http://dx.doi.org/10.zillow.forthcoming

Uncertainty in Automated Valuation Models

Error- vs Model-Based Approaches

ABSTRACT

Point estimates from Automated Valuation Models (AVMs) represent the most likely value from a distribution of possible values. The uncertainty in the point estimate – the width of the range of possible values at a given level of confidence – is a critical piece of the AVM output, especially in collateral and transactional situations. Estimating AVM uncertainty, however, remains highly unstandardised in both terminology and methods. In this paper we present and compare two of the most common approaches to estimating AVM uncertainty – model-based and error-based prediction intervals. We also present a uniform language and framework for evaluating the calibration and efficiency of uncertainty estimates. Based on empirical tests on a large, longitudinal dataset of home sales, we show that model-based approaches outperform error-based ones in all but cases with very highest confidence level requirements. The differences between the two methods are conditioned on model class, geographic data partitions and data filtering conditions.

Introduction

All predictions or estimates have an inherent uncertainty to them; valuations of real property are no exception. Despite this, traditional real estate valuations practices – the manual valuations done by appraisers, valuers and surveyors – often report a single value with no notion or acknowledgement of the uncertainty of that value (French and Gabrielli 2005). This may be driven by a number of causes such as lender requirements to provide a single point estimate only, difficulty in formulating uncertainty estimates from small sample valuation methods or valuers' belief that their estimate is the exact value and has no uncertainty. Regardless of the reason, the reporting of a single value has been the standard for decades and the industries surrounding real property valuation have become accustomed to this approach.

This is, perhaps, the most curious tradition or custom in the real property valuation space. For many users of valuation estimates there is considerable financial risk being undertaken based on the value estimate. Possessing some idea of how certain that value is would be useful in negotiations, insuring and general risk avoidance measures that parties to a real property transaction may look to undertake (Bellotti 2017). Yet, reporting of uncertainty remains an afterthought in the traditional real property valuation methods.

With the advent and steady incorporation of statistical methods and, more recently, fully automated valuation models (AVMs) into valuation practice, uncertainty is now much more easily, and likely more accurately, estimated. In fact, the ability to quantify uncertainty is one of the defining improvements that AVMs offer over traditional approaches (Mortgage Bankers Association 2019). Unfortunately, no standard on how to generate and report the uncertainty of real estate value estimates has arisen. The diversity in the output from current AVM producers makes it difficult for users of valuations that do provide measures of uncertainty to both fully understand what these estimates mean and to compare between outputs from different providers/valuations. In short, confusion around and/or lack of uncertainty estimates represents an information loss in regards to the potential risk involved in real estate transactions that rely on value estimates.

In this paper we profile and compare the various methods used to generate uncertainty estimates from automated valuation models. Focusing on prediction intervals, we test for the most reliable approach to measuring and reporting uncertainty. Our empirical tests use a deep longitudinal dataset of home sales from the Seattle Metropolitan Area (Washington, USA) to compare prediction intervals derived from error- and model-based approaches.

Our findings suggest that model-based approaches to generating uncertainty estimates in the form of prediction intervals are better calibrated than those from an error-based approach at moderate to high levels of confidence, but the reverse holds at very high confidence levels (95%). In most cases, the error-based prediction intervals tend to be too conservative (too wide). These general findings hold across comparisons from linear and non-linear models and for models estimated county-wide versus those done at the submarket level. A sensitivity test on a set of data filtered to the sales near the middle 80% of the price distribution shows that when less variable data is used, the performance difference between the two approaches narrows, but the ordering remains the same. While much of the literature around AVM uncertainty methods suggests that prediction intervals (and by relation, Forecast Standard Deviations)) should be derived directly from known error distributions, the findings in this paper suggest that model-based approaches, instead, produce more calibrated and more efficient measures of value uncertainty in all cases but those demanding the highest levels of confidence.

Terminology

In general, the terminology around the concept of uncertainty in property valuation is poorly standardised. Except where purposefully pointing out differences in terms

[1] .

How to cite: Krause, A., Martin, A. and Fix, M.(2020). Uncertainty in Automated Valuation Models. Forthcoming

or directly quoting or paraphrasing from existing work, we use the following terms as:

- Uncertainty: General concept representing the fact that valuations (all predictions) have a distribution of possible or likely values. It is intended to represent the concept itself, not the measurement thereof (French and Gabrielli 2004).
- Risk: Direct measure of the potential losses due to a decision or event (French and Gabrielli 2004).
- Uncertainty Estimate: A measure of uncertainty of an estimate*. Commonly, this is an FSD, a prediction interval, or a confidence score. Uncertainty Estimates should provide a measure of general comparability that gives model users an indication of when one estimate is expected have greater uncertainty than another estimate.
- Confidence Interval: A range of possible values for a parameter estimate, such as a mean, at a given confidence level. A confidence interval for a regression function surrounds a conditional mean. In the context of a real estate valuation, it represents the estimate of an interval around the market value (the expected mean) (James et al 2013).
- Prediction interval: A range of possible values for a single prediction at a given confidence level (James et al 2013). For an AVM, this can be expressed as a pair of low and high values (a range) bounding the point estimate and provides exact probabilistic statements about how often subsequent observations are expected to fall inside the interval. In the context of a real estate valuation, it represents an interval around the realizations of sale prices.
- Confidence Level: A number from 0% to 100% giving a probability for some statement about the uncertainty estimate. Ex. A prediction interval of $100,000 to $150,000 with a confidence level of 50% means that the true value is likely to fall within this interval 50% of the time. (Papadopoulos et al 2011).
- Capture Percentage: The percentage of validation points (ground truth observations) falling within the prediction interval. The metric is only meaningful when measured out-of-sample and in aggregate. Each validation point is measured in binary form; either it is within the prediction interval or not – it is 'captured' or not. Averaging the capture over all validation points provides the capture percentage (Cumming and Maillardet 2006).
- Uncertainty Calibration: A measure of how well the probability statement from a set of uncertainty estimates matches a target probability on an out-of-sample validation set. Allows probability statements to, on aggregate, be falsifiable. For example, a calibrated model that is providing prediction intervals at an 80% confidence level will have a capture percentage of 80% – will see 80% of actual observed observations from a validation set fall within those ranges (Bellotti 2017).
- Interval Efficiency: A measure of the narrowness of the prediction interval standardized by the point prediction interval (Bellotti 2017). A measure of interval efficiency for a single estimate is calculated as:

$$\frac{pi_{hi} - pi_{lo}}{point\_estimate}$$

  where pi_{hi} is the upper prediction interval, pi_{lo} is the lower prediction interval and $point\_estimate$ is the point estimate. A summary measure of individual interval efficiency across a set of predictions can be made by taking a summary measures such as the mean or median of the individual efficiencies.

Terms within AVM Industry:

- Confidence Score: A value produced by AVM providers indicating a measure of their internal confidence in their point estimate value. Maybe a numeric or relative measure.
- Prediction Error: Difference between the point estimate and the actual observed sale price. Prediction errors are often standardized by the observed sale price to present the error in percentage terms.
- Forecast Standard Deviation (FSD): A aggregate measure of accuracy calculated as the standard deviation of the Prediction Errors for a set of estimates. Most uses of FSD are based on percentage prediction errors. The industry commonly uses the FSD (a point statistic) to create an interval estimate by multiply the FSD the standard Z-statistic related to the desired interval confidence level (ex: FSD * 1.28 = 80% Prediction Interval; FSD * 1.64 = 90% Prediction Interval). W Doing so, assumes that errors are symmetric and normally distributed. The difference between Forecast Standard Deviation and regular Standard Deviations is that Standard Deviation measures deviations from the mean, whereas Forecast Standard Deviation measures deviations of the errors from 0.
- Accuracy: Ability of the model to produce estimates with low prediction error. This has two components, bias and dispersion
- Bias: Closeness to 0 of the central tendency of the prediction errors. An unbiased model over-values as much as it under-values.
- Dispersion: The width of the prediction errors. A tight dispersion means small errors with many predictions near the eventual sale price, a wide dispersion suggests the opposite. May be measured a number of ways, including but not limited to: median absolute percentage error, mean absolute percent error, percent of errors within X%

There is confusion in the literature and the industry at large around confidence and prediction intervals. Much of the literature on uncertainty in valuation refers to the

*As opposed to uncertainty around measurement (ISO 2008)[R3.2]

Table 1. Examples of Uncertainty Estimates

	Prediction Interval	**FSD**	**Confidence Scores**
Reported Format	Interval and a confidence level	Numeric value, typically reported as a percentage of the home value estimate	Varied formats. Often Reported as a letter grade A-F
Uncertainty Estimate Example	A 90% prediction interval is $175,000 to $210,000	FSD = 0.3	B
Confidence Level Interpretation Example	There is a 90% chance that if this home sells it will sell within the specified range.	There is a 68% chance that if this home sells, it will sell for +- 30% the estimated value	None
Uncertainty Calibration (aggregate measure)	The percentage of out-of-sample validation points falling inside the interval. Target 90% capture percentage	The percentage of out-of-sample validation points falling within 1 SD, 2SD, etc. Approximate targets correspond to the standard normal distribution capture percentage	None
Uncertainty Efficiency (aggregate measure)	The typical width of an interval	Varied. May be the typical width of the implied intervals for a particular target capture percentage or the percentage of estimates below some threshold	The share of estimates receiving each grade

range around a single predicted value as a 'confidence interval' instead of a 'prediction interval'. Confidence intervals express the likely interval around a measure of a parameter or an summary measure, not a single prediction instance. More simply put, confidence intervals relate to model parameter estimates, prediction intervals to estimates of the dependent variable (Wood 2005). As most users of AVMs are interested in the single predicted value of one or a small number of properties, 'prediction intervals' are of greater concern to the industry than true 'confidence intervals' that are formed around the expected prediction value mean. Therefore, we consider 'prediction interval' to be the correct metric to analyse and, likewise, the proper terminology to use. Moreover, prediction intervals are usually considerably wider than a confidence interval as they incorporate the variance of all terms in a model relevant to make a prediction, including any global error terms. Adding to the confusion here is the use of 'Confidence Scores' in standard AVM outputs and the general use of the term 'confidence' to represent the opposite of uncertainty. We use 'prediction intervals' to indicate the a range of values around a single predicted value and 'confidence level' to indicate how much confidence (and a scale of 1% - 100%) that the prediction interval represents.

Literature Review

A key defining feature of algorithmically-driven predictions – be it of house values, the weather or medical diagnoses – is the ability to produce estimates of uncertainty (Ghahramani 2015, Scher and Messori 2018). A forecast of sunny and 28 degrees with high certainty of clear skies versus the same forecast with high uncertainty of the chance of precipitation will likely mean a different choice of clothing, or at the very least opting to bring an umbrella. Likewise, high certainty around a house price estimate may create greater leverage in negotiations and/or a more streamlined mortgage origination process as opposed to a situation where the value is highly uncertain. In short, the point estimate prediction may often grab the headline, but the extent of uncertainty should inform decisions and prescribe policy.

As context, Automated Valuation Modeling in North America occurs in a variety of forms and for a multitude of purposes. The standard use case is in mortgage lending, whereby a bank or mortgage originator relies on an automated valuation of a home to assist in assessing the value of the underlying collateral. The AVM estimate is most often used as a supporting piece of evidence alongside a more traditional appraisal, especially for mortgage originations. As of October 2019, however, originations on loans for home sales up to $400,000 in price in the U.S. may rely solely on an AVM estimate (U.S. Department of Treasury 2019). The use of AVMs in place of appraisals is also more common in refinance situations where overall risk of default is lower. Additionally, many financial organizations that hold large numbers of mortgages will occasionally request AVM estimate for all collateral in their portfolio to understand overall performance.

The second common usage of AVMs is for tax assessment purposes. Though often overlooked as a central use case of AVMs, the property tax profession has the longest history of using statistical models and automated process to produce annual values for thousands of homes. A third and newer use of automated valuation models fall broadly under marketing or informational concerns. Marketing firms may use home value estimates for large sets of addresses in order to better target their products to customers based on perceived wealth and demand for particular products. Likewise, many online listing portals – like Zillow.com and Redfin.com –

offer AVM estimates of all homes with an eye towards drawing users and informing potential buyers and sellers as to local market conditions. It is helpful to keep these varied use cases in mind when considering the differences in the approaches to uncertainty offered by the various agencies, researchers and practitioners reviewed below.

The literature on AVMs and statistical real property valuation more generally is spread broadly across three different forms of publications: 1) regulatory standards, guidelines and policy papers; 2) traditional academic research, and; 3) industry white papers. Among these sources discussions of uncertainty vary widely. With the numerous entities engaged in property valuation together with their divergent use cases for valuations, the lack of unified thinking around uncertainty is not surprising. We begin this literature review by gathering the diversity of positions. This, then, serves as the input for our efforts to develop a taxonomy of methods for deriving uncertainty estimates in the latter half.

Regulatory Standards

Regulatory standards in the AVM space break down generally into two sectors – those for tax assessment and those for collateral valuation (mortgage lending). Overall, the tax assessment sector has a more unified and consistent set of standards. Driven by the International Association of Assessing Officers (IAAO), the assessment community shares a common set of up-to-date and well documented guidelines that cover nearly all aspects of tax assessment in considerable detail. However, concrete guidelines on reporting valuation uncertainty are noticeably absent from this corpus of standards (IAAO 2017; IAAO 2018). The IAAO's discussion of uncertainty is limited to notions of parameter uncertainty in aggregated metrics (IAAO 2018) and not single point prediction.

On the collateral valuation side, there are traditional valuers (appraisers) and the AVM producers. For traditional valuation and appraisal practice, the Royal Institute of Chartered Surveyors' (RICS) Valuation Global Standards (2017) and the Appraisal Institute's Uniform Standards of Professional Appraisal (USPAP) (2018) are the guiding documents. USPAP has two specific sections dedicated to mass appraisal, neither of which discuss the possibility of anything but point estimates. USPAP does often make reference to the requirement that mass appraisal of properties should be done under 'recognised testing procedures' but does not specify what those might be or provide a reference thereto. The RICS Standards themselves do not address uncertainty either, however, it is mentioned in the International Valuation Standards (IVS) (2020) referenced by RICS (2017). The main IVS standards make passing mention of valuation uncertainty (p. 117) pointing to the market, the model or input data as its source.

In response to calls from RICS-related working groups (Mallison 1994; Carsberg 2003; French and Gabrielli 2004), the International Valuation Standards Council (IVSC) released a technical report on valuation uncertainty in 2013 (IVSC 2013). In it, they identify three drivers of uncertainty in valuations: 1) Those due to basic market forces that are exogenous to valuation; 2) Those due to the use of different modeling techniques or approaches to value; and 3) Those due to uncertainty in the inputs to valuation models. Additionally, this work (IVSC 2013) suggests that uncertainty should be both material – economically significant – and necessary to or desired by the client or user in order to warrant consideration by a valuer. When both conditions are met, valuers are encouraged to provide qualitative and descriptive, as opposed to quantitative, measures of uncertainty.

Within the AVM industry there are a collection of industry bodies which do or did offer support but not binding oversight – the Europeoan AVM Alliance (EAA), Collateral Assessment Technology Committee (CATC) and the Mortgage Banker's Association (MBA). All three have issued guidance on the measurement and reporting of uncertainty in AVMs in the recent past (CATC 2009; EAA 2019; MBA 2019). The core message in the recommendations from these bodies is that measures of confidence should correlate with actual model performance (observed model prediction errors). There is no prescriptive stance on how uncertainty should be measured and reported, only on the relationship of the output to reality.

Overall, the regulatory standards in place vary on the level of prescriptive guidance provided in regards to uncertainty. The tax assessment regulators generally ignore the issue. Organizations related to appraisers and valuers such as RICS and IVS focus on identifying the source of uncertainty and providing a narrative around it. This is a marked contrast from the advice of AVM industry bodies who are concerned solely with the output – the correlation between the measure of uncertain and the variation in observed market activities. None of these bodies offer a standardized terminology around uncertainty nor advice on constructing uncertainty measures.

Academic Research

The voluminous set of academic research into property valuation modeling can be broadly categorised into three forms: 1) Conceptual discussions of valuation methods and issues; 2) Research into the predictive accuracy of given valuation methods (often in a comparative sense); and 3) Research into the impact of a given variable(s) or feature(s) on price and/or rent. The latter – and likely larger (Knight et al 1992) – corpus of the three are highly concerned with the concept of uncertainty, but almost exclusively in the context of statistical parameter estimate uncertainty for their chosen variable(s) of interest. We ignore this body of research in this work.

We begin by reviewing the conceptual discussion of uncertainty in valuation broadly. Following this, we sample the set of comparative performance literature – the second category above – to measure how frequently uncertainty, often expressed as valuation ranges, are used as a performance criteria. It should be noted that some research has both feature impact (category 3) and performance criteria goals; so long as there is a focus on predictive accuracy we consider it in our review below.

Conceptual Work

The RICS-funded Mallinson Report (1994) catalyzed some of the earliest conceptual work around uncertainty in property valuation (Mallinson and French 2000, French and Gabrielli 2004). Both variations in model inputs and difficulties in measuring the current market and predicting futures ones

are seen as the key driving factors of uncertainty (French and Gabrielli 2004). These unknowns and unknowable factors are contrasted against the concept of risk, which is defined as a direct measure of the potential loss due to a decision or event. Another way to cast this distinction is that uncertainty is the inherent imprecision in making an estimate, risk is what one party stands to lose as a result of this imprecision (Kucharska-Stasiak 2013). The valuation profession – as impartial measurers of value(s) – are almost exclusively concerned with uncertainty and not risk.

In a series of papers, French and Gabrielli (2004, 2005, and 2006) argue that a probabilistic framework for thinking about the uncertainty of valuation inputs and market behavior is key. Their work progresses from a broad discussion of reporting and input distributional assumptions (2004), to an example cash flow simulation (2005) on through to a full case study (2006). Additional conceptual work on valuation uncertainty is limited. Meszek (2013) extends the work of French and Gabrielli (2005) using Crystal Ball by adding in a game theory component to supplement uncertainty measures obtained solely by simulation. Directly related to AVMs, Lipscomb (2017) argues for the usefulness of deriving single estimate prediction intervals via a bootstrapping approach, however, no details or examples are provided.

The initial conceptual and applied research focuses heavily on the context of a single valuation; French and Gabrielli (2005; 2006) in a commercial real estate framework and Meszek (2013) in a purely hypothetical one. While these efforts set a solid foundation for thinking about uncertainty in practice, looking only at a single valuation does not allow for those measures of uncertainty to be validated.

Uncertainty Calibration

The industry guidelines from the EAA, MBA, CATC and others all agree that confidence estimates should 'correlate' with actual model performance. To test for this form of correlation or agreement, multiple valuations need to be analyzed. Automated Valuation Models (AVMs) present an ideal use case for exploring the agreement between reported uncertainty and actual results.

Bolletti (2017) offers the first substantial test of uncertainty calibration in an AVM context. Borrowing from the Conformal Predictors (CP) literature (Shafer and Vovk 2008), Bolletti conducts a test of the relationship between uncertainty measures and model predictions. The CP approach uses two metrics to measure calibration, or the agreement of uncertainty measures with model performance: 1) Validity and 2) Efficiency. As an example, if a model provides prediction intervals at an 80% confidence level and 80% of the observed values (sale prices) fall within these ranges then the uncertainty measures are valid. If only 75% do, then it is not valid. Within the CP framework, Validity is a binary measure. If the capture percentage meets or exceeds the confidence level it is valid, if not it is invalid. The overall width of the prediction intervals is referred to as Efficiency. All else equal, smaller intervals are more informative or efficient than larger ones.

The empirical work by Bellotti (2017) is offered as a proof of concept only. In it, he compares a model that uses point estimates for market adjustments and one that uses a probabilistic approach. The value ranges (referred to as 'regions' in CP) from the probabilistic approach offer validity (at a 90% confidence) but less efficiency than those from the point estimate approach. The innovation in this work is the application of the conformal predictor approach to an AVM and in expressing a framework for testing for uncertainty calibration. The empirical results do not allow us to generalise much about different approaches to estimating prediction intervals.

Performance Comparison Research

There is a wide set of academic research that offers comparisons of varying valuation methods and techniques. Often, this work takes the form of testing a new and/or improved statistical or machine learning model against a more commonly used or benchmark model. We review this work with an eye toward understanding how often, if ever, uncertainty calibration is considered when judging model performance.

We took a convenience sample of 42 published studies that presented at least one comparison of valuation models. We sampled from both traditional real estate-focused journals as well as newer publications aimed at the broader machine learning discipline. Our sample leveraged a recent literature review of hedonic pricing studies by Wang and Li (2019) to generate this sample. Each reviewed paper made some use of accuracy metrics that compared a predicted value to an actual sales price.

Within this sample, not a single paper presented an analysis of uncertainty or prediction intervals. In each case, the only model output that was analyzed was the point prediction. Given the discussion of the importance of uncertainty calibration within AVM industry guidelines, it is surprising to see value ranges, and uncertainty in general, overlooked by comparative AVM studies within the academic research.

Industry White Papers

The actual details of how industry AVMs operate are usually a closely guarded trade secret. Many AVMs and firms related to the AVM industry do publish white papers that provide some insight into how each company's models work. Within the professional AVM space, there are two broad types of AVM producers – those that focus on providing single valuations to banks for lending purposes and those that do not, or at the very least, have a broader focus usually on marketing or information provision. This distinction is important as lending-focused providers commonly follow the industry standard approach of representing uncertainty as a combination of forecast standard deviation and Confidence Score, while non-lending focused AVMs tend to be more creative.

The standard FSD and Confidence Score approach to reporting uncertainty is standard in name only. In fact, the variation in how FSDs and Confidence Scores are created between providers is a major source of difficulty throughout the entire AVM industry (MBA 2019; Clear Capital 2020). Broadly, forecast standard deviations are produced to represent the 68.2% confidence level of possible values and are intended to provide a 'statistical degree of certainty' (Corelogic 2017). This much is agreed

[†]See Appendix A for the complete list

on by most producers. How this value is calculated varies widely and is often left poorly explained.

Confidence scores are even less standardised than FSDs. Confidence scores do not have an agreed upon representation, other than a subjective level of confidence about the valuation. Some providers give this in letter format, grades of A to F, or from High to Low. Others represented it purely as a reciprocal of the FSD, a practice that provides no additional information to the user. Still others provide a numeric confidence score, say of 60 to 100 (Corelogic 2017), but are not clear on how a value of, say 85 might differ from 65.

Moving away from the lending-focused providers additional innovation is present in the reporting of uncertainty. HouseCanary maintains the FSD terminology but does not assume normality and, instead computes FSDs from the empirical error distribution that they observe through their validation tests (HouseCanary 2018). GeoPhy – a provider of AVMs for office and retail properties – opts for Robustness and Interpretability scores as measures of the confidence in and understandability of its AVM results (GeoPhy 2019).

Summary of Literature

There are few common threads among the three bodies of literature. The more traditional valuation bodies – USPAP, RICS and IAAO – agree that there is uncertainty in markets and in aggregate measures, but do not directly address the fact that any particular estimate of value has, itself, a level of uncertainty about that value. The reporting of standard property appraisals/valuations and tax assessments mirror this as only single point value estimates are provided to customers or taxpayers. Valuers are encourage to discuss uncertainty in a narrative and not quantitative format. On the other hand, AVM practitioners and industry bodies more closely aligned with the financial industry as well as academics concur on the inherent uncertainty around point estimates and the importance of reporting in quantitatively; however, they have not been able to agree on a common language or process for measuring and reporting uncertainty.

Each of the three groups – Regulators, Academics and Industry Practitioners – generally offer very myopic contributions to the overall discussion on uncertainty in property valuation. The industry regulators that do mention uncertainty provide broadly interpretable guidelines that leave definitions and methods of measurement rather vague. For instance, the AVM organizations all agree that uncertainty measures and actual model performance should be correlated; i.e. model uncertainty should be calibrated. However, these suggestions are lacking in more prescriptive measures as to how uncertainty should be calculated, measured or reported. The academic research community has laid a conceptual basis for uncertainty in valuation, but has put forth little energy towards empirically testing for calibration in uncertainty estimates. Additionally, most academic research aimed at improving model predictions completely ignores issues of uncertainty. This absence makes it difficult for practitioners to fully leverage this work and/or understand the likely performance of these models in an applied environment where uncertainty is a critical component of model performance. Finally, industry white papers provide a brief peek into how the major AVM practitioners are representing uncertainty in their AVM outputs, but here, too, there is little agreement and few details. Providers offer varying approaches to measuring and reporting uncertainty. This together with methodological opaqueness driven by protecting trade secrets leaves AVMs customers under-informed and the reduces the ability of regulators and academics to collaborate with industry on unified standards.

Approaches to AVM Uncertainty

It can be difficult to completely disentangle FSDs and Confidence Scores when attempting to understand and categorise the methods used to measure and report uncertainty in AVMs. As a result, we survey the existing reported approaches to AVM uncertainty providers by considering FSDs and Confidence Scores jointly. In doing so, we find two fundamental approaches to calculating uncertainty are generally applied:

- Error-Based
- Model-Based

Error-Based

Error-based approaches use a distribution of known prediction errors from the model to directly create the uncertainty estimates for new predictions that are made. Ecker et al (2019) offer a clear explanation of this within a toy example using a very basic AVM with a training set of 30 sales used to value one sample home. They begin by calculating the predictive error on each of the 30 training sales through leave one out cross validation approach on a linear model. They then compute the standard deviation of the 30 cross-validated errors and label this the forecast standard deviation of the model. Next the 30 sales in the same linear model specification are used to generate a point prediction estimate of the single example home. To create a low and high value range (prediction intervals) they multiply the FSD by the factor that creates a normal distribution coverage that is desired – a 68% range would be a factor of 1.0, a factor of 1.28 for 80% and 2.0 for 95%.

The process described in HouseCanary's white paper follows a similar path by using the empirical distribution of the errors but does not force normality on the range. The approach by which they then directly tie observed errors to individual home prediction intervals is not elaborated. Other vendors and consultant reports appear to suggest an error-based approach, but also do not provide definitive descriptions (Gordon 2005; Connected Analytics 2015; Freddie Mac 2020). GeoPhy's (2019) 'robustness score' also falls into this category. This score is created by summing the reciprocals of a number of standard error metrics, though again, no exact mapping to individual predictions is provided.

Model-Based

Model-based approaches to uncertainty are generally derived one of two ways. The first is in cases where multiple

[‡]Corelogic (2011, 2017) presents a possible third approach; one based on information quality, noting that their uncertainty measure is [a] 'range of estimates based on consistency of information'. As construction of this metric is proprietary we do not consider this approach in this paper

different models are calculated – an ensemble (Clear Capital 2020). In such cases, a measure of uncertainty derives from looking at the distribution of values that are produced by the different models and creating uncertainty estimates (at a certain confidence levels). This can be done through different model classes in a traditional ensemble or through a bootstrapping, or re-sampling, approach with the same model class and specification (Lipscomb 2017). As an example, Miller and Sklarz (2017) offer a method for using variation in the sale prices of comparable properties used within traditional appraisal framework to derive prediction intervals.

A second model based approach is to use standard prediction interval calculations derived from parametric model assumptions. This approach is generally limited to linear based models that have coefficient and standard error outputs from which to calculate prediction intervals. Intervals from this latter approach end up being symmetric, while those from the first can take any distributional shape.

We see advantages and disadvantages to both of the above approaches. Conceptually, we can compare these across three components: 1) Directly tie to past model performance; 2) Consideration of model class used; and 3) Recognition of data completeness, density and quality. Each approach can be categorised as addressing these components either directly, indirectly or not at all (Table 2).

Table 2. Prediction Interval Method Classification

Method	Ties to Observed Errors	Considers Model Class	Recognises Data Quality Differences
Error-Based	Directly	Indirectly	No
Model-Based	Indirectly	Directly	Indirectly

One additional issue to consider is the symmetry of the prediction interval. A symmetric prediction interval means that the upper interval value is always the same distance from the point estimate as the lower interval value. Methods that are based on normal distributions and multipliers of FSD values (Gordon 2005; Freddie Mac 2020) necessitate symmetric ranges. Other approaches are not limited to symmetric ranges. Ecker et al (2019) argue that symmetric ranges are always desirable; we disagree. As many (most?) real estate pricing decisions are influenced, at least in part, by comparison of local and recent comparable sales, it is unlikely that all price points of reference are evenly distributed around the point estimate in a symmetric fashion. Related, home prices, like other financial assets, often fall into approximately log-normal distributions. This means that certain model classes and/or data transformations can inherently generate symmetric prediction intervals and errors; however, this symmetry will not be represented in standard dollars, but rather in log dollars. Finally, other drivers of uncertainty such as measurement errors may also not show symmetric variation. In sum, we hold that the desirability of symmetric uncertainty estimates should be an empirical question.

Overall, we concur with many of the industry bodies and commentators that the variation in the terminology around and approach to uncertainty in property valuation is a disservice to the industry and consumers in general. Further, we believe the question of which approach to use should be answered empirically. In the remainder of this paper, we set out to test these two approaches to generating prediction intervals on a deep, longitudinal data sample of homes sales from King County WA (Seattle Metro).

Methodology

Our review of the literature shows two primary methods for creating prediction intervals: error-based and model-based. In this section we detail the approach we use to test the performance of different methods to generate calibrated prediction intervals. We begin by discussing the model classes and geographic scales at which we perform our empirical tests. Next, we outline the two approaches to prediction interval creation that we will use. We follow this by describing the metrics used to evaluate calibration. Finally, we conclude this section by outlining the full experimental approach, including a sensitivity test.

Model Classes and Geographic Partitions

The goal of this paper is to compare the quality of error- and model-based approaches to creating prediction intervals. To provide the most generalizable set of findings, we make these comparisons across two different model classes – linear and non-linear – and at two different geographic partitions – global and local.

Our linear model specification uses a simple ordinary least squares (OLS) estimator in which the natural log of the sale price is expressed as:

$$ln(price) = f(S, L, T)$$

Where S are structural features, L are locational features and T are temporal features (Fik et al 2003). More specifically, these variables are:

- Townhome (*binary: Is property a townhome?*)(S)
- Year Built (S)
- Home Size in SqFt (S)
- Home Quality (S)
- Home Condition (S)
- # of Bedrooms (S)
- # of Bathrooms (to the .25 bath) (S)
- Lot Size in Square Feet (S)
- Waterfront Location (*binary: Is property waterfront?*) (L)
- View Score (L)
- Latitude (L)
- Longitude (L)

- Month of Sale (*monthly dummy variables*) (T)

While we recognise that it is unlikely that commercial-grade AVMs use a specification this simple, our goal is to explore prediction interval calibration, not model accuracy, and a relatively parsimonious OLS specification provides a useful baseline.

As a complement, we also estimate a non-linear model via a random forest. To keep the comparisons as equal as possible, we use a very similar model specification:

$$ln(Price_{timeadj}) = f(S, L)$$

Where *ln*(*Price_timeadj*) is the natural log of the time adjusted sale price, S are structural features and L are locational features. There are a few slight differences from the linear specification:

1. We time adjust[§] the sale prices before estimating the model as random forests do not handle temporal control, or fixed effects, variables as well as linear models.
2. We use the default random forest hyperparameters set in the R `ranger` package: 500 trees, a minimum node size of 5; and an 'mtry' of the square root of the number of features in the model (Wright and Ziegler 2017).

We also conduct our comparisons with two different geographic partitions. The primary reason for doing so is that overall data size or density may influence the quality of prediction intervals. By estimating models with all data and then again with subsets of it, we can test for changes to calibration and efficiency based on sample size.

The first partitioning scheme is the entire county, whereby we use all of our data in the same model. As the test county is rather large, nearly 2.2 million residents in 2019, we can break the area into many smaller 'submarkets' and still have enough data to estimate and evaluate model performance. We manually create 19 residential submarkets based on existing county tax assessor assessment areas. Most of these areas contain roughly 1/19 of the total volume of transactions in the county, though there are a few larger and smaller submarkets. In the results below we refer to the full county models as 'Global' and those partitioned at the submarket level as 'Local'.

Prediction Interval Methods

The example presented by Ecker et al (2019) represents an error-based approach. In their example all properties, regardless of their characteristics, receive the same prediction interval width. These widths are expressed in a percentage sense; i.e. relative to the point prediction. As an example, if the model has a known FSD of 15%, then all properties would get a prediction +- 15% if the provider were giving an 68.2% prediction interval. If 90% confidence level is desired, then the prediction intervals would be be 24.7% +/- (15% * 1.645). Regardless of the confidence level, the error approach will produce symmetric and normally distributed prediction intervals.

Producing an error-based uncertainty measure requires knowledge of the known error distribution of the model. The errors can come from any type of class of model, so long as they can be summarized in a single measure of dispersion, such as FSD. Following Ecker et al. (2019) we calculate these via cross-validation (5-fold) of the training sample with both classes of models discussed below. The FSD is derived from the standard deviation of the cross-validated percentage prediction errors from the model. An error-based approach produces low and high deviations from the point prediction in percentage terms – ex. prediction of $100,000 +/- 15% – which are then converted into low and high range values – $85,000 and $115,000 in this example.

Unlike the error-based method, the model-based approach varies by model class. Additionally, and also in contrast to the error-based method, model-based approaches directly estimate the low and high range values, not the width of the range itself. As such, the model-based approaches can be both asymmetric and non-normally distributed.

Linear Model

For the linear model, we create prediction intervals by resampling errors (Davidson and Hinkley 1997). We assume the error term in our linear model has *iid* errors, but we do not require the assumption of normality. We then mimic samples from the distribution : $(\widehat{\beta} * X) - (\beta * X + \epsilon)$. We run 100 bootstrap trials on each model and select our prediction intervals from the distribution of the predictions resulting from the 100 trials. In each trial, the dependent variable is the predicted value of that observation perturbed by a variance adjusted residual, sampled from the residuals of the base model, with replacement (Davidson and Hinkley 1997). For example, for an 80% confidence level, we would extract the 10th and the 90th percentiles from the distribution of bootstrapped predictions as our prediction interval.[¶]

The full algorithm for a single prediction, *i*, with an 80% prediction interval is:

1. Estimate the base model with the full set of training data, length *N*
2. Make a point prediction for for observation *i* using the base model
3. Variance adjust the residuals from the base model where the adjusted residual is the residual divided by the square root of 1 - the leverage of that observation
4. Repeat the following 100 times:
   (a) Sample from the adjusted residuals in #3 *N* times
   (b) Create a perturbed predicted value for each training observation by adding the sampled

[§]We use a robust linear model to create a house price index and then adjust all sales to a single point in time (the most recent month in the data). We use the `hpiR` R package to create the house price indexing (Krause 2020).

[¶]It should be noted that this approach is a considerably more complex approach to prediction intervals than the standard linear model extension common in textbooks, whereby the distributions of parameter estimates and model standard errors are combined leveraged to produce prediction intervals.

residual from 4(a) to the predicted value of each training observation from the base model

(c) Re-estimate the model with new dependent variable values from step 4(b)

(d) Extract the residuals from this new model

(e) Variance adjust these with the formula from #3

(f) Sample 1 of the residuals from 4(e) and add to the predicted value for observation *i* from model in 4(c) to create a new predicted value

5. From the 100 prediction from step 4(f), sort in order and select the desired quantiles. In the case of a 80% confidence level, the 10th and 90th percentile.

Non-Linear Model

Generating prediction intervals from the non-linear model leverages the quantile random forest approach first offered by Meinhausen (2006). For each predicted value, we examine the entire distribution of sales prices that occupy shared terminal leaf nodes from each individual tree in the random forest. We order these sale prices, and then, like the linear bootstrap approach above, we select our low and high prediction interval values from the corresponding quantiles in the distribution of values extracted from the leaf nodes. Again, for a 80% confidence level, we choose the estimated values at the 10th and 90th percentiles of the distribution of values from the leaf nodes.

Evaluation Metrics

Calibration is the primary metric for evaluating prediction interval reliability. In this context, calibration refers to the agreement between the capture percentage of the prediction intervals – the proportion of validation points which fall within the interval range – and the desired confidence level (Leathart and Polaczuk 2020). As an example, for a model producing prediction intervals at an 80% confidence level 80% of the actual observed validation points would need to fall within the prediction intervals. For the tests below we examine confidence levels at 50%, 68%, 80% and 95%, but in practice any set of confidence levels could be used.

Previous work (Shafer and Vovk 2008; Bellotti 2017) treats the measure of calibration as a binary metric; the capture percentage of a model's prediction intervals at a given confidence level either hit the intended proportion (are calibrated) or they don't (are not calibrated). As such, it is a one-sided measure whereby a capture percentage of 84% for an 80% confidence interval is more desirable than a 79% capture percentage. Here, we treat mis-calibration as a two-sided, continuous measure in which intervals that are too conservative – capture percentage greater than confidence level – are equally wrong as intervals that are too aggressive (too tight). We adopt this differing interpretation as the one-sided approach offered by Shafer and Vovk (2008) is certainty pertinent in a risk-related framework, but may not be in other situations. To present the broadest and most widely operable evaluation framework, we opt for a two-sided metric.

All else equal, narrower prediction intervals are more informative and more efficient than wider ones. A measure of this efficiency is the second common metric relating to the evaluation of prediction intervals. We term this 'Interval Efficiency' and measure it as the mean of the widths of all prediction intervals at a given confidence level. To standardize the widths across predictions of different valuation amounts, we convert these to relative measures where the width of the prediction interval is divided by the point estimate.

Data

To empirically test the performance of the error- and model-based uncertainty methods we use a large, longitudinal dataset of home sales from King County, WA. These data include all single family residential home sales (detached and townhomes only) in King County (Seattle Metro) from January 1, 1999 through December 31, 2019. These data originate with the King County Tax Assessor but have been collected, cleaned and is available online in an open sourced R package.[‖] The raw data include 485,044 transactions from the 21-year time period. There are 38 different features or variables about each home. The spatial extent of the sales within the county is shown in Figure 1.

Table 3 below shows a summary of sales data along with the variables used in the models tested below. There is a wide variety of prices in King County as the housing stock in Seattle and its suburbs are highly variable. Multi-million dollar homes are common on the waterfront areas and in the central, older neighborhoods, with smaller and more affordable residences in the far south and north. The variables in the data are, generally, self explanatory with the exception of view score. The King County Assessor assigns a rating on a scale of 0 (no view) to 4 (excellent view) for a variety of different views (water bodies, mountains and city skylines) in the region. We sum all the view score ratings to create a composite view score for each observation in data.

In addition to the variables provided in the raw data, we add one more, a submarket label. The existing data have an Assessment Area field that indicates the small assessment zones designated by the local tax officials. There are 95 of these, many of which may be too small in which to estimate separate valuation models. To remedy this we aggregate these 95 areas into 19 core 'submarkets' based on geographic proximity and natural boundaries. A useful feature of these assessment areas is that they are not always geographically contiguous. Some, especially those covering special properties such as high-end waterfront homes, are spread over a wide area, interspersed with other zones. This highlights the advantage of using these assessment zones over, say, ZIP codes designations as these zones are specially created to capture local market effects.

To create the local models in the analysis below, we estimate separate valuation models in each of the 19 submarkets using the identical model specification. We take this approach, as opposed to the common approach of using a fixed-effect approach, to allow for our intercept and coefficients to vary by submarket in the linear model and to simplify the tree-splitting depth needed in the random forest approach.

In order to best represent the 'noisiness' of data in a real-world AVM context, we maintain the majority of outlying data observations in our primary empirical analysis. There

[‖]See https://github.com/andykrause/kingCoData for instructions on how to access this data as well as details on its construction.

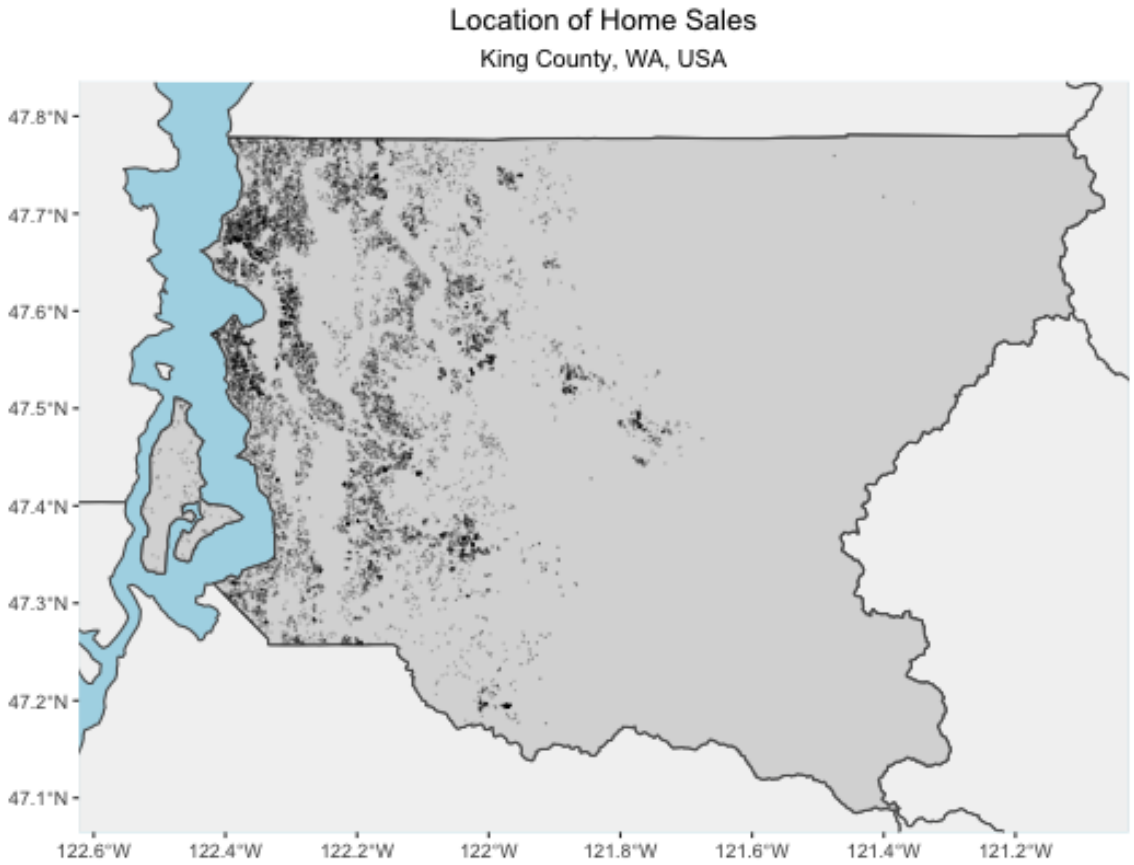

Figure 1. Location of Home Sales

Table 3. Summary Statistics

Variable	Min	25th	Median	Mean	75th	Max
Sale Price	$50,293	$280,000	$400,000	$503,894	$600,000	$30,000,000
Townhome (y/n)	0.00	0.00	0.00	0.06	0.00	1.00
Year Built	1900	1952	1977	1972	1999	2019
Home Size (sq.ft.)	402	1440	1940	2090	2570	20140
Home Quality	1.00	7.00	7.00	7.64	8.00	20.00
Home Condition	1.00	3.00	3.00	3.48	4.00	5.00
Bedrooms	1.0	3.0	3.0	3.4	4.0	13.0
Bathrooms	0.50	1.75	2.25	2.15	2.50	12.75
Lot Size (sq. ft.)	375	5100	7500	14239	10380	2380118
View Score	0.00	0.00	0.00	0.50	0.00	20.00
Waterfront (y/n)	0.00	0.00	0.00	0.01	0.00	1.00
Latitude	47.16	47.45	47.56	47.55	47.67	47.78
Longitude	-122.53	-122.32	-122.22	-122.21	-122.12	-121.16

are a number of records that do appear to be either data errors or properties that are not single family residences or townhomes. We employ the following filters to the data, which remove 10,714, or 2% of the total observations.

- Home size (sq. ft.) greater than 400 square feet (37 sq m.)
- At least one half bathroom
- At least one bedroom
- Fewer than twenty bedrooms
- First floor square footage greater than or equal to lot size
- Year built less than or equal to year of sale

Empirical Approach

To test the performance of the two uncertainty methods we employ the following 'walk-forward' approach (Stein 2007) across the entire time frame of our study, January 1999 through December 2019.

1. Isolate one year's worth of sales (training data). Ex. January 1999 through December 1999
2. Use the training data to make point predictions on sales in the month following - Ex. January 2000

3. For both of the uncertainty approaches – error-based and model-based – calculate prediction intervals at confidence levels of 50%, 68%, 80% and 95%.
4. Compute the valuation error by comparing the predicted value to the known sale price
5. Move the 'window' of analysis ahead one month and repeat steps 1-4 for the entire time period. With 21 years of data, minus the initial one year training period this amounts to 240 different 'windows' of analysis.
6. Combine the results – the predicted values, prediction errors and the two sets of prediction intervals – from all 240 periods
7. Calculate the model's predictive accuracy for the point prediction values
8. Calculate uncertainty calibration and efficiency for each of the two sets of uncertainty methods at each of the 4 confidence levels.

As discussed above, we evaluate the uncertainty methods for two different model classes – linear and non-linear – and for two different partitioning schemes – global and local. This equates to four different comparative situations.

- Linear Model, Global Partition
- Linear Model, Local Partition
- Non-Linear Model (Random Forest), Global Partition
- Non-Linear Model (Random Forest), Local Partition

When analyzing the locally partitioned models, we aggregated the results from all 19 submarkets in order to offer an equal comparison to the global models.

Results

We begin by providing an overview of the predictive accuracy of the four comparative modeling setup. The focus of this work is on prediction intervals and their calibration and efficiency; we present these point estimate accuracy figures as context only. Table 4 shows a summary of the predictive accuracy of the point predictions for the greater than 400,000 sales in King County over the 2000 through 2019 time frame. The metrics are as follows: MdAPE, median absolute percentage error; MdPE, median percentage error; PE10, percent within 10 percent of sale price; and PE30, percent within 30 percent of sale price. These findings are relevant in interpreting the calibration and efficiency results below.

Table 4. Accuracy Results

Model	Partition	MdAPE	MdPE	PE10	PE30
Linear	Global	0.135	0.005	0.388	0.826
Linear	Local	0.094	0.000	0.525	0.899
Non-Lin	Global	0.073	-0.002	0.616	0.923
Non-Lin	Local	0.071	-0.003	0.623	0.926

First, the non-linear models – the random forests – are considerably more accurate than the linear approach. This is especially true at the global (entire county) scope. Moving to locally partitioned models substantially improves the linear model, while it has limited impact on the accuracy of the non-linear model as the flexibility in the non-linear approach can better handle spatial heterogeneity. This is not a surprising result given the heterogeneity of the housing market in King County.

Second, all four models are relatively unbiased with median percentage errors (MdPEs) near 0. The slight under-prediction bias is likely due to the experimental design whereby we generate predictions in, for example, January 2000 using data from January 1999 through December 1999. The overall level of appreciation in the prediction month likely contributes to this small bias figure as there were more months of positive price gains during this 20-year period than months with price declines. Again, given our focus on prediction intervals and on the relatively low level of bias we are not concerned with any negative impacts on our generalizability due to the experimental design.

Calibration

A comparison of calibration results for both methods across all four combinations of model class and geography are shown in Table 5. A number of general findings can be made.

Across all four partition-model combinations, the error-based prediction intervals are usually overly conservative – that is, the capture percentages are higher than the confidence levels. For example, in the Global partition, Linear model experiment 62% of observed sale prices fell within the 50% prediction intervals. The level of conservative-ness falls as the confidence level rises and, once we get to a 95% confidence level the prediction intervals from the error-based approach are very well calibrated. The trend of over conservative-ness that decreases with increasing confidence levels holds across model classes and partition schemes.

Conversely, the model-based prediction intervals show differing calibration based on model class. For the linear models, there is a slight over-confidence in the prediction intervals – capture percentage is less than the confidence level. This mis-calibration is small, generally around 1-2% (ex. capture percentage of 49% for a confidence level of 50%) and holds regardless of confidence levels. For the non-linear models, the model-based prediction intervals are overly conservative, though generally less so than error-based approach and less so at higher confidence levels.

Comparing between the error- and model-based shows some mixed results. For confidence levels 80% and lower the model-based prediction intervals offer better agreement between capture percentages and confidence levels (are less mis-calibrated). Within these lower confidence scenarios, the primacy of the model-based approach to prediction intervals is more evident in linear models as opposed to the non-linear approach. The difference is also larger the smaller the required confidence level. Once looking at a 95% confidence level, the error-based approach become the preferred approach. The differences are relatively small – 2% or less – but are consistent across the four different partition and model class combinations.

Table 5. Calibration Results

Model	Conf. Level.	Error-Based	Model-Based
Global			
Linear	50%	0.62	0.49
Linear	68%	0.79	0.66
Linear	80%	0.87	0.78
Linear	95%	0.95	0.93
Non-Lin	50%	0.69	0.57
Non-Lin	68%	0.81	0.75
Non-Lin	80%	0.88	0.86
Non-Lin	95%	0.95	0.97
Local			
Linear	50%	0.64	0.50
Linear	68%	0.80	0.67
Linear	80%	0.88	0.79
Linear	95%	0.95	0.93
Non-Lin	50%	0.67	0.56
Non-Lin	68%	0.80	0.74
Non-Lin	80%	0.87	0.85
Non-Lin	95%	0.95	0.96

Values in this table denote the capture percentage of each model at each confidence level

Table 6. Interval Efficiency Results

Model	Conf. Level.	Error-Based	Model-Based
Global			
Linear	50%	0.37	0.26
Linear	68%	0.54	0.40
Linear	80%	0.69	0.53
Linear	95%	1.06	0.97
Non-Lin	50%	0.25	0.19
Non-Lin	68%	0.37	0.29
Non-Lin	80%	0.47	0.39
Non-Lin	95%	0.73	0.69
Local			
Linear	50%	0.27	0.19
Linear	68%	0.39	0.29
Linear	80%	0.51	0.39
Linear	95%	0.78	0.71
Non-Lin	50%	0.22	0.18
Non-Lin	68%	0.33	0.27
Non-Lin	80%	0.42	0.37
Non-Lin	95%	0.65	0.64

Values in this table denote the median relative width of the prediction intervals for each model at each confidence level

A likely reason for the poorer performance of the error approach at lower confidence levels is that it assumes a normal and symmetrical error distribution. This assumption may not hold as well in the 'middle' distribution of the data as when the prediction intervals are expanded to capture 95% of the likely values. Additionally, the error-based approach requires an *a priori* error distribution drawn from a previous assessment of model performance. Inherently, this assumes that the types of homes that sell in the prediction month are similar to those in the previous model assessment period; deviations from this will cause less reliable prediction intervals under the error approach as opposed to the model approach. The model-based approach is able to account for changing samples as it uses the characteristics of the homes being valued to generate prediction intervals and not just past model performance (as the error-based method does).

Interval Efficiency

Next, we look at the efficiency, or narrowness of the intervals. Given the same level of calibration, narrower prediction intervals are preferred to wider ones. The figures shown in Table 6 are relative measures of interval efficiency. For example, the value of 0.35 indicates that the mean relative prediction interval for this model at this confidence level is 35% of the predicted price. On an example home with a predicted price of $100,000, this would equate to a prediction interval of $82,500 to $117,500.

As expected, lower confidence levels result in narrower prediction intervals, all else equal. For both approaches, interval widths at 95% confidence are about three times those at 50%, with width increasing sharply when going from 80% confidence to 95%. In other words, that added level of confidence is quite expensive from an interval efficiency sense.

When using all transactions in the same model (Global partition), the non-linear intervals are about 30% narrower than the those from the linear model. This holds for both the error- and the model-based approaches. Moving to local models this differences diminishes suggesting that, much like the accuracy results in Table 3, a localized partitioning scheme can narrow the performance gaps between linear and non-linear modeling approaches.

The model-based prediction intervals are consistently narrower or more efficient than those from the error-based approach at each model and confidence level combination. The error-based intervals range from 5% to 35% wider, with smaller differences at the higher levels of confidence. We would expect this convergence at the high end as confidence near 100% the intervals from either method converge around very wide widths nearly equivalent to the price itself. Simply put, there is less ability for the modeling approach to differentiate at those very high levels of confidence.

Calibration vs Efficiency

Both metrics – calibration and efficiency – highlight related, but distinct, components of measuring and reporting uncertainty in automated valuation modeling. In the best case, a model would be both calibrated and have efficient prediction intervals. In reality, there is often a trade-off to be had between the two.

Figure 2 illustrates the relationship between calibration (x-axis) and prediction interval efficiency (y-axis). In this example comparison, we see that the model-based approach is both more efficient (lower on the y-axis) and more closely calibrated (closer to the vertical gray line on the x-axis). Note that points to the left of the gray vertical line represent overly aggressive prediction intervals (too narrow) and those to the right overly conservative ones. As the scale indicate, we see more conservatism in the analyses we ran. There are notable

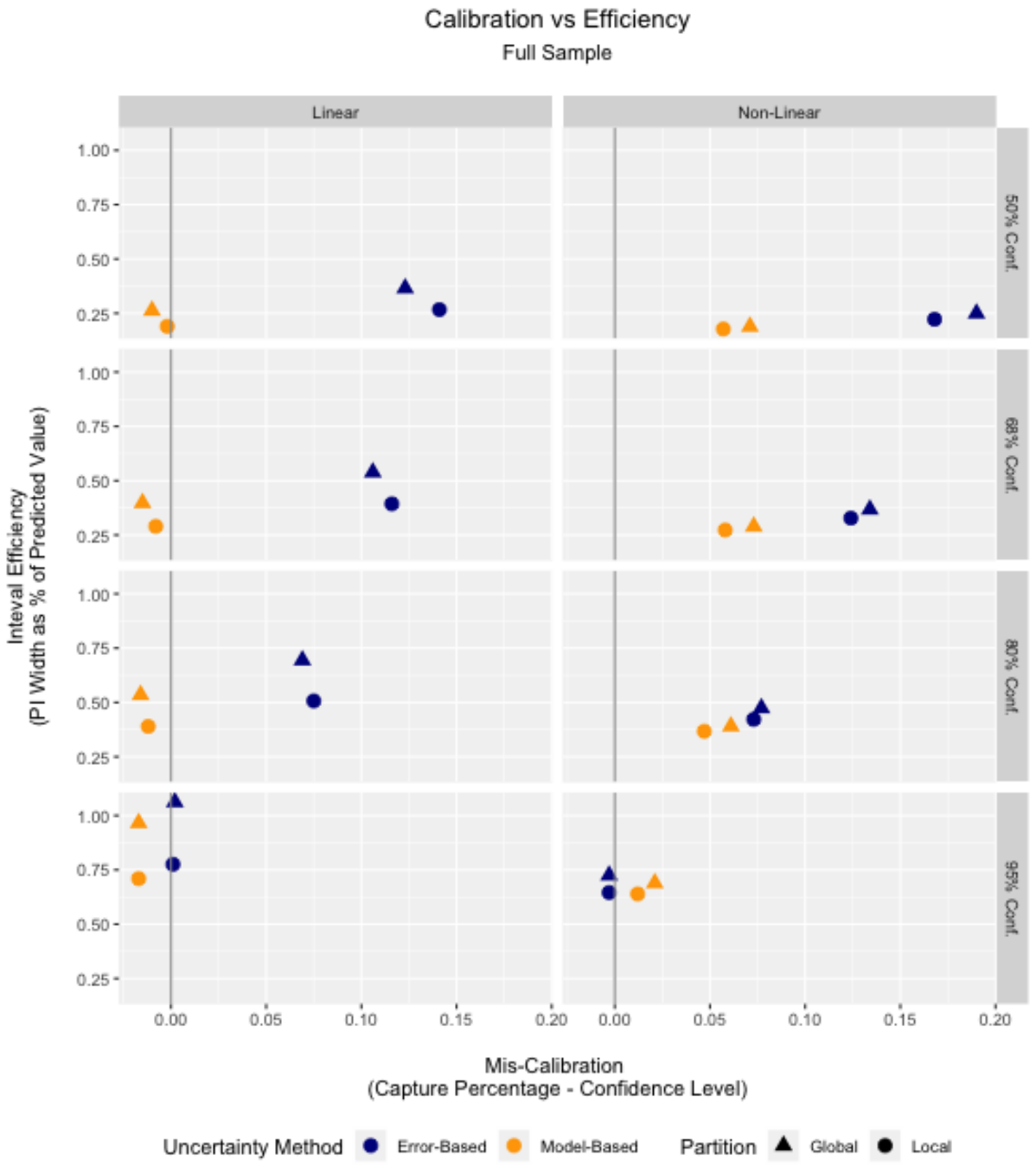

Figure 2. Calibration vs Efficiency

differences between the calibration-efficiency relationship between model- and error-based approaches depending on confidence level. Intervals at the highest confidence level – 95% – indicate that error-based approach is better calibrated though slightly less efficient. For the three lower confidence levels, the model-based intervals clearly dominate the error-based ones. These differences are greater for the linear model (left panel) than the non-linear. The relationship holds relatively similar between the Global and Local partitions, though with Local modeling showing much better efficiency, though often worse calibration for the linear model.

It is also important to note that, though in most cases the linear models (left panels) are more closely calibrated than the non-linear models, they are also considerably less accurate (see Table 3). Accuracy becomes an additional consideration that users or model developers may take into account when calibration and efficiencies are relatively close.

The major takeaway here is that the model-based intervals are more appropriately calibrated and as or more efficient for each model-partition-confidence interval pair, except for measures of calibration at the very high 95% confidence level. At the highest level, the error-based model does appear comparable, though the differences are smaller.

Sensitivity Test

The above results were derived using most (98%) of the observations from the original dataset. These findings are based on a large, longitudinal dataset from which we've removed very few outlying observations. In practice, an

AVM provider or researcher may apply more stringent filters to their data. To examine the sensitivity of our initial findings to the data quality, we perform a sensitivity test where we remove a much larger set of data. We do this to expand the generalizability of our findings to a wider swatch of potential AVM use cases.

For this 'filtered data' test, we employ two filters. First, we remove all transactions that had a major change in home facts over the twenty-year time period. We remove these as there is some doubt over whether or not the sale was observed before or after the renovation or rebuild. Next, we limit our data to the middle 80th percentile of sales prices in each of the 21 years, essentially eliminating both the low and high end decile of sales from the sample. This creates a more homogeneous set of homes to value. Employing these two filters reduces our observations down to 348,407, a reduction of 27% from the dataset used in the original analysis above.

To put the differences in the datasets in context, Table 7 shows the accuracy improvement due to the filtered sample. All model classes and partition combination gain accuracy from removing outlying sale prices, with the improvements relatively even across model classes. It should be cautioned that this doesn't mean that the filtered data results in a better model, only that the set of sales for which the model is built on and, more importantly, evaluated on is 'easier' in the sense that they are more likely to be less unique homes.

Table 7. Accuracy Results - Sensitivity Tests

Model	Sample	Full	Filtered	Change
Linear	Global	0.136	0.116	-14.1%
Linear	Local	0.093	0.076	-19.1%
Non-Lin	Global	0.074	0.061	-16.4%
Non-Lin	Local	0.072	0.059	-16.9%

The results of the sensitivity test along with the results from the original analysis (grayed out) are shown in Figure 3. In all cases, across interval type, model class, confidence level and partition, the intervals from the filtered sample are as or more reliably calibrated than in the full dataset. This is indicated by horizontal distance to the gray vertical line that denotes perfect calibration.

For the linear models (left panels) the model-based prediction intervals showed little difference from those found in the full sample. Interestingly, regardless of confidence level and partition, the model-based for the linear models all cluster around 0.02 percentage points too aggressive in terms of calibration. The error-based intervals for the linear models show considerable improvement with the filtered data sample for all of the confidence levels except 95%. At the lower confidence levels, the global partitions are better calibrated though less efficient than the locally partitioned models.

With the non-linear models (right panels), there are big gains in calibration reliability for the error-based model with the filtered sample, while the model-based approach sees more modest improvements. At the 95% confidence level we see a near parity in calibration reliability between the the error- and model-based intervals as well as between global and local partitions. In general, there is little difference in calibration between the global and local partitioning schemes for the non-linear models.

Looking into the efficiency measures (y-axis) we see improvements across the board due to the filtering of the data. This is expected as removing transactions at the tails of the price distribution should result in a more homogeneous set of validation sales and, overall, more confidence in the valuations. The accuracy results in Table 7 highlight the predictive benefits removing outlying observations from the set of sales on which to evaluate model accuracy. In both the error and the model-based approaches, the increases in efficiency (reduction in interval width) are greater for the higher confidence levels. This is especially apparent for the 95% confidence level where the prediction interval widths are halved for linear models and reduced about 30% for the non-linear ones. For the lower confidence levels, the filtered data benefits the error-based model slightly more than the model-based approach, though this finding is reversed at 95% confidence.

Finally, whereas in the full data sample the model-based approaches were generally more efficient than the error-bssed ones at the lower confidence level, the findings from the filtered sample show parity in interval efficiency. Additionally, the 95% confidence level seems to favor the model-based approach for the linear model but the error-based approach for the non-linear, though the differences are fairly small.

Overall, this sensitivity test is generally supportive of our earlier findings; 1) Model-based methods of determining uncertainty (via prediction intervals) are more reliably calibrated then error-based ones, except at the highest levels of confidence; 2) Model-based methods tend to offer slightly more efficient prediction intervals – though conditioned by data quality – again, with some exceptions at the highest level of confidence.

Discussion

The uncertainty of point estimates from AVMs is highly useful information for many users. Despite this, the existing literature and guidelines around how to estimate, measure and report uncertainty is rather limited. Industry and organizational guidelines for the AVM providers do emphasise that uncertainty calibration – the agreement between confidence level and capture percentage – is critically important, however, they offer few details on the mechanics of creating prediction intervals and measuring calibration. The academic literature has been producing research on mass valuation of real estate for nearly fifty years, yet the discussion of uncertainty estimates in general, and prediction intervals specifically, is limited and empirical tests of methods to do so are practically non-existent. Adding to the difficulties here is a lack of established terminology around uncertainty in general.

In short, there is very little shared language, limited practical advice on interpreting uncertainty for users of AVMs and a lack of specific instruction for AVMs producers on creating, measuring and reporting AVM uncertainty. In this paper we clarify the language around discussing uncertainty and outline an evaluation framework for measuring the quality or reliability of uncertainty estimates. Using a twenty-year dataset of sales (over 480,000 observations)

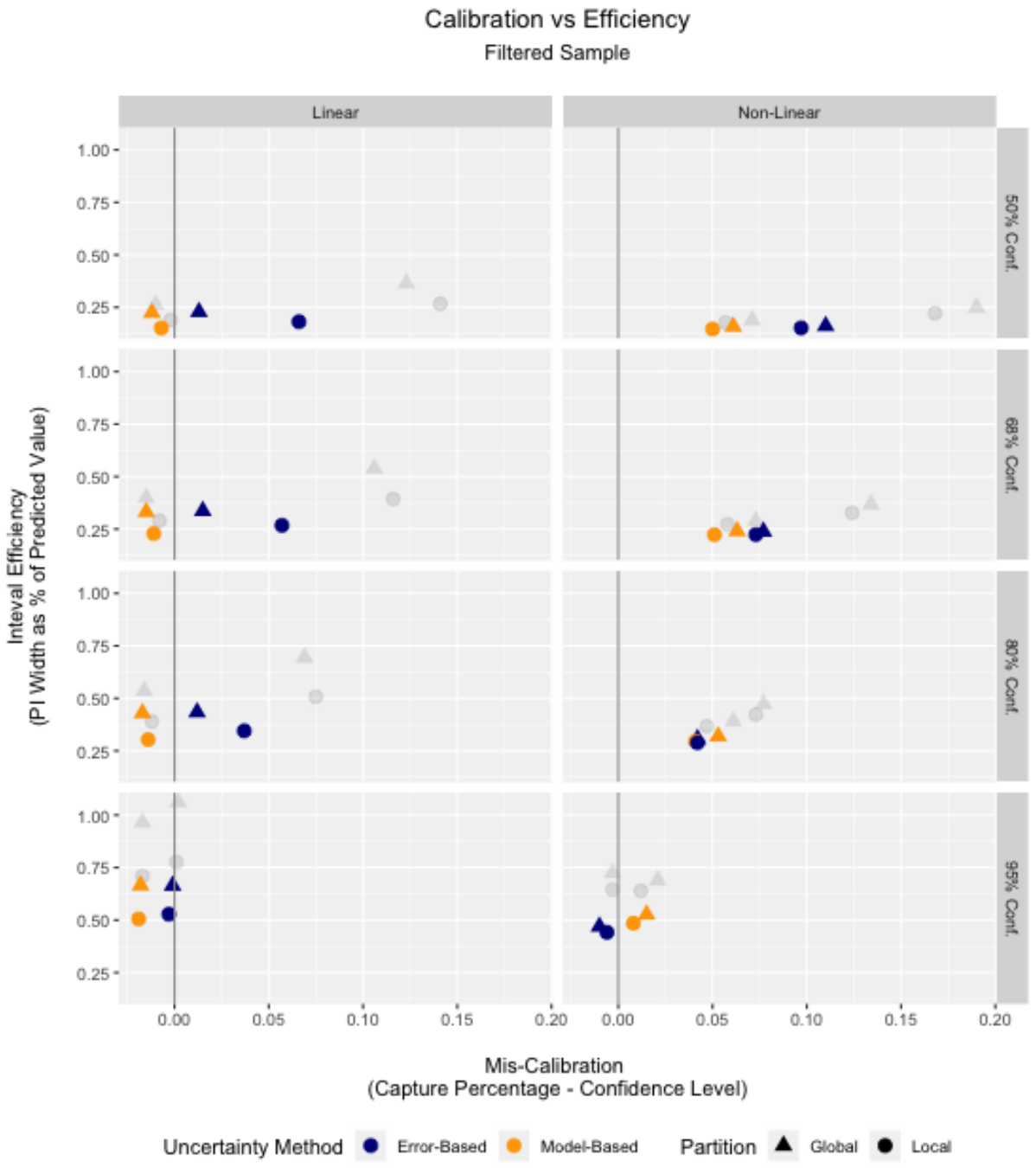

Figure 3. Calibration vs Efficiency, Filtered Data

from King County, WA, USA, we then empirically test the performance of two opposing methods for generating uncertainty estimates (prediction intervals) across two different model classes and two different data partitioning schemes.

Our results show that the model-based approaches to generating prediction intervals dominate the error-based method by being more reliably calibrated as well as more efficient for moderate confidence levels from 50% to 80%. The lower the confidence level, the greater the advantage of the model-based approach. At the very high end of the confidence scale (95% in our study), the error-based method shows superiority in terms of calibration reliability. These preferences holds across both model classes (linear and non-linear) and for both partition schemes. Given that the error-based approach is based off of the assumptions of normality and symmetry, the resulting mis-calibrations at the low end are not surprising. The errors from the full sample tests are highly leptokurtic – kurtosis measurements ranging from 10 to 15. As such, the FSD calculations are likely influenced by these outlying observations and, correspondingly, produce overly wide intervals at lower confidence levels. Once the confidence level get high enough to capture tail observations (around 95%) the leptokurtic nature of the distribution becomes an asset not a hindrance. In short, real estate prices – like many economic measures – and their related modeling errors are not normally distributed and computations relying on assumptions of normality may suffer.

A look at interval efficiency – the width of the prediction intervals relative to the predicted point estimate – tell a very similar story. Model-based approaches dominate at lower confidence levels, with a reversal at the high end of the confidence range scale. The model-based intervals range from about 30% more efficient at the lower confidence intervals up to 5% as we approach 80% confidence. Here too, the violation of the normality assumption is likely playing a key role.

An additional factor shaping interval efficiency is the fact that model-based intervals more heavily leverage existing data observations (sales prices, in this case) when setting interval limits. This puts natural bounds on extreme interval values. By anchoring to actual observations and by allowing for asymmetry in the interval, the model-based approaches can provide more efficiency, especially toward the low end of the range. Conversely, error-based approaches leverage a multiplier of the point predictions, which can lead to prediction interval values outside of actual observed sale prices.

Our sensitivity analysis using a filtered data set with fewer extreme values confirms the general trends from the full sample analysis, with a few minor differences. Most importantly, a more normally distributed set of model errors – kurtosis figures ranging from 3 to 7 – does offer considerable improvements to the error-based model as opposed the model-based approach. This finding supports our concern about normality assumptions in the error-based approach. We specifically filtered the sensitivity set to a greater extent than a practitioner or AVM provider would likely do solely to test our assumptions and provide generalizations. Though the accuracy, calibration and efficiency figures are superior in the filtered data, we are not suggesting that such filter is advised.

These findings suggest that the confidence level desired should be considered when choosing an approach to developing prediction intervals. For moderate levels of confidence (less than 95%), our findings contradict some advice given in the various industry guidelines, namely that uncertainty (often expressed as FSDs) should be derived directly from the known distribution of errors from the model. For moderate levels of confidence we find that model-derived measures of uncertainty offer more closely calibrated prediction intervals. Users and modelers interested in very high confidence levels – around 95% – either approach appears valid. However, we offer caution that prediction intervals at those levels of confidence can be very high, ranging from 75% to over 100% of the predicted value. Intervals with that width may not be very useful in practical applications. Finally, we offer the caveat that we have only tested two approaches in this paper, we leave future improvement and expansion to follow-up work.

Our contribution is three-fold: 1) Consolidation of the literature from diverse sources such academic research, professional standards and industry white papers in order to create a standard set of terms with which to discuss uncertainty; 2) Adoption and application of a framework for evaluating the quality of uncertainty measures (prediction intervals) and 3) Execution of empirical tests of the most commonly discussed approaches for generating prediction intervals.

Reproducibility

The results shown in this paper are fully reproducible. The King County data are hosted in an R package at www.github.com/ andykrause/kingCoData. All code used above, again compiled in an R package, can be found at www.github.com/ andykrause/avmUncertainty. Both the data and code links have instructions for installation, use and full reproduction.

References

Acciani, C., Fuccilli, V., and Sardaro, R. (2011) Data mining in real estate appraisal: A model tree and multivariate adaptive regression spline approach. *AESTIMUM*, 58, 27-45

Antipov, E. and Pokryshevskaya, E. (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(2), 1772-1778.

Appraisal Institute. (2018) *Uniform Standards of Professional Appraisal Practice* (*USPAP*), 2018-2019.

Bao, H., and Wan, A. (2004) On the use of spline smoothing in estimating hedonic housing price models: Empirical evidence using Hong Kong data. Real Estate Economics, 32(3), 487-507.

Bao, H, and Wan A (2007) Improved Estimators of Hedonic Housing Price Model. Journal of Real Estate Research, 29(3) 267-304

Belotti, A. (2017) Reliable region predictions for automated valuation models. Annals of Mathematics and Artificial Intelligence. 81, 71-84.

Bidanset, P. and , Lombard, J. (2014) Evaluating spatial model accuracy in mass real estate appraisal: A comparison of geographically weighted regression and the spatial lag model. Cityscape: A Journal of POlicy Development and Rse

Bin, O. (2004) A prediction comparison of housing sales by parametric versus semi-parametric regressions. Journal of Housing Economics, 13(1), 68-84.

Bourassa, S., Cantoni, E., and Hoesli, M. (2007) Spatial dependence, housing submarkets and house price prediction. The Journal of Real Estate Finance and Economics. 35(2), 143-160. et al 2007

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(2), 139-159.

Carsberg Report (2002) *Property Valuation*, Royal Institute of Chartered Sureveyors, London, UK

Case, B, Clapp, J., Dubin, R., and Rodriquez, M. (2004) Modeling spatial and temporal house price patterns: A comparison of four models. The Journal of Real Estate Finance and Economics. 29(2), 167-191.

Chen, J., Ong, C., Zheng, L., and Hsu, S. (2017) Forecasting spatial dynamics of the housing market using Support Vector Machines. International Journal of Strategic Property Management. 21(3), 273-283.

Cheung, S (2017) Localized model for residential property valuation. International Real Estate Review. 20(2), 221-250.

Chico-Olmo, J (2007) Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging. Journal of Real Estate Research. 29(1), 91-114.

Clapp, J., Kim, H., and Gelfand, A. (2002) Predicting spatial patterns of House Prices using LPR and Bayesian Smoothing. Real Estate Economics, 30(4), 505-532.

Clapp, J (2003) A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation. Journal of Real Estate Finance and Economics, 27(3) pp 303-320.

Clear Capital (2020) The modern lender's guide to the world of AVMs. Available at: https://www.clearcapital.com/resources/ebooks/ebook-the-modern-lenders-guide-to-the-world-of-avms/

Collateral Assessment Technology Committee (CATC) (2009) Best Practices in Automated Valuation Model (AVM) Validation.

Connected Analytics (2015) Best Practice Validation and Comparison for Automated Valuation Models (AVMs). Available at: https://www.corelogic.com.au/sites/default/files/2018-03/20151028-CL-RP_AVM.pdf

Corelogic (2011) Automated Valuation Testing. Available at: https://www.corelogic.com/downloadable-docs/automated-valuation-model-testing.pdf

Corelogic (2017) Forecast standard deviation and AVM confidence scores. Available at: https://www.corelogic.com/downloadable-docs/fsd-and-avm-confidence.pdf

Cumming, G. and Maillardet, R. (2006) Confidence intervals and replication: Where will the next mean fall? Psychological Methods, 11(3), 21-227.

Davidson, A. and Hinkley, D. (1997) Bootstrap Methods and Their Application, Cambridge University Press.

Demetriou, D (2017) A spatially based artificial neural network mass valuation model for land consolidation. Environment and Planning B: Urban Analytics and City Science. 44(5) 864-883.

Ecker, M., Isakson, H., and Kennedy, L. (2019) An exposition of AVM Performance Metrics. Working Paper. Available at: http://www.math.uni.edu/ ecker/research

European AVM Alliance (EAA) (2019) European Standards for Statistical Valuation for Residential Properties.

Feng, Y. and Jones, K (2015) Comparing multilevel modelling and artificial neural networks in house price prediction. IEEE International Conference on Spatial Data Mining and Geographical Knowledge. 108-114.

Fik, T, Ling, D. and Mulligan, G. (2003). *Real Estate Economics, 31(4), 623-646.*

Freddie Mac (2020) Freddie Mac Website, available at: http://www.freddiemac.com/hve/fsd.html

French, N. and Gabrielli, L. (2004) The uncertainty of valuation. Journal of Property Investment and Finance. 22(6) 484-500

French, N. and Gabrielli, L. (2005) Discounted Cash Flow: Accounting for Uncertainty. Journal of Property Investment and Finance. 23(1), 76-89

French, N. and Gabrielli, L. (2006) Uncertainty and Feasibility Studies: An Italian Case Study. Journal of Property Investment and Finance. 24(1), 49-67

Gao, G., Bao, Z., Cao, J., Qin, A., Sellis, T., and Wu, Z. (2019) Location-centered house price prediction: A multi-task learning approach. arXiv:1901.01774v1

GeoPhy (2019) Inside the Geophy AVM: The Evolution of Commercial Real Estate (CRE) Valuations. Version 1.2.6, February 7, 2019. Available at www.geophy.com.

Ghahramani, Z. (2015) Probabilistic machine learning and artificial intelligence. Nature 521:452{459. https://www.nature.com/articles/nature14541

Glennon, D., Kiefer, H, and Maycock, T. (2018). Measurement error in residential property valuation: An application of forecast combination. Journal of Housing Economics, 41, 1-29.

Gilley, O and Pace, R (1990) A Hybrid Cost and Market-Based Estimator for Appraisal. Journal of Real Estate Research. 5(1) pp 75-88.

Gonzalez, M. and Formoso, C (2006) Mass appraisal with genetic fuzzy rule-based systems. Property Management. 24(1), 20-30.

Gordon, D. (2005) Metric Matter, Freddie Mac Blog. Available at: http://www.freddiemac.com/hve/pdf/ dougwhitepaper_metricsmatter.pdf

Hannonen, M (2008) Predicting urban land prices: A comparison of four approaches. International Journal of Strategic Property Management. 12(4) 217-236.

HouseCanary (2018) HouseCanary Valuation Whitepaper, July 2018, Available at www.housecanary.com

International Association of Assessing Officials (IAAO). (2017). Standard on Mass Appraisal of Real Property. Kansas City, MO. https://www.iaao.org/media /standards/StandardOnMassAppraisal.pdf

International Association of Assessing Officials (IAAO). (2018). Standard on Automated Valuation Models (AVMs). Kansas City, MO. https://www.iaao.org/media /standards/AVM_STANDARD_2018.pdf

International Organization for Standardization (ISO) (2008) Uncertainty of Measurement - Part 3: Guide to the expression of uncertainty measurement), 98-3, 1-120

International Valuation Standards Council (2020) International Valuation Standards

International Valuation Standards Council (2013) Valuation Uncertainty. https://www.ivsc.org/files/file/download/id/296

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning, Springer, New York.

Janssen, C., Soderberg, B., and Zhou, J. (2001). Robust estimation of hedonic models of price and income for investment property. Journal of Property Investment and Finance. 19(4) 342-360.

Kang, H. and Reichert, A. (1987) An evaluation of alternative estimation techniques and functional forms in developing statistical appraisal models. Journal of Real Estate Research, 2(1), 1-29.

Kang, H. and Reichert, A. (1991) An emprical analysis of hedonic regression and grid-adjustment techniques in real estate appraisal. Real Estate Economics, 19(1), 70-91.

Knight, J., Hill, R, and Sirmans, C. (1992) Biased Prediction of Housing Values. Journal of the American Real Estate and Urban Economics Association. 20(3), 427-456.

Krause, A. (2020) hpiR: An R package for House Price Indexes. Version 0.3.1. Accessible at https://cran.r-project.org/web/packages/hpiR/index.html

Kucharska-Stasiak, E. (2013) Uncertainty of Property Valuation as a Subject of Academic Research. Real Estate Management and Valuation, 21(4) 17-25.

Leathart, T and Polaczuk, M (2020) Temporal Probability Calibration. arXiv: 2002.02644

Lozano-Garcia, N. and Anselin, L (2012) Is the Price Right? Assessing estimates of cadastral values for Bogota, Columbia. Regional Science Policy and Practice, 4(4), 495-508.

Lui, X. (2013) Spatial and Temporal Dependence in House Price Prediction. The Journal of Real Estate Finance and Economics. 47(2) 341-369.

Lipscomb, C. (2017) The next generation of AVMs. Fair and Equitable, March, 29-32.

Mallinson Report (1994), Commercial Property Valuations, Royal Institution of Chartered Surveyors.

Mallinson, M. and French, N. (2000) Uncertainty in property valuation–The nature and relevance of uncertainty and how it might be measured and reported. Journal of Property Investment and Finance. Vol. 18 No. 1, pp. 13-32. https://doi.org/10.1108/14635780010316636

Mayer, M., Bourassa, S., Hoesli, M, and Scognamiglio, D. (2019) Estimation and Updating Methods for Hedonic Valuation. Journal of European Real Estate Research

Meinhausen, N. (2006) Quantile Regression Forests, Journal of Machine Learning Research, 7, 983-999

Meszek, W. (2013) Property Valuation under Uncertainty: Simulation vs Strategic Model. International Journal of Strategic Property Management, 17(1), 79-92.

Mimis, A., Rovolis, A., and Stamou, M. (2013) Property valuation with artificial neural network: The case of Athens. Journal of Property Research, 30(2), 128-143.

Mortgage Bankers Association. (2019). The State of Automated Valuation Models in the Age of Big Data. January. pp 1-31. Available at http://www.mba.org/documents/ MBA _Real_Estate_Appraisals_(0).pdf

Nguyen, N. and Cripps, A. (2001) Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. Journal of Real Estate Research, 22(3), 313-

Pace, R. and Gilley, O. (1990) A hybrid cost and marked-based estimator for appraisal, Journal of Real Estate Research, 5(1), 75-88.

Pace, R., Barry, R., and Gilley, O. (2000) A method for spatial temporal forecasting with an application to real estate. International Journal of Forecasting, 16, 229-246.

Paez, A., Long, F., Farber, S. (2008) Moving window approaches for hedonic price estimation: An empirical comparison of modelling techniques. Urban Studies. 45, 1565-1582.

Papadopoulos, H., Vovk, V., and Gammerman, A. (2011) Regression Conformal Prediction with Nearest Neighbors. Journal of Artificial Intelligence Research, 40, 815-840.

Pavlov, A. (2000) Space-varying regression coefficients: A Semi-parametric approach applied to real estate markets. Real Estate Economics, 28(2), 249-283

Peterson and Flanagan (2009) Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research, 31(2), 147-164

Reichert and Kang (1987) An evaluation of alternative estimation techniques and functional forms in developing statistical appraisal models. Journal of Real Estate Research, 2(1), 1-29

Royal Institute of Chartered Surveyors (RICS) (2017) RICS Valuation { Global Standards [Red Book].

Scher, S. and Messori, G. (2018) Predicting weather forecast uncertainty with machine learning. Quarterly Journal of the Royal Meteorological Society, 144(717), pp 2830-2841. https://doi.org/10.1002/qj.3410

Shafer, G and Vovk, V. (2008) A Tutorial on Conformal Prediction. Journal of Machine Learning Research, 9, 371-421.

Shi, D., Guan, J., Zurada, J., Levitan (2015) An innovative clustering approach to market segmentation for improved price prediction. Journal of International Technology and Information Management , 24(1).

Stein, R. (2007) Benchmarking default prediction models: pitfalls and remedies in model validation. Journal of Risk Model Validation, 1(1), 77-113.

Thibodeau, T. (2003) Marking Single-Family Properties to market. Real Estate Economics. 31(1) 1-22

United States Department of Treasury (2019) Real estate appraisals. 12 CFR Part 34. Docket No. OCC-2019-0038. RIN 1557-AE57. Available at: https://www.federalregister.gov/documents/2019/ 10/08/2019-21376/real-estate-appraisals

Valente, J., Wu, S., Gelfand, A., and Sirmans, C. (2005) Apartment Rent Prediction Using Spatial Modeling. Journal of Real Estate Research, 27(1), 105-136.

Wang, D. and Li, V. (2019) Mass appraisal models of real estate in the 21st century: A systematic literature review. Sustainability, 11(24), 7006

Wood, G. (2005) Confidence and prediction intervals for generalised linear accident models. Accident Analysis & Prevention. 37(2), 267-273.

Wright MN, Ziegler A (2017). \ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." Journal of Statistical Software, 77(1), 1{17. doi: 10.18637/jss.v077.i01.

Yacim, J and Boshoff, D. (*2018*) *Impact of Artificial Neural Networks Training Algorithms on Accurate Prediction of Property Values. Journal of Real Estate Research.*

Yakubovskyi, V, Bychkov, O., Dimitrov, G., and Panayotova, G. (*2017*) *Combined neural network model for real estate market estimation. Proceedings of the Fourth International Conference on AI and Pattern Recognition, Lodz, Poland.*

Yang, B. and Cao, B. (*2018*) *Research on Ensemble Learning-Based Housing Price Prediction Model. Big Geospatial Data and Data Science. 1, 1-8*

Zurada, J, Levitan, A, and Guan, J. (*2011*) *A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. Journal of Real Estate Research, 33*(*3*), *pp 349-387.*

Appendix

Table 8. Summary of Empirical Papers Reviewed)

Authors	Year	Publication	Tests for Calibration
Kang and Reichert	1987	Journal of Real Estate Research	No
Pace and Gilley	1990	Journal of Real Estate Research	No
Kang and Reichert	1991	Real Estate Economics	No
Knight et al	1992	Journal of Real Estate Research	No
Pace et al	2000	International Journal of Forecasting	No
Pavlov	2000	Real Estate Economics	No
Nguyen and Cripps	2001	Journal of Real Estate Research	No
Janssen et al	2001	Journal of Property Investment and Finance	No
Clapp et al	2002	Real Estate Economics	No
Thibodeau	2003	Real Estate Economics	No
Clapp	2003	Journal of Real Estate Finance and Economics	No
Case et al	2004	Journal of Real Estate Finance and Economics	No
Bao and Wan	2004	Real Estate Economics	No
Bin	2004	Journal of Housing Economics	No
Valente et al	2005	Journal of Real Estate Research	No
Gonzalez and Mormoso	2006	Property Management	No
Bourassa et al	2007	Journal of Real Estate Finance and Economics	No
Bao and Wan	2007	Journal of Real Estate Research	No
Chica-Olmo	2007	Journal of Real Estate Research	No
Hannonen	2008	Int'l Journal of Strategic Property Management	No
Paez et al	2008	Urban Studies	No
Bourassa et al	2010	Journal of Real Estate Research	No
Peterson and Flanagan	2009	Journal of Real Estate Research	No
Acciani et al	2011	AESTIMUM	No
Zurada et al	2011	Journal of Real Estate Research	No
Antipov and Pokryshevska	2012	Expert Systems with Applications	No
Lozano-Garcia and Anselin	2012	Regional Science Policy and Practice	No
Mimis et al	2013	Journal of Property Research	No
Liu	2013	Journal of Real Estate Finance and Economics	No
Bidanset et al	2014	Cityscape	No
Shi et al	2015	Journal of Int'l Tech. and Information Management	No
Feng and Jones	2015	IEEE Conference on Spatial Data Mininge	No
Demetriou	2017	Environment and Planning B	No
Cheung	2017	International Real Estate Review	No
Yakubovskyi et al	2017	4th International Conference on AI and Pattern Recognition	No
Chen et al	2017	International Journal of Strategic Property Management	No
Glennon et al	2018	Journal of Housing Economics	No
Yang and Cao	2018	Big Geospatial Data and Data Science	No
Yacim and Boshoff	2018	Journal of Real Estate Research	No
Gao et al	2019	arXiv	No
Mayer et al	2019	Journal of European Real Estate Research	No