Proceedings of Machine Learning Research 230:1–21, 2024 Conformal and Probabilistic Prediction with Applications

Clustered Conformal Prediction for the Housing Market

Anders Hjort ANDERDH@MATH.UIO.NO
Eiendomsverdi AS & Department of Mathematics, University of Oslo, Oslo, Norway
Jonathan P. Williams JWILLI27@NCSU.EDU
Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA
Johan Pensar JOHANPEN@MATH.UIO.NO
Department of Mathematics, University of Oslo, Oslo, Norway

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Conformal prediction (CP) is a framework for constructing confidence sets around predictions from machine learning models with finite sample guarantees with few assumptions on both the prediction model and the data. In practice, the construction of CP sets typically relies on quantile estimates from an empirical distribution of non-conformity scores. When the data set consists of predefined, non-overlapping classes such as geographical regions, a common technique for improving the confidence sets is to calculate a different quantile for each class. However, the classwise quantile estimate suffers from high variance when the number of observations in each class is low. To circumvent this, one can share calibration data between classes with similar empirical distributions of non-conformity scores to reduce the variance of the quantile estimate. We study this approach for the application of house price prediction in the Norwegian housing market, where 286 different municipalities serve as the initial classes of the data. We find that clustering together municipalities based on non-conformity score distributions, agnostic of the spatial distance between them, leads to CP sets that achieve, on average, a lower coverage gap in each municipality, in particular for the municipalities with few observations.

Keywords: Mondrian Conformal Prediction, Automated Valuation Models, Real Estate, Prediction Intervals

1. Introduction

Non-parametric machine learning methods like random forests and gradient-boosted trees are increasingly popular for prediction tasks with tabular data due to their flexibility and accuracy. To increase the trustworthiness and usefulness of the predictions in practical applications, it is often desirable to provide not only a point prediction but also a set of possible values with a confidence level, referred to as a confidence set. Conformal prediction (CP; Vovk et al. [2005]) is a distribution-free uncertainty quantification tool to construct such confidence sets with finite sample guarantees with very few assumptions on the underlying prediction model. The CP framework uses a non-conformity measure to quantify how unusual (or non-conforming) a data point is compared to a set of previously made observations. Quantiles of an empirical distribution of non-conformity scores are then used to construct confidence sets for test observations. Under the assumption of exchangeability between previously observed data and a new test instance, the CP sets are theoretically valid, i.e., for any user-specified confidence level $(1 - \alpha) \in (0, 1)$ the probability of a new observation being excluded from the CP set is upper bounded by α.

The exchangeability assumption might be violated in many real-world scenarios due to, for instance, distributional drifts over time, spatial trends in the data, or because the non-conformity scores behave differently in different subsets of the feature space. In particular, when the data set of interest consists of several classes or categories, for instance, different age groups or geographical regions, it might be reasonable to assume exchangeability within a given class but not necessarily globally. To account for this, class-conditional approaches have been suggested (Vovk et al. [2005]; Vovk [2012]). In these approaches, the prediction sets are calibrated per class rather than globally, leading to more adaptive confidence sets with theoretical guarantees also within each class.

A challenge with the direct applications of classwise approaches occurs when the number of observations in some or all of the classes is low, making it necessary to share data between similar classes to reduce the variance of the estimated quantile used to calibrate the confidence sets. At the same time, sharing data between classes with different empirical distributions of non-conformity scores induces a bias in the estimator, making it desirable to cluster classes with a similar distribution of non-conformity scores. This approach, referred to as Clustered CP in Ding et al. (2023), allows us to share data between classes with the distributional similarity between the empirical non-conformity scores being used as the similarity measure guiding the clustering.

In this work, we study the Clustered CP approach in the realm of the housing market, where the classes are municipalities in Norway. The overarching goal is to use CP to construct a confidence set for a sale price, given a set of features describing the dwelling. It is known that both the nominal level of house prices and also the number of transactions vary significantly between different geographical regions. These factors make the construction of CP sets challenging, as regular CP tends to struggle to capture the complex spatial patterns in the prices, whereas classwise approaches, on the other hand, fail in regions with few observations.

The sparsity of data in certain parts of the considered region also makes locally weighted quantile estimates (Guan [2022]; Tibshirani et al. [2019]; Mao et al. [2023]) less robust, even though it is shown to work well on housing data exclusively from urban areas in Hjort et al. (2023). Furthermore, in the study of an entire national housing market, it is not obvious that a locally weighted CP version is the optimal choice, as it discards data from regions that are geographically far away, albeit still similar in many ways due to economic, cultural, or demographic reasons. For instance, the housing market dynamics might be similar in two university towns, even if they are spatially far apart. Patterns like this will be overlooked by a locally weighted CP but might be captured by a more data-driven approach based on the distribution of non-conformity scores in each class. The rest of the paper is structured as follows. We introduce the CP framework in Section 2. In Section 3, we conduct a simulation study, and in Section 4, we study a data set collected from the Norwegian housing market. In Section 5 we provide some concluding remarks.

1.1. Related methods and contributions

The seminal work on CP is the book by Vovk et al. (2005), which was complemented in Shafer et al. (2008). Split CP was introduced in Papadopoulos et al. (2002) to reduce the computational complexity of the method. Several methodological advances have been made

to adjust the CP framework for breaches of the exchangeability assumptions, including Mondrian CP (Vovk et al. [2005]), classwise CP (Vovk [2012]), approaches that account for known covariate shift (Tibshirani et al. [2019]), unspecified distribution drift (Foygel Barber et al. [2023]), spatial trends (Mao et al. [2023]), and for a time series setting (Xu et al. [2023]). Clustered CP was proposed in Ding et al. (2023) for classification tasks with many classes and limited data per class and serves as the starting point for our research.

An application of CP to the housing market is presented in Bellotti (2017), which demonstrates that the CP sets are calibrated when applied to a data set from the London housing market. Furthermore, Lim et al. (2021) build upon this work and develop several non-conformity scores tailored to account for the strong correlation between the sale price and absolute residuals that is often observed in housing data. It is also demonstrated that combining multiple prediction models yields narrower CP sets, as the absolute residuals are, on average, lower. A case study from the San Francisco (US) area is conducted in Bastos et al. (2024), where conformal methods based on quantile regression is found to perform best. In Hjort et al. (2023), the focus is on applying various versions of weighted CP to a data set from Oslo, Norway. The primary motivation for this is to account for spatial trends in the non-conformity scores to ensure a low coverage gap in various geographical subsets.

Our main contribution is to adopt the Clustered CP approach to a new setting, namely a regression setting where the initial classes are geographical regions rather than an image recognition task, as studied in Ding et al. (2023). We demonstrate on a data set from the Norwegian housing market that Clustered CP, in many scenarios, constructs confidence sets that yield a lower Mean Absolute Coverage Gap per municipality, effectively making the confidence sets more trustworthy. Although the clustering in this research is merely a means to achieve better-calibrated confidence sets, the clustering itself can be of independent interest to the housing market literature, in particular, related to the literature on housing markets segmentation and identification of homogeneous submarkets (Å. Sommervoll et al. [2019]; D. E. Sommervoll [2023]; Goodman et al. [1998]).

2. Conformal prediction

We consider a supervised regression setting with features $X \in \mathcal{X}$ and response $Y \in \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. Furthermore, we assume that we have access to a (pre-trained) prediction model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that can be used to make predictions about Y given X.

Conformal prediction (CP; Vovk et al. [2005]) is a distribution-free method to construct a confidence set around the point prediction $\hat{f}(X)$ with limited assumptions on the choice of the prediction model. We will present Split CP, a computationally efficient version proposed in Papadopoulos et al. (2002). We refer to Shafer et al. (2008) for a tutorial on Full CP, which is the initial formulation of the CP idea. CP uses a non-conformity function that quantifies any data point's strangeness or non-conformity compared with a bag of already observed examples. The non-conformity score can be used to calibrate the expectations for a new unobserved example and, in turn, construct a prediction set.

While many choices of non-conformity functions can be made depending on the particular problem, an intuitive and widely used choice of non-conformity score for regression is simply the absolute residual, i.e., $\Psi(X, Y) = |\hat{f}(X) - Y|$ or the normalized residual,

$\Psi(X,Y) = |\hat{f}(X) - Y|/\hat{\sigma}(X)$ for some suitable normalization function $\hat{\sigma}(X)$. Another popular choice, presented in Romano et al. (2019), is to construct the non-conformity score based on quantile estimates from a quantile regression model. In particular, they suggest

$$\Psi(X,Y) = \max\{\hat{q}_{\alpha/2}(X) - Y, Y - \hat{q}_{1-\alpha/2}(X)\},$$

where $\hat{q}_{\alpha/2}(X)$ and $\hat{q}_{1-\alpha/2}(X)$ are estimates of the $(\alpha/2)$:th and $(1-\alpha/2)$:th quantiles of $Y|X$.

We assume that we have access to a calibration set consisting of $(X_i, Y_i), i = 1, ..., N$ that have not been used to train the prediction model, and have calculated the corresponding non-conformity scores $\hat{s}_1, ..., \hat{s}_N$, where $\hat{s}_i = \Psi(X_i, Y_i)$. Our goal is to use this to construct a confidence set for predictions on a test set $\mathcal{D}_{\text{test}}$. For concreteness, consider a test data point $(X_{N+1}, Y_{N+1}) \in \mathcal{D}_{\text{test}}$. If the test data is exchangeable with the calibration set, the rank of the corresponding non-conformity score $\hat{s}_{N+1}$ is uniformly distributed among $\{1, 2, ..., N, N+1\}$. We can use this to construct a CP set in the following way. Let $\hat{q}_{1-\alpha}$ be the empirical $(1-\alpha)$th percentile of $\hat{s}_1, ..., \hat{s}_N$, then

$$C_{1-\alpha}(X_{N+1}) = \{y \in \mathcal{Y} : \quad \Psi(X_{N+1}, y) \leq \hat{q}_{1-\alpha}\},$$

for any confidence level $1-\alpha \in (0,1)$. If we construct the CP set in this fashion, they come with marginal coverage guarantees,

$$P\Big(Y_{N+1} \in C_{1-\alpha}(X_{N+1})\Big) \geq 1 - \alpha, \tag{1}$$

for any $\alpha \in (0,1)$ (Papadopoulos et al. [2002]; Vovk [2012]) as long as we assume exchangeability between the test data and calibration data.

The CP sets are evaluated by their empirical coverage on a test set. While the expected coverage is $(1-\alpha)$, as per (1), it is known (Vovk [2012]; Angelopoulos et al. [2023]) that the empirical coverage on a test set of size N_{test} given a calibration set of size N follows a beta-binomial distribution,

$$\begin{aligned} \text{Cov}(\mathcal{D}_{\text{test}}) &\sim \frac{1}{N_{\text{test}}}\text{Binom}(N_{\text{test}}, \mu) \\ \mu &\sim \text{Beta}(N+1-l, l) \end{aligned} \tag{2}$$

where $l = \lfloor (N+1)\alpha \rfloor$. This is important since it highlights the direct link between the calibration set size and the expected variation in coverage: when N_{test} and N is low, the variance in the observed coverage increases.

2.1. Mondrian conformal prediction

It is often desirable to make stronger statements than the marginal coverage guarantee. Mondrian CP (MCP) is proposed in Vovk et al. (2005) for scenarios where the data consists of several known categories. It is assumed that each data point belongs to a specific class, which is known and specified through a Mondrian taxonomy. In practice, the Mondrian taxonomy can be a function of the feature space $\mathcal{X}$, the label space $\mathcal{Y}$, or both. The motivation behind MCP is to achieve correct coverage on average but also conditioned on class membership.

Let g_i denote the class membership of observation i, and let $g_i \in \{1, ..., K\}$ for every i. In other words, there are K different non-overlapping classes. Let N_k be the number of calibration observations in class k. Furthermore, assume that the non-conformity scores in class k are sampled from a distribution F_k. We denote $\hat{F}_k$ to be the Empirical Cumulative Distribution Function (ECDF) of the scores in class k, defined to be

$$\hat{F}_k(t) = \frac{1}{N_k} \sum_{\hat{s}_i \in k} \mathbb{I}(\hat{s}_i \leq t).$$

The principal idea behind MCP is to estimate the empirical quantile separately in each class. Thus, for any new X_{N+1} with known class membership g_{N+1}, we construct the CP set as

$$C^k_{1-\alpha}(X_{N+1}) = \{y \in \mathcal{Y} : \quad \Psi(X_{N+1}, y) \leq \hat{q}^k_{1-\alpha}\},$$

where $\hat{q}^k_{1-\alpha}$ is the $(1-\alpha)$th percentile of $\hat{F}_k$. It is known (Vovk et al. [2005]) that the MCP sets are valid within each class, that is,

$$P\Big(Y_{N+1} \in C^k_{1-\alpha}(X_{N+1}) | g_{N+1} = k\Big) \geq 1 - \alpha$$

for every k, although it is noted in Ding et al. (2023) that this requires $N_k > 1/\alpha - 1$.

2.2. Clustered conformal prediction

A challenge with the MCP approach arises when N_k is low for some or all of the classes. A direct application of MCP might then lead to significant variance in the observed coverage, as evident from the distribution of coverage described in (2). In order to reduce the variance, we can enrich the calibration set, for example, by sharing data between classes. For this purpose, Clustered CP is introduced in Ding et al. (2023) in the context of image classification, where the classes were defined by the image labels. Clustered CP aims to map the K classes to a set of $M << K$ clusters. Let

$$h : \{1, ..., K\} \to \{1, ..., M\}$$

be a mapping function that performs this clustering, such that $h(k) = m$ means that class k belongs to cluster m. Importantly, many classes will be mapped to the same cluster. For a test point belonging to class k, Clustered CP constructs confidence sets in the same manner as MCP, but using the empirical distribution of cluster $h(k)$, i.e., of the cluster that class k belongs to.

The cluster function is related to the expected coverage gap in the following way. Let $\mathcal{I}_m$ be the set of classes that are mapped to cluster m, that is, $h(k) = m$ for every class $k \in \mathcal{I}_m$. Define ε_m to be the largest Kolmogorov-Smirnov-distance (Kolmogorov [1933]) between the ECDFs of any two classes that are mapped to cluster m:

$$\varepsilon_m = \max_{i,j \in \mathcal{I}_m} D_{\text{KS}}(\hat{F}_i, \hat{F}_j),$$

where

$$D_{\text{KS}}(\hat{F}_i, \hat{F}_j) = \sup_t |\hat{F}_i(t) - \hat{F}_j(t)|.$$

The expected coverage gap in any class in cluster m is then

$$P\Big(Y_{N+1} \in C(X_{N+1}) | g_{N+1} = k\Big) \geq 1 - \alpha - \varepsilon_m, \quad \forall k \in \mathcal{I}_m. \tag{3}$$

We refer to Ding et al. (2023) for a proof of this (Appendix A, Proof of Proposition 3). The coverage guarantee in (3) should be interpreted as follows. If the mapping function h clusters together two classes where the corresponding distribution of non-conformity scores are similar (measured by the Kolmogorov-Smirnov distance between the ECDFs in each class), then the coverage gap is close to $(1-\alpha)$. The bias in the expected coverage of cluster m is thus of magnitude at most ε_m.

Some practical matters must be considered when applying the Clustered CP approach. The first, and perhaps most important, is the choice of clustering function h. To reduce the bias ε_m, classes with similar ECDFs should be clustered together. The approach proposed in Ding et al. (2023) is to run a k-means clustering with Euclidean distance between discretized representations of the ECDFs, for instance, the 50th, 60th, ..., 90th percentile. Outside of Clustered CP, other distance measures have also been proposed for the task of clustering ECDFs. Examples include EP-Means (Empirical Probability-Means; Henderson et al. [2015]), which is based on Earth Movers distance, and an approach presented in Zhu et al. (2021) based on Kolmogorov-Smirnov. Furthermore, the calibration set should be split into two parts: one part to train the clustering function h and another part to estimate the quantile of interest, with neither of these two data sets being used during model training. Another technical detail proposed in Ding et al. (2023) is to use a special rule for handling classes with very few observations. These classes are automatically assigned to a NULL cluster, in which the globally calculated $\hat{q}_{1-\alpha}$ is utilized to construct the CP sets.

2.3. Evaluation Metrics

The coverage gap over N_{test} test observations quantifies the deviance between the empirical coverage and the expected coverage, i.e.,

$$\text{CovGap}(\mathcal{D}_{\text{test}}) = (1-\alpha) - \widehat{\text{Cov}}(\mathcal{D}_{\text{test}}),$$

where $\widehat{\text{Cov}}(\mathcal{D}_{\text{test}})$ is the empirical coverage. We also study the Mean Absolute Coverage Gap (MACG) over the classes $1, ..., K$,

$$\text{MACG}(\mathcal{D}_{\text{test}}) = \frac{1}{K}\sum_{k=1}^{K} |\text{CovGap}(\mathcal{D}_{\text{test}}^k)|,$$

where $\mathcal{D}_{\text{test}}^k$ is the subset of $\mathcal{D}_{\text{test}}$ that belongs to class k. This is a more informative performance measure for the considered context, as a low value indicates low coverage gaps not only marginally but also conditionally across the given classes.

2.4. Determining the optimal number of clusters

Determining the optimal number of clusters for a given data set has been the subject of much research in the statistical literature **??**. One widely used tool is the Caliński-Harabasz (CH) index **?**, which scores a clustering based on an adjusted ratio of the between-cluster sum-of-squares (BCSS) and within-cluster sum-of-squares (WCSS).

In this context, we consider the ECDF of a single class to be one data point represented by $\boldsymbol{q}_k$, the vector consisting of the 10th, 20th, ..., 90th quantile of the ECDF $\hat{F}_k$. To calculate the sum-of-squares between two classes j and k, we thus use the quantity $\|\boldsymbol{q}_j - \boldsymbol{q}_k\|^2$. For a proposed clustering of the K initial classes into M clusters, the CH index thus becomes

$$\text{CH}(M) = \frac{\text{BCSS}/(M-1)}{\text{WCSS}/(K-M)},$$

with

$$\text{BCSS} = \sum_{m=1}^{M} N_m \|\boldsymbol{q}^{(m)} - \boldsymbol{q}\|^2 \quad \text{and} \quad \text{WCSS} = \sum_{m=1}^{M} \sum_{k \in \mathcal{I}_m} \|\boldsymbol{q}_k - \boldsymbol{q}^{(m)}\|^2,$$

where $\boldsymbol{q}^{(m)}$ is the mean of the individual classes in cluster m, N_m is the number of classes in cluster m, and $\boldsymbol{q}$ is the mean of the entire data set. The BCSS calculates the average distance from each cluster center to the global centroid, whereas the WCSS calculates the distance from each individual class ECDF to the cluster centroids. The optimal number of clusters, M_{opt}, is the number that yields the highest CH index.

A heuristic that is more specific to the Clustered CP method is proposed in Ding et al. (2023). They suggest deriving both the clustering fraction γ and the number of clusters M from $\tilde{n}$, which denotes the maximum of $1/\alpha - 1$ and the number of calibration points in the smallest class. They then set $M = \lfloor \tilde{n} \cdot \gamma / 2 \rfloor$. Furthermore, they suggest to set $\gamma = \frac{\tilde{K}}{\tilde{K}+75}$, where $\tilde{K}$ is the number of classes with $N_k > \tilde{n}$. Their motivation is to ensure that there are at least 150 observations per cluster.

3. Simulation study

We now study a toy example where $\hat{F}_1, ..., \hat{F}_K$ are drawn from known distributions with different mean values. This ensures that the non-conformity scores are exchangeable within each class but with clear differences between classes. We study the following data-generating mechanism,

$$\begin{aligned} G &\sim U(1, ..., K) \\ \mu_k &\sim U(1, 2, ..., \sqrt{K}) \\ S|G = k &\sim \mathcal{N}(\mu_k, \sigma^2). \end{aligned}$$

The simulation setup first draws one of K classes with equal probability and then draws the non-conformity scores from $\mathcal{N}(\mu_k, \sigma^2)$. The mean μ_k takes one of $\sqrt{K}$ values, ensuring that multiple classes will indeed be drawn from the same distribution despite being assigned a different class. We are interested in studying if Clustered CP is able to recognize which

classes are drawn from the same distribution. We set $K = 100$, yielding 10 different possible values for μ_k. The σ parameter effectively determines the between-class differences. We perform simulations with N_k varying in $(10, 25, 50)$ for every k, and also vary σ in $(0.1, 1, 3)$ to study the sensitivity to changes in between-class differences. We set aside a fraction $\gamma \in (0, 1)$ of the calibration data for the clustering task, while the rest is used for calibration. In this simulation, we omit a NULL cluster since we keep N_k the same for each class.

We run Clustered CP with $M = 1, ..., 100$ clusters. We use a k-means clustering to conduct the clustering, where the distance metric is the Euclidean distance between the 10th, 20th, ..., 90th percentile of the ECDFs.[1] The quantiles are calculated based on the native `quantile` function in `R`, which applies interpolation if necessary. We also tested using the EP-Means algorithm (Henderson et al. [2015]) for the clustering and found that the results were highly similar to the results when using the k-means.

All the results reported are the mean of 50 simulations with the synthetic data redrawn each time. A benchmark for our experiments is MCP and a regular CP approach with $\gamma = 0$, i.e., we use the entire calibration set for calibration since no clustering is required. Furthermore, we also compare with an Oracle method that knows which classes are drawn with the same μ_k and uses this as a clustering function.

Figure 1 displays the results for $\gamma = 0.5$ for different combinations of N_k and σ. Each figure displays MACG per class as a function of the number of clusters. All the figures display a similar pattern, with MACG decreasing quickly when we increase the number of clusters up to a certain point before increasing, or in some cases, being relatively flat if the number of clusters is increased further. When N_k is high, the performance of MCP is better than when N_k is low, which is intuitive: calibrating completely classwise without sharing data between classes is sufficient when N_k in each class is high. When σ increases, the CP benchmark improves relative to the other benchmarks, as the distributions $\mathcal{N}(\mu_k, \sigma^2)$ with different μ_k effectively become more similar. In Appendix 1 we repeat the analysis for $\gamma = 0.25$ and $\gamma = 0.75$. The results display a similar "hockey stick" trend as those presented here, with MACG quickly decreasing when the number of groups is increased to around $M \approx 20$ before slowly increasing with higher M. The increase with higher M is most notable when $\gamma = 0.75$, that is, when we set aside fewer data points for calibration. The results indicate some sensitivity to the choice of γ, as setting a too high γ is unfavorable, particularly when the number of calibration points per class is already low.

Figure 2 displays the ECDFs from the original classes (in grey) overlayed with the identified clusters with $M = 10$ for one simulation. The plot indicates whether or not the clustering function can correctly identify the clusters among the classwise ECDFs. For $\sigma = 0.1$, the ECDFs for the original classes are very distinct, making it relatively easy for the Clustered CP to identify the different clusters. It should, however, be pointed out that the Clustered CP is not perfect, as some clusters display saddle points in their ECDFs, indicating that classes have been clustered together despite being drawn from different underlying distributions. As σ increases, the ECDFs of the initial classes become less distinct.

1. The k here is not the same as previously used to denote class membership. To be consistent with our notation of forming M clusters, we should refer to the method as M-means clustering but use the term k-means clustering since this is the established term in the literature.

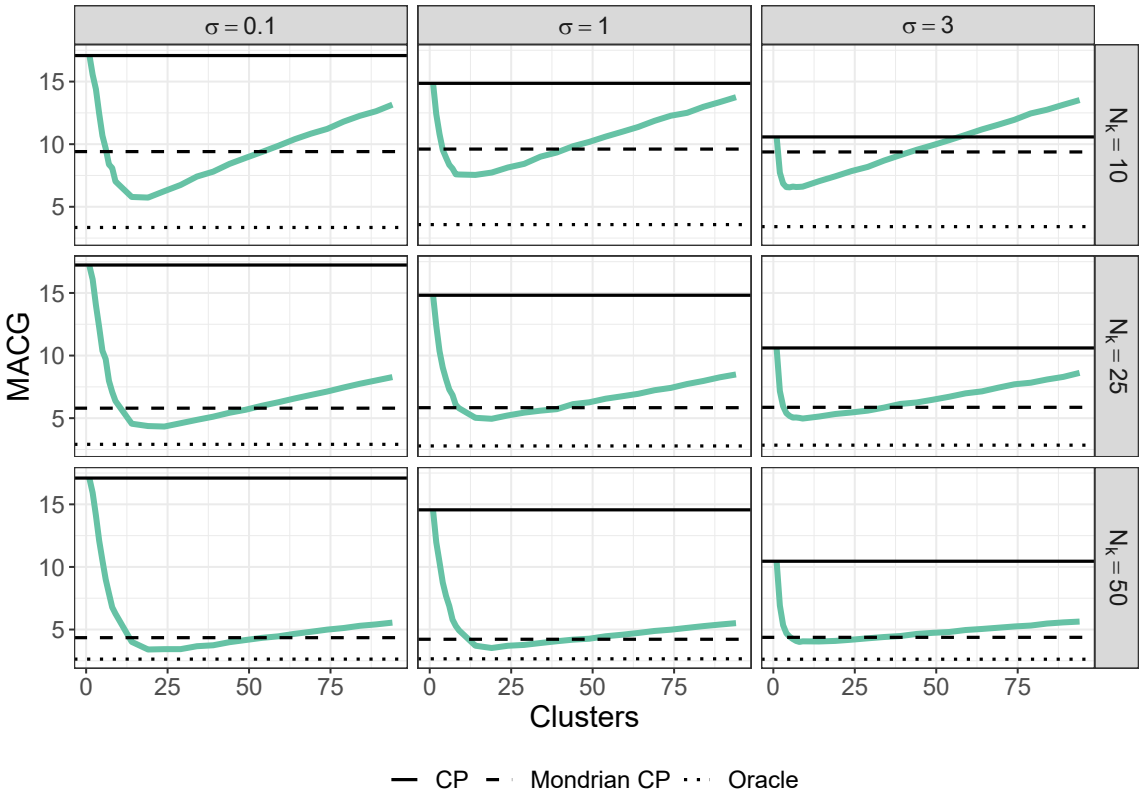

Figure 1: Mean Absolute Coverage Gap per class when half of the data is used for clustering and half for calibration. We vary $\sigma \in (0.1, 1, 3)$ and $N_k \in (10, 25, 50)$. The plots show the average of 50 simulations.

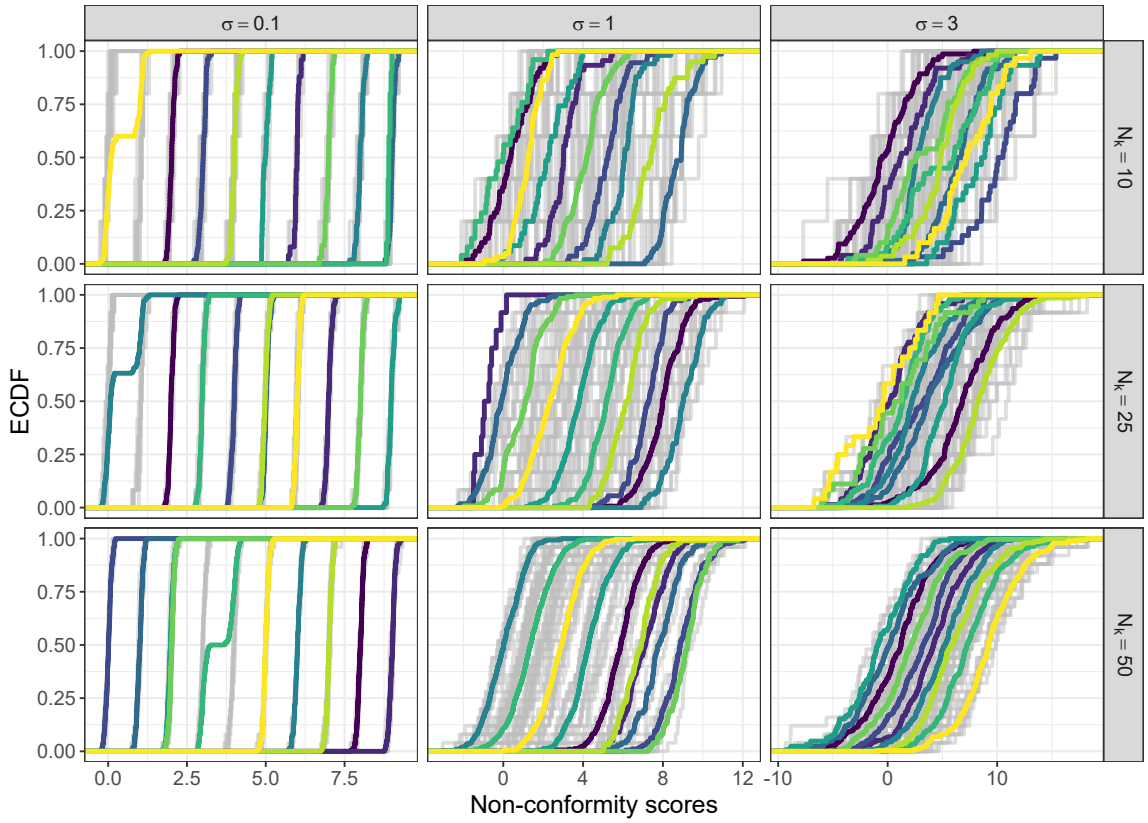

Figure 2: The ECDF identified by Clustered CP for $\sigma \in (0.1, 1, 3)$ and $N_k \in (10, 25, 50)$. Each color represents one of the $M = 10$ clusters identified. Each grey line represents one of the initial $K = 100$ classes.

The simulation study demonstrates the idea behind Clustered CP, namely that sharing calibration data between similar classes decreases the MACG. As we decrease σ, the MACG approaches the performance of the Oracle, as expected when the between-class differences are higher in the data-generating mechanisms. However, even with $\sigma = 3$, when the ECDFs are more challenging to distinguish visually, the Clustered CP performs better than the CP and MCP benchmarks.

The MACG curves presented in Figure 1 reach a minimum between 10 and 20 clusters. As a comparison, we calculate the CH index for every value of M (see Figure 10 in Appendix 2). A general trend is that M_{opt} decreases when σ increases, as a more noisy data-generating process leads to less distinct clusters. For $\sigma = 0.1$, the CH curve increases steadily until $M \approx 15$ and remains stable if M increases beyond this. We observe the same trend for $\sigma = 1$, albeit with a clearer decrease in the CH index if M is increased past the peak around 15 clusters. Finally, for $\sigma = 3$, we see a peak around three or four clusters, with a steeper decline in the CH index after this. In conclusion, higher σ leads to a lower estimated M_{opt} based on the CH index, and a higher N_k leads to a more defined peak in the CH index.

In comparison, the heuristics developed in Ding et al. (2023) with $\gamma = 0.5$ yields $M_{\text{opt}} = 2$, $M_{\text{opt}} = 6$, and $M_{\text{opt}} = 12$ when $N_k = 10$, $N_k = 25$, and $N_k = 50$, respectively. This is a somewhat lower value than the CH index and the MACG curves.

4. Norwegian housing data

We now turn to a real-world data set consisting of transactions from the Norwegian housing market. Predicting the value of a dwelling, given the location and additional characteristics, is of interest to banks, homeowners, and other financial institutions. The models used for this purpose are referred to as automated valuation models (AVMs). Several studies demonstrate the accuracy of non-parametric machine learning models as AVMs, including in the US (Park et al. [2015]), South-Korea (Ho et al. [2020]), and Australia (Gao et al. [2022]). For a thorough review of the use of AVMs, we refer to d'Amato et al. (2017) or Steurer et al. (2021).

We study a data set of $N = 84\,975$ transactions conducted in the open housing market in 2015. The response variable is a sale price measured in million NOK that a seller and buyer have agreed upon after an open auction. Each sale contains features that are known to correlate with the sale price, such as the size of each dwelling in square meters, the number of bedrooms, and, importantly, the location of the dwelling. The location is encoded both through coordinates and via a municipality dummy variable. The data contains sales from $K = 286$ in different municipalities, which will serve as our initial classes. Summary statistics are presented in Table 1.

Variable	Unit	Mean	St. Dev.	Min	Max	Type
Sale Price	NOK (mill.)	3.07	1.72	0.02	28.7	Numerical
Size	m^2	100	54	0	819	Numerical
Gross Size	m^2	112.42	67.48	0	1131	Numerical
Longitude	degrees	9.82	2.90	4.79	30.47	Numerical
Latitude	degrees	60.71	2.37	57.99	70.72	Numerical
Altitude	m	101.69	136.49	0	1151	Numerical
Bedrooms	-	2.56	1.20	0	15	Numerical
Municipality	-	-	-	-	-	Categorical

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

The number of observations per municipality varies from $N_k = 10$ to $N_k = 18\,028$ in Oslo, the largest city in Norway. Among the $K = 286$ municipalities there are 167 that have $N_k < 100$ and 16 municipalities with $N_k > 1\,000$. A histogram of N_k is displayed in Figure 3

4.1. Experimental setup

We split the data set into three parts: 25% of the data is used to train the prediction model, 25% as a test set, and 50% of the data is used for calibration, including the clustering task. We vary γ, the proportion of the calibration data used for clustering, in the range $(0.25, 0.5, 0.75)$.

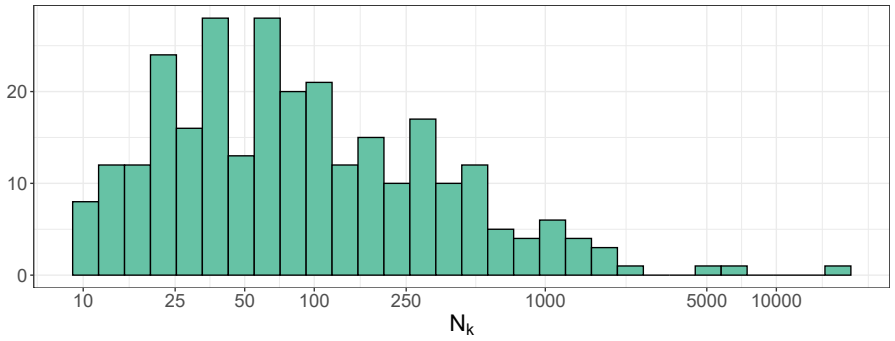

Figure 3: A histogram of N_k on log scale for the $K = 286$ municipalities considered in this study. The three largest municipalities are Oslo ($N_{\text{Oslo}} = 18\,028$), Bergen ($N_{\text{Bergen}} = 5\,890$), and Trondheim ($N_{\text{Trondheim}} = 4\,735$).

We employ three non-conformity scores: Conformalized Quantile Regression (Romano et al. [2019]), the absolute residuals $|Y_i - \hat{f}(X_i)|$, and the normalized absolute residuals, $|Y_i - \hat{f}(X_i)|/\hat{f}(X_i)$, as suggested in Lim et al. (2021) when studying house prices. We refer to the three scores as CQR, CP, and Normalized CP, respectively. We use a random forest model (Breiman [2001]) as our point predictor $\hat{f}$ for the latter two scores and a Quantile Regression Forest (Meinshausen [2006]) when constructing the CQR scores. We assign all classes with $N_k < 10$ to a NULL cluster, which uses the globally calculated $\hat{q}_{1-\alpha}$ in the construction of the confidence sets.

We calculate all the CP sets at confidence level $\alpha = 0.1$ and report the mean MACG over 50 simulations, each time with a new split into training, calibration, and test set. MCP and CP with $\gamma = 0$ serve as benchmarks, but we also compare compare with a spatial clustering, where we employ a k-means clustering based on the geographical centroid of the municipalities.

4.2. Results on Norway housing data

Figure 4 displays the MACG for Clustered CP with $M = 1, ..., 50$ clusters based on the ECDFs of the municipalities. The results are fairly similar for the CP and Normalized CP score functions, with Clustered CP achieving significantly better than spatial clustering. Interestingly, Clustered CP also performs better than the benchmarks, indicating that there is indeed room for improvement over both CP and MCP for the right number of clusters M.

The results are not as clear for the CQR, although we see a slight improvement for Clustered CP over the spatial clustering. The nominal level of MACG is also much lower for the CP benchmark with CQR. This indicates that CQR does better at constructing exchangeable non-conformity scores than the other choices of non-conformity measure. Still, the Clustered CP improves over CQR with $M \approx 15$.

In the simulation study, we had an idealized setting with the same N_k for every class, whereas there is a significant imbalance in the housing data set. It is reasonable to assume that the most significant improvement in the coverage gap with Clustered CP comes from

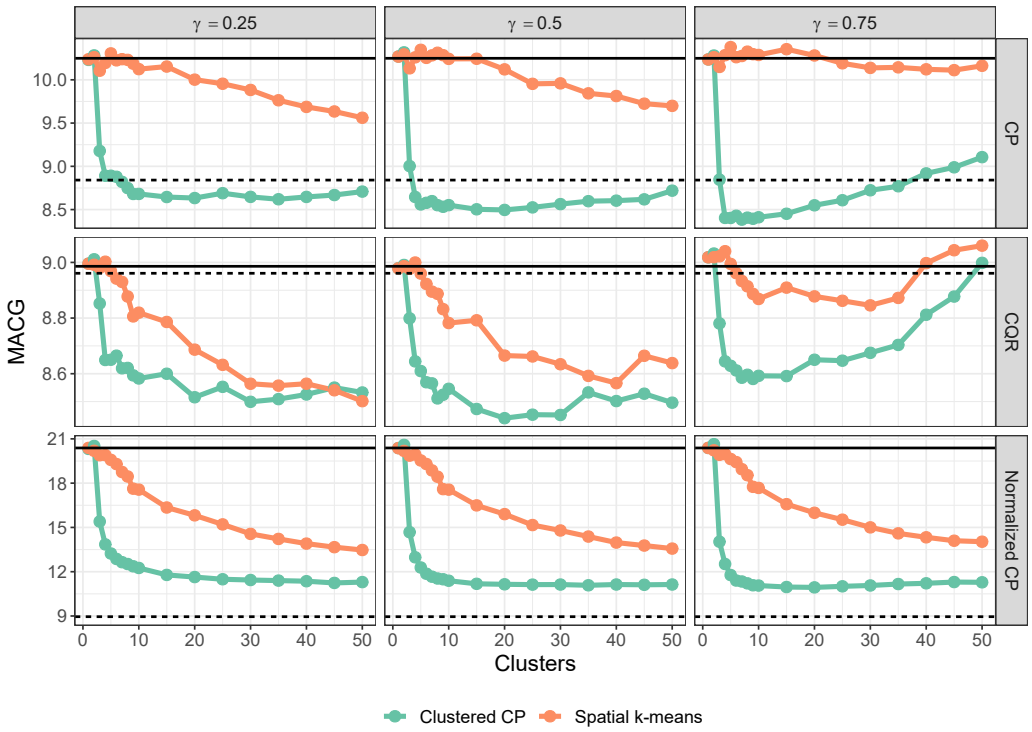

Figure 4: Results of Clustered CP for $M = 1, ..., 50$, compared against a Spatial k-means clustering. We run the analysis for different non-conformity scores (vertical) and γ values (horizontal). The dotted line is MCP, and the straight line is CP, with $\gamma = 0$, i.e., all of the calibration data used to estimate the empirical quantile. Note that the range of MACG is different for the different non-conformity scores.

classes where N_k is low. In Figure 5, we study the coverage gap as a function of N_k, where Clustered CP with $M = 10$ is compared with MCP and CP.

Regardless of the calibration method, the variance of the observed coverage gap is higher when N_k is low, as expected. As N_k increases, most box plots are centered closer to a coverage gap of zero. As expected, the most significant improvement of Clustered CP compared with CP and MCP comes from the bins with the fewest observations. Interestingly, the largest bins, containing the bigger cities, exert slightly wider variance with Clustered CP than for MCP, albeit the Clustered CP improves upon the CP approach, where the coverage gap in the larger cities (where prices are higher) is lower when $\hat{q}_{90}$ is calculated based on non-conformity scores from outside the urban areas (where prices are lower).

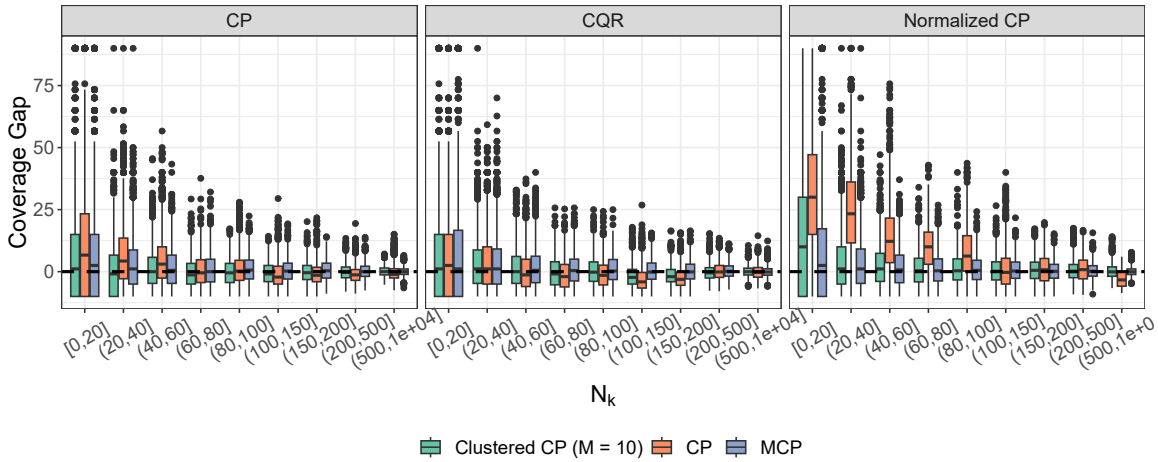

Figure 5: Coverage gap for different bins of N_k for MCP, CP, and Clustered CP with $M = 10$. The results are for confidence level $\alpha = 0.1$ with a fraction $\gamma = 0.5$ set aside for clustering in Clustered CP.

Finally, we briefly consider the question of identifying the optimal number of clusters in the Norwegian housing data. The MACG curves imply an M_{opt} between 10 and 20 for most combinations of non-conformity scores and simulation parameters. For comparison, we display the CH index in Figure 11 in Appendix 2. The CH index indicates an M_{opt} value between four and six for most combinations of γ and non-conformity score, with a slightly higher value when $\gamma = 0.25$. This estimate of M_{opt} is somewhat lower than what we get using the MACG as a criterion. Using the Clustered CP heuristic from Ding et al. (2023), we obtain an M_{opt} between two and three, depending on the choice of γ. Both the CH index and the heuristic yield a slightly lower number compared with the MACG curves in Figure 4.

4.3. A closer look at the identified clusters

The identified clusters can be of independent interest to real estate practitioners. For concreteness, we study some properties of the clusters constructed by the Clustered CP method when $M = 6$ for one particular simulation. A map of the identified clusters is seen in Figure 6 with corresponding cluster-wise ECDFs in Figure 7. The grey areas in

the map belong to the NULL cluster, i.e., municipalities with $N_k < 10$. This also includes municipalities without any observations in the data set. The coloring of the clusters in the map does not carry any significance other than the fact that municipalities with the same color are assigned to the same cluster.

Extracting information from the maps alone is not trivial, indicating that the identified clusters do not follow any clear patterns in space. This observation is noteworthy, as empirical evidence exists that real estate prices exhibit spatial autocorrelation (Basu et al. [1998]; Ismail [2006]). If this were also true about the non-conformity scores, we would expect neighboring municipalities to be clustered more frequently, given a sufficiently effective clustering function. However, spatial autocorrelation might be present on a smaller scale but not necessarily between adjacent municipalities.

For the Normalized CP non-conformity score, every municipality with $N_k > 1\,000$ is clustered together in one cluster, revealing that the normalized absolute residuals in the larger cities display a similar ECDF regardless of spatial distance. Upon closer inspection of each group's summary statistics, some interesting patterns are revealed. Studying the size (in square meters) of the dwellings in each cluster also indicates the clustering. For all the non-conformity scores, five out of the six clusters have a mean dwelling size that is between approximately 100 and 120 square meters, while there is one cluster with significantly lower mean size (87 square meter for CP, 86 square meter for CQR, and 92 square meter for Normalized CP).

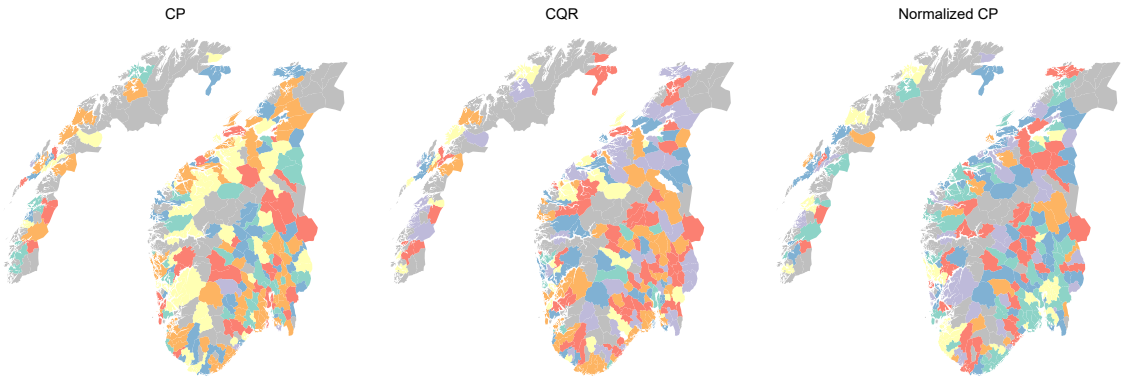

Figure 6: An example of the identified clusters with the Clustered CP methodology for $M = 6$ clusters. The grey municipalities either have no observations or are part of the NULL cluster.

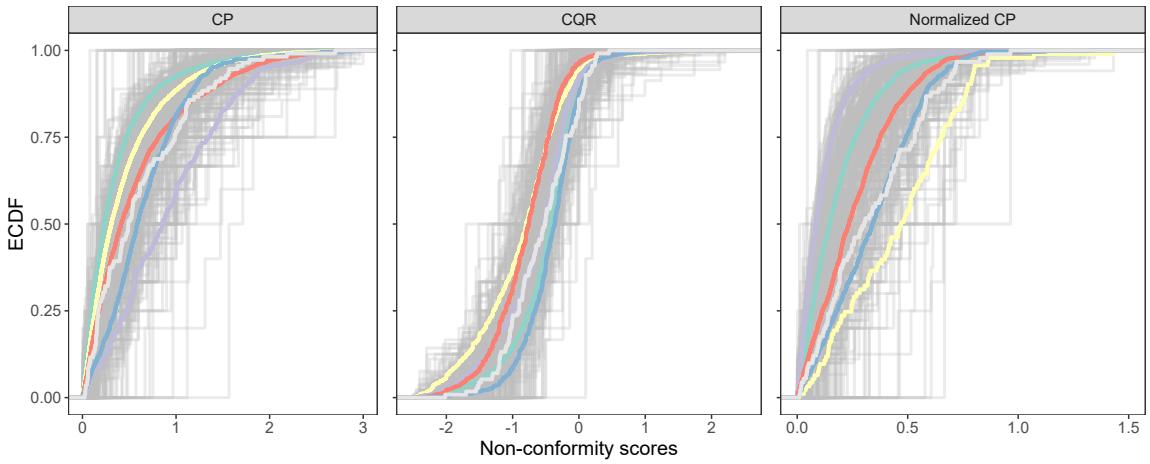

Figure 7: The ECDF of the identified clusters with Clustered CP for $M = 6$, overlaying the individual ECDFs for each municipality.

5. Discussion and conclusion

Achieving class-conditional coverage guarantees in CP is often desirable, but it is challenging when many classes have few observations. We are thus inclined to share calibration data between classes to reduce the empirical coverage variance. In this research, we investigate the use of Clustered CP to identify clusters of municipalities in the Norwegian housing market and use these clusters to calibrate the CP sets. We show that Clustered CP is a viable option in the considered context when a classwise calibration approach fails due to insufficient calibration data, yet a global calibration approach is too simplistic to account for between-class differences.

The Clustered CP approach does not encode any spatial information when clustering the municipalities but still improves over both CP and MCP, as well as a clustering approach based on spatial location. The clusters identified based on the ECDFs can potentially be used to gain new insights about submarkets in the Norwegian housing market. It is worth reiterating the role of exchangeability, or lack thereof, in these results. In an ideal scenario where the non-conformity scores are globally exchangeable regardless of the class membership, neither MCP nor Clustered CP is necessary, as a global estimate of the quantile $\hat{q}_{1-\alpha}$ is a reasonable estimate of the quantile of interest also in subsets of the data. We observe that the CQR method yields the smallest improvement in MACG for Clustered CP compared with spatial approaches but also a lower nominal level of MACG compared with the CP and Normalized CP non-conformity scores. This indicates that the CQR creates non-conformity scores that are closer to exchangeable in space.

While there is extensive literature on identifying the optimal number of clusters in other statistical frameworks, this remains an open question in the considered context of distribution-free methods like CP. Our research indicates a slight discrepancy between the optimal number of clusters found when comparing the popular and conventionally used CH index against the target coverage metric estimated on the test set. An interesting direction for future research is to develop novel clustering algorithms that iteratively partition the

data set until the data in each cluster is internally exchangeable rather than relying on a user-specified number of clusters.

Acknowledgements

We thank Eiendomsverdi AS for providing the data set and valuable domain knowledge about the Norwegian housing market. We are grateful to the two anonymous reviewers for valuable comments that improved the paper. Thanks also to Gudmund Horn Hermansen for insightful discussions during the early phase of the work leading up to this paper.

References

Angelopoulos, Anastasios N. and Bates, Stephen (2023). "Conformal Prediction: A Gentle Introduction". In: *Foundations and Trends in Machine Learning* 16.4, pp. 494–591.

Bastos, JA and Paquette, J (2024). *On the uncertainty of real estate price predictions*. Preprint downloaded from https://rem.rc.iseg.ulisboa.pt/wps/pdf/REM_WP_0314_2024.pdf.

Basu, S. and Thibodeau, T.G. (1998). "Analysis of Spatial Autocorrelation in House Prices". In: *The Journal of Real Estate Finance and Economics* 17.1, pp. 61–85.

Bellotti, A (2017). "Reliable region predictions for automated valuation models". In: *Annals of Mathematics and Artificial Intelligence* 81.1, pp. 71–84.

Breiman, L. (2001). "Random forests". In: *Machine Learning* 45, pp. 5–23.

d'Amato, M. and Kauko, T. (2017). *Advances in Automated Valuation Modeling: AVM After the Non-Agency Mortgage Crisis*. Studies in Systems, Decision and Control. Springer International Publishing. ISBN: 9783319497440.

Ding, Tiffany, Angelopoulos, Anastasios N., Bates, Stephen, Jordan, Michael I., and Tibshirani, Ryan (2023). "Class-Conditional Conformal Prediction with Many Classes". In: *Thirty-seventh Conference on Neural Information Processing Systems*.

Foygel Barber, Rina, Candès, Emmanuel, Ramdas, Aaditya, and Tibshirani, Ryan J. (2023). "Conformal prediction beyond exchangeability". In: *The Annals of Statistics* 51.2, pp. 816–845.

Gao, Qishuo, Shi, Vivien, Pettit, Christopher, and Han, Hoon (2022). "Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia". In: *Land Use Policy* 123, p. 106409.

Goodman, Allen C. and Thibodeau, Thomas G. (1998). "Housing Market Segmentation". In: *Journal of Housing Economics* 7.2, pp. 121–143.

Guan, Leying (2022). "Localized conformal prediction: a generalized inference framework for conformal prediction". In: *Biometrika* 110.1, pp. 33–50.

Henderson, Keith, Gallagher, Brian, and Eliassi-Rad, Tina (2015). "EP-MEANS: an efficient nonparametric clustering of empirical probability distributions". In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. SAC '15, pp. 893–900.

Hjort, Anders, Hermansen, Gudmund Horn, Pensar, Johan, and Williams, Jonathan P. (2023). *Uncertainty quantification in automated valuation models with locally weighted conformal prediction*. arXiv: 2312.06531 [stat.ML].

Ho, Winky K. O., Tang, Bo-Sin, and Wong, Sui Wai (2020). "Predicting property prices with machine learning algorithms". In: *Journal of Property Research*.

Ismail, Suriatini (Jan. 2006). "Spatial autocorrelation and real estate studies: A literature review". In: *Regional Science and Urban Economics* 35.

Kolmogorov, A (1933). "Sulla determinazione empirica di una legge didistribuzione". In: *Giorn Dell'inst Ital Degli Att* 4, pp. 89–91.

Lim, Zhe and Bellotti, Anthony (2021). "Normalized nonconformity measures for automated valuation models". In: *Expert Systems with Applications* 180, pp. 115–165.

Mao, Huiying, Martin, Ryan, and Reich, Brian J. (2023). "Valid Model-Free Spatial Prediction". In: *Journal of the American Statistical Association* 0.0, pp. 1–11.

Meinshausen, Nicolai (2006). "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7.35, pp. 983–999.

Papadopoulos, Harris, Proedrou, Kostas, Vovk, Volodya, and Gammerman, Alex (2002). "Inductive Confidence Machines for Regression". In: *Machine Learning: ECML 2002*, pp. 345–356.

Park, Byeonghwa and Bae, Jae Kwon (2015). "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data". In: *Expert Systems with Applications* 42.

Romano, Yaniv, Patterson, Evan, and Candes, Emmanuel (2019). "Conformalized Quantile Regression". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Shafer, Glenn and Vovk, Vladimir (2008). "A tutorial on conformal prediction". In: *Journal of Machine Learning Research* 9, pp. 371–421.

Sommervoll, Åvald and Sommervoll, Dag Einar (2019). "Learning from man or machine: Spatial fixed effects in urban econometrics". In: *Regional Science and Urban Economics* 77.

Sommervoll, Dag Einar (July 2023). *Price and Hedonic Heterogeneity Measures in Local Housing Markets*. DOI: 10.13140/RG.2.2.29861.68321.

Steurer, Miriam, Hill, Robert J., and Pfeifer, Norbert (2021). "Metrics for evaluating the performance of machine learning based automated valuation models". In: *Journal of Property Research* 38.2, pp. 99–129.

Tibshirani, Ryan J, Foygel Barber, Rina, Candes, Emmanuel, and Ramdas, Aaditya (2019). "Conformal Prediction Under Covariate Shift". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Vovk, Vladimir (2012). "Conditional Validity of Inductive Conformal Predictors". In: *Proceedings of the Asian Conference on Machine Learning*. Vol. 25. Proceedings of Machine Learning Research, pp. 475–490.

Vovk, Vladimir, Gammerman, Alex, and Shafer, Glenn (2005). *Algorithmic Learning in a Random World*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387001522.

Xu, Chen and Xie, Yao (2023). "Conformal Prediction for Time Series". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10, pp. 11575–11587.

Zhu, Yingqiu, Deng, Qiong, Huang, Danyang, Jing, Bingyi, and Zhang, Bo (2021). "Clustering Based on Kolmogorov–Smirnov Statistic with Application to Bank Card Transaction Data". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 70.3, pp. 558–578.

1. Additional simulation results

Here, we present additional results for the simulation studies in Section 3. The results are similar to those presented in Figure 1 but with different choices of γ, the fraction used for clustering. We display $\gamma = 0.25$ in Figure 8 and $\gamma = 0.75$ in Figure 9.

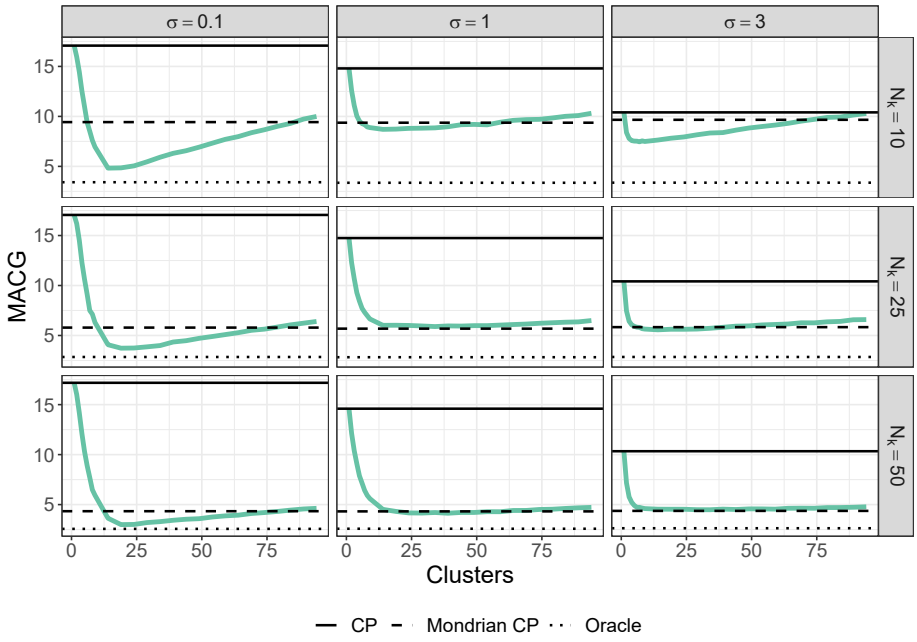

Figure 8: Mean Absolute Coverage Gap for $\gamma = 0.25$.

2. Caliński-Harabasz index

Figure 10 displays the Caliński-Harabasz (CH) index for the simulated data with different number of calibration points (vertically) and σ in the data-generating process (horizontally). Each plot displays one grey line for each of the 50 simulations with different split into calibration and clustering. The red lines mark the mean of the simulations, with the highest CH index marked with a dotted line. The CH values are normalized so that the highest in each plot is 1.

A similar plot is displayed in Figure 11 for the Norwegian housing data, with different non-conformity scores (vertically) and cluster proportions γ (horizontally).

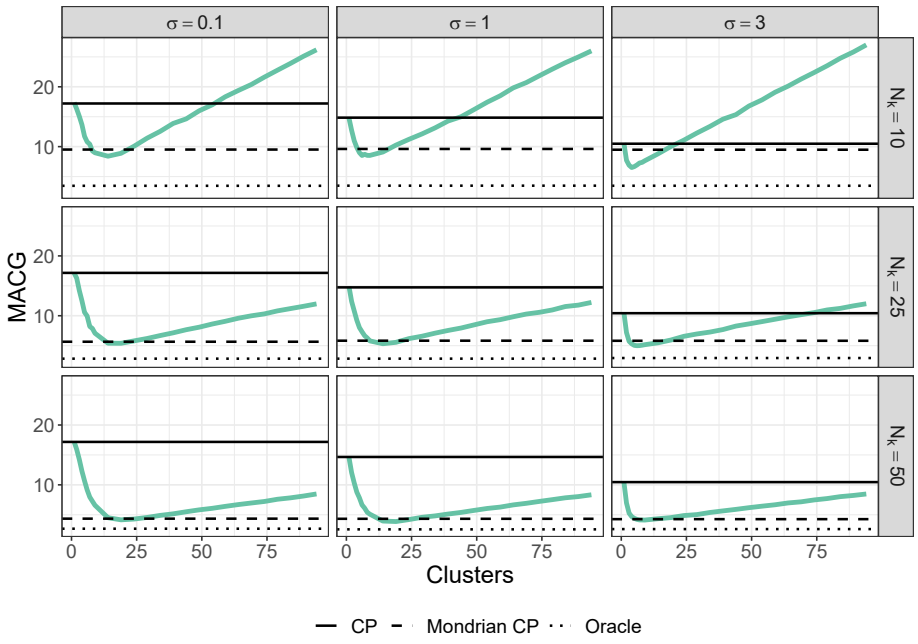

Figure 9: Mean Absolute Coverage Gap for $\gamma = 0.75$.

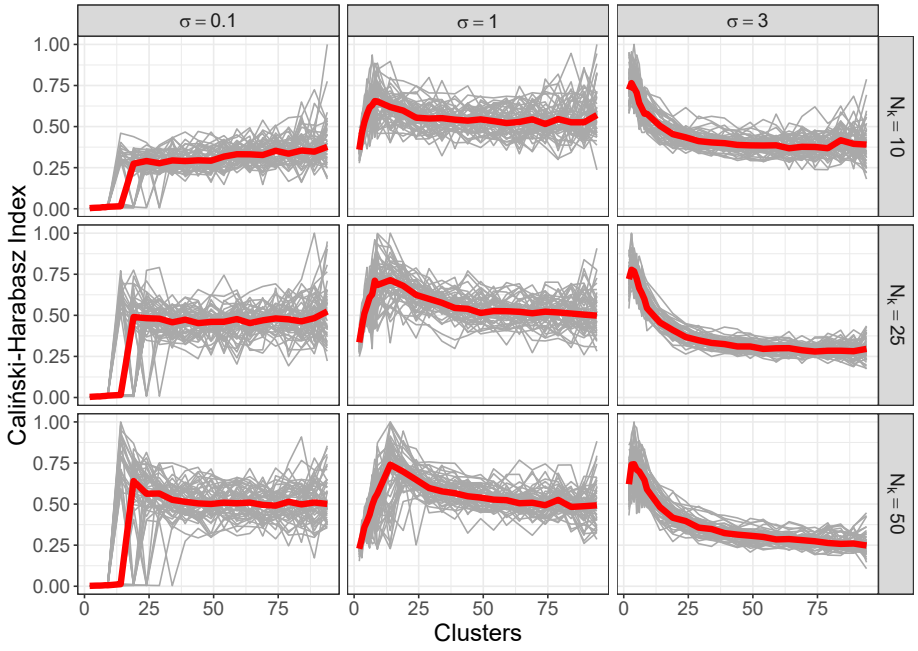

Figure 10: Caliński-Harabasz index for the simulated data.

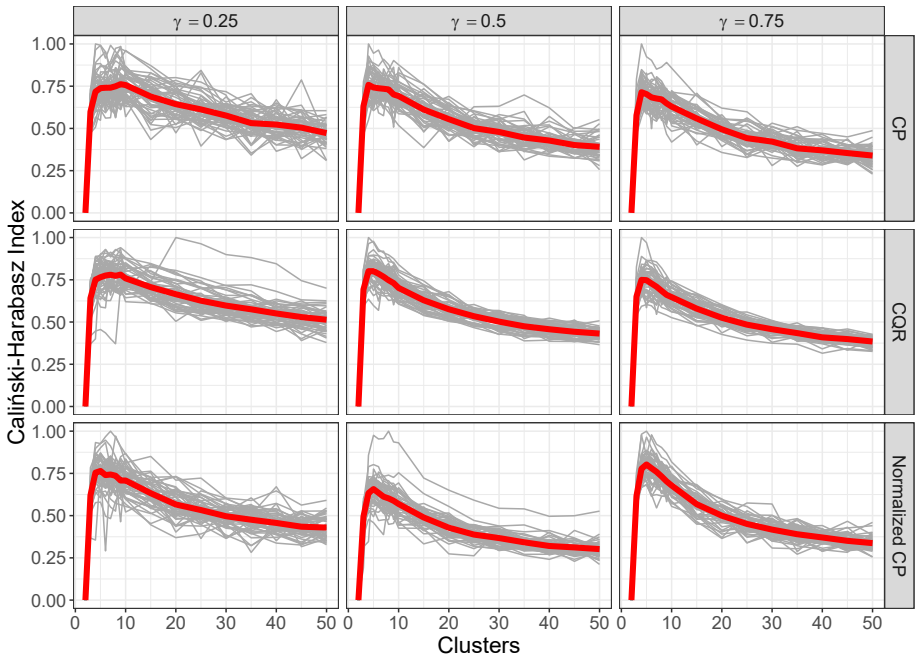

Figure 11: Caliński-Harabasz index for the Norwegian housing data.