WORKING PAPER

RELIABLE PREDICTION INTERVALS FOR AUTOMATED RENTAL VALUATIONS

Maarten Van Besien

February 2026
2026/1136

GHENT UNIVERSITY

Department of Economics

Faculty of Economics and Business Administration
Campus Tweekerken, St.-Pietersplein 5, 9000 Ghent - BELGIUM

Reliable Prediction Intervals for Automated Rental Valuations

Maarten Van Besien [a]

[a] *Department of Economics, Ghent University*

Abstract

Automated valuation models (AVMs) are widely used for large-scale residential rent appraisal, yet standard models do not provide predictive uncertainty measures with guaranteed out-of-sample coverage at prespecified nominal levels, creating risks for institutional decision-making in valuation, risk management, and policy design. Using a transaction-level dataset covering the Flemish rental market in Belgium, we study AVM performance and uncertainty quantification in a large-scale, heterogeneous, and feature-poor setting, where only location, property type, energy performance, number of bedrooms, and rent prices are observed. We show that industry-standard point-prediction accuracy can be achieved by exploiting non-linear spatial structure using coarse geospatial units such as boroughs. For uncertainty quantification, we compare ensemble quantile regression and Inductive Conformal Prediction (ICP). While both improve empirical coverage, ICP is preferred as it guarantees finite-sample marginal coverage without distributional assumptions at substantially lower computational cost. Conditioning ICP calibration on bedroom count (Mondrian ICP) yields the largest efficiency gains, reducing 95% coverage prediction interval width by up to 5.3% relative to absolute residual split conformal prediction. Overall, our results demonstrate that valuation uncertainty can be materially reduced in large-scale, feature-poor housing data with minimal additional modeling complexity.

Keywords: Conformal Prediction, Automated Rent Valuation, Rental Uncertainty Quantification.

1. Introduction

Automated Valuation Models (AVMs) have long evolved from the traditional hedonic regression frameworks (Rosen, 1974) towards increasingly complex machine learning systems (e.g., Tay and Ho, 1992; Fan et. al., 2006; Selim, 2009; Antipov and Pokryshevskaya, 2012), driven by the growth of rich property datasets and advances in algorithms and computational power. Yet, despite these advances, evaluation methods remain centered on point prediction metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Average Percentage Error (MAPE). These metrics address accuracy but do not quantify the uncertainty surrounding each estimate, a critical omission in valuation and risk-sensitive domains.

In applied settings, rental values are often used as inputs in decision-making processes where the distribution of possible outcomes, rather than merely the expected value, drives risk assessment, regulatory compliance, and strategic behavior. Quantifying predictive uncertainty is therefore central to applying AVM information in at least four settings. First, public authorities rely on rental valuations for taxation and fiscal policy design. Poorly quantified uncertainty can result in systematic mispricing, legal disputes, or inequitable tax burdens, particularly when assessments are performed at scale. Second, rental estimates are often capitalized into asset values through yield-based valuation frameworks. In this context, uncertainty in rents directly propagates into uncertainty in asset prices. Third, real-estate professionals and platforms use rental valuations for market placement and pricing strategies. Prediction intervals around estimated rents inform negotiation margins, listing strategies, and time-on-market expectations. Finally, institutional and private investors increasingly rely on AVMs to guide portfolio allocation, stress testing, and investment strategy. Without reliable uncertainty estimates, downside risk may be systematically underestimated, affecting risk-weighted capital requirements. In all these applications, the ability to quantify risk hinges on the availability of well-calibrated prediction intervals. As AVMs are increasingly deployed in large-scale, possible feature-restricted settings, robust uncertainty quantification becomes not a secondary refinement, but a core requirement for responsible and effective rent appraisal.

Prediction intervals for complex AVM models such as CatBoost, LightGBM, and Random Forests are most commonly constructed using quantile regression (QR) (Koenker and Basset, 1978), through direct estimation of conditional quantiles. While QR-based intervals are attractive due to their flexibility and ease of implementation, they do not provide valid uncertainty quantification in the sense of guaranteed coverage. Firstly, quantile regression estimates conditional quantile functions rather than prediction intervals with explicit coverage guarantees. Even when conditional quantiles are consistently estimated, the resulting intervals attain nominal coverage only asymptotically and under strong assumptions, including correct model specification and sufficient data density throughout the feature space. In finite samples, particularly in high-dimensional or feature-restricted settings common in real-estate applications, these assumptions are rarely satisfied. Secondly, QR-based intervals are highly sensitive to model misspecification and to the regularization and aggregation mechanisms inherent in flexible learners. In tree-based ensemble models, quantile estimates may become unstable in regions with sparse data, exhibit quantile crossing, or be excessively smoothed by the learning algorithm. These effects are algorithmic in nature and persist even when overall predictive accuracy is high, rendering the resulting intervals difficult to interpret and unreliable at the local level. Consequently, while QR-based prediction intervals may appear reasonable, they cannot be interpreted as reliable measures of predictive uncertainty. This limitation motivates the use of calibration frameworks, such as conformal prediction, that explicitly target and guarantee coverage by construction.

Conformal prediction (CP) provides a modern framework for constructing prediction intervals with formal coverage guarantees, regardless of the underlying data distribution (Gammerman et al., 1998; Vovk et al., 2005; Shafer and Vovk, 2008). At its core, CP leverages any underlying point predictor, e.g., linear regression, random forests, gradient boosted decision

trees or neural networks, and constructs a prediction interval by measuring how nonconforming a new data point is with respect to the training data. This nonconformity is typically assessed using residual errors. Provided the data are exchangeable, CP guarantees that the true value will fall within the constructed interval at a user-specified nominal level (e.g., 90%) with exact finite-sample validity. This property is a key advantage over traditional prediction intervals that often rely on strong distributional assumptions. Another practical advantage of the CP-framework is its simple implementation as a wrapper to any pre-trained machine learning model. Additionally, the computational scalability to deliver any desired level of coverage without retraining the underlying model on both necessary quantiles as in QR, makes CP a highly suitable choice for production level settings.

Conformal prediction has been increasingly applied in real estate research to construct prediction intervals for property prices with guaranteed finite-sample coverage. Early work by Bellotti (2017) introduced conformal prediction for housing AVMs, demonstrating reliable region-based prediction intervals for UK property sale values. Subsequent studies have focused on improving interval efficiency through refinements of the nonconformity measure. For example, Lim and Bellotti (2021) apply normalized nonconformity scores to housing data from the Ames dataset, yielding narrower yet valid prediction intervals. More recent contributions extend conformal methods by incorporating quantile-based structure. Bastos and Paquette (2025) implement conformalized quantile regression (Romano et al., 2019) for house price prediction in the San Francisco Bay Area, showing that adaptive, covariate-dependent intervals can improve efficiency relative to standard split conformal approaches. Along similar lines, Hjort et al., (2024) evaluate a range of conformal prediction techniques, including split conformal prediction (Papadopoulos et al., 2002), conformalized quantile regression, and Mondrian conformal regression (Boström and Johansson, 2020), on Norwegian housing data, demonstrating that methods tailored to local market structure can substantially enhance interval efficiency and local validity.

We are, to our knowledge, the first to apply the conformal prediction (CP) framework to a large-scale, transaction-level rental dataset. We demonstrate that CP constructs prediction intervals that empirically attain the desired coverage levels using a simple Inductive Conformal Prediction (ICP) wrapper around standard rent AVMs. Additionally, we show that prediction intervals derived from quantile regression applied to a stacked ensemble of these AVMs can achieve comparable empirical coverage; however, at a substantially higher computational cost and without accompanying theoretical finite-sample coverage guarantees. The remainder of the paper is structured as follows: In section 2 we give a short overview QR-based prediction intervals, the theoretical framework underpinning CP and the framework for implementing ICP. Section 3 discusses the dataset construction as well as key variables. Section 4 discusses AVM models and their scoring criteria as well as CP-scoring criteria and the Non-Conformity Measures (NCM). In section 5 we discuss the empirical results by assessing the AVM point prediction accuracy and comparing the out of sample test coverage of standard prediction intervals with ICP-prediction intervals. Section 6 concludes.

2. Quantifying Uncertainty

In this section, we explain the use of Quantile Regression (Koenker and Basset, 1978) for prediction interval construction in ML-driven AVMs. We also provide a brief overview of conformal prediction and the commonly used method of split-conformal prediction (Papadopoulos et al., 2002) for large datasets and computationally intensive models. For excellent introductory material for conformal prediction and its applications we refer to Angelopoulos and Bates (2021).

2.1 Quantile Regression

To quantify uncertainty in nonlinear, ML-driven automated valuation models, this study employs quantile regression as a widely used benchmark for constructing prediction intervals (Taylor, 2000; Friedman, 2001; Meinshausen and Ridgeway, 2006; Takeuchi et al., 2006). Quantile regression provides a flexible framework for estimating conditional quantiles of the outcome distribution and is commonly applied in empirical settings where heteroskedasticity and non-Gaussian errors are present. The motivation for using quantile regression in this context is both practical and methodological. First, the predictive models considered in this study (Random Forest, CatBoost, LightGBM) do not natively provide well-calibrated measures of predictive uncertainty or Bayesian posterior intervals. Quantile regression offers a model-agnostic approach by directly targeting conditional quantiles of the target distribution, thereby enabling the construction of central prediction intervals around the conditional median without imposing parametric assumptions on the error distribution. Second, quantile regression serves as a natural comparison point for conformal prediction methods, as both aim to characterize uncertainty in predictions rather than parameter estimates. Formally, a (1-α) prediction interval (e.g., 90%) can be obtained by training two separate models targeting the lower and upper quantiles $\tau = \alpha/2$ and τ = 1-$\alpha/2$, respectively (e.g., τ = 0.05 and τ = 0.95 for a 90% interval). The models are trained minimizing the pinball loss (also known as the quantile loss). We denote random variables as (X, Y) and their realizations as (x, y). Pinball loss L_τ is defined as:

$$L_\tau(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}), & \text{if } y \geq \hat{y}, \\ (1 - \tau)(\hat{y} - y), & \text{if } y < \hat{y}. \end{cases} \quad (1)$$

This asymmetric loss penalizes under- and over-predictions differently, depending on the target quantile τ. By minimizing the expected pinball loss, QR estimates the conditional quantile function $Q_\tau(\mathrm{Y}|\mathrm{X} = x)$ via an estimator $\hat{Q}_\tau(x)$, defined as the value $\hat{y}$ such that a fraction τ of the conditional outcome distribution falls below it. Quantile regression can represent heteroskedasticity through covariate-dependent quantiles, allowing interval widths to vary with observable characteristics. However quantile-based prediction intervals are model-dependent: their empirical coverage depends on the accuracy and calibration of the estimated conditional quantiles. In finite samples and high-dimensional settings, estimated quantiles may be biased or miscalibrated, and quantile crossing may occur. As a result, unlike conformal prediction, quantile regression does not provide finite-sample, distribution-free guarantees on prediction interval coverage.

2.2 The Conformal Prediction Framework

Conformal prediction provides a general, model-agnostic framework for constructing prediction sets with finite-sample coverage guarantees under minimal assumptions. Assume we have a training dataset ($\mathcal{D}$) of n past observations $Z_1, \dots, Z_n$ (e.g., rental contracts), where each data point $Z_i = (X_i, Y_i)$ consists of features $X_i \in \mathcal{X}$ (e.g., property attributes such as location, number of rooms, energy efficiency) and a response $Y_i \in \mathcal{Y}$ (e.g., rent price). We assume that the joint distribution of $Z_1, \dots, Z_n$ is invariant under permutations, that is, the data are exchangeable. No specific distributional assumptions are imposed on X or Y, nor any particular form assumed for Y given X. Given a new covariate vector X_{n+1}, our task is to construct a prediction set $C(X_{n+1}) \subseteq \mathcal{Y}$ based on the training data ($\mathcal{D}$) and X_{n+1}, such that

$$\Pr\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha \tag{2}$$

holds for a user specific error level $\alpha \in (0,1)$. This coverage guarantee is marginal, in the sense that it holds on average over the joint distribution of the data and the test point, rather than conditionally on specific covariate values. Under exchangeability, conformal prediction achieves this coverage exactly or up to a discretization error of order $1/(n+1)$, without relying on parametric assumptions (Vovk et al., 2005).

The core idea of conformal prediction is to assess how nonconforming (atypical) a new observation would be relative to the previously observed data ($\mathcal{D}$). This is implemented via a user-chosen nonconformity measure, defined as a function $A(Z_i, \mathcal{D})$ that assigns a real-valued score to a data point Z_i given a reference dataset ($\mathcal{D}$). In a regression setting, a very common choice is to use the absolute residual based on a fitted prediction model ($\hat{f}_{\mathcal{D}}$), trained on the reference dataset. We define the nonconformity score in this setting for each test point (i) as:

$$r_i := A(Z_i, \mathcal{D}) = \left| y_i - \hat{f}_{\mathcal{D}}(x_i) \right|. \tag{3}$$

In full (transductive) conformal prediction (Gammerman et al., 1998), for every test covariate vector (X_{n+1}), we incorporate new hypothetical examples to decide whether candidate labels y for X_{n+1} are consistent with our past data. To determine if a given y should be in the prediction set $C(X_{n+1})$, we examine (X_{n+1}, y) as an extra data point and evaluate how it "conforms" with the training sample. In full conformal prediction, the prediction model $\hat{f}_{\mathcal{D}}$ is refit on the augmented dataset that includes the candidate pair (X_{n+1}, y), and nonconformity scores (see Equation 3) for all $n+1$ points are computed. We then compute the rank (empirical p-value) of the test point's nonconformity score among all $n+1$ scores. To ensure finite-sample validity when nonconformity scores are tied, we adopt a conservative tie-breaking rule by using a weak inequality ($\geq$) instead of a strict one to ensure valid coverage in the presence of ties as:

$$p(y) := \frac{\#\{i = 1, \dots, n+1 : r_i \geq r_{n+1}(y)\}}{n+1}. \tag{4}$$

This p-value measures the fraction of data points (including the test point itself) whose nonconformity score is at least as large as that of the test point. A high $p(y)$ means the candidate y is not particularly nonconform (many points have equal or worse fit), whereas a low $p(y)$ means y is atypical relative to the training sample. The CP-set for X_{n+1} at error rate α is then defined as all candidate labels whose p-value exceeds α:

$$C(X_{n+1}) \coloneqq \{y \in \mathcal{Y}: p(y) > \alpha\}. \tag{5}$$

Equivalently, we include y in the prediction set if the test point's nonconformity score is not among the top α fraction of largest nonconformity scores. Given exchangeability, for any α holds

$$\Pr\{Y_{n+1} \notin C(X_{n+1})\} \leq \alpha \tag{6}$$

implying $\Pr\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$, which guarantees marginal coverage. The procedure described here is transductive in the sense that each test point is treated symmetrically with the training data and evaluated individually. Conceptually, this requires retraining the prediction model and recomputing nonconformity scores for any possible candidate y for every individual test point X_{n+1}.

For large datasets, continuous outcome spaces, or computationally intensive machine learning models, full conformal prediction quickly becomes computationally impractical. To address these computational limitations, inductive conformal prediction (ICP), also known as split conformal prediction, was introduced by Papadopoulos et al., (2002). The key idea is to split the available training data in two disjoint subsets: a proper training set used to fit the underlying predictive model ($\hat{f}$) once, and a separate calibration set used to compute the non-conformity scores (e.g., absolute residuals for the hypothetical y values). Due to the construction of a separate calibration set, to match the desired error rate α, a correction term is now required on the rank of non-conformity scores (see Equation 4) of the hypothetical y with respect to the training set $\mathcal{D}$. We outline the inductive conformal procedure used in the rest of this paper in steps below:

I. Split the training data into a proper training and a calibration set (size m), ensuring the observations in the calibration set are exchangeable with the training set.

II. Fit any point prediction model $\hat{f}$ on the proper training set.

III. Apply $\hat{f}$ to each observation (j) in the calibration set to obtain predictions $\hat{y}_j = \hat{f}(x_j)$. Compute the nonconformity scores (e.g., absolute residuals $r_j = |y_j - \hat{y}_j|$) on the calibration set. This yields a collection of m nonconformity scores $r_j: j = 1, \ldots, m$.

IV. Rank the obtained nonconformity scores in ascending order. Let $k = \lceil (m+1)(1-\alpha) \rceil$ and set the conformal threshold at Q, corresponding to the empirical k/m quantile of the

calibration residuals with finite-sample rank correction included. Intuitively, Q is chosen such that approximately $(1-\alpha)\times 100\%$ of past residuals are less than or equal to Q.

V. Given a new covariate vector from the test set X_{n+1}, compute $\hat{y}_{n+1} = \hat{f}(X_{n+1})$, and form the inductive conformal prediction interval as:

$$C(X_{n+1}) \coloneqq \left[\hat{f}(X_{n+1}) - Q,\ \hat{f}(X_{n+1}) + Q\right]. \tag{7}$$

Under the assumption of exchangeability, the following finite-sample guarantee holds:

$$\Pr\left\{Y_{n+1} \in \left[\hat{f}(X_{n+1}) - Q,\ \hat{f}(X_{n+1}) + Q\right]\right\} \geq 1-\alpha, \tag{8}$$

ensuring the desired finite-sample marginal coverage. The inductive conformal prediction procedure outlined is computationally efficient and can be applied to any point prediction model without modification of the underlying learning algorithm. The main limitation is that the calibration threshold is applied uniformly across all test points and does not result in adaptive prediction interval width across the feature space. However, due to its computational efficiency, ICP has become the standard CP procedure, especially for large-scale problems or for any computationally intensive model (e.g., neural networks).

Conformal methods span a spectrum of computational and statistical trade-offs. Full conformal prediction treats each test point transductively and can be computationally demanding, whereas split (inductive) conformal prediction is computationally efficient but may yield less efficient (wider) prediction intervals, as only a subset of the data is used for calibration. Between these approaches, several variants aim to reuse data more effectively while retaining distribution-free marginal coverage under exchangeability. Examples include conformalized quantile regression (Romano et al., 2019) and the jackknife+ (Barber et al., 2021b), which leverage additional structure or resampling to improve interval efficiency relative to basic split conformal prediction. More generally, when a conformal scheme comes with a formal guarantee, the primary impact of the chosen scheme is on efficiency (e.g., average interval width) rather than on marginal validity, provided the data satisfy exchangeability and the implementation avoids leakage. In the remainder of this paper, we implement split (inductive) conformal prediction, as the scale of our dataset and the computational cost of our models render full transductive conformal prediction impractical.

3. Data

Our empirical analysis relies on a dataset covering the Flanders region of Belgium where geo-data, land and building registry information is publicly accessible. The final dataset comprises 52,222 signed rental leases for unique residential properties, provided by the Flemish Realtor Lobby association (CIB) and a large insurer (Korfine). The first subsection describes the data augmentation, cleaning and filtration procedure applied to obtain a final sample of unique rental properties with sufficient data quality for the construction of automated valuation

models. The second subsection presents summary statistics and graphical evidence for key categorical variables.

3.1 Data Augmentation, Cleaning and Filtration

Our initial dataset consists of 880,000+ signed rental leases for the entire country of Belgium spanning the period January 2014 to March 2025. The data contains standard lease information including, but not limited to signature/commencement date, monthly rental charges, service fees as well as the type of residential unit (e.g., apartment, (terraced/ (semi)-detached) houses, villas, student housing, etc.). We exclude commercial and social housing (removing 98,947 entries) and match the rental contracts to the Flemish land registry via the contract postal code, street name, house and letterbox numbers (retaining 536,677 entries). Boudt et al., (2025) outline the data cleaning steps and address the matching procedure to identify unique rental properties in depth. We then restrict the rental leases to the most recent occurrence of each rental unit via this unique identifier (postal code, street name, house and letterbox number) which results in a total of 191,847 rental leases for unique properties.

In a first data augmentation step, we extend the signed leases with raw estimates of the building ground surface and lot surface from the official building/land-registry. We extract the registry building centroid as the geo-coordinate for each listing. In a second step, we augment our dataset by parsing the unit-description section of all physical (written) rental contracts. Where available, we extract the number of bedrooms, the energy efficiency of the building (EPC-score), the presence of balconies, terraces, courtyards, gardens, parking spaces and indication of ground or top floor located apartments. In a third step we include publicly available spatial data regarding mobility scores, a summary measure for locational accessibility, and we also calculate distance to the Belgian shoreline. Next, we apply 3 simple data cleaning steps sequentially, retaining as much data as possible. In a first step we remove 15,434 observations explicitly identified as non-primary residence units in the dataset (e.g., student housing, studio and short-term rentals). These units exhibit rental dynamics that differ structurally from the residential market of interest and fall outside the scope of this study. In a second step, we apply DBSCAN (Density-Based Spatial Clustering of Applications with Noise; Ester et al., 1996) to a set of pairwise combinations of log EPC, log CPI deflated rent (including service fees), log ground surface area, and log lot surface to identify and remove outliers. The procedure is run separately for apartment units and the other residential types (houses and villas), due to apartments having smaller surfaces, absence of gardens, etc. Including the outcome variable (log CPI-deflated rents) is common practice when working with real estate data (Baur et al., 2023). In total we exclude 4,379 rental units (2.5%), aiming to remove only datapoints that exhibit abnormal feature combinations. Similar studies (e.g., Lenaers and De Moor, 2023) often remove a substantial share of observations during the data-cleaning stage (approximately 16%) or mention data cleaning without reporting the number of observations excluded from the initial dataset (e.g., Lenaers et al., 2024; Lorenz et al., 2023; Zhou et al., 2019; Baur et al., 2023). In a final filtering step, we retain only observations for which both the number of bedrooms and the energy performance of the rental unit are observed.

In total, 52,222 unique rental units are retained. The absence of several key real estate characteristics (e.g., habitable surface, year of construction, and build quality) motivates the requirement that at least these two features be observed for each residential unit. As a final remark, we note that the dataset does not include observations paired with registered or appraised sale values, nor does it contain property tax information, and is restricted to the most recent observation for each individual rental unit. Not enforcing these restrictions would make estimating rents considerably less challenging and the obtained AVMs not applicable for valuing genuinely out-of-sample observations**,** that is, newly observed properties for which only information available at valuation time is known, as in online or operational valuation settings. Including past observations for the same property in the training dataset would make prediction for such out-of-sample prices trivial.

3.2 Summary Statistics

Table 1 reports summary statistics for the numerical variables in our dataset. Reported CPI deflated rental prices (base year 2013), improving temporal comparability, are heavily skewed with the ratio between the minimum and maximum exceeding a factor 7. The majority of the rental stock in our dataset meets the mandatory threshold for energetic renovations placed at a score of 400 (Flemish Government, 2023). The median mobility score (mobiscore) indicates that most rental units are situated in highly connected urban areas close to public transportation. The lot surface indicates highly skewed data, explained by the presence of both apartments and houses/villas in the dataset. The minimum observed lot surface of 4 m² indicates the presence of measurement errors in the official land registry and illustrates that the preceding DBSCAN-based outlier removal was applied conservatively rather than being aggressively tuned to eliminate all implausible observations. Distance to the Belgian coastline is included as a locational covariate capturing the sharp rental premium associated with walkable proximity to the coast. Given the highly localized nature of this effect, distance is measured as straight-line distance and capped at 1 km, corresponding approximately to a 20-minute walking distance, beyond which the marginal coastal premium dissipates. The divergence between mean and median distance indicates a concentration of units very close to the beach. Figure 1 visualizes these findings, as well as the observed CPI-deflated rental prices exhibiting substantial spatial dispersion across the region and the presence of densely clustered city centers and sparse country sides, consistent with the Belgian pattern of ribbon development.

Figure 2 displays the distribution of key categorical variables in the dataset. The upper panel reports unit-specific characteristics, while the lower panel reports the spatial classification of properties at progressively finer geographic levels. The distribution of the number of bedrooms indicates 2 bedrooms-units to be the most prevalent, units with more than 5 reported bedrooms are extremely rare and omitted from this figure. The building topology is dominated by apartments (Ap, 70.7%), explaining the majority presence of 2-bedroom rental units, followed by Terraced, Semi-Detached and Detached-Houses. The share of Villa's (0.7%) is negligible, with the distinction between classification of a unit as a Detached-House and Villa at the discretion of the realtor. Terraces, gardens, garages and car-spots are common occurrences while balconies (2,084 cases) and courtyards (1,614 cases) are relatively rare occurrences. On

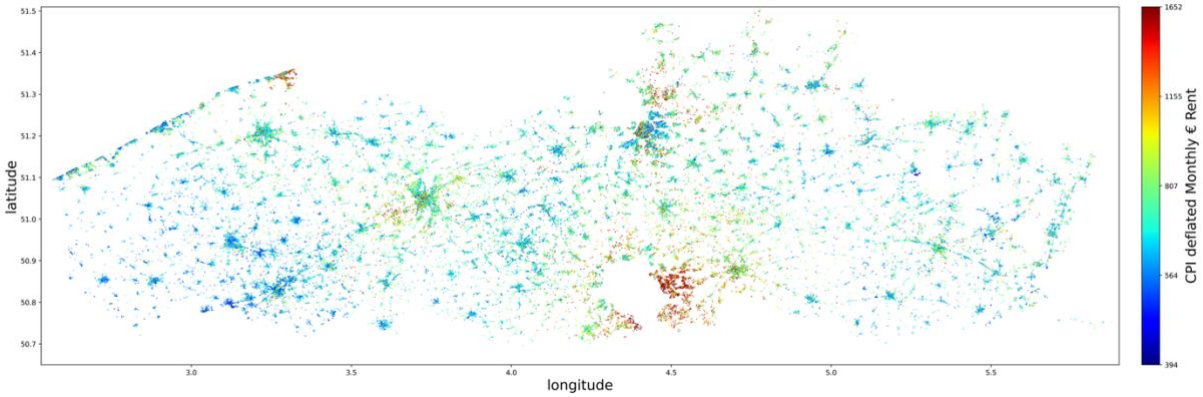
the lower panel, we showcase the number of rental units aggregated per geographical agglomeration in increasing detail, from County to Postal code to Borough. Our dataset covers 299 Counties, 507 Postal codes and 2,148 Boroughs. Given the size of our dataset, 52,222 unique rental units, and the geographical clustering indicated per Figure 1, a large share of Boroughs (250+) contains a single rental unit, limiting the performance of highly localized models.

Table 1. Summary Statistics for Rental Units Containing both Bedrooms and EPC-scores.

variable	count	min	p10	median	mean	p90	max	std
rent	52,222	309	504	643	682	901	2,437	188
rent + service fee	52,222	344	525	673	711	935	2,478	192
energy score	52,222	0	84	193	231	435	1,137	151
mobiscore	52,101	3.4	6.3	8.3	8.0	9.3	9.7	1.1
lot surface	52,080	4	123	500	1,132	2460	108,413	2,317
distance beach	2,540	7	60	216	310	718	999	256

Table 1 shows key variables for rent price prediction. Rents and service fees are displayed in a CPI-deflated format to facilitate comparability between prices spanning an entire decade. Surfaces are measures in m^2 and distances in meters. The mobiscore ranges theoretically between 0 and 10 and the energy score (EPC) measures the energy efficiency of a building (the lower the number the less energy is used for heating).

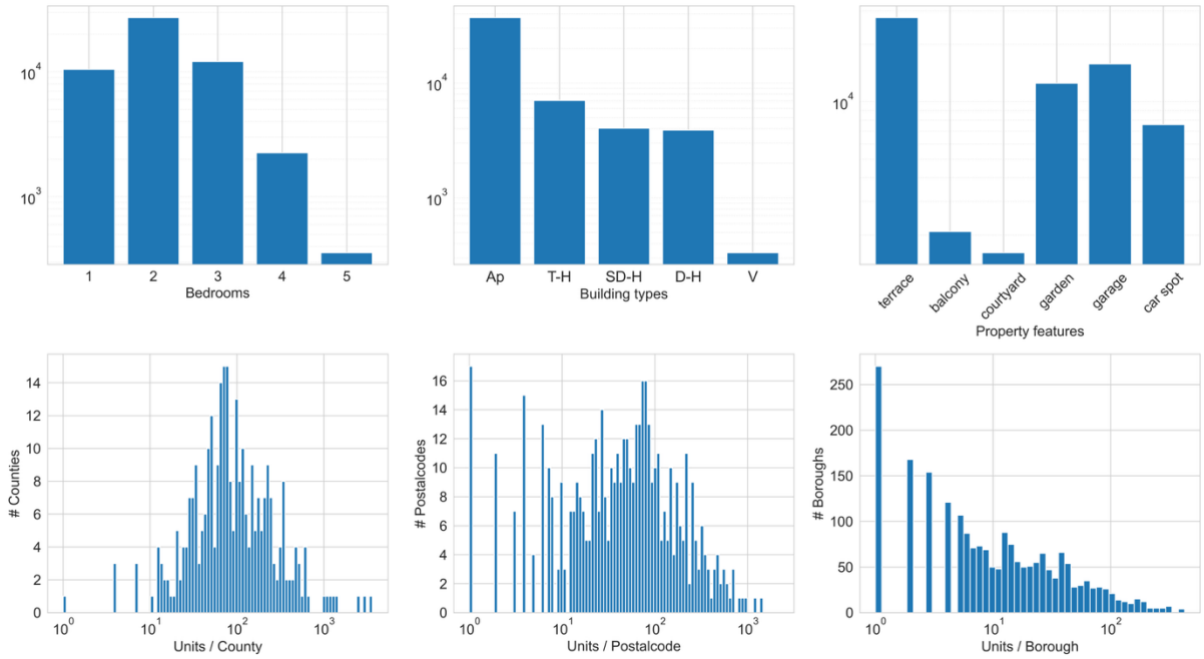
Figure 1. Geographical Concentration of CPI-Deflated Rents Across Flanders.



Overview of the geographically matched 52,222 datapoints spanning from January 2014 until March 2025. Prices are denoted in CPI deflated monthly rent prices (base year 2013). The map indicates the spread of buildings across the entire region, the presence of ribbon development (a particular Flemish feature), and the clear visibility of the Belgian shoreline (top left).

In summary, our dataset comprises a substantial number of rental units (52,222), spanning a large region (13,626 km²) and a long period (January 2014–March 2025). The absence of several commonly used residential real estate characteristics (e.g., building quality, age, and habitable surface area) reflects realistic information constraints in operational rental valuation settings and motivates our restriction to units for which both the number of bedrooms and the energy performance score (EPC) are observed. These data characteristics make rent prediction non-trivial while ensuring that the resulting AVMs remain applicable for large-scale, real-world implementation.

Figure 2. Categorical Variables for Rental Units Containing both Bedrooms and EPC-scores.



Key categorical variables for rent price prediction are displayed. The top row presents the number of bedrooms, type of rental units and the presence of unit-specific features. The view for the number of bedrooms is restricted to 5. Real estate type category includes Apartment, (T)erraced House, (Semi-Detached) House, (D)etached House and Villa. Property features display the presence of each given feature for rental units. The bottom row presents rental units aggregated at different geographical levels, from left to right the level of spatial detail increases.

4. Methodology

This section describes the methodology underlying our automated valuation models, including both point prediction models and the metrics used to evaluate the performance of prediction intervals. Subsection 4.1 discusses the use of log CPI-deflated rents as the target variable for training the point prediction and AVM models, as well as the loss functions and evaluation metrics employed. Subsection 4.2 introduces the metrics used to assess empirical coverage. Subsection 4.3 concludes with a brief discussion of the nonconformity measures (NCMs) used within the conformal prediction framework.

4.1 Real Estate AVMs, Training and Performance Metrics

Empirical research on rental price prediction typically models either nominal rents (e.g., Lorenz et al., 2023; Lenaers & De Moor, 2023; Zhou et al., 2019; Clark and Lomax, 2018) or rents normalized by floor area, such as rent per square meter or square foot (e.g., Baur et al., 2023). In contrast, we adopt logarithmically transformed **(**Yoshida et al., 2024**),** CPI-deflated rents as the target variable for all point prediction models.

This choice is motivated by several practical and methodological considerations. First, deflating rents using the Consumer Price Index (CPI) removes a substantial component of nominal price drift over time, allowing the models to focus on real rental price dynamics rather than inflationary variation. Given the long observation window (January 2014 – March 2025), this transformation improves temporal comparability of rental prices without requiring an explicit time-series specification. As the primary objective of this paper is to evaluate the use

of Inductive Conformal Prediction (ICP) for uncertainty quantification, rather than to model temporal dynamics per se, CPI deflation provides a pragmatic normalization of the target variable. Second, while conformal prediction methods rely on an assumption of exchangeability between calibration and test observations, this assumption is unlikely to hold strictly in housing market data due to temporal evolution, market segmentation, and changes in the distribution of covariates. CPI deflation does not restore exchangeability, nor does it address potential covariate or distributional shift in a formal sense. However, by mitigating systematic nominal trends in rents, it reduces one important source of non-stationarity in the marginal distribution of the target variable. Addressing distributional shift explicitly through weighted or adaptive conformal methods (Xu and Xie, 2023) is beyond the scope of this paper, which focuses on the baseline ICP framework. Third, the distribution of rental prices in our dataset exhibits pronounced right skewness and is well approximated by a log-normal distribution, consistent with findings in the real estate literature. Logarithmic transformation therefore stabilizes variance, reduces the influence of extreme observations, and improves the performance of regression-based prediction models. Finally, the absence of key structural housing characteristics, most notably habitable surface area, precludes reliable normalization of rents on a per-square-meter basis, reinforcing the decision to model (log-transformed) rent levels directly.

To generate point predictions, we employ three widely used automated valuation models that have demonstrated strong empirical performance in real estate applications: Random Forests (Breiman, 2001), Light Gradient Boosting Machines (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018). These models are well suited to housing data due to their ability to capture nonlinear relationships, high-order interactions, and heterogeneous effects across market segments, while remaining scalable to large datasets.[1] In addition, we construct a stacked ensemble that combines the predictions of the three base learners in order to assess whether the ICP framework yields consistent improvements in uncertainty quantification across both individual models and their combination. All models are trained using the Root Mean Squared Error (RMSE) as the primary optimization objective. Model performance is evaluated using a standard set of regression metrics, including RMSE, explanation of variance (R^2), Mean Average Percentage Error (MAPE), Mean Absolute Error (MAE) and Price Percentage Error (PPE) for a percentage λ.[2] Let Y_i denote the true observed value, $\hat{Y}_i$ the model predicted value and n the number of test observations. The metrics are defined as:

[1] Beyond tree-based ensembles, related approaches include hedonic regression (Rosen, 1974), ridge regression (Hoerl & Kennard, 1970), Lasso (Tibshirani, 1996), elastic net (Zou & Hastie, 2005), geographically weigthed regression (Brunsdon et al., 1996), and support vector machines (Cortes & Vapnik, 1995). These approaches typically require imputation of missing covariates, which substantially complicates model construction. Moreover, these methods are generally not considered as top-performing in settings characterized by complex nonlinear interactions, spatial heterogeneity and pervasive missing data.

[2] Root Mean Squared Error (RMSE) is commonly used as the primary optimization and evaluation metric in regression-based automated valuation models. In applied real estate valuation, additional accuracy measures such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are frequently reported to facilitate interpretability and benchmarking. Price Percentage Error, or hit rate, measure the proportion of predictions for which the absolute percentage error falls below a predefined threshold. Common benchmark thresholds in the AVM literature include 5%, 10%, and 20%.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2} \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \times 100 \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right| \tag{12}$$

$$PPE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \leq \frac{\lambda}{100}\right). \tag{13}$$

4.2 Evaluating Conformal Predictors

In the previous section we introduced the metrics for evaluating AVM point prediction accuracy. We now introduce the metrics used for evaluating the performance of a conformal prediction interval along three key dimensions: validity, efficiency and adaptivity.

Validity refers to the agreement between the nominal coverage level of a prediction interval and its empirically observed out-of-sample coverage. A prediction interval constructed at a given nominal level (e.g., 90%) should contain approximately the same proportion of true outcomes in an out-of-sample test set. Empirical coverage on the test sample is measured as:

$$cov = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(Y_i \in \mathcal{C}(X_i)\right) \tag{14}$$

where $\mathbb{1}(.)$ denotes the indicator function, taking the value one if the realized outcome Y_i lies within the prediction set $\mathcal{C}(X_i)$ constructed for the out-of-sample covariate vector X_i, and zero otherwise. The deviation between nominal and empirical coverage, referred to as the coverage gap, provides a direct measure of interval miscalibration. Efficiency captures the informativeness of prediction intervals conditional on achieving the required marginal coverage. In a regression setting, efficiency is measured by the mean width of the prediction intervals over the test data. Given $\mathcal{C}(X_i) = [a_i, b_i]$, where a_i and b_i respectively indicate the lower and upper bound of the prediction interval, mean interval width is defined as:

$$\bar{W} = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)\,. \tag{15}$$

For a given level of attained marginal coverage, narrower average prediction interval widths (see Equation 15) indicate greater efficiency, as they provide more informative uncertainty estimates. By contrast, a trivial prediction interval, such as $(-\infty, +\infty)$, achieves 100% coverage but offers no practical utility. While the conformal algorithm itself does not optimize interval length, a more accurate point prediction AVM will naturally yield smaller residuals and hence tighter conformal intervals. Adaptivity refers to the ability of prediction intervals to reflect heterogeneous uncertainty across the feature space. Well-calibrated intervals should be narrow in regions where predictions are relatively easy and wider in regions where predictions are inherently more difficult. While adaptivity does not admit a single scalar summary statistic, it is closely related to the concept of conditional coverage.

Exact conditional coverage is unattainable under the exchangeability condition alone (Barber et al., 2021a). Nevertheless we highlight two metrics which indicate how coverage varies across subsets of the data. First, Hjort et al., (2024) define a Mean Absolute Coverage Gap (MACG) with respect to a test set $\mathcal{D}_{test}$, defined over *K* classes, where α denotes the nominal miscoverage level as:

$$MACG(\mathcal{D}_{test}) = \frac{1}{K}\sum_{k=1}^{K}\left|(1-\alpha) - cov\left(\mathcal{D}_{test}^{k}\right)\right|, \tag{16}$$

with lower scores indicates better conditional coverage. Second, Angelopoulos et. al., (2020) introduce Feature-Stratified Coverage (FSC) which evaluates the minimum empirical coverage across groups defined by covariates ($|\mathcal{J}_g|$) defined as:

$$\text{FSC} = \min_{g\in\{1,\ldots,G\}} \frac{1}{|\mathcal{J}_g|}\sum_{i\,\in\,\mathcal{J}_g} 1\left\{Y_i^{(test)} \in \mathcal{C}\left(X_i^{(test)}\right)\right\}. \tag{17}$$

For the FSC metric a value closer to $(1-\alpha)$ indicates better conditional coverage.

4.3 The Choice of Non-Conformity Measure

An additional factor influencing the informativeness of conformal prediction intervals is the choice of non-conformity measure (NCM). The NCM evaluates how strange (non-conform) a hypothetical observation is relative to the already observed data (the calibration set under ICP) and determines the distribution of conformity scores used to construct the prediction intervals. While the validity of the constructed conformal prediction intervals is not affected by the choice of NCM under exchangeability, careful construction can lead to more efficient and adaptive prediction intervals. We consider a basic NCM variation. Papadopoulos et al., (2002) introduce the Normalized NCM in general form as:

$$r_i = \left| \frac{Y_i - \hat{Y}_i}{\sigma_i} \right| \tag{18}$$

with σ_i a normalizing factor intended to proxy the inherent difficulty of predicting the outcome for an observation X_i. With $\sigma_i = 1$ this metric reduces to the absolute error yielding the absolute error NCM (see Equation 3). Lim and Bellotti (2021) propose a NCM weighted by the AVM predicted price motivated by the empirical observation that valuation uncertainty tends to increase with property sale prices. Two alternative normalizing factors are considered, defined as:

$$\sigma_i = \gamma + \hat{Y}_i \; or \; \sigma_i = exp(\gamma + \hat{Y}_i), \tag{19}$$

where γ in both cases controls the sensitivity of the NCM to the change in the predicted price. However, the literature lacks a clear framework for tuning the hyperparameter γ.

In this study, we adopt a simpler and more transparent approach by defining non-conformity measures on log CPI-deflated rental prices. Working on the log scale implicitly induces heteroskedasticity proportional to price levels when prediction intervals are mapped back to nominal rents, without requiring the introduction of additional tuning parameters. Importantly, conformal prediction intervals are invariant to monotonic transformations of the target variable during construction; however, when intervals are transformed back to the original price scale, their absolute width naturally increases with the level of the predicted rent. This mechanism yields price-dependent uncertainty behavior qualitatively analogous to that proposed by Lim and Bellotti (2021), while avoiding ad-hoc parameterization and additional calibration complexity.

5. Results

The objective of this section is to evaluate the benefits of applying an inductive conformal prediction (ICP) wrapper to standard real estate automated valuation models. The analysis proceeds in three steps. First, we train point prediction models including CatBoost, LightGBM, Random Forest, and a stacked ensemble, on our dataset and construct prediction intervals using quantile regression. We then evaluate the resulting intervals by examining their empirical coverage gaps on an out-of-sample test set. Second, we demonstrate that augmenting these models with an ICP wrapper, using a standard absolute-residual nonconformity measure, yields empirically satisfactory coverage across all model specifications, with out-of-sample coverage rates closely aligned with the nominal miscoverage level (see Equation 14). Third, we show that a simple Mondrian conformal prediction approach, using stratification by the number of bedrooms, a primary determinant in rent setting, produces more efficient prediction intervals and narrows the gap toward conditional coverage.

5.1 Real Estate AVMs and Quantile Regression Prediction Intervals

All automated valuation models (CatBoost, LightGBM, Random Forest, and stacked Ensemble) are trained to predict log CPI-deflated rents, as described in Section 4.1. The

empirical evaluation follows a rolling, expanding-window design with four temporally disjoint test periods. In each split, the final three months of observations are held out as a test set, ensuring temporal separation between training and testing data.[3] In the final split (February–April 2025), the test set comprises approximately 2.5% of the full dataset. Model training uses five-fold cross-validation (CV) with rolling time windows. Each fold preserves temporal ordering, maintains disjoint training and validation samples, and contains approximately equal numbers of observations. This design mitigates overfitting while preserving the time structure inherent in rental market data.[4]

Table 2 reports standard point-prediction accuracy metrics, Root Mean Squared Error (RMSE), coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), averaged across the four test periods. The metrics are reported on the CPI-deflated rent scale for interpretability while the models are trained to predict log CPI-deflated rents. In addition, we report Price Percentage Error (PPE) statistics, which summarize the proportion of test observations whose predicted rents fall within ±5%, ±10%, and ±20% of the observed values. All results are averaged across the four disjoint test quarters. Across models, predictive performance is broadly comparable. CatBoost, LightGBM, Random Forest, and the stacked ensemble achieve test-set MAPE values close to 10%, which is consistent with reported performance levels for large-scale real estate AVMs operating under limited feature availability.[5] The moderate R^2 values observed across specifications reflect the absence of key hedonic characteristics, such as habitable surface area, construction or renovation year, and building condition, which are known to account for a substantial share of rental price variation. PPE results indicate that approximately one third of predictions fall within ±5% of observed rents, while roughly 60% fall within a ±10% band.

Table 2. Point Prediction Accuracy and Price Percentage Errors for Real Estate AVMs

Model	Prediction Accuracy				Price Percentage Error		
	RMSE	R^2	MAE	MAPE	5%	10%	20%
CatBoost	121	0.67	78.42	10.4%	33.3%	58.9%	87.3%
LightGBM	122	0.66	80.02	10.6%	32.0%	58.5%	86.6%
RF	130	0.62	82.24	10.7%	32.1%	58.4%	85.8%
Ensemble	**118**	**0.68**	**76.44**	**10.2%**	**34.3%**	**60.2%**	**87.5%**

Results are shown for each model, averaged over the disjoint last 4-quarterly test sets of the dataset, and reported on CPI-deflated rents. In the left section we display standard point prediction accuracy metrics: Root Mean Squared Errors (RMSE), Variance explanatory power R^2, Mean Absolute Error (MAE) and Mean Average Percentage Error (MAPE). In the right section we show Price Percentage Error (PPE). For each % (5, 10, 20) the number of rent estimates that fall within *n*-% of the true range are displayed. The best figures per metric are displayed in bold.

[3] Holding out a full calendar quarter as the test set ensures a sufficiently large sample to reliably assess empirical coverage, as defined in Equation 14. In an operational setting, the model could be recalibrated at a higher frequency (e.g., daily) as new observations become available, without affecting the validity of the evaluation framework.

[4] Hyperparameters are tuned via random search over standard parameter spaces by minimizing mean five-fold cross-validated RMSE, with early stopping to control overfitting. Although more extensive tuning could further improve point-prediction accuracy, the focus of this study is on evaluating conformal prediction–based uncertainty quantification rather than on optimizing predictive performance.

[5] Reported point-prediction accuracy is in line with existing evidence from the real estate AVM literature. For example, Baur et al., (2023), Lenaers & De Moor (2023), Lenaers et al., (2024), Lorenz et al., (2023), Gyger et al., (2026) and Zhou et al., (2019) report MAPE values of approximately 11% for their best-performing models, often using substantially larger datasets and/or richer sets of explanatory variables, including net habitable surface area, building age, building condition, transaction prices, or tax-related information. Yoshida et al., (2024) achieve MAPE values around 9.5% on a dataset of roughly twice the size, incorporating more extensive feature information and restricted to apartment units. Clark and Lomax (2018) report median percentage prediction errors exceeding 12% in a large-scale application involving more than one million observations over a two-year period.

For each model in Table 2, prediction intervals are constructed using quantile regression (QR), as described in Section 4.2. Table 3 reports the performance of these QR-based prediction intervals at nominal coverage levels of 80%, 90%, and 95%, averaged across the four temporally disjoint test periods. The first column reports the mean absolute empirical coverage gap, defined as the absolute difference between empirical test coverage (see Equation 14) and the nominal level. Absolute gaps are reported to prevent systematic under- and over-coverage across test periods from offsetting one another, which is particularly relevant for the stacked ensemble. Across the individual AVMs, QR-based prediction intervals exhibit systematic negative coverage gaps across confidence levels, indicating undercoverage relative to the nominal targets. For the stacked ensemble, empirical coverage is closer to nominal levels. As a diagnostic check, we assess whether observed coverage is statistically consistent with the nominal level using a two-sided binomial test applied separately to each test period and confidence level. For the individual models, the null hypothesis of equality between empirical and nominal coverage is rejected at the 1% level in the majority of cases. For the ensemble, empirical coverage is statistically indistinguishable from the nominal level across test periods (see Appendix: Table A).

Importantly, this result for the ensemble reflects empirical alignment rather than a formal coverage guarantee and is contingent on the residual behavior of the ensemble in the present dataset, which aggregates errors across heterogeneous base learners. Quantile regression does not provide finite-sample marginal coverage guarantees, and failure to reject the binomial test should therefore be interpreted as descriptive rather than inferential evidence of calibration. The second column of Table 3 reports the median prediction interval width for each model and confidence level averaged over the test sets. Interval widths are presented descriptively and are not ranked across models, as differences in width are not directly comparable when empirical coverage differs. The results in Table 3 provide a baseline characterization of QR-based prediction intervals against which conformal prediction methods are evaluated in the subsequent section.

Table 3. Empirical Coverage Gap and Prediction Interval Width

Model	Empirical CovGap			Med. Bandwidth		
	80%	90%	95%	80%	90%	95%
CatBoost	4.9%	2.9%	2.5%	185	260	336
LightGBM	8.5%	5.3%	2.5%	176	257	356
RF	7.4%	4.7%	4.7%	181	258	319
Ensemble	**0.4%**	**0.3%**	**0.4%**	208	288	372

Results are shown for each model, averaged over the disjoint last 4-quarterly test sets of the dataset, and reported on CPI-deflated rents. In the left section we display mean Empirical coverage gap over the 4 last disjoint test quarters with respect to different confidence levels. The coverage gap is reported in absolute percentage points as all models exhibit undercoverage. The best figures per scoring metrics are displayed in bold. In the right section we show mean median bandwidth of the prediction intervals for the disjoint test quarters for different confidence levels.

5.2 The Inductive Conformal Prediction Wrapper

The results in Table 3 demonstrate that prediction intervals obtained via quantile regression (QR) do not generally achieve the stated nominal empirical coverage in out-of-sample testing. This is consistent with the theoretical properties of QR, which estimates conditional quantiles but does not guarantee finite-sample marginal coverage. In this section, we examine whether applying an inductive conformal prediction (ICP) wrapper to the same point prediction models yields empirically calibrated prediction intervals with coverage closer to the nominal level.

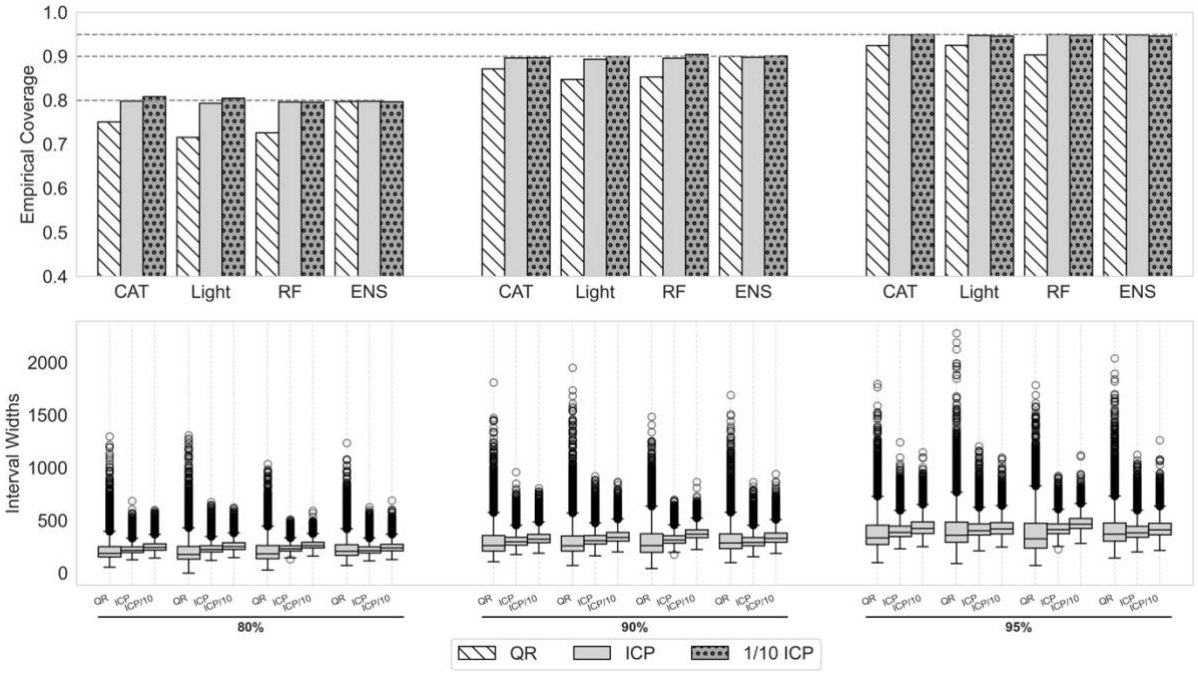
Figure 3 compares the empirical coverage of QR-based prediction intervals with ICP-based prediction intervals at nominal coverage levels of 80%, 90%, and 95% for CPI-deflated rents. All models are trained using five-fold cross-validation with strictly time-ordered folds, and empirical coverage is evaluated on the same disjoint out-of-sample test sets as those used in Section 5.1. For the ICP implementation, two variants are reported. In the first, ICP is applied using the full training dataset, where a subset (10%) of the training observations is set aside for calibration. In the second, denoted ICP/10, a randomly drawn 10% subset of the training data is first selected, and this reduced dataset is then split into a proper training sample and a calibration sample (again 10%). In both cases, prediction intervals are evaluated on the full out-of-sample test sets, ensuring direct comparability with the QR-based intervals. This design implies that the QR models are trained on more data than the corresponding ICP/10 models, which makes the latter a conservative benchmark for assessing finite-sample conformal validity.

The upper panel of Figure 3 reports empirical out-of-sample coverage for QR-based and ICP-based prediction intervals at nominal levels of 80 %, 90 %, and 95 %. Across all models and confidence levels, the QR-based intervals display systematic empirical under-coverage. By contrast, ICP-based prediction intervals achieve coverage that is much closer to the stated nominal levels, including in the specification where ICP is implemented using only 10 % of the available training data. This pattern is consistent with the finite-sample marginal coverage property of ICP under exchangeability. The comparison between QR and ICP constructed on the 10 % training subsample is particularly informative: despite the considerably smaller calibration set, ICP still attains empirical coverage that is closer to nominal targets than QR estimated on the full dataset. The lower panel of Figure 3 shows the corresponding distribution of interval widths. QR-based intervals are generally narrower than their ICP counterparts; however, this reduction in width coincides with the systematic under-coverage documented above, illustrating the trade-of between interval efficiency and validity. ICP-based intervals are wider on average, reflecting the adjustments required to satisfy marginal coverage in finite samples. When ICP is implemented using a reduced training sample, interval widths increase further, in line with the intuition that smaller calibration samples imply greater uncertainty and hence wider prediction sets.

A noteworthy exception arises for the stacked ensemble model. For this specification, QR-based prediction intervals deliver empirical coverage that is close to the nominal level. This finding suggests that improvements in the underlying point predictor may enhance the accuracy

of the conditional quantile estimates on which QR relies. Nevertheless, ICP continues to provide finite-sample marginal coverage guarantees under weaker conditions. Overall, the evidence indicates that QR-based prediction intervals applied to standard AVMs tend to be relatively narrow and exhibit systematic under-coverage of realized rents. By contrast, ICP-based prediction intervals, constructed here using a simple absolute-residual non-conformity measure, consistently achieve empirical coverage close to the corresponding nominal levels across models and training set sizes. Moreover, ICP can be implemented as a post-estimation procedure for any underlying prediction model, in contrast to QR, which requires re-estimation for each quantile level.

Figure 3. Quantile Regression vs. ICP



This figure reports empirical out-of-sample coverage and median prediction-interval width for quantile-regression (QR) and inductive conformal prediction (ICP) intervals at nominal coverage levels of 80%, 90% and 95%. All results are computed on CPI-deflated rents. The underlying point prediction models are trained using five-fold cross-validation with strictly time-ordered folds. ICP results are shown both for the models trained on the full dataset and for the models trained on a randomly drawn 10% subsample of the training data (ICP/10) while coverage is always evaluated on the same out-of-sample test sets to ensure comparability across methods.

5.3 Beyond Marginal Coverage.

The previous section established ICP-based prediction intervals to deliver empirical out-of-sample marginal coverage across all AVM specifications. While marginal validity is a necessary condition for credible automated valuation, applied use depends on the efficiency of the resulting intervals and on the degree to which estimated uncertainty varies with observable heterogeneity in the housing stock. Although exact conditional coverage is infeasible under mild regularity conditions (Barber et al., 2021a), it is nonetheless informative to examine whether prediction intervals contract in market segments where rents are relatively predictable and widen where predictive uncertainty is structurally greater. To examine these issues, Mondrian-ICP is implemented as an extension of the baseline ICP procedure on the CatBoost model. Mondrian-ICP partitions the sample into mutually exclusive subsets and conducts the

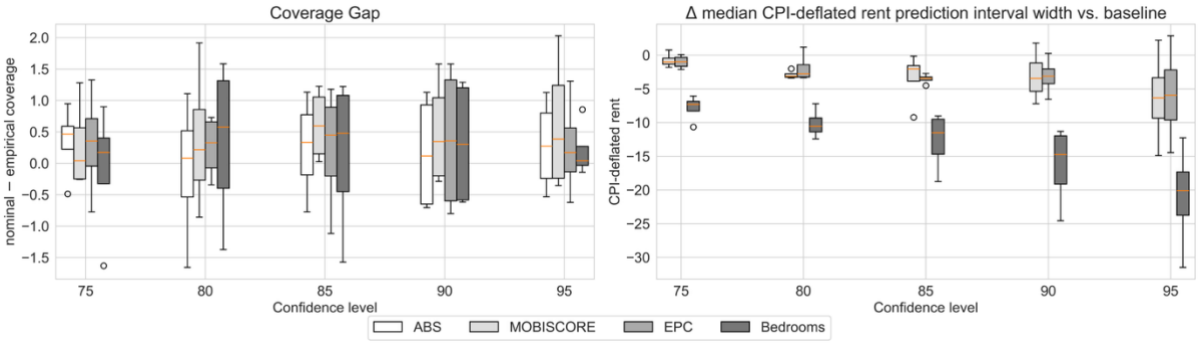
conformal calibration step separately within each partition. Provided that exchangeability holds within partitions, marginal coverage is preserved, while efficiency may improve when variation in predictive uncertainty aligns with the partition structure.

Given the constraints imposed by the available covariates, the scope for defining partitions that are simultaneously statistically robust and economically meaningful is restricted to three attributes that represent fundamental structural dimensions of rental markets and that can be interpreted consistently across empirical contexts. First, the number of bedrooms serves as a primary indicator of dwelling size and use; departures from the modal two-bedroom category may be associated with thinner sets of comparable observations and greater heterogeneity, implying systematic variation in predictive uncertainty. This approach is well-suited to partitioning by building typology, as bedroom count provides a more parsimonious and behaviorally relevant measure of dwelling size that aggregates information across structural categories while retaining flexibility in capturing within-type heterogeneity. Due to the small number of observations, all units containing more than 3 bedrooms are grouped into a single class. Second, the EPC label is the principal energy-performance classification communicated to market participants and proxies for building quality, renovation state, and construction period, none of which are directly observed here despite their relevance for rent formation. Third, mobiscore deciles summarize accessibility to employment, services, and transport infrastructure; geographic variation in these fundamentals plausibly contributes to dispersion in rents and, consequently, in predictive accuracy. These attributes are structural rather than dataset-specific and therefore represent natural candidates for Mondrian partitioning.

Figure 4 reports empirical coverage gaps and median prediction interval widths for Mondrian-ICP under each partition at nominal levels of 80%, 90%, and 95% across all out of sample test sets. Across all specifications (#bedrooms, EPC-labels, mobiscore-deciles), empirical coverage remains close to the nominal target, consistent with finite-sample validity. However, interval widths differ across implementations. The bedroom-based ICP-Mondrian specification yields the narrowest prediction intervals among the procedures, reducing the median width of CPI-deflated rent prediction intervals with 5.3% (€20) relative to the absolute residual split conformal prediction intervals (median CPI-deflated prediction interval width of €377, see Figure 3). The results are consistent with bedroom count capturing residual distributional differences in prediction errors that remain after conditioning on the full feature set, making it an effective Mondrian partitioning variable in feature poor large scale AVM setting.

Figure 5 provides a more in-depth analysis of the results, evaluating conditional coverage performance at confidence levels of 80%, 90%, and 95% across the test sets conditioned on EPC label, building typology, mobiscore decile, bedroom category, and true out-of-sample CPI-deflated rent decile. The left-hand column reports the Mean Absolute Coverage Gap (MACG), which summarizes the average absolute deviation between empirical and nominal coverage across groups. The right-hand column reports the Feature-Stratified Coverage (FSC), defined as the lowest coverage attained across the groups within each partition.

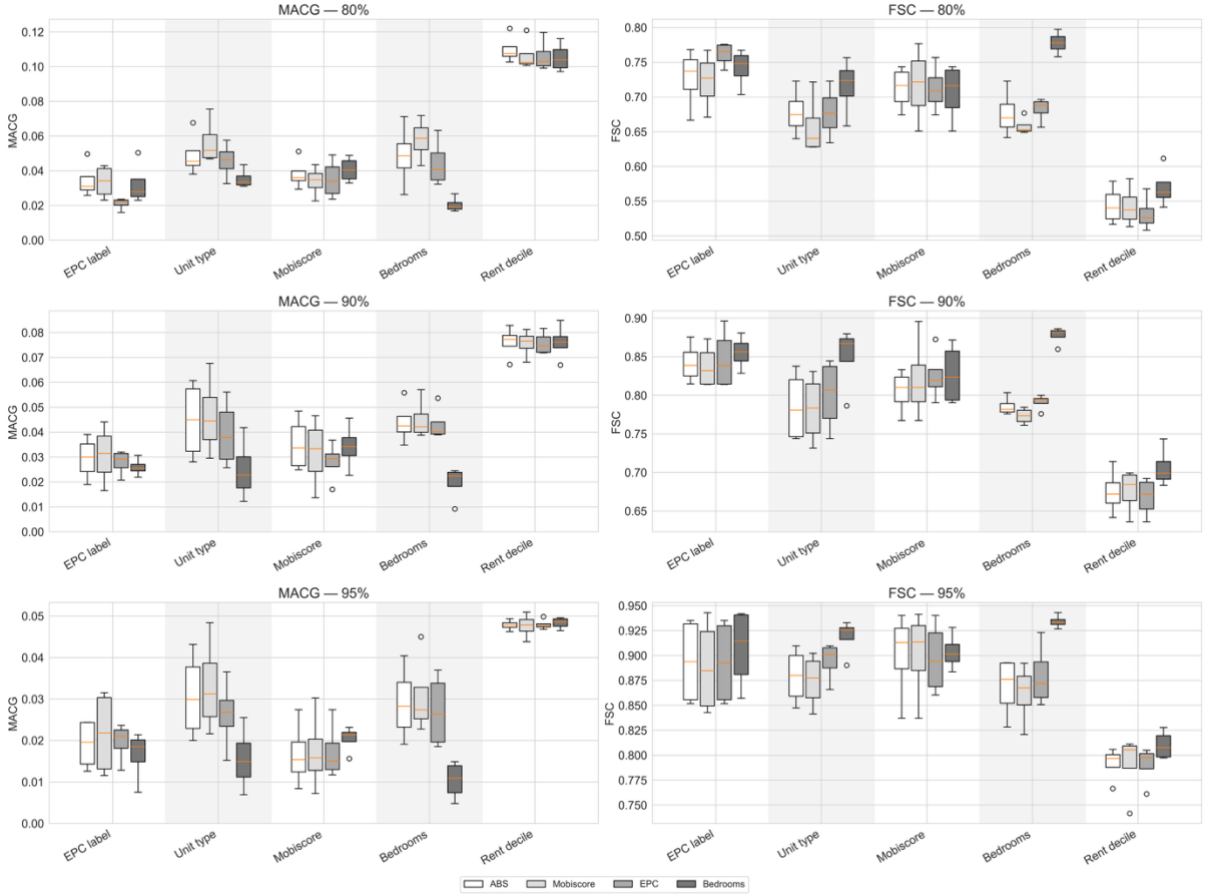
Figure 4. Marginal Coverage and ICP Interval Widths



We show the results for different ICP implementations with respect to the Coverage gap and the difference in median interval width with the absolute residual NCM ICP implementation for CatBoost for different levels of coverage (80%, 90% and 95%). Results are displayed for the last 4 out of sample temporal test quarters. The ICP implementations includes an absolute (ABS) NCM and a Mondrian-ICP implementation on the mobiscore, the Energy efficiency label (EPC) and the number of Bedrooms.

Across EPC labels, dwelling type, and mobiscore deciles, the MACG and FSC statistics are broadly similar across ICP implementations. This indicates that Mondrian partitioning along these dimensions does not materially improve conditional calibration relative to the baseline ICP procedure. Notably, this remains true even when the Mondrian split is aligned directly with the conditioning variable: neither the EPC-based nor the mobiscore-based Mondrian implementation yields discernible gains in conditional coverage when coverage is evaluated conditional on those same attributes. These results suggest that, conditional on the covariates included in the underlying AVM, systematic variation in predictive uncertainty is only weakly associated with energy-efficiency or accessibility segmentation. By contrast, more pronounced differences emerge when conditioning the coverage across test sets on bedroom category and building typology. Here, bedroom-based Mondrian-ICP consistently achieves lower MACG and higher FSC than the baseline specification. This pattern indicates improved conditional stability and is consistent with the view that the number of bedrooms captures a central structural dimension of rental heterogeneity. Conditional performance is weakest when the test sample is stratified by rent deciles. Across all nominal coverage levels, both MACG and FSC exhibit materially higher dispersion than in the other partitions, indicating substantial heterogeneity in group-level coverage. This implies that, although marginal validity is preserved, coverage errors are not evenly distributed across the rent distribution. Inspection of the decile-specific results shows that the largest deviations from nominal coverage arise in the first, second, ninth, and tenth deciles, while coverage for the intermediate deciles remains close to target levels. This pattern is consistent with the behavior of the underlying point predictors, whose absolute and relative errors are also largest in the tails of the rental distribution. A plausible mechanism is covariate incompleteness. In the absence of key hedonic attributes, such as habitable floor area, construction or renovation year, and explicit indicators of dwelling quality, the models cannot separate intrinsically low-quality from high-quality units. The conditional distribution of rents given observables is therefore substantially more dispersed in the tails than in the center of the distribution. As a result, the conformity scores exhibit greater variability in the extreme deciles, and neither baseline ICP nor Mondrian-ICP fully offsets this form of latent heterogeneity.

Figure 5. Conditional Coverage Across Different ICP Implementations



We show the results for different ICP implementations with respect to the Mean Absolute Coverage Gap (MACG) and the Feature Stratified Coverage (FSC) for different levels of coverage (80%, 90% and 95%). Results are displayed for the last 4 out of sample temporal test quarters. The ICP implementations include an absolute (ABS) and relative (REL) NNCM and a Mondrian ICP implementation on the mobiscore, the Energy efficiency label (EPC) and the number of Bedrooms.

Overall, the evidence indicates that Mondrian-ICP yields modest but meaningful improvements in conditional coverage only when partitions align with structural attributes that materially shape valuation uncertainty, in this dataset, the number of bedrooms. In contrast, segmentation by spatial accessibility or energy performance does not materially affect conditional coverage, suggesting that these attributes play a lesser role in explaining predictive uncertainty given the available feature space. Nonetheless, these findings should be regarded as context-specific rather than universal: in richer datasets with more granular structural and locational information, alternative partition dimensions may prove more informative.

6. Conclusions

In current automated valuation model (AVM) practice, uncertainty quantification is typically implemented through quantile regression (QR), including in widely used machine-learning frameworks such as CatBoost, LightGBM, and Random Forests. Since QR estimates conditional quantiles, nominal coverage is only guaranteed when the conditional distribution of rents is correctly specified. In applied housing-market settings, where covariates are incomplete and residual heterogeneity is substantial, this assumption rarely holds. Consistent

with this expectation, the empirical analysis shows that QR-based prediction intervals systematically under-cover realized rents across standard rental AVM specifications.

Inductive Conformal Prediction (ICP) provides a simple and conceptually robust remedy. Applied as a wrapper around pre-trained AVMs, ICP delivers finite-sample *marginal* coverage guarantees under weak exchangeability assumptions and achieves empirical coverage close to nominal levels at negligible additional computational cost. Although a highly tuned stacked ensemble combined with QR can, in this application, produce empirical coverage comparable to ICP, this outcome depends on dataset-specific residual behavior and running rather than on formal finite-sample guarantee. By contrast, ICP offers a model-agnostic and theoretically grounded framework for uncertainty quantification in AVMs. It is important to note, however, that ICP does not guarantee conditional coverage; rather, it ensures that predictive intervals are valid on average across the population.

The results are obtained in the rental segment, where valuation is closely tied to income flows. They are therefore particularly relevant for institutional users who depend on reliable uncertainty estimates: banks and financial institutions valuing collateral based on rental income, governments determining property tax assessments derived from imputed rents, and real-estate professionals communicating expected rental yields to market participants. A Mondrian-ICP extension based on bedroom count yields modest gains in efficiency and conditional stability, consistent with the structural role of dwelling size in rent setting. Conditional coverage remains weakest in the extreme rent deciles, reflecting omitted heterogeneity associated with unobserved dwelling quality and unit size. CPI-deflation and logarithmic transformation provide a simple and robust preprocessing strategy for long-horizon rental data, improving temporal comparability and stabilizing the conditional error distribution.

Finally, it is worth emphasizing that the implementation adopted here is deliberately conservative: a naïve ICP procedure is applied without explicit adjustments for covariate or distributional shift. That this approach already delivers marked improvements in empirical coverage underscores the practical value of conformal prediction in rental AVMs. Overall, conformal prediction should be viewed as a natural benchmark for uncertainty quantification in large-scale valuation systems, combining theoretical validity, empirical robustness, and operational feasibility. Future work could extend these methods to richer feature environments and explicitly address conditional validity and distributional change.

Acknowledgement

I gratefully acknowledge financial support from the Flemish Science Foundation (S006721N) and CIB-Flanders. I thank Oris NV and Korfine for providing access to their datasets as well as Pieter Decelle, Sander Franck, Bart Van Der Schueren and Jonas Zaman for valuable feedback as well as participants at Ghent University internal seminars. I am greatly indebted to Kris Boudt and Koen Inghelbrecht for their insightful comments and suggestions.

References

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Angelopoulos, A., Bates, S., Malik, J., & Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772-1778.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021, a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, *10*(2), 455-482.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021, b). Predictive inference with the jackknife+. *The Annals of Statistics*, *49*(1), 486-507.

Bastos, J. A., & Paquette, J. (2025). On the uncertainty of real estate price predictions. *Journal of Property Research*, *42*(1), 1-19.

Baur, K., Rosenfelder, M., & Lutz, B. (2023). Automated real estate valuation with machine learning models using property descriptions. *Expert Systems with Applications*, *213*, 119147.

Bellotti, A. (2017). Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, *81*(1), 71-84.

Boudt, K., Inghelbrecht, K., & Van Besien M. (2025). Stable and reliable monthly repeat rent indices: a robust approach. *Ghent University, Faculty of Economics and Business Administration,* (Working Paper No. 25/1126).

Boström, H., & Johansson, U. (2020). Mondrian conformal regressors. *Proceedings of Machine Learning Research, 128,* 114-113.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, *28*(4), 281-298.

Clark, S. D., & Lomax, N. (2018). A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. *Journal of Big Data*, *5*(1), 43.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining, 96*, 226-231.

Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies*, *43*(12), 2301-2315.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Gammerman, A., Vovk, V., & Vapnik, V. (1998). Learning by transduction, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence,* 148-156.

Gyger, T., Hauri, S., Bühlmann, S., Lehner, M., Schlesinger, J., & Sigrist, F. (2026). Explainable spatial machine learning for hedonic real estate modeling. *Real Estate Economics*, 1-46.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hjort, A., Williams, J. P., & Pensar, J. (2024). Clustered conformal prediction for the housing market. *Proceedings of Machine Learning Research*, *230*, 1-21.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances Neural Information Processing Systems*, *30*.

Koenker, R., & Basset, G. (1978). Regression results. *Econometrica*, *46*(1), 33-50.

Lenaers, I., Boudt, K., & De Moor, L. (2024). Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques. *International Journal of Housing Markets and Analysis*, *17*(1), 96-113.

Lenaers, I., & De Moor, L. (2023). Exploring XAI techniques for enhancing model transparency and interpretability in real estate rent prediction: A comparative study. *Finance Research Letters*, *58*, 104306.

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real estate economics*, *51*(5), 1178-1208.

Lim, Z., & Bellotti, A. (2021). Normalized nonconformity measures for automated valuation models. *Expert Systems with Applications*, *180*, 115-165.

Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(35), 983-999.

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. *Proceedings of the 13th European Conference on Machine Learning,* 345-356.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances Neural Information Processing Systems*, *31*.

Romano, Y., Patterson, E., & Candes, E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, *32*.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy,* 82(1), 34–55.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert systems with Applications*, *36*(2), 2843-2852.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*, 371-421.

Takeuchi, I., Le, Q., Sears, T., & Smola, A. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research, 7*(45): 1231-1264.

Tay, D. P., & Ho, D. K. (1992). Artificial Intelligence and the Mass Appraisal of Residential Apartments. *Journal of Property Valuation and Investment*, *10*(2), 525-540.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of forecasting*, *19*(4), 299-311.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. *Springer US.*

Xu, C., & Xie, Y. (2023). Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 11575-11587.

Yoshida, T., Murakami, D., & Seya, H. (2024). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. *The Journal of Real Estate Finance and Economics*, *69*(1), 1-28.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320.

Zhou, X., Tong, W., & Li, D. (2019). Modeling housing rent in the Atlanta metropolitan area using textual information and deep learning. *ISPRS International Journal of Geo-Information*, *8*(8), 349

Appendix

Table A. Empirical Coverage Across All Test Folds

		Nominal Coverage Level								
Model	Fold	80%			90%			95%		
		n	Emp. Cov.	p-value	n	Emp. Cov.	p-value	n	Emp. Cov.	p-value
CATBOOST	0	1119	0.723	0.0000	1119	0.875	0.0000	1119	0.937	0.0635
	1	1750	0.743	0.0000	1750	0.869	0.0000	1750	0.919	0.0000
	2	2108	0.763	0.0000	2108	0.877	0.0000	2108	0.925	0.0000
	3	2219	0.777	0.0000	2219	0.874	0.0000	2219	0.918	0.0000
LIGHTGBM	0	1119	0.731	0.0000	1119	0.853	0.0000	1119	0.932	0.0000
	1	1750	0.691	0.0000	1750	0.827	0.0000	1750	0.925	0.0000
	2	2108	0.706	0.0000	2108	0.852	0.0000	2108	0.918	0.0000
	3	2219	0.732	0.0000	2219	0.863	0.0000	2219	0.926	0.0000
RF	0	1119	0.730	0.0000	1119	0.829	0.0000	1119	0.883	0.0000
	1	1750	0.743	0.0000	1750	0.866	0.0000	1750	0.904	0.0000
	2	2108	0.719	0.0000	2108	0.857	0.0000	2108	0.897	0.0000
	3	2219	0.712	0.0000	2219	0.854	0.0000	2219	0.928	0.0000
ENSEMBLE	0	1119	0.799	0.9762	1119	0.905	0.5839	1119	0.953	0.4507
	1	1750	0.801	0.9846	1750	0.901	0.9682	1750	0.951	0.8264
	2	2108	0.786	0.7023	2108	0.893	0.3093	2108	0.948	0.6175
	3	2219	0.802	0.7704	2219	0.900	1.0000	2219	0.943	0.1190

Empirical coverage is computed on temporally disjoint out-of-sample test sets (folds 0–3). For each model and nominal coverage level, the table reports the number of observations in the test fold (n), the empirical coverage (Emp. Cov.), and the p-value from a two-sided binomial test of equality between empirical and nominal coverage. Small p-values indicate statistically significant under- or over-coverage relative to the nominal rate. For the CatBoost, LightGBM, and Random Forest models the null hypothesis of nominal coverage is rejected in all cases, consistent with systematic under-coverage of QR-based prediction intervals. By contrast, for the stacked ensemble the empirical coverage is not statistically distinguishable from the nominal level across folds and confidence levels. This alignment should be interpreted as an empirical property of the ensemble residuals in the present application rather than as evidence of a general coverage guarantee.