Uncertainty quantification in automated valuation models with locally weighted conformal prediction

Anders Hjort[1,2*], Gudmund Horn Hermansen[1], Johan Pensar[1], Jonathan P. Williams[3]

[1*]Department of Mathematics, University of Oslo, Oslo, Norway.
[2]Eiendomsverdi AS, Oslo, Norway.
[3]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA.

*Corresponding author(s). E-mail(s): anderdh@math.uio.no;
Contributing authors: gudmunhh@math.uio.no; johanpen@math.uio.no; jwilli27@ncsu.edu;

Abstract

Non-parametric machine learning models, such as random forests and gradient boosted trees, are frequently used to estimate house prices due to their predictive accuracy, but such methods are often limited in their ability to quantify prediction uncertainty. Conformal Prediction (CP) is a model-agnostic framework for constructing confidence sets around machine learning prediction models with minimal assumptions. However, due to the spatial dependencies observed in house prices, direct application of CP leads to confidence sets that are not calibrated everywhere, i.e., too large of confidence sets in certain geographical regions and too small in others. We survey various approaches to adjust the CP confidence set to account for this and demonstrate their performance on a data set from the housing market in Oslo, Norway. Our findings indicate that calibrating the confidence sets on a *locally weighted* version of the non-conformity scores makes the coverage more consistently calibrated in different geographical regions. We also perform a simulation study on synthetically generated sale prices to empirically explore the performance of CP on housing market data under idealized conditions with known data-generating mechanisms.

Keywords: confidence sets, distribution-free calibration, model-free calibration, prediction region, safe learning

1 Introduction

The housing market plays a crucial role in most modern economies, and buying a home is the most significant financial investment many people make throughout their lives. Homeowners, real estate professionals, and financial institutions like banks and insurance companies, therefore, need to monitor the value of a home regularly. While manual appraisal often results in an accurate valuation, performing it regularly on a portfolio of thousands of homes is infeasible. Therefore, participants in the real estate market rely on Automated Valuation Models (AVMs), statistical models that estimate the value of a home given its location, characteristics, and previous sale price. Modern AVMs often rely on non-parametric machine learning approaches like gradient boosted trees, random forests, or a combination of both (Steurer et al [42]). These models tend to deliver superior predictive accuracy compared with more classical approaches, such as linear regression, due to their flexible and non-parametric nature while also being able to handle a combination of numerical and categorical features, which are often present in housing market data sets (Kim et al [21]).

Despite their predictive accuracy, the above-mentioned models often lack inherent mechanisms for quantifying the uncertainty in their predictions. Practitioners are often interested in not only a point prediction but a range of plausible values, referred to as a confidence set. Returning a confidence set in addition to the point prediction makes it easier to evaluate the trustworthiness of the AVM and use it for risk-management purposes. Empirical data indicates that house prices display significant heteroscedasticity and spatial heterogeneity (Marques et al [31]) and that even the state-of-the-art AVMs yield very different accuracy across different spatial regions and price levels (Hjort et al [17], Lim and Bellotti [26]), further motivating the need for calibrated uncertainty quantification in the AVMs.

Conformal prediction (CP) is one such method to quantify uncertainty in machine learning models. The method was popularly introduced in Vovk et al [46] as a distribution-free test of exchangeability but has recently gained popularity within the machine learning community as a general-purpose method for constructing confidence sets around a prediction (see Lei and Wasserman [24], Lei et al [25] and the references therein). The recent popularity has seen CP being used in a wide range of scenarios, including in medical applications (Vazquez and Facelli [44]), survival analysis (Candès et al [9]), time series analysis (Stankeviciute et al [41]), image recognition tasks (Angelopoulos et al [2]), and natural language processing (Dey et al [11], Maltoudoglou et al [29]).

The fundamental idea behind the CP framework is to perform tests to assess whether a test data point is exchangeable with a set of training observations, relying on a non-conformity score that quantifies how dissimilar or non-conforming each data point is to the training set. From the training set, we can use these scores as a reference to calibrate our expectations about the uncertainty of a new, unobserved instance. Specifically, we create a confidence set for the test instance by only including values that conform sufficiently with the training set at some user-specified confidence level. An appealing trait of the CP framework lies in its generality. The confidence sets created by the CP procedure are distribution-free, with few assumptions on the point prediction used in the regression. Furthermore, the confidence sets are *marginally valid*

in the following sense: A 90% confidence set created by a CP algorithm on a training set has a 90% probability of covering a test instance under the assumption that the test instance is exchangeable with the training data. We thus expect 90% of exchangeable test observations to be covered by the confidence sets created by the CP framework without knowing the distribution of the data.

In addition to the marginal validity guarantees, it is often desirable in practice to have theoretical guarantees also in subsets of the feature space or geographic space. Recent years have seen a surge in such research directions. Many of these methods accomplish this through different modifications of the original CP framework, including CP under covariate shift (Tibshirani et al [43]), CP under non-exchangeable data (Foygel Barber et al [13]), Localized CP (Guan [15]), and Spatial CP (Mao et al [30]). While these methods differ in multiple ways, they all rely on the idea of *weighting* the non-conformity scores. Instead of treating all the non-conformity scores as equally important when calibrating the confidence sets, higher weight is assigned to non-conformity scores that stem from training observations close to the test instance, either in geographic space or feature space. The motivation behind this is typically to achieve some form of local validity in addition to the marginal validity that the original CP framework comes with.

While the last years have seen a substantial number of theoretical contributions to the field of CP with weighted non-conformity scores, few guidelines exist on how to apply weighting functions in practical prediction problems with multidimensional data. Our study aims to shed light on this question for the specific application of house price prediction by implementing the available weighting schemes in the CP literature. In addition to the regular CP framework of Vovk et al [46], we employ *locally weighted* CP, where the non-conformity scores in a close geographic neighborhood of a test instance are assigned higher weight (Foygel Barber et al [13], Mao et al [30]). We also compare with feature weighted CP, where the features of each dwelling are used to calculate the weight of the non-conformity scores (Tibshirani et al [43]). While regular CP has been applied to housing data previously (Bellotti [4], Lim and Bellotti [26]), we are, to the best of our knowledge, the first to study how to design the weighting function in an AVM setting. Our research is motivated by the pursuit of confidence sets that are more adaptive to changes in price level and price variability in different spatial regions. We use a data set from the housing market in Oslo (Norway) as our case study. To strengthen the analysis, we analyze all the different weighting methods with multiple non-conformity scores and point prediction models.

In summary, our paper makes the following contributions to the literature of applied CP:

- We provide a survey on recent methodological advances in scenarios where the non-conformity scores are not necessarily globally exchangeable, but where a weighted version of the scores can adjust for this.
- We conduct an extensive case study comparing five different variants of CP, with up to five different weighted non-conformity scores per method, for the application of house price prediction on a data set from Oslo, Norway. The case study demonstrates that assigning higher weight to non-conformity scores in a spatial neighborhood

yields confidence sets that adapt better to local spatial patterns than the standard unweighted CP.

- We conduct a simulation study on a synthetic data set, based on the real-world data, exploring the limits of the different methods in a scenario where the underlying data-generating mechanism is known.

The rest of the paper is structured as follows. Section 2 introduces the application of AVMs, Section 3 introduces the data set from Oslo that will be studied, Section 4 presents the necessary theoretical framework on CP and its extensions based on weighted non-conformity scores. Section 5 presents the results of all the methods in two scenarios: First with synthetically generated sale prices and then on the real transaction prices observed in the Oslo data set. Concluding remarks are made in Section 6.

2 Automated valuation models

Buying a home is often among the most significant financial decisions people make throughout their lives. Most home buyers need to take out a mortgage to finance the acquisition, and The Organisation for Economic Co-operation and Development (OECD) reports that the average Loan-to-Income rate in OECD countries was approximately 4.4 in 2018 (van Hoenselaar et al [19]). It is, therefore, necessary for both banks and homeowners to regularly monitor the value of a home or a portfolio of homes. While manual appraisal through physical inspections is one way of estimating the home value, these are often time-consuming and hard to conduct regularly. For these reasons, many financial institutions rely on AVMs to provide statistical estimates of housing values.

While multiple approaches to constructing AVMs exist, the most common models are *hedonic pricing models* that estimate the price of a dwelling given its characteristics, such as the location, size, number of bedrooms, etc. Early examples of regression models being used for this purpose can be seen in Bailey et al [3] and Rosen [37], and recent years have seen an increased use of non-parametric machine learning models like random forests (Breiman [7]) and gradient boosted trees (Chen and Guestrin [10]) as hedonic price models in AVMs, including in Hong Kong (Ho et al [18]), South Korea (Kim et al [21]), Malaysia (Mohd et al [33]), Norway (Hjort et al [17]), and Virginia, USA (Wang and Wu [47]).

The popularity of non-parametric tree models can be attributed to, amongst other things, their ability to detect higher-order interactions and their ability to work with both categorical and numerical data. A decision tree (Breiman et al [8]) recursively partitions the feature space into a set of non-overlapping regions $\mathcal{R}_1, ..., \mathcal{R}_L$, where each region is represented by a leaf or terminal node in the tree structure. Each region takes a value that serves as the prediction for any test instance whose feature values place it in $\mathcal{R}_l$. Although a single decision tree might be limited in its predictive power, ensemble methods like random forests or gradient boosted trees often achieve superior predictive accuracy in many prediction tasks on tabular data. Certain tree-based regression methods can be interpreted as a form of nearest neighbor regression

(Lin and Jeon [27]). For a thorough review of tree-based regression models, we refer to Hastie et al [16] or Chen and Guestrin [10]).

AVMs typically suffer from considerable omitted-variable bias, with vital information such as the floor plan, the degree of interior luxury, or the dwelling's exterior design typically being unknown or hard to quantify (Zhou et al [49]). There are examples of the use of image recognition software to enrich the available data for AVMs (Kiyota [22], Nouriani and Lemke [34]), but this is rarely reliably available for an entire portfolio of dwellings. Other sources of influence on the final sale price might include weather patterns around the day of the open house (Gourley [14]) or bidding strategies in the auction leading up to the sale (Sommervoll [40]), which both might be challenging to account for a priori. As a consequence of the inherent uncertainty and noise in the prediction task, AVMs are often evaluated using the Percentage Error Range (PER), which measures the fraction of predictions on a test set that is within either $\pm 10\%$ or $\pm 20\%$ of the actual sale price (Steurer et al [42]). This industry-specific performance metric indicates that even predictions that under- or overestimate by more than 10% may be deemed acceptable due to the challenging nature of the prediction task.

The increased deployment of machine learning methods in AVMs also calls for increased attention to uncertainty quantification in AVMs. The non-parametric nature of the models mentioned above comes with the consequence that they are rarely equipped with straightforward methods for creating prediction intervals. A dichotomization between *model-based* and *error-based* uncertainty models for AVMs is introduced in Krause et al [23]. Model-based uncertainty measurements are derived directly from the underlying point prediction, often relying on parametric models (like linear regression) with some inherent uncertainty quantification. In contrast, the error-based uncertainty models use historical errors to assign confidence to new predictions. Although several industry white papers exist on evaluating confidence sets for AVMs, there is a lack of universally agreed-upon metrics to quantify and assess the different uncertainty quantification methods (Krause et al [23]).

Academic research on specific methods for creating prediction intervals in AVMs is also scarce, with Bellotti [4] and Lim and Bellotti [26] among the few case studies. These studies find that CP creates calibrated prediction intervals around a variety of different machine learning methods. Furthermore, Bellotti [4] develop non-conformity functions specifically tailored to create smaller CP sets in the house price prediction context. These studies do not address weighting of the non-conformity scores, or other approaches to offer local validity guarantees.

3 The Oslo data set

We study a data set of transactions from the housing market in Oslo, Norway's capital and largest city. The data set consists of $N = 26\,362$ transactions, each representing a single sale between a buyer and a seller. All transactions are from the years $2016-2017$ and only include transactions of apartments, as opposed to detached or semi-detached homes. The data set consists of the sale price, measured in million Norwegian kroner

(NOK)[1]. The sale price is the transaction price agreed upon by a seller and buyer after negotiations, not the initial ask price. The data set also includes $p = 16$ dwelling-specific covariates. Examples of such covariates are size, number of bedrooms, and floor. There are two variables encoding information about the size of the dwellings: size and gross size. The first encodes the livable area, which includes bedrooms, bathrooms, kitchens, and living rooms. The latter, gross size, refers to the usable area, which includes the livable area and additional areas such as basements or storage rooms. Both are measured in square meters.

The location of the dwelling is encoded via the coordinates (longitude and latitude) but also by a city district variable. There are 15 city districts in Oslo. In addition, the "Lake Distance" and "Coast Distance" variables encode information about the relative location. Similarly, the variables "Homes Nearby" and "Other Buildings Nearby" provide information about the number of residential and commercial buildings close to each dwelling, based on a discretization of Norway into grids of 250×250 meter. A data summary is provided in Table 1.

Variable	Unit	Mean	St. Dev.	Min	Max	Type
Sale Price	NOK (mill.)	4.28	1.84	0.35	27.9	Numerical
Size	m^2	65.68	24.43	12	343	Numerical
Gross Size	m^2	67.20	25.29	12	368	Numerical
Longitude	degrees	10.78	0.06	10.62	10.95	Numerical
Latitude	degrees	59.92	0.03	59.82	59.98	Numerical
City District[1]	-	-	-	-	-	Categorical
Altitude	m	94.93	62.21	0	480	Numerical
Bedrooms	-	1.78	0.76	0	8	Numerical
Floor[2]	-	3.03	2.07	-4	99	Numerical
Age	years	58.55	36.71	0	266	Numerical
Coast Distance	m	3,261	2,453	5	12,201	Numerical
Lake Distance	m	993	504	26	3,018	Numerical
Balcony[3]	-	0.75	0.43	0	1	Categorical
Elevator[3]	-	0.36	0.48	0	1	Categorical
Units On Address[4]	-	20.86	29.12	0	274	Numerical
Homes Nearby[5]	-	2,721	1,586	98	6,746	Numerical
Other Buildings Nearby[5]	-	2,721	1,585	98	6,746	Numerical

[1]There are 15 distinct city districts in the data set.

[2]If the dwelling has multiple floors, this variable will be the lowest floor.

[3]In cases where the information is missing, this is set to 0.

[4]In apartment buildings, multiple dwellings might have the same address.

[5]Norway is divided into squares of 250m×250m. This variable counts the number of homes or other buildings (stores, schools, churches) in all the adjacent squares to the square where the targeted dwelling is, i.e., the 8 neighboring squares.

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

[1]1 NOK ≈ 0.09 USD per December 2023.

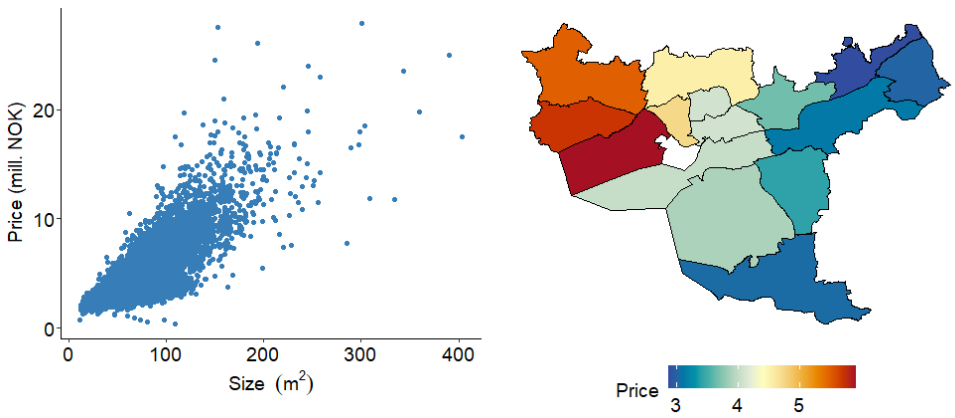

Fig. 1 Visualizations of the Oslo 2016 – 2017 data set. **Left:** Sale price plotted against the size in m^2. **Right:** Mean price per in million NOK per city district. The white city district has no observations.

Figure 1 displays how the sale price varies with the size of the dwelling (left panel) and how the average price varies per city district in Oslo (right panel). These two attributes, the size and the location, are often among the most important characteristics for a home buyer, and typically serve as significant predictors for the final sale price. Not surprisingly, the mean price increases with the size of the dwelling. It should be noted that while homes up to 400 square meters are visualized in Figure 1, more than 90% of the dwellings are 100 square meters or less, and the median size is 62 square meters.

Significant variability is observed across regions, as the northwestern city districts display prices that are, on average, above NOK 5 million. In comparison, the city districts located further east have mean prices around NOK 3 million. The city district colored in white in Figure 1 is Sentrum, a city district without observations in the data set. The absence of transactions in this area reflects that office buildings and tourist attractions dominate this city district with little residential housing.

4 Conformal prediction

CP is a distribution-free and model-agnostic framework for the construction of confidence sets at a pre-specified confidence level $\alpha \in (0, 1)$ for unobserved test data, given a sequence of already observed training data from the same albeit unknown distribution. This section introduces how CP can be used to create confidence sets in regression settings, and presents a literature survey on ways to adjust the CP framework when the test instance is believed to come from a different distribution as the training data.

4.1 Simple illustration

Assume that we have observed a sequence of random variables $Z_{1:N} := (Z_1, ..., Z_N)$ with $Z_i \in \mathbb{R}$ for every i, and seek to create a confidence set for an unobserved $Z_{N+1} \in \mathbb{R}$ given that it comes from the same distribution as $Z_{1:N}$. The CP approach (Vovk et al

[46]) quantifies uncertainty by hypothesizing a value $Z_{N+1} = z$, and then calculating a conformal p-value for this hypothesis. To calculate a p-value without making distributional assumptions about the data-generating process, CP utilizes a *non-conformity* measure that quantifies how dissimilar the hypothesized value z is from the previously observed $Z_{1:N}$. Consider, for instance, the non-conformity measure

$$\Psi_i(Z_i) := |Z_i - \bar{Z}_{-i}|, \tag{1}$$

where $\bar{Z}_{-i}$ is the mean of $\{Z_{1:(N+1)}\}\backslash\{Z_i\}$. A high value of the non-conformity measure for Z_{N+1} indicates that Z_{N+1} is dissimilar to the training data, while a low value indicates a high degree of similarity. Next calculate a non-conformity score for each data point in the sequence $(Z_1, ..., Z_{N+1} = z)$. We denote the realized non-conformity scores $s_1, ..., s_{N+1}$, i.e., $s_i := \Psi_i(Z_i)$ as in (1). Under the null hypothesis that the hypothesized value z is compatible with $Z_{1:N}$, then the rank of its non-conformity score s_{N+1} is uniformly distributed among $s_1, ..., s_n$. A p-value for the hypothesis that z comes from the same distribution as $Z_{1:N}$ can be constructed as

$$p_{N+1} = \frac{1}{N+1} \sum_{i=1}^{N+1} \mathbb{1}\{s_{N+1} \leq s_i\}.$$

We accept the hypothesis if $p_{N+1} > \alpha$ for some confidence level $\alpha \in (0, 1)$. Define $\hat{q}_{1-\alpha}$ to be the $(1-\alpha)$th percentile of $s_1, ..., s_N$, i.e.,

$$P(s_{N+1} \leq \hat{q}_{1-\alpha}) \geq 1 - \alpha$$

under the null hypothesis. We can now invert the conformal p-value over all possible values z and construct a CP set $C_{1-\alpha} \subset \mathbb{R}$ that contains all the hypothesized z values that are accepted at level α based on the conformal p-value. Formally, the CP set is thus

$$C_{1-\alpha} = \left\{z : \Psi_{N+1}(z) \leq \hat{q}_{1-\alpha}\right\}.$$

4.2 Inductive conformal prediction

The procedure described until now, where the mean $\bar{Z}_{-i}$ is recalculated for every i through a hold-out approach, is referred to as Transductive CP (Vovk et al [46]) or Full CP (Lei et al [25]). This method is often computationally infeasible since the mean must be recomputed $N + 1$ times for every hypothesized $Z_{N+1} = z$. A solution to this was suggested in Papadopoulos et al [35] under the name Inductive CP and later re-introduced as Split CP (Lei et al [25]). Split CP circumvents the recalculation of the mean (i.e., in the case of our simple illustration) for every score by splitting the full data set into two parts: A training set of size m and a calibration set of size n, such that $m + n = N$. In this setting, the training set can be used to calculate the mean $\bar{Z}_m$ once, and the non-conformity measure is rewritten as

$$\Psi_i(Z_i) := |Z_i - \bar{Z}_m|,$$

with the training mean $\bar{Z}_m$ replacing the hold-out mean $\bar{Z}_{-i}$ used in the Transductive CP non-conformity score from (1). The reduction in computational complexity comes at the cost of a reduction in statistical efficiency, however, potentially resulting in larger CP sets (Lei et al [25]), since the number of non-conformity scores used in the calibration step is reduced. The mean estimate $\bar{Z}_m$ is is also less efficient since $m < N$. Since smaller CP sets are typically preferred, the Transductive CP approach is generally favorable, but the Inductive CP provides a more practical and computationally quicker alternative that is reasonable in situations where the data set is large enough to set aside a sizeable calibration set.

4.3 Extension to multiple regression case

We now expand to a regression context where the goal is to create a CP set for an unobserved $Y_{N+1} \in \mathbb{R}$ given a feature vector $X_{N+1} \in \mathbb{R}^d$. For notational simplicity we denote $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, and only present the Split CP method. For a detailed review of Full CP for regression, we refer to the tutorial in Shafer and Vovk [39].

As in Split CP, we assume the existence of a training set $(Z_1, ..., Z_m)$ and a calibration set $(Z_{m+1}, ..., Z_N)$. The training set is used to train a prediction model $\hat{f} : \mathbb{R}^d \to \mathbb{R}$, with no particular restrictions on the class of models or the algorithm used to train $\hat{f}$ other than that it is invariant to the order of the training data (Lei et al [25]). The canonical non-conformity measure in this context is $\Psi_i : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, with

$$\Psi_i(Z_i) := |Y_i - \hat{f}(X_i)|$$

being a widely used choice (Vovk et al [46]). Although X_i is now of d dimensions, the non-conformity score is still defined in $\mathbb{R}$.

When a model is trained, and a non-conformity score is defined, we then use these to score all the observations in the calibration set, and obtain realized non-conformity scores $s_1, ..., s_n$. When given an instance $X_{N+1} \in \mathbb{R}^d$ and a pre-specified confidence level $\alpha \in (0, 1)$, the CP set $C_{1-\alpha}(X_{N+1})$ is defined as

$$C_{1-\alpha}(X_{N+1}) = \left\{ y : \Psi_{N+1}(X_{N+1}, y) \leq \hat{q}_{1-\alpha} \right\} \tag{2}$$

where $\hat{q}_{1-\alpha}$ is the $(1 - \alpha)$th empirical quantile of $s_1, ..., s_n$.

In practice, inverting the conformal p-value for every value of y is a computational bottleneck. It is thus demonstrated in Lei et al [25] that when the non-conformity measure is the absolute residual (as described here) and we use the Split CP algorithm, the CP set can be expressed as

$$C_{1-\alpha}(X_{N+1}) = [\hat{f}(X_{N+1}) \pm \hat{q}_{1-\alpha}], \tag{3}$$

reducing the computational burden.[2]

[2] To see how this reduces, observe that $P(s_{N+1} \leq \hat{q}_{1-\alpha}) \geq 1 - \alpha$ under the null hypothesis that Z_{N+1} is exchangeable with $(Z_1, ..., Z_N)$. In the Split CP setting, s_{N+1} does not rely on a recomputation of $\hat{f}$, so this is equivalent to writing $P(|Y_{N+1} - \hat{f}(X_{N+1})|) \leq \hat{q}_{1-\alpha}) \geq 1 - \alpha$, which equates to $P(\hat{f}(X_{N+1}) - \hat{q}_{1-\alpha} \leq Y_{N+1} \leq \hat{f}(X_{N+1}) + \hat{q}_{1-\alpha}) \geq 1 - \alpha$.

4.4 Coverage guarantees

If we create $C_{1-\alpha}(X_{N+1})$ as described in (2), the CP set comes with a *marginal coverage guarantee*, formally expressed as

$$\mathbb{P}\big\{Y_{N+1} \in C_{1-\alpha}(X_{N+1})\big\} \geq 1 - \alpha. \tag{4}$$

CP sets that yield this guarantee are referred to as being marginally valid (Vovk et al [46]). The guarantee in (4) holds regardless of the underlying distribution, as long as the data is exchangeable. A sequence $(Z_1, ..., Z_{N+1})$ is exchangeable if its joint probability density is invariant to permutations on the order.

Although the marginal coverage guarantee is useful, the guarantee is averaged across a whole test set. A more desirable guarantee is *conditional coverage*, i.e.,

$$\mathbb{P}\big\{Y_{N+1} \in C_{1-\alpha}(X_{N+1})\big|X_{N+1} = x\big\} \geq 1 - \alpha.$$

where the coverage holds conditionally for every possible value $X_{N+1} = x$. From a practical point of view, a conditional coverage guarantee is typically more useful than a marginal coverage guarantee, but it is demonstrated in Lei and Wasserman [24] that conditional coverage is only possible if the prediction sets $C_{1-\alpha}(X_{N+1})$ are infinitely large. Since exact conditional coverage guarantees are infeasible, a milder requirement is *approximate conditional coverage*,

$$\mathbb{P}\big\{Y_{N+1} \in C_{1-\alpha}(X_{N+1})\big|X_{N+1} \in \mathcal{X}_k\big\} \geq 1 - \alpha \quad \text{for} \quad k = 1, ..., K \tag{5}$$

where $\mathcal{X}_1, ..., \mathcal{X}_K$ is a partition of the feature space. Variations of (5) are presented in various ways in the literature with slight differences in how the subsets $\mathcal{X}_k$ are constructed. We refer to Foygel Barber et al [12] for a detailed discussion of approximate conditional coverage, including lower bounds on the size of the subsets $\mathcal{X}_k$.

4.5 The role of the non-conformity measure

The choice of non-conformity measure plays an important role in the construction of the CP sets. If $Z_{1:N}$ are exchangeable, as required by the CP procedure, any choice of non-conformity measure leads to marginally valid CP sets (Shafer and Vovk [39]). However, a well constructed non-conformity measure can lead to more *efficient* sets, i.e., sets that are smaller on average. Given the choice between two different procedures for CP set construction that have the same theoretical coverage guarantees, more efficient sets are preferred.

The non-conformity measure proposed so far leads to CP sets of constant size in the Split CP regression setting. In the regression context of (3), a potentially more efficient more adaptive non-conformity measure proposed in Papadopoulos et al [35], is

$$\Psi(Z_i) = \frac{|Y_i - \hat{f}(X_i)|}{\hat{\sigma}(X_i)},$$

where $\hat{\sigma}$ is a normalizing function. With this non-conformity measure, the CP sets take the form

$$C_{1-\alpha}(X_{N+1}) = [\hat{f}(X_{N+1}) \pm \hat{\sigma}(X_{N+1}) \cdot \hat{q}_{1-\alpha}],$$

that is, it is adjusted based on the value of X_{N+1} (Lei et al [25], Papadopoulos et al [35]). As opposed to (3), which creates CP sets of constant width, the $\hat{\sigma}$ function makes CP sets of adaptive size. The function serves as an estimate of the Mean Absolute Deviance of the quantity $Y - \hat{f}(X)|X = x$ (Lei et al [25]), and is referred to as the *difficulty* function in Boström et al [5]. Estimation of the difficulty function may rely on the inherent properties of the considered point predictor (Johansson et al [20], Mao et al [30]) or characteristics of the data set and application at hand (Lim and Bellotti [26]). Alternatively, the function has also been learned directly by fitting a model from the unnormalized non-conformity scores from the training set (Papadopoulos et al [35]).

Instead of using the absolute residual from the point prediction to build the non-conformity score, scores based on conditional quantiles have been proposed in Romano et al [36]:

$$\Psi_i(Z_i) := \max\left\{\hat{Q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{Q}_{1-\alpha/2}(X_i)\right\},$$

where $\hat{Q}_{\alpha/2}$ and $\hat{Q}_{1-\alpha/2}$ are the conditional quantiles from a quantile regression. The CP sets for a test instance are then constructed as $C_{1-\alpha}(X_{N+1}) = [\hat{Q}_{\alpha/2}(X_{N+1}) - \hat{q}_{1-\alpha}, \hat{Q}_{1-\alpha/2}(X_{N+1}) + \hat{q}_{1-\alpha}]$, where $\hat{q}_{1-\alpha}$, as usual, is an empirical percentile of the non-conformity scores. Unlike the other non-conformity scores presented, $\hat{q}_{1-\alpha}$ might take either positive or negative values, resulting in an expansion or shrinking of the conditional quantiles returned from the quantile regression method.

4.6 Approximate conditional coverage through weighted quantiles

Recent years have seen many methodological improvements that seek to achieve approximate conditional guarantees under different circumstances. A straightforward approach was presented in the original works on CP (Vovk et al [46]), where the authors suggest Mondrian CP. The Mondrian CP framework relies on a *Mondrian taxonomy* that maps each data point to one of K classes $\mathcal{X}_k$, and then runs the CP procedure separately on the non-conformity scores in each class $\mathcal{X}_k$. The Mondrian taxonomy is often an inherent feature of the data set, such as different geographical regions. However, it can also be defined in more sophisticated ways by binning feature values or clustering observations. The Mondrian CP procedure achieves class-conditional coverage guarantees when the $\mathcal{X}_1, ..., \mathcal{X}_K$ are disjoint (Vovk et al [46]).

Weighted CP is introduced in Tibshirani et al [43], motivated by a scenario with covariate shift between the training set and test set, i.e., where the feature values are drawn from a different distribution in the training set versus test set. Under this scenario, the test instance Z_{N+1} is no longer exchangeable with $(Z_1, ..., Z_N)$ and not

even the marginal coverage guarantee in (4) holds if CP is applied. The solution presented in Tibshirani et al [43] is to reweight the non-conformity scores based on the likelihood ratio between the training and test set. The authors demonstrate that this results in *weighted exchangeability*, in which case regular CP is marginally valid. A significant drawback with this approach, however, as pointed out in Foygel Barber et al [13], is its dependence on specifications of the exact likelihood ratio, which is not always available.

It is also demonstrated in Tibshirani et al [43] how their approach can be used to achieve approximate conditional coverage in some local neighborhood of around any value $X_{N+1} = x$ by weighing the non-conformity scores with a kernel function that assigns higher weights to non-conformity scores s_i where X_i is close to x (the observed value of the test feature). A similar idea is presented in Guan [15] under the term Localized CP, where the non-conformity scores are also reweighted based on similarity in feature space. A challenge with this framework arises if the number of observations around x is small, which is solved by adjusting the confidence level α individually for each test instance.

A more general framework called Non-exchangeable CP is presented in Foygel Barber et al [13], where the CP method is adjusted for unknown violations of the exchangeability assumption. This method also relies on weighting the non-conformity scores, but an important distinction from the work presented in Tibshirani et al [43] is that the weights are fixed rather than data-dependent. Examples of fixed weights are weights decaying over time (older observations receive less weight) or in space (observations far from the test location receive less weight). Under these conditions, the authors prove that an upper bound on the deviation from the nominal coverage $(1 - \alpha)$, referred to as the coverage gap, can be formulated as

$$\text{Coverage-Gap} \leq \frac{\sum_{i=1}^{N} w_i \cdot \mathrm{d_{TV}}(Z, Z^{(i)})}{1 + \sum_{i=1}^{N} w_i}$$

where $Z^{(i)} := (Z_1, ..., Z_{i-1}, Z_{N+1}, Z_{i+1}, Z_N, Z_i)$, i.e., a permuted version of the original data sequence where the position of Z_i and Z_{N+1} is swapped, and $\mathrm{d_{TV}}$ measures the total variation distance. By assigning low weight to observations that are breaching the exchangeability assumption (evident by a high total variation distance between the original data and the permuted data), we can reduce the coverage gap. Consequently, developing an appropriate weighting scheme is important for the theoretical properties of the CP method to hold.

Spatial CP is another modification of CP, presented in Mao et al [30]. The authors show that, given certain assumptions on the data-generating process, non-conformity scores are asymptotically *locally exchangeable* in a neighborhood around a test instance. Exact conditional coverage properties of the Spatial CP method is established in Mao et al [30] under an infill asymptotics regime. When constructing the CP sets, the Spatial CP method uses a spatially weighted version of the non-conformity scores, with weights $w_i := \exp(-d_{i,N+1}^2/\eta)$, where $d_{i,N+1}$ is the Euclidean distance between non-conformity score i and a test instance, and η is a hyperparameter. The

authors also present a non-smooth version of the same idea, with $w_i = 1$ if $d_{i,N+1}$ is less than some threshold, and $w_i = 0$ otherwise.

Although Weighted CP, Localized CP, Non-exchangeable CP, and Spatial CP differ in the methodology, motivation, and the underlying assumptions, all of the strategies can be formulated as modifications of the original CP procedure in (2) where the quantiles are replaces by weighted quantiles of the non-conformity scores. Mondrian CP also fits into the framework by letting $w_i := 1$ if both Z_i and Z_{N+1} come from the same subset $\mathcal{X}_k$, and $w_i := 0$ otherwise.

4.7 Evaluation metrics

When evaluating CP sets on a test set consisting of N_{test} data points, we evaluate the empirical coverage as

$$\text{Coverage}(Z_{N+1}, ..., Z_{N_{\text{test}}}; \alpha) := \frac{1}{N_{\text{test}}} \sum_{i=N+1}^{N_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\},$$

or the coverage gap between the empirical and theoretical coverage as

$$\text{Coverage-Gap}(Z_{N+1}, ..., Z_{N_{\text{test}}}; \alpha) := (1-\alpha) - \text{Coverage}(Z_{N+1}, ..., Z_{N_{\text{test}}}; \alpha),$$

such that a positive coverage gap indicates too low empirical coverage and a negative coverage gap indicates too high empirical coverage.

We will expect to see greater variability in the empirical coverage when either the number of calibration instances and/or the number of test instances is low. In an extreme scenario with only one test instance, the empirical coverage is either 1 or 0. It is known (Angelopoulos and Bates [1], Vovk [45]) that the coverage follows the beta-binomial distribution:

$$\begin{aligned} \text{Coverage}(Z_{N+1}, ..., Z_{N_{\text{test}}}; \alpha) &\sim \frac{1}{N_{\text{test}}} \text{Binom}(N_{\text{test}}, \mu) \\ \mu &\sim \text{Beta}(n+1-l, l), \end{aligned}$$

where $l = \lfloor \alpha(n+1) \rfloor$. This formulation might seem abstract at first, but is a useful tool to quantify the variability in the coverage gap as a function of the size of the calibration set n, the size of the test set N_{test}, and the confidence level α. Accordingly, for practical prediction problems, assuming the data are exchangeable, the coverage follows a known distribution, which can be used as a diagnostic tool when comparing the empirical coverage with the theoretical level of $1-\alpha$.

Additionally, practitioners are often interested in studying the size of the CP sets, sometimes referred to as the *efficiency*. The efficiency of a CP set can either be measured in absolute terms, or as a percentage of the true response, referred to as *relative efficiency*.

5 Results

The theoretical guarantees of the CP framework rely on the exchangeability of the non-conformity scores, which is not always satisfied for real data. In this section, we design and present the results of a simulation study with synthetic sale prices to compare how the various CP and weighted CP methods perform in an idealized setting where the data-generating mechanism is known.

5.1 Simulation setup

We consider a data-generating mechanism where the sale prices $Y = (Y_1, ..., Y_N)$ are drawn simultaneously from a joint Gaussian distribution with a covariance matrix that encodes covariance between two sale prices based on their spatial distance. Formally, let $Y \sim \mathcal{N}_N(X\beta, \Sigma)$, where $X \in \mathbb{R}^{N \times p}$, $\beta \in \mathbb{R}^p$, and $\Sigma \in \mathbb{R}^{N \times N}$ with components $\Sigma_{ij} = \sigma_\varepsilon^2 + \sigma^2 \cdot e^{-\rho|L_i - L_j|}$, for $i, j \in \{1, \ldots, N\}$, where L_i and L_j are the spatial location of observation i and j, and σ_ε^2, σ^2, and ρ are parameters that determine the range and magnitude of the covariance, as well as observational noise.

We obtain maximum likelihood estimates (MLE) of the parameters following the approach presented in Schabenberger and Gotway [38] (Section 5.2.2). We find an MLE of $\hat{\rho} = 4.0$, resulting in a covariance kernel that halves the covariance between spatial locations after 170 meters. Accordingly, the spatial covariance is limited to a relatively small local neighborhood.

We then generate synthetic responses based on the MLE parameters and the actual feature values and locations in the Oslo data. To ease the computational burden we study a subset of $N = 6\,000$ data points that are split into equally sized sets for training, calibration, and testing. This process is repeated 10 times for our simulation study. For each of the 10 synthetic data sets in our simulation study, we construct non-conformity scores using predictions from fitting a gradient boosted trees model (via the XGBoost package presented in Chen and Guestrin [10]) with $M = 1\,000$ trees with a maximum depth of 4 in each tree and a learning rate of 0.03, a random forests model (via the ranger implementation presented in Wright and Ziegler [48]) with $M = 500$ trees, a linear regression, and a spatial regression. The latter method is an Oracle prediction method in the sense that it is based on the true data-generating mechanism, however, it does not have access to the parameter values used to generate the data but estimates them as part of the training process.

Our simulation study compares the following non-conformity scores, as described in Subsection 4.5:

- Standard: The non-conformity score is the absolute residual, $|Y_i - \hat{f}(X_i)|$.
- Normalized: The non-conformity score is the normalized absolute residuals. We experiment with three different normalization functions $\hat{\sigma}$: The prediction $\hat{f}(X_i)$ itself, as proposed in Lim and Bellotti [26], the city district average price, and $\hat{\sigma}$ as a linear function of the features, learned from the training data. We refer to these as Normalized 1, Normalized 2, and Normalized 3, respectively.
- CQR: The non-conformity scores are derived from a quantile regression forests model (Meinshausen [32]). We only display this non-conformity score with the random forests model.

For the spatial regression we also construct an Oracle non-conformity score that utilize the covariate matrix, i.e., $\Psi_i = \{|\Sigma^{-1/2}(Y - X^\top \beta)|\}_i$. Furthermore, we consider five variants of CP:

- **CP**: The regular unweighted version of CP.
- **Mondrian CP (MCP)**: The unweighted version of CP, but calibrated per city district.
- **Spatial CP (SCP)**: Weighted CP with weights $w_i = \exp(-d^2_{i,N+1}/\eta)$. We choose η such that the weight is halved after approximately 800 meters.
- **Nearest Neighbor CP (NNCP)**: A non-smoothed version of SCP, where $w_i = 1$ only in a local neighborhood of size $D = 1$ kilometer around X_{N+1}.
- **Feature Weighted CP (FWCP)**: Weighted CP with w_i derived from a kernel in feature space. Calculating feature distances via kernel methods in high-dimensional space is computationally challenging (Liu et al [28]). We, therefore, approximate the feature distance via a random forests kernel (Breiman [6], Lin and Jeon [27]). This estimates feature distance by utilizing the partitions of the feature space found in a random forests model. We train a separate model that only uses the feature values, not the spatial coordinates, and use a maximum tree depth of six to ensure sufficient observations in each partition.

5.2 Results on synthetic data

Before considering the calibration of the CP sets, Table 2 displays the prediction accuracy of the point prediction models on a data set with synthetically drawn sale prices. The metrics displayed are Root Mean Squared Error (RMSE), Median Absolute Error (MdAE), and the Percentage Error Range (PER), which measures the fraction of errors that are within some user-specified range, for instance ±10% or ±20%. This metric is widely used in the AVM industry (Steurer et al [42]). High values of the PER metric are desirable.

The table shows that the spatial and linear regression performs better than the tree-based models on all metrics. Given the low spatial covariance used in the data-generating process, it is not surprising that the linear regression performs similarly to the spatial regression.

Model	RMSE	MdAE	PER10 (%)	PER20 (%)
Gradient boosted trees	1.33	0.89	25.1	47.5
Random forests	1.33	0.89	25.2	48.3
Linear regression	1.25	0.85	27.2	50.2
Spatial regression	1.25	0.85	27.2	50.3

Table 2 Accuracy of the point predictions in the scenario with synthetically generated sale prices from a linear model with spatial covariance with $\rho = 4.0$. The results are averaged over 10 simulations, each time with $N = 6\,000$ observations in the total data set. The table displays Root Mean Squared Error (RMSE), Median Absolute Error (MdAE), and Percentage Error Range (PER10 and PER20).

Figure 2 summarizes the coverage and efficiency of CP sets at confidence level 0.9 for these four point prediction methods, averaged over 10 simulations. The upper plot shows the empirical coverage gap on a test set, with a positive value indicating that the empirical coverage is slightly below 90%. The lower plot shows the relative efficiency as a percentage of the observed sale price, with lower values indicating smaller CP sets. As a benchmark for the empirical coverage, we include dotted lines indicating the 5th and 95th percentile of the theoretical coverage distribution of a CP set, which follows a known beta-binomial distribution.[3] Figure 2 indicates that all methods construct marginally valid CP sets, as the coverage gap for all methods are within the theoretical beta-binomial coverage distribution bounds. The coverage gap of the Oracle non-conformity score is highlighted in the rightmost panel. The linear regression and spatial regression methods tend to have the best relative efficiency, which is not surprising since these methods are close to the underlying data-generating mechanism.

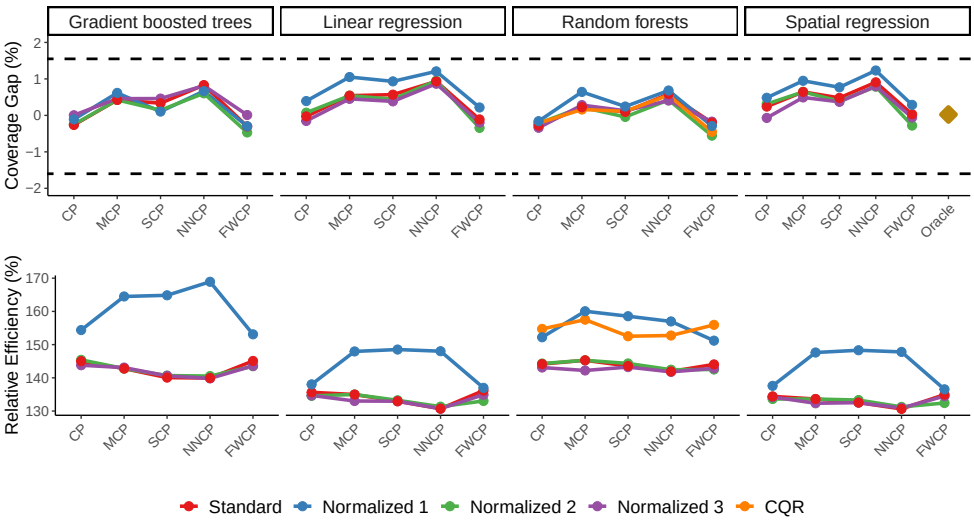

Fig. 2 The performance of CP sets at $1 - \alpha = 0.9$ on synthetic data set. These results are averaged over 10 simulations, each on a subset of $N = 6\,000$ observations. **Top**: Empirical coverage gap from the desired $1 - \alpha$. **Bottom**: Size of the CP set as a percentage of the (synthetic) sale price.

Figure 3 shows the conditional coverage gap per city district for the spatial regression model. Each city district is colored with its empirical coverage gap on a test set for a combination of non-conformity scores (horizontal) and weighting methods (vertical). This gives more insight into how calibrated the CP sets are within different parts of Oslo. It is desirable to achieve a coverage gap close to zero not only marginally but also within each city district. The maps demonstrate that the Standard CP method in the upper left corner achieve a coverage gap close to zero consistently

[3] Bounds are based on the $\frac{1}{2\,000}\text{Binom}(2\,000, \mu)$, where $\mu = \text{Beta}(1\,801, 200)$, but shifted by $1 - \alpha$ since we display the coverage gap rather than the coverage.

across the city districts, while the NORMALIZED 1 and NORMALIZED 2 non-conformity scores display some spatial variability. However, the CP methods that use some form of geographically local calibration, like MCP, SCP, or NNCP, smooth out these differences and achieve a more consistent coverage gap across city districts for all the non-conformity scores. The fact that the STANDARD non-conformity score does so well can be attributed to the low spatial covariance in the synthetic data.

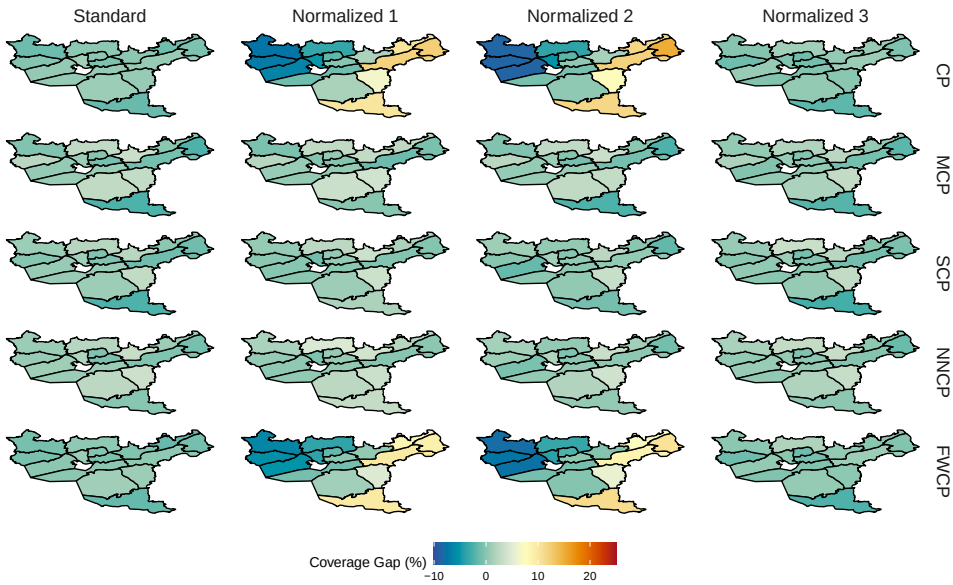

Fig. 3 Coverage gap per city district for the spatial regression. The map shows the performance for different non-conformity measures (horizontally) and weighting methods (vertically).

To summarize the map shown in Figure 3 and similar maps for every point prediction method, we present in Figure 4 a box plot of the city district conditional coverage gaps for each method. The box plot shows the interquartile range of the coverage gap in each of the 15 city districts, with a low range and a center around zero being preferable. For comparison, we include dotted lines that mark a beta-binomial distribution's first and third quantiles with a calibration set of 2 000 observations and a test set of 2 000/15 observations.

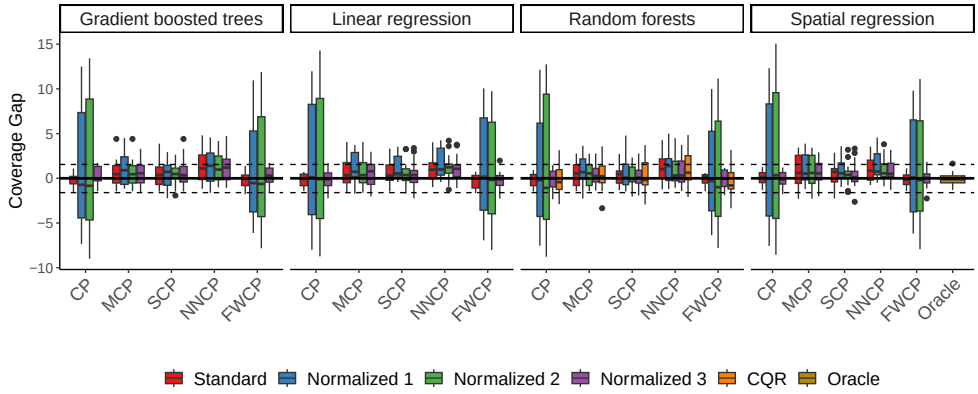

Fig. 4 Box plots of the coverage gap observed in each of the 15 city districts. The box marks the first and third percentile of observations, while the dotted lines show the first and third quantiles of the corresponding beta-binomial distribution.

Overall, the results on the synthetic data demonstrate little apparent improvement in performance with the use of weighting methods versus the Standard non-conformity score, indicating that the absolute residuals are close to globally exchangeable when the sale prices are synthetically generated from the spatial regression model first fit to the real Oslo data set. Since the fitted covariance matrix is so close to diagonal, it is likely that this canonical spatial model is not able to adequately capture geographic dependencies in the real Oslo data set. Accordingly, this is further evidence for the potential utility of weighted CP approaches, i.e., when an adequate spatial model is not known. This simulation study also indicates that the Oracle method performs as expected, lending confidence to the results presented in the next section.

5.3 Results on the Oslo data set

We now repeat the analysis in Section 3 with actual sale prices instead of synthetically generated ones, such that 10 subsets of size $N = 6\,000$ observations are each randomly sampled from the real Oslo data set. Results on the complete Oslo data set for a subset of the methods are available in Appendix A.

Table 3 displays the predictive accuracy for the actual sale prices. Here, the tree-based models perform significantly better than the linear regression and spatial regression, possibly due to the non-linearity exhibited between variation in sale prices and covariates in the actual data set. Further, with the RMSE being significantly smaller for all models for the real data versus for the synthetic data sets, we anticipate a higher efficiency of the CP sets.

The coverage gap and relative efficiency are displayed in Figure 5. The coverage gaps are slightly above zero for certain methods but within the theoretical bounds for all methods. The relative efficiency is, as expected, generally improved here versus the synthetic data case, and the tree-based models have the most efficient CP sets in this setting. Again, the spatial regression and linear regression display quite similar

Model	RMSE	MdAE	PER10 (%)	PER20 (%)
Gradient boosted trees	0.73	0.31	59.2	88.2
Random forests	0.75	0.32	58.0	87.3
Linear regression	0.83	0.42	47.0	76.7
Spatial regression	0.84	0.43	46.8	76.3

Table 3 Accuracy of the point predictions on the real data. The table displays Root Mean Squared Error (RMSE), Median Absolute Error (MdAE), and Percentage Error Range (PER10 and PER20).

performance, most likely due to the structure of the estimated spatial covariance matrix. For every point prediction, the Standard CP has a coverage gap close to zero.

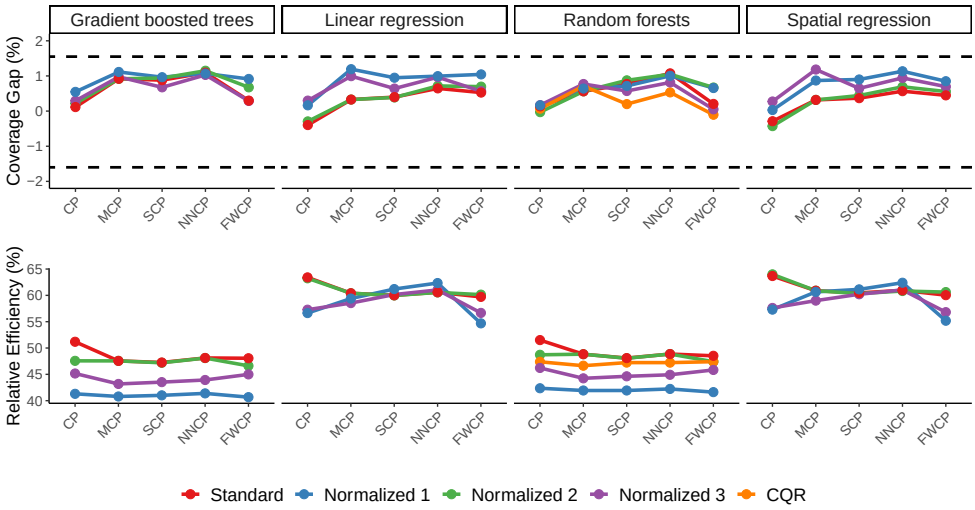

Fig. 5 The performance of the CP sets at $1 - \alpha = 0.9$ on the real data set. The results are averaged over 10 simulations, each on a subset of $N = 6\,000$ observations. **Top**: Empirical coverage gap from the desired $1 - \alpha$. **Bottom**: Size of the CP set as a percentage of the actual sale price.

In Figure 6 we present heat maps of the coverage gap in each city district, but this time for the random forests point prediction since the tree-based models achieve the highest accuracy on the real data set. The Standard CP shows substantial spatial variability in the coverage gap, with the city districts in the west having a positive coverage gap and the city districts in the east having a negative coverage gap. This phenomenon reflects that in the actual data the magnitude of the absolute residuals systematically differ over spatial regions, with the west having larger non-conformity scores on average than the east. Normalizing the non-conformity scores via the Normalized 1, 2, 3 non-conformity measures helps to improve this somewhat, but not entirely, and similar geographical patterns are observed.

The methods that rely on calibrating the CP sets based on a local neighborhood around each test instance – MCP, SCP, and NNCP – display a much more consistent coverage gap between the different city districts. This makes it reasonable to assume that the non-conformity scores on the real data set have an element of *local exchangeability*, even if the scores are not globally exchangeable. Interestingly, these weighted CP methods achieve a consistent coverage gap across city districts even for the STANDARD non-conformity scores.

Again, we summarize the city district conditional coverage in box plots in Figure 7. The regular CP method consistently has the largest variance in the city district conditional coverage, regardless of the non-conformity score. The three CP methods that rely on spatially weighted CP all display less variability, indicating a more consistent coverage gap across city districts. This is further evidence that the non-conformity scores are not globally exchangeable when we use actual sale prices, and that spatially weighted scores are effective for constructing consistently calibrated CP sets for each city district.

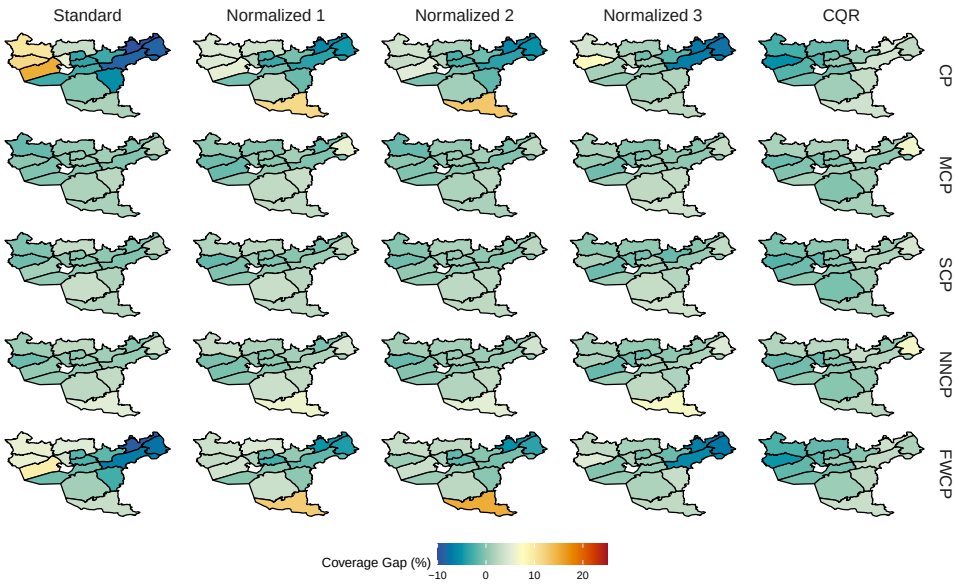

Fig. 6 Coverage gap per city district for the random forests method on the actual sale prices from the Oslo data set. The map shows the performance for different non-conformity measures (horizontally) and weighting methods (vertically).

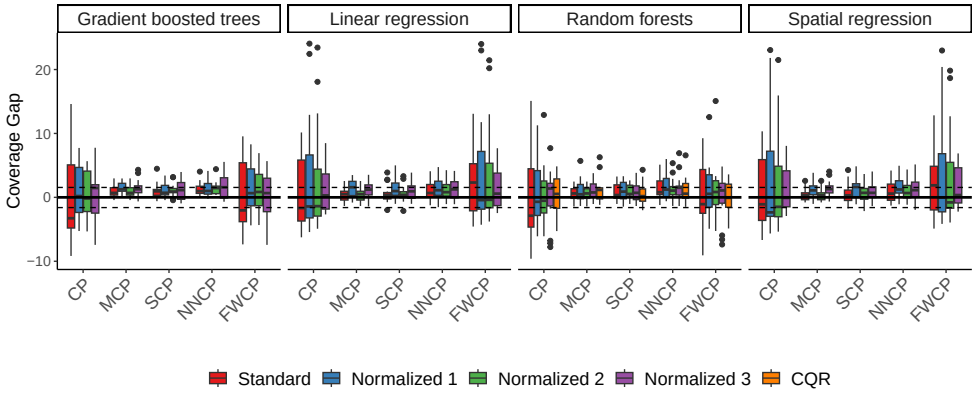

Fig. 7 Box plots of the coverage gap observed in each of the 15 city districts. The box marks the first and third percentile of observations, while the dotted lines show the first and third quantiles of the corresponding beta-binomial distribution.

6 Discussion

We have studied various point predictions, non-conformity scores, and weighting methods in this research, leaving practitioners with many engineering choices. Should practitioners focus on crafting a better non-conformity score for the application at hand so that the non-conformity scores become as close as possible to exchangeable, or should the weighting method adjust for the (potential) lack of exchangeability in the scores? Under ideal conditions, weighting the non-conformity scores should not be necessary since the CP framework, by definition, is only valid if the non-conformity scores are exchangeable. However, in practical applications with unknown data-generating mechanisms, the non-conformity scores are not necessarily exchangeable despite our best efforts at normalizing them, and weighting the non-conformity scores can serve as a solution to make the CP sets valid.

The relationship between the choice of non-conformity score and weighting method remains largely unanswered in the literature, but the results presented here might give some indications in practice. A good choice of non-conformity score will drastically improve the performance of the CP sets on a test set. However, making the non-conformity scores exchangeable is not always possible or straightforward in practical applications where the data-generating mechanism is unknown. The locally weighted CP methods perform well in the scenarios we consider, making them valuable tools.

References

[1] Angelopoulos AN, Bates S (2023) Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591

[2] Angelopoulos AN, Bates S, Jordan M, et al (2021) Uncertainty sets for image classifiers using conformal prediction. In: International Conference on Learning Representations

[3] Bailey MJ, Muth RF, Nourse HO (1963) A regression method for real estate price index construction. Journal of the American Statistical Association 58(304):933–942

[4] Bellotti A (2017) Reliable region predictions for automated valuation models. Annals of Mathematics and Artificial Intelligence 81(1):71–84

[5] Boström H, Linusson H, Löfström T, et al (2017) Accelerating difficulty estimation for conformal regression forests. Annals of Mathematics and Artificial Intelligence 81:125–144

[6] Breiman L (2000) Some infinity theory for predictor ensembles. Tech. Rep. 579, Statistics Department, UC Berkeley

[7] Breiman L (2001) Random forests. Machine Learning 45:5–23

[8] Breiman L, Friedman J, Stone C, et al (1984) Classification and Regression Trees. Taylor & Francis, Wadsworth, New York

[9] Candès E, Lei L, Ren Z (2023) Conformalized survival analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology 85(1):24–45

[10] Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

[11] Dey N, Ding J, Ferrell J, et al (2022) Conformal prediction for text infilling and part-of-speech prediction. The New England Journal of Statistics in Data Science 1(1):69–83

[12] Foygel Barber R, Candès EJ, Ramdas A, et al (2020) The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA 10(2):455–482

[13] Foygel Barber R, Candès E, Ramdas A, et al (2023) Conformal prediction beyond exchangeability. The Annals of Statistics 51(2):816 – 845

[14] Gourley P (2021) Curb appeal: how temporary weather patterns affect house prices. The Annals of Regional Science 67(1):107–129

[15] Guan L (2022) Localized conformal prediction: a generalized inference framework for conformal prediction. Biometrika 110(1):33–50

[16] Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA

[17] Hjort A, Pensar J, Scheel I, et al (2022) House price prediction with gradient boosted trees under different loss functions. Journal of Property Research

39(4):338–364

[18] Ho WKO, Tang BS, Wong SW (2020) Predicting property prices with machine learning algorithms. Journal of Property Research

[19] van Hoenselaar F, Cournède B, Pace FD, et al (2021) Mortgage finance across oecd countries. Tech. Rep. 1693, Organization for Economic Cooperation and Development

[20] Johansson U, Boström H, Löfström T, et al (2014) Regression conformal prediction with random forests. Machine Learning 97:1–22

[21] Kim J, Won J, Kim H, et al (2021) Machine-learning-based prediction of land prices in Seoul, South Korea. Sustainability 13(23):202–211

[22] Kiyota Y (2021) Frontiers of Computer Vision Technologies on Real Estate Property Photographs and Floorplans, Springer Nature Singapore, Singapore, pp 325–337

[23] Krause A, Martin A, Fix M (2020) Uncertainty in automated valuation models: Error-based versus model-based approaches. Journal of Property Research 37(4):308–339

[24] Lei J, Wasserman L (2014) Distribution-free prediction bands for non-parametric regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76

[25] Lei J, Rinaldo A, Tibshirani RJ, et al (2018) Distribution-free predictive inference for regression. Journal of the American Statistical Association 113(523):1094–1111

[26] Lim Z, Bellotti A (2021) Normalized nonconformity measures for automated valuation models. Expert Systems with Applications 180:115–165

[27] Lin Y, Jeon Y (2006) Random forests and adaptive nearest neighbors. Journal of the American Statistical Association 101(474):578–590

[28] Liu H, Lafferty J, Wasserman L (2007) Sparse nonparametric density estimation in high dimensions using the rodeo. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, pp 283–290

[29] Maltoudoglou L, Paisios A, Papadopoulos H (2020) Bert-based conformal predictor for sentiment analysis. In: Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research, pp 269–284

[30] Mao H, Martin R, Reich BJ (2023) Valid model-free spatial prediction. Journal of the American Statistical Association 0(0):1–11

[31] Marques J, Batista P, Castro E, et al (2021) Spatial Automated Valuation Model (sAVM) - From the Notion of Space to the Design of an Evaluation Tool, vol 12952, pp 75–90

[32] Meinshausen N (2006) Quantile regression forests. Journal of Machine Learning Research 7(35):983–999

[33] Mohd T, Masrom S, Johari N (2019) Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia. Int J Recent Technol Eng 8(2):542–546

[34] Nouriani A, Lemke L (2022) Vision-based housing price estimation using interior, exterior & satellite images. Intelligent Systems with Applications 14:200081

[35] Papadopoulos H, Proedrou K, Vovk V, et al (2002) Inductive confidence machines for regression. In: Machine Learning: ECML 2002, pp 345–356

[36] Romano Y, Patterson E, Candes E (2019) Conformalized quantile regression. In: Advances in Neural Information Processing Systems

[37] Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy 82(1):34–55

[38] Schabenberger O, Gotway C (2004) Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, Boca Raton, Florida

[39] Shafer G, Vovk V (2008) A tutorial on conformal prediction. Journal of Machine Learning Research 9:371–421

[40] Sommervoll DE (2020) Jump bids in real estate auctions. Journal of Housing Economics 49:101713

[41] Stankeviciute K, M. Alaa A, van der Schaar M (2021) Conformal time-series forecasting. In: Advances in Neural Information Processing Systems

[42] Steurer M, Hill RJ, Pfeifer N (2021) Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research 38(2):99–129

[43] Tibshirani RJ, Foygel Barber R, Candes E, et al (2019) Conformal prediction under covariate shift. In: Advances in Neural Information Processing Systems

[44] Vazquez J, Facelli J (2022) Conformal prediction in clinical medical sciences. Journal of Healthcare Informatics Research 6(3):241–252

[45] Vovk V (2012) Conditional validity of inductive conformal predictors. In: Proceedings of the Asian Conference on Machine Learning, pp 475–490

[46] Vovk V, Gammerman A, Shafer G (2005) Algorithmic Learning in a Random World. Springer-Verlag, Berlin, Heidelberg

[47] Wang C, Wu H (2018) A new machine learning approach to house price estimation. New Trends in Mathematical Sciences 6(4)

[48] 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

[49] Zhou T, Clapp JM, Lu-Andrews R (2022) Examining omitted variable bias in anchoring premium estimates: Evidence based on assessed value. Real Estate Economics 50(3):789–828

Appendix A Results on the full data set

Here we report the most important results for the full data set of $N = 26\,362$. We only report the results with linear regression, random forests, and gradient boosted trees as point predictors, as the optimization of the spatial regression model is a computational bottle neck with large N. We split the data set into three equally large sets for training, calibration, and testing, and run the CP procedures in the same manner as in the main paper.

Table A1 Accuracy of the point predictions on the full Oslo data set ($N = 26\,362$), averaged over 10 simulations.

Model	RMSE	MdAE	PER10 (%)	PER20 (%)
Gradient boosted trees	0.66	0.28	61.3	90.0
Random forests	0.68	0.29	61.7	89.8
Linear regression	0.81	0.39	50.5	79.8

Table A1 shows the prediction accuracy of the point predictors on the full data set. The results are aligned with the results presented in Section 5 on a subset of the real data, with the tree-based models outperforming ordinary least squares regression.

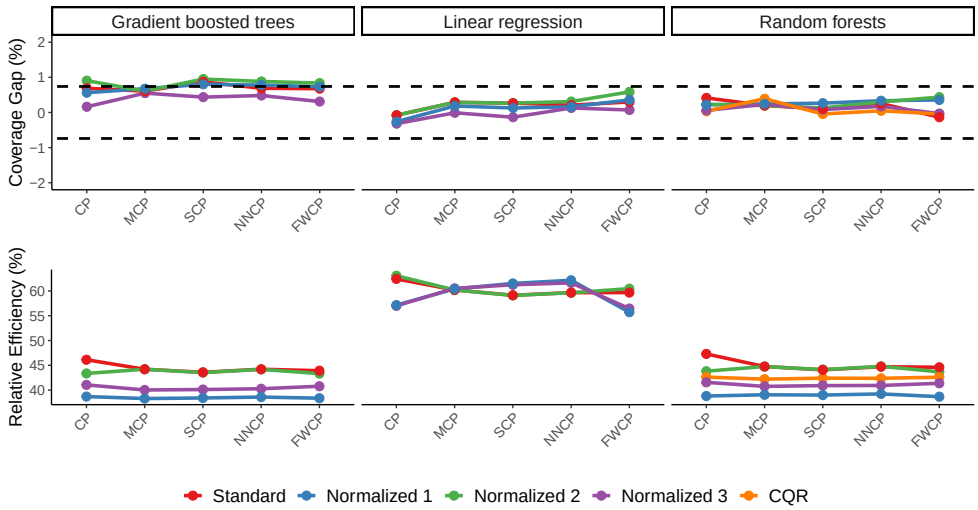

Fig. A1 The performance of CP sets at $(1-\alpha) = 0.9$ on the real data set. These results are averaged over 10 simulations, each with a new split of training, calibration, and test set. **Top**: Empirical coverage gap from the desired $(1-\alpha)$. **Bottom**: Size of the CP set as a percentage of the actual sale price.

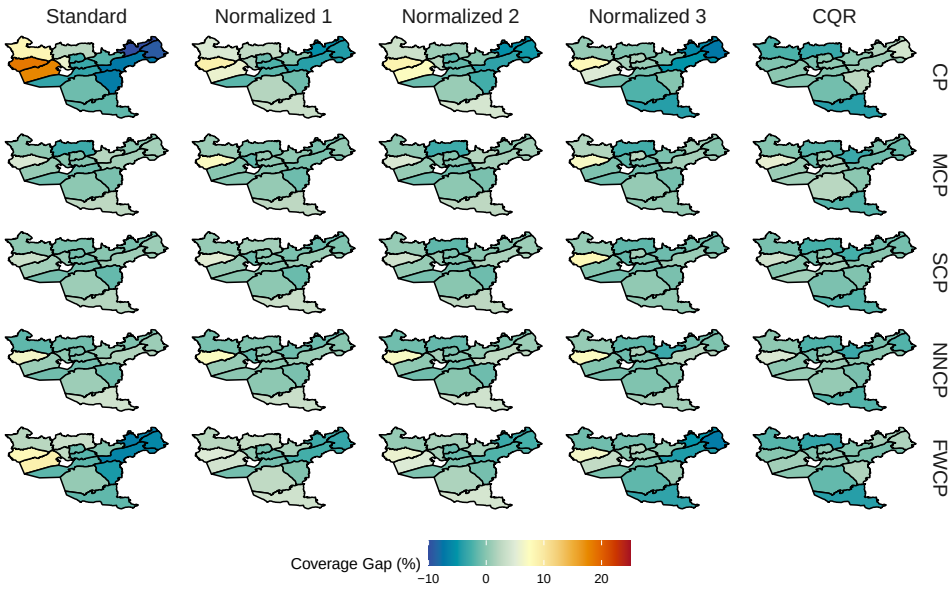

Fig. A2 Map of the coverage gap per city district for the random forests on the full data set.

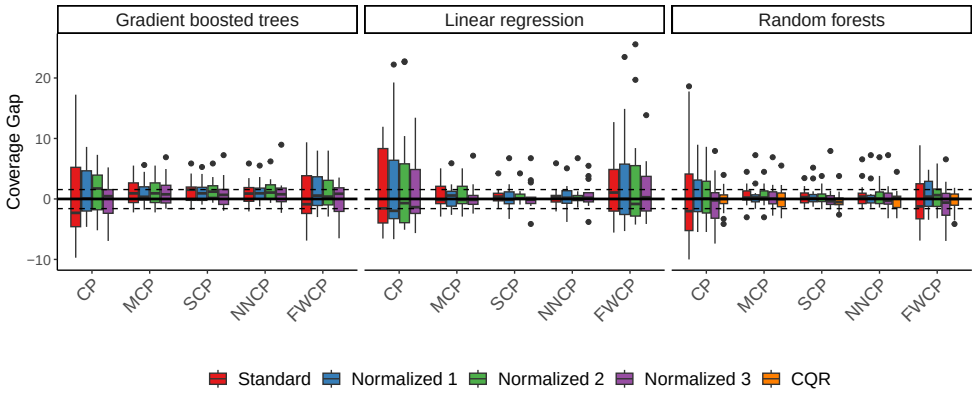

Fig. A3 Box plots of the coverage gap observed in each of the 15 city districts. The box marks the first and third percentile of observations, while the dotted lines show the first and third quantiles of the corresponding beta-binomial distribution.

Figure A2 shows the coverage gap per city district for the random forests model. A box plot with all city district conditional coverages are displayed in Figure A3. Figure A1 shows the coverage gap and relative efficiency of the methods studied on the full Oslo data set. The upper plot has dotted lines indicating the 5th and 95th percentile of a beta-binomial with a third of the full data set set in both calibration and test set. The results are consistent with the results presented in Section 5.