REV	MSRE	MARE	RER(0.1)	RER(0.2)
Detached ($N_T = 1,019$)							
Unconditional	0.0053	-0.0903	0.1618	0.1618	0.2842	0.2100	0.4289
Polynomial (Ols)	0.0087	-0.0033	0.0513	0.0514	0.1448	0.4759	0.7709
Polynomial (Ridge)	-0.0065	-0.0139	0.0365	0.0365	0.1428	0.4553	0.7630
Polynomial (Lasso)	-0.0075	-0.0164	0.0354	0.0354	0.1390	0.4553	0.7699
Penalised splines	0.0055	0.0018	0.0380	0.0381	0.1338	0.4956	0.8191
Random forest	-0.0390	-0.0426	0.0244	0.0259	0.1199	0.5270	0.8142
Boosting machine	-0.0214	-0.0159	0.0278	0.0282	0.1205	0.5408	0.8201
Spatial autoregressive	-0.0014	-0.0030	0.0675	0.0675	0.1390	0.5261	0.8043
Non-detached ($N_T = 3,632$)							
Unconditional	-0.0041	-0.1549	0.2108	0.2108	0.3273	0.1602	0.3475
Polynomial (Ols)	-0.0062	-0.0200	0.0369	0.0369	0.1407	0.4438	0.7541
Polynomial (Ridge)	-0.0114	-0.0282	0.0348	0.0350	0.1488	0.4053	0.7219
Polynomial (Lasso)	-0.0098	-0.0268	0.0351	0.0351	0.1475	0.4182	0.7313
Penalised splines	-0.0087	-0.0082	0.0187	0.0188	0.1037	0.5796	0.8744
Random forest	-0.0389	-0.0382	0.0201	0.0216	0.1126	0.5394	0.8425
Boosting machine	-0.0144	-0.0126	0.0185	0.0187	0.1043	0.5785	0.8673
Spatial autoregressive	-0.0089	-0.0084	0.0201	0.0201	0.1055	0.5845	0.8764
Flats ($N_T = 4,344$)							
Unconditional	-0.0145	-0.1100	0.1973	0.1975	0.3366	0.1694	0.3745
Polynomial (Ols)	-0.0146	-0.0173	0.0744	0.0746	0.1881	0.3512	0.6218
Polynomial (Ridge)	-0.0306	-0.0269	0.0595	0.0604	0.1923	0.3211	0.5956
Polynomial (Lasso)	-0.0234	-0.0221	0.0571	0.0577	0.1902	0.3312	0.5983
Penalised splines	-0.0151	-0.0178	0.0460	0.0462	0.1502	0.4357	0.7384
Random forest	-0.0679	-0.0650	0.0373	0.0419	0.1587	0.4019	0.6978
Boosting machine	-0.0240	-0.0226	0.0364	0.0370	0.1446	0.4498	0.7471
Spatial autoregressive	-0.0214	-0.0213	0.0448	0.0453	0.1573	0.4065	0.7047

Table 7: Mean squared relative error 2016Q3-2019Q4. Shows MSRE for different implementations. Log market value predictions are re-transformed to natural scale using the smearing estimator of Duan (1983). Number of observation is 8,995.

	LQ1	LQ2	NQ1	NQ2
Unconditional	0.3488	0.1998	0.3446	0.1986
Polynomial (Ols)	0.0449	9e+22	0.0479	0.1270
Polynomial (Ridge)	0.0566	0.5792	0.0470	0.0474
Polynomial (Lasso)	0.0478	0.0455	0.0473	0.0460
Penalised splines	0.0268	0.0271	0.7112	0.0342
Random forest	0.0301	0.0321	0.0326	0.0319
Boosting machine	0.0279	0.0294	0.0274	0.0286
Spatial autoregressive	0.0339	2e+25	0.0343	0.0376

Table 8: Comparison of prediction intervals. Shows summary statistics for standard and conformal prediction intervals for penalised splines model. Nominal coverage is $(1 - \alpha)$. Penalised splines model is fitted to log prices with quarterly rolling window, samples split by property type. Length reports the interval length relative to the market value estimate.

	Standard		Conformal	
	Coverage	Length	Coverage	Length
Detached ($N = 1,019$)				
$\alpha = 0.1$	0.929	0.600 [0.085]	0.898	0.517 [0.086]
$\alpha = 0.2$	0.881	0.465 [0.065]	0.826	0.394 [0.062]
Non-detached ($N = 3,632$)				
$\alpha = 0.1$	0.935	0.507 [0.061]	0.891	0.428 [0.051]
$\alpha = 0.2$	0.869	0.394 [0.047]	0.782	0.310 [0.032]
Flats ($N = 4,344$)				
$\alpha = 0.1$	0.933	0.756 [013.6]	0.871	0.595 [0.096]
$\alpha = 0.2$	0.866	0.583 [0.103]	0.762	0.437 [0.063]

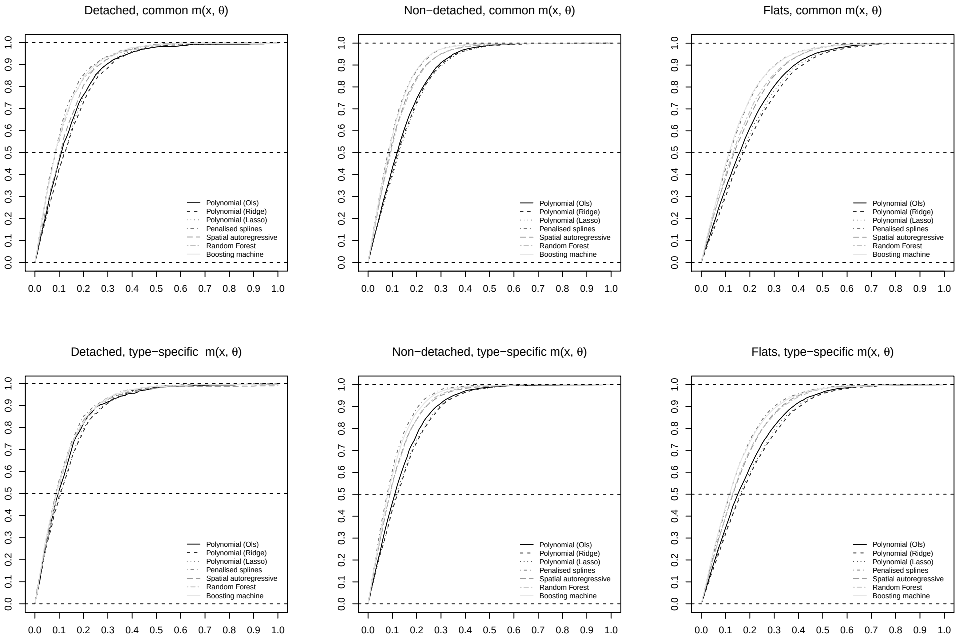

Figure 1: Relative error rate for 2016Q3-2019Q4, implementation LQ1 and LQ2. Shows the relative error rate for market value models fitted to log prices with quarterly rolling window, samples are (not) split by property type in lower (upper) panel. Log market value predictions are re-transformed to natural scale using the smearing estimator of Duan (1983). Relative error rate is truncated at 1.0.

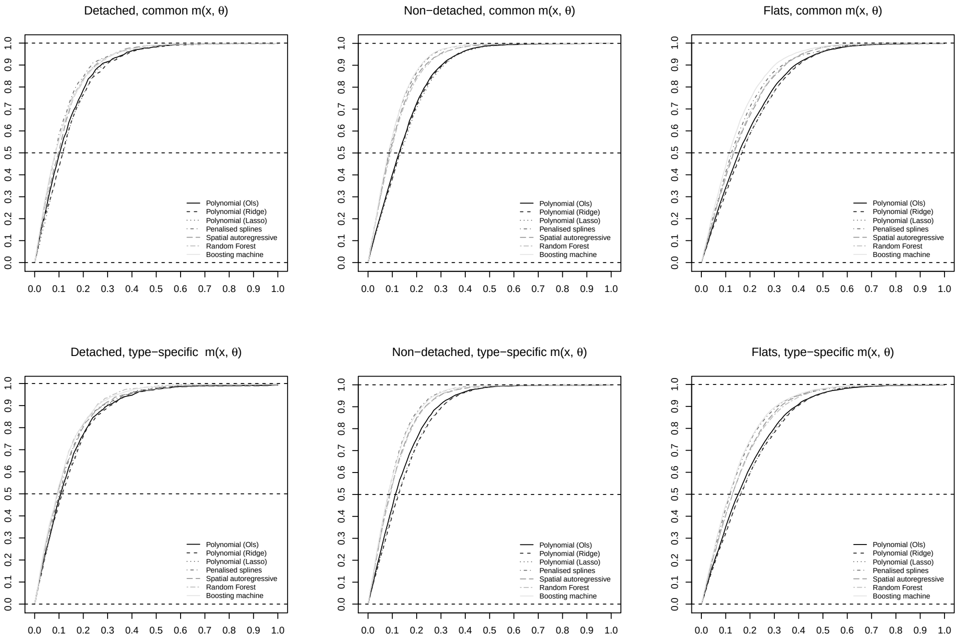

Figure 2: Relative error rate for 2016Q3-2019Q4, implementation NQ1 and NQ2. Shows the relative error rate for market value models fitted to prices with quarterly rolling window, samples are (not) split by property type in lower (upper) panel. Relative error rate is truncated at 1.0.

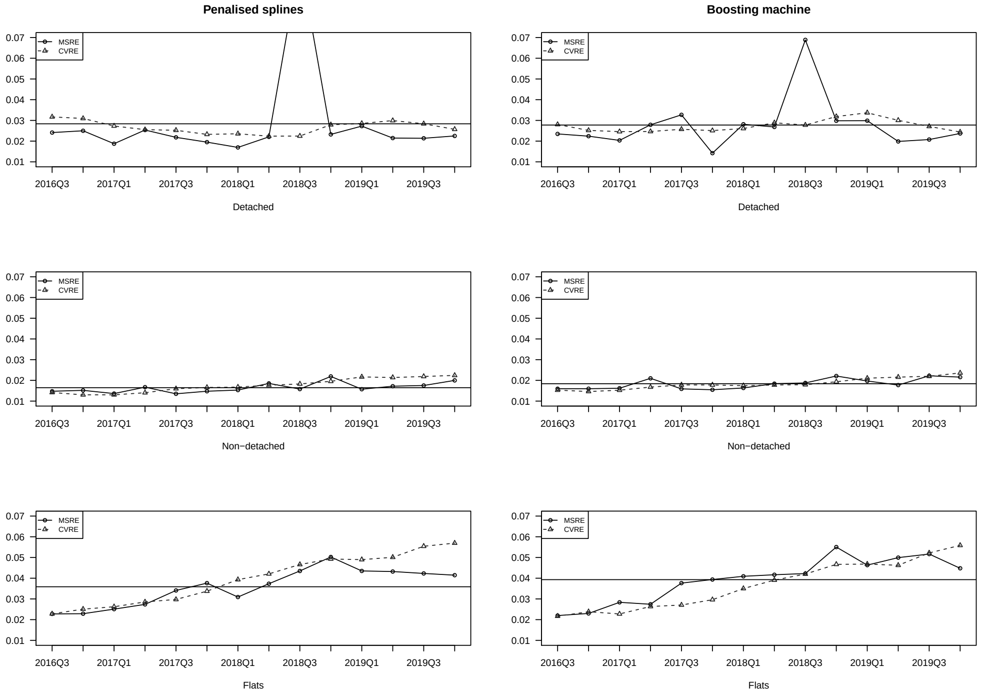

Figure 3: MSRE and CVRE for each quarter in 2016Q3-2019Q4, implementation LQ2. Shows the MSRE and CVRE for penalised splines model and boosting machine. Horizontal line is MSRE for the whole period. Models are fitted to log prices with quarterly rolling window, samples are split by property type in lower (upper) panel. Log market value predictions are re-transformed to natural scale using the smearing estimator of Duan (1983).