Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rough-set regions as a local reliability signal for conformal prediction: when does Pawlak boundary membership help?

Balaraju M

dveloperbbobby@gmail.com

Acharya Institute of Technology

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11481472/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Rough-set regions as a local reliability signal for conformal prediction: when does Pawlak boundary membership help?

Balaraju M[1*]

[1*]Acharya Institute of Technology, Bengaluru, Karnataka, India.

Corresponding author(s). E-mail(s): dveloperbbobby@gmail.com;

Abstract

Conformal prediction is a widely used framework for distribution-free uncertainty quantification. It guarantees that prediction sets contain the true label with a prescribed probability on average, but the average guarantee may hide groups of inputs on which the prediction sets fail. Rough set theory divides the input space into information granules without training any model. Granules whose training labels agree form the positive region, granules with conflicting labels form the boundary region, and attribute combinations absent from training form an unseen region. In this paper, a region-aware framework is proposed to examine whether this three-way partition is a useful reliability signal for conformal prediction. The framework uses the regions in two ways: as predictors of conformal miscoverage, and as a Mondrian taxonomy for group-wise calibration. Experiments are conducted on eight OpenML classification datasets, three levels of granulation and ten random splits. Pooled over all settings, region indicators reduced the held-out log loss of a miscoverage model by 0.0035 (95% interval -0.00002 to 0.0081) and raised its area under the ROC curve (AUC) by 0.0054 (0.0006 to 0.0116). The effect was concentrated on Nursery and Car (AUC gains 0.023 and 0.014) and weakened as the share of unseen inputs grew (Spearman correlation -0.60). Region calibration restored coverage on boundary and unseen inputs at a few percent larger sets, and region-by-class calibration used 18.3% smaller sets than confidence-bin-by-class calibration at 8.3 points lower worst confidence-bin coverage. The results indicate that rough-set regions are a local, granulation-dependent reliability signal rather than a general one.

Keywords: Rough sets, Information granules, Conformal prediction, Mondrian calibration, Conditional coverage, Uncertainty quantification

1 Introduction

Uncertainty quantification is an essential requirement when machine learning models are used for decision making. A classifier that reports a single label gives no indication of how reliable that label is. Conformal prediction addresses this problem by converting the output of any classifier into a prediction set that contains the true label with probability at least $1-\alpha$, under the sole assumption that the data are exchangeable [1, 2]. Owing to its simplicity and its finite-sample guarantee, conformal prediction has been applied to a wide range of classification and regression problems.

However, the guarantee of standard conformal prediction is marginal. It holds on average over the input distribution and does not ensure comparable performance on every group of inputs. A prediction set procedure can therefore satisfy its nominal coverage while failing badly on a particular subgroup. Barber et al. [3] showed that exact coverage conditional on every input is impossible without further assumptions. For this reason, practical methods calibrate separately

within chosen groups of inputs, an approach known as Mondrian conformal prediction [1, 4]. The usefulness of a Mondrian predictor depends entirely on how the groups are chosen.

Rough set theory, introduced by Pawlak [5, 6], provides a natural way to form such groups without training any model. Objects that take identical values on a selected set of attributes are indiscernible and form an information granule. Granules whose training labels agree constitute the positive region, and granules whose training labels conflict constitute the boundary region. When an attribute combination is absent from the training data, the corresponding input falls into a third, unseen group. The boundary region identifies inputs that cannot be separated by the selected attributes, and it is therefore reasonable to expect that a predictor is less reliable there.

In recent years, several methods have been proposed to construct groups for conformal calibration. Martinez Gil et al. [7] identified interpretable groups whose nonconformity scores are homogeneous. Kiyani et al. [8] learned partitions of the feature space that target conditional coverage. Gibbs et al. [9] and Jung et al. [10] guaranteed coverage over rich families of overlapping groups. Campagner et al. [11] established formal links between three-way decisions and conformal prediction. Although rough-set regions can be used as Mondrian groups directly, it remains unclear whether they carry information about conformal failures that is not already captured by standard signals or learned groupings.

To address this question, a region-aware framework for conformal reliability analysis is proposed in this paper. The framework uses the rough-set regions in two complementary ways. First, region membership is used as a predictor of the inputs that a marginal conformal predictor misses. Second, the regions are used as a Mondrian taxonomy for group-wise calibration. The framework is evaluated against class-based, confidence-based, decision-tree-based and randomised groupings on eight benchmark datasets.

The main contributions of this paper are as follows:

1. A region-aware framework is proposed that uses Pawlak's positive, boundary and unseen regions both as a diagnostic signal for conformal miscoverage and as a training-free Mondrian taxonomy.
2. A controlled evaluation is carried out on eight datasets, three levels of granulation and ten random splits, against a learned interpretable grouping and twenty randomised groupings per run.
3. The experimental results show that the region signal is local rather than general. It is strong on two datasets, small on three and absent on three, and it is largest at coarse granulation with a moderate boundary share and few unseen inputs.
4. The calibration trade-offs of the regions are characterised. Region calibration restores coverage on ambiguous inputs at a small cost in set size, and region-by-class calibration yields smaller sets than confidence-bin-by-class calibration at the price of weaker coverage on low-confidence inputs.

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 introduces the preliminaries of rough sets and conformal prediction. Section 4 presents the proposed framework. Section 5 reports the experimental results and their analysis. Section 6 concludes the paper and outlines future work.

2 Related work

2.1 Conformal prediction and conditional coverage

Split conformal prediction fits a model on one part of the data and calibrates its nonconformity scores on a held-out part [1, 2]. Sadinle et al. [12] proposed the least ambiguous set-valued classifier (LAC), which yields the smallest prediction sets under marginal coverage. Romano et al. [13] proposed adaptive scores that produce more even coverage across inputs. Class-conditional calibration is the best-known form of Mondrian conformal prediction [4, 14], and equalised coverage across protected groups is another [15]. Because exact conditional coverage is impossible in general [3], the choice of calibration groups determines how well a conformal predictor behaves on subgroups.

2.2 Learned and overlapping groups

A number of recent studies construct calibration groups from data. Martinez Gil et al. [7] searched for interpretable groups with homogeneous nonconformity distributions and calibrated within them. Kiyani et al. [8] learned feature-space partitions that target conditional coverage. Gibbs et al. [9] derived guarantees over a function class of possibly overlapping groups, and batch multivalid methods [10] provide

coverage over many groups simultaneously. These methods learn their groups, whereas the regions studied in this paper are fixed by the training data alone. The decision-tree grouping used in Section 5 is a simple comparator inspired by learned-group methods and is not an implementation of any of them.

2.3 Rough sets and three-way decisions

Rough set theory approximates a concept by a lower approximation, an upper approximation and the boundary between them [5, 6]. Rough membership functions measure the degree to which a granule belongs to a class [16]. The original model has been extended to variable-precision rough sets [17], probabilistic rough sets and three-way decisions [18], fuzzy rough sets [19] and covering-based rough sets [20], and rough-set attribute reduction is widely used for feature selection [21]. Campagner et al. [11] showed that three-way decisions and conformal prediction share several formal properties as cautious learning approaches. Nearest-neighbour conformal predictors [22, 23] are also closely related to rough membership. In the literature searched for this study, Pawlak's positive, boundary and unseen regions have not been evaluated as a reliability signal for conformal prediction against learned and randomised groupings.

3 Preliminaries

This section introduces the basic notions of rough sets and conformal prediction used in the rest of the paper.

Definition 1 (Decision table [6])**.** *A decision table is a tuple $DT = (U, C \cup \{d\})$, where U is a non-empty finite set of objects, C is a non-empty finite set of condition attributes, and d is a decision attribute that assigns a class label $d(x)$ to each $x \in U$. Every attribute $a \in C$ takes values in a finite set V_a.*

Numeric attributes are discretised before the definitions below are applied (Section 5.2).

Definition 2 (Indiscernibility relation and granule [5])**.** *For a subset $B \subseteq C$, the indiscernibility relation is*

$$IND(B) = \{(x,y) \in U \times U : a(x) = a(y),\ \forall a \in B\}. \tag{1}$$

The equivalence class of x under $IND(B)$, denoted $[x]_B$, is called the information granule of x. Its key is the tuple $\kappa_B(x) = (a(x))_{a \in B}$.

Definition 3 (Lower and upper approximations [5])**.** *For a class $X \subseteq U$, the lower and upper approximations with respect to B are*

$$\underline{B}X = \{x \in U : [x]_B \subseteq X\}, \tag{2}$$

$$\overline{B}X = \{x \in U : [x]_B \cap X \neq \varnothing\}. \tag{3}$$

For the partition $U/d = \{X_1, \ldots, X_K\}$ induced by the decision attribute, the positive region is $POS_B(d) = \bigcup_j \underline{B}X_j$ and the boundary region is $BND_B(d) = \bigcup_j \overline{B}X_j \setminus POS_B(d)$.

Definition 4 (Region of a new input)**.** *Let $D_{\text{fit}} \subseteq U$ be a training sample and let $G_v = \{x \in D_{\text{fit}} : \kappa_B(x) = v\}$ be the training granule with key v. For any input x, possibly outside D_{fit}, the region of x is*

$$\rho_B(x) = \begin{cases} \text{POS}, & G_{\kappa_B(x)} \neq \varnothing,\ |d(G_{\kappa_B(x)})| = 1, \\ \text{BND}, & G_{\kappa_B(x)} \neq \varnothing,\ |d(G_{\kappa_B(x)})| \geq 2, \\ \text{NEW}, & G_{\kappa_B(x)} = \varnothing, \end{cases} \tag{4}$$

where $d(G)$ denotes the set of class labels occurring in G.

POS and BND coincide with Pawlak's positive and boundary regions computed on the training sample. NEW extends the observed universe to keys that are absent from training. It means unseen in this sample and does not imply that the input is out of distribution. Inputs in BND or NEW are called non-positive. The number of attributes $k = |B|$ sets the level of granulation: a small k gives coarse granules with many boundary inputs and few unseen ones, while a large k gives fine granules with few boundary inputs and many unseen ones.

Definition 5 (Split conformal prediction [1])**.** *Let $D_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$ be a calibration sample disjoint from D_{fit}, and let $s(x,y)$ be a nonconformity score fixed independently of D_{cal}. The conformal p-value of a candidate label y and the prediction set at level α are*

$$p(x,y) = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{s(X_i,Y_i) \geq s(x,y)\}}{n+1}, \tag{5}$$

$$C_\alpha(x) = \{y :\ p(x,y) > \alpha\}. \tag{6}$$

In this paper, the LAC score $s(x,y) = 1 - \widehat{p}(y \mid x)$ is used [12], where $\widehat{p}$ is the class probability estimated by the trained classifier.

Definition 6 (Mondrian conformal prediction [1, 4])**.** *Let $g(x,y)$ be a taxonomy that maps each pair to a group. A Mondrian conformal predictor computes*

the p-value (5) of a candidate pair using only the calibration pairs with $g(X_i, Y_i) = g(x, y)$.

Theorem 1 (Group-wise coverage [1, 4])**.** *If the calibration and test pairs are exchangeable and the taxonomy and score are fixed independently of the calibration and test outcomes, then for every group c with positive probability,*

$$\Pr\{Y_{n+1} \in C_\alpha(X_{n+1}) \mid g(X_{n+1}, Y_{n+1}) = c\} \;\geq\; 1 - \alpha. \tag{7}$$

Remark 1. *Since ρ_B in Definition 4 depends only on x and on D_{fit}, the region taxonomy, and its product with the candidate label, satisfy the conditions of Theorem 1. This guarantee is a known result and is not a contribution of this paper. The open question addressed below is empirical: whether the regions are informative.*

4 Proposed framework

The proposed region-aware framework is shown in Fig. 1. It consists of three stages: region construction, region-based Mondrian calibration, and region-based miscoverage diagnosis.

4.1 Region construction

Given a training sample, numeric attributes are discretised into a small number of quantile bins and the k attributes with the largest mutual information with the class label are selected. The training objects are then grouped by their keys $\kappa_B(x)$ into a granule table, which records the class labels occurring in each granule. For a new input, the region is obtained by a single lookup of its key, following Definition 4.

4.2 Region-based Mondrian calibration

The region taxonomy is used to calibrate a conformal predictor separately within each region, as described in Algorithm 1. Two variants are considered. The Region variant uses $g(x, y) = \rho_B(x)$, and the Region $\times$ class variant uses $g(x, y) = (\rho_B(x), y)$, which combines region-wise and class-wise calibration. When a group has no calibration points, the corresponding candidate label is included, so that the procedure never loses coverage in an empty group.

4.3 Region-based miscoverage diagnosis

To test whether the regions carry information about conformal failures, a marginal conformal predictor is first applied, and the miss indicator $M = \mathbf{1}\{Y \notin C_\alpha(X)\}$ is recorded for every input in an auxiliary sample and a test sample. A logistic model of M is fitted on the auxiliary sample and evaluated on the test sample. The base model uses five standard signals: the predicted class, the largest predicted probability, the gap between the two largest probabilities, the predictive entropy and $\log(1 + |G_{\kappa_B(x)}|)$, the size of the training granule. The region model adds two indicators for BND and NEW, with POS as the reference category. All of these predictors are available for a new input before its label is known. The improvement of the region model over the base model on the test sample measures the additional information that the regions provide.

Algorithm 1 Region-based Mondrian conformal prediction

Require: D_{fit}, D_{cal}, attribute set B, level α, flag ByClass
Ensure: prediction set $C_\alpha(x)$ for a test input x

```
Train p̂ on D_fit; let s(x,y) = 1 − p̂(y | x)
Build {G_v} from D_fit; define ρ_B by Eq. (4)
for (X_i, Y_i) ∈ D_cal do
    g_i ← ρ_B(X_i), or (ρ_B(X_i), Y_i) if ByClass
end for
For each group c: S_c ← sorted {s(X_i, Y_i) : g_i = c}
C_α(x) ← ∅
for each label y do
    c ← ρ_B(x), or (ρ_B(x), y) if ByClass
    if S_c = ∅ then C_α(x) ← C_α(x) ∪ {y}
    else
        p ← (1 + |{t ∈ S_c : t ≥ s(x,y)}|)/(|S_c| + 1)
        if p > α then C_α(x) ← C_α(x) ∪ {y}
        end if
    end if
end for
return C_α(x)
```

4.4 Computational complexity

Let n_f, n_c and m denote the sizes of the training, calibration and test samples, and let K be the number of classes. Building the granule table requires $O(n_f k)$ time with hashing, and each region lookup requires $O(k)$ time. Sorting the calibration scores within groups requires $O(n_c \log n_c)$ time, and computing the prediction set of a test input requires $O(K \log n_c)$ time by binary search. The regions therefore add only negligible cost to standard split conformal prediction. In the experiments, all 230 runs required 371 seconds of computation in total.

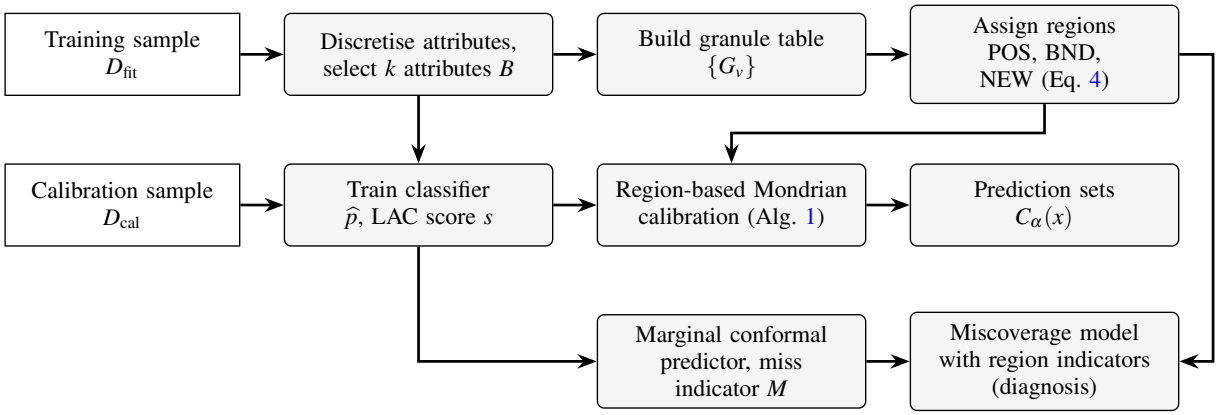


Fig. 1 Framework of the proposed region-aware conformal reliability analysis

Table 1 Datasets used in the experiments

Dataset	OpenML ID	Version	Rows
Adult	1590	2	48,842
Bank Marketing	1461	1	45,211
Covertype	1596	4	50,000
Nursery	26	1	12,960
Mushroom	24	1	8,124
Kr-vs-kp	3	1	3,196
Car	21	1	1,728
Credit-g	31	1	1,000

5 Experiments

The experiments are designed to answer three research questions:

RQ1. Does region membership predict which held-out inputs a marginal conformal predictor misses, beyond the predicted class, model confidence and granule size?

RQ2. What does region-based calibration gain and cost, compared with class, confidence-bin, decision-tree and randomised groupings?

RQ3. At which levels of granulation does the region signal appear or vanish?

5.1 Datasets

Eight classification datasets from OpenML [24] are used, as summarised in Table 1. The Covertype dataset was capped at 50,000 rows, sampled once with a fixed seed.

5.2 Experimental settings

Each run divides the data at random into a training sample (55%), a calibration sample (15%), an auxiliary sample (15%) and a test sample (15%), using scikit-learn [25] with seeds 42 to 51. Numeric attributes with more than four distinct values are imputed with training medians and discretised into four quantile bins whose edges are fitted on the training sample. Categorical attributes are mapped to integer codes. The k attributes with the largest mutual information on the training sample are retained, for $k \in \{4, 6, 8\}$. Since Car has six attributes, its $k = 8$ setting coincides with $k = 6$ and is counted once, which gives 23 dataset-by-k settings and 230 runs.

The classifier is an L2-regularised logistic regression (regularisation constant 1, lbfgs solver, at most 2,000 iterations) trained on one-hot encodings of the selected attributes. The LAC score is used with $\alpha = 0.10$.

The following groupings are compared for RQ2. All of them are calibrated on the calibration sample and evaluated on the test sample.

- Marginal: a single group.
- Class: one group per candidate label.
- Confidence bin $\times$ class: tertiles of the largest predicted probability, with cut points fitted on the auxiliary sample, crossed with the candidate label.
- Tree-3 and Tree-6: the leaves of regression trees with at most three or six leaves, fitted on the auxiliary sample to predict the true-label score.
- Randomised maps: twenty per run, with as many groups as there are regions and calibration counts matched to those of the regions. Since they are built

from calibration inputs, they serve as descriptive controls rather than valid alternative predictors.
- Region and Region $\times$ class: Algorithm 1.

For RQ1, the region indicators are compared with Tree-6 leaf indicators and with the group indicators of the randomised maps, reported as the mean over twenty maps and, as a stress test, as the best of twenty. Since the best map is selected on the test outcomes, the latter comparison is deliberately biased against the regions.

5.3 Evaluation metrics and statistical analysis

For RQ1, the primary metric is the reduction in test log loss of the miscoverage model when the region indicators are added, and the change in AUC is reported as a secondary metric. For RQ2, test coverage, mean set size, coverage on boundary, unseen and non-positive inputs, the lowest coverage over true classes and the lowest coverage over confidence bins are reported.

Within a dataset, 95% intervals are obtained from 2,000 bootstrap resamples of the ten seeds. Across datasets, a two-level bootstrap resamples datasets and then seeds. Comparisons between groupings are paired within dataset, seed and k. A setting is called degenerate when region membership does not vary in the auxiliary or test sample. Four of the 23 settings are degenerate (Car with $k = 6$, Covertype with $k = 4$, Mushroom with $k = 8$ and Nursery with $k = 8$), and they are included in the primary analysis.

5.4 Results for RQ1: miscoverage prediction

Table 2 and Fig. 2 report the results of the miscoverage diagnosis. When all datasets are pooled, adding the region indicators reduced the held-out log loss by 0.0035, with a hierarchical 95% interval from -0.00002 to 0.0081. Since this interval includes zero, the pooled improvement is not established. The AUC increased by 0.0054 (0.0006 to 0.0116).

As shown in Table 2, the region indicators reduced the log loss on seven of the eight datasets and outperformed the mean randomised map on seven of the eight. Against the best of the twenty maps, they performed better on only three datasets, with a pooled difference of -0.0001 (-0.0050 to 0.0053). Tree-6 outperformed the regions on Covertype, Kr-vs-kp and Mushroom, while the regions outperformed Tree-6 on the remaining five datasets. Only on Nursery and Car did the regions raise the AUC by more than 0.01. When the four degenerate settings are excluded, the pooled log-loss reduction is 0.0056 (0.0001 to 0.0129) and the AUC gain is 0.0086 (0.0007 to 0.0187).

5.5 Results for RQ2: calibration trade-offs

Table 3 reports the results at $k = 6$ for Nursery and Bank Marketing, and Fig. 3 plots coverage on non-positive inputs against mean set size. Under marginal calibration, coverage on boundary inputs dropped to 0.507 on Nursery and 0.833 on Bank Marketing. Region calibration restored it to 0.900 on both datasets, while the mean set size increased only from 0.960 to 1.018 on Nursery and from 1.005 to 1.060 on Bank Marketing. It should be noted that part of this improvement is by construction, since region calibration targets exactly these inputs. The informative quantities are therefore the set-size cost and the behaviour of the other groupings on the same inputs.

Table 4 summarises the paired differences over all datasets and values of k. Region calibration raised non-positive coverage above the randomised maps by 0.127 at 3.1% larger sets. Compared with Tree-6, its coverage advantage of 0.024 was not distinguishable from zero, while its sets were 4.9% smaller. Region $\times$ class calibration produced 18.3% smaller sets than confidence bin $\times$ class calibration, and its sets were smaller in every seed in 14 of the 23 settings. However, its lowest confidence-bin coverage was 8.3 points lower and its worst-class coverage 1.1 points lower. Therefore, neither grouping dominates the other.

5.6 Results for RQ3: effect of granulation

The region signal depends on the level of granulation. The mean log-loss reductions were 0.0059 at $k = 4$, 0.0031 at $k = 6$ and 0.0005 at $k = 8$. Fig. 4 plots the reduction of each of the 23 settings against the share of unseen and boundary test inputs. The reduction decreased as the unseen share increased (Spearman correlation -0.60), and it was almost unrelated to the boundary share (Spearman correlation -0.04).

The largest gains were obtained at coarse granulations with a moderate boundary share and almost no unseen inputs: Nursery with $k = 4$ (40% boundary inputs, no unseen inputs, reduction 0.0304, AUC gain 0.045), Nursery with $k = 6$ (13% boundary inputs, reduction 0.0228) and Car with $k = 4$ (25% boundary inputs, reduction 0.0151, AUC gain 0.028). At

Table 2 Results for RQ1: reduction in held-out log loss of the miscoverage model when each grouping is added to the base model (positive values indicate improvement), and AUC change of the region model. Means over seeds and k; brackets give 95% seed-bootstrap intervals, which are shown for the comparators in Fig. 2. The randomised-map column is the mean over twenty maps

Dataset	Region	Region AUC change	Randomised maps	Tree-6
Nursery	+0.0177 [0.0154, 0.0201]	+0.023 [0.020, 0.026]	+0.0002	+0.0090
Car	+0.0075 [0.0010, 0.0135]	+0.014 [0.008, 0.020]	−0.0009	−0.0020
Kr-vs-kp	+0.0026 [0.0008, 0.0047]	+0.002 [−0.000, 0.005]	+0.0010	+0.0045
Mushroom	+0.0014 [0.0004, 0.0024]	+0.000 [−0.000, 0.000]	+0.0003	+0.0024
Bank Marketing	+0.0009 [0.0008, 0.0011]	+0.001 [0.001, 0.002]	−0.0001	+0.0005
Covertype	+0.0001 [−0.0003, 0.0004]	+0.001 [0.001, 0.002]	+0.0001	+0.0028
Adult	+0.0000 [−0.0000, 0.0001]	+0.000 [0.000, 0.000]	+0.0001	−0.0000
Credit-g	−0.0025 [−0.0055, 0.0006]	+0.001 [−0.007, 0.009]	−0.0084	−0.0165

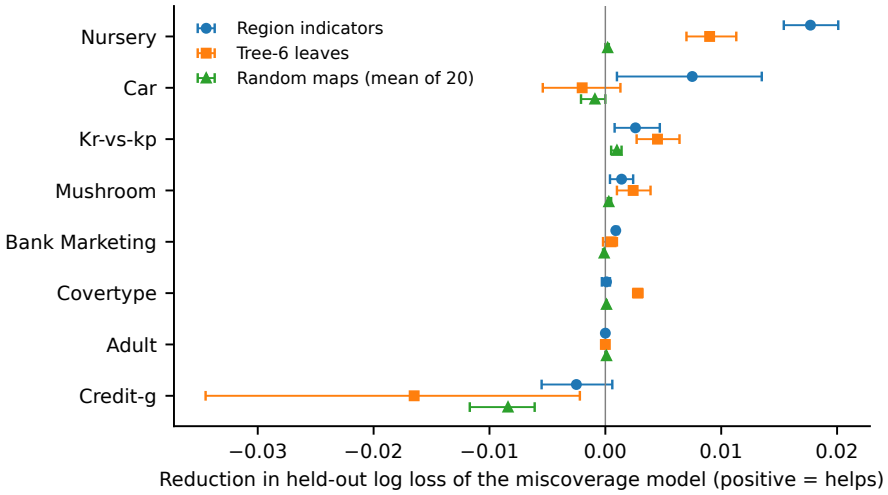


Fig. 2 Reduction in held-out log loss of the miscoverage model for region indicators, Tree-6 leaves and randomised maps (mean of twenty), with 95% seed-bootstrap intervals

the two extremes, the regions carried little information. On Credit-g with $k = 6$ and $k = 8$, 62% and 88% of the test inputs were unseen and the reductions were negative. On Covertype, 98% to 100% of the inputs were boundary inputs, so that region membership hardly varied. These correlations are descriptive, because the 23 settings are not independent and changing k changes the granules, the region shares and the granule sizes simultaneously.

5.7 Sensitivity analysis

The randomised maps could not always match the calibration counts of very small regions. The median, over runs, of the largest relative mismatch was 0.007 for Car, 0.024 for Nursery, 0.067 for Adult, 0.116 for Credit-g and 0.129 for Bank Marketing, but 1.0 for Covertype, Kr-vs-kp and Mushroom, whose smallest regions contained only about one to nine test inputs per run.

A post hoc sensitivity analysis was therefore carried out on the non-degenerate runs whose twenty maps all matched within 25%. This retained 88 of the 230 runs, from Adult (21), Bank Marketing (19), Car (10), Credit-g (24) and Nursery (14). On these runs, the regions outperformed the mean randomised map by 0.0105 in log loss (0.0015 to 0.0203) but not the best of twenty (0.0034; -0.0068 to 0.0144). Region calibration raised non-positive coverage over the randomised maps by 0.117 (0.028 to 0.205) at 4.5% larger sets (2.2% to 6.7%). Since the filter was defined after the results were seen and removes three datasets, this analysis supports, but does not replace, the primary results.

Table 3 Results for RQ2 at $k = 6$ (test means over ten seeds). Cov.: overall coverage; Size: mean set size; BND and NEW: coverage on boundary and unseen inputs; Worst class and Worst bin: lowest coverage over true classes and over confidence bins. Each Nursery test sample contains 1,944 inputs, of which about 255 are boundary and 6 unseen; each Bank Marketing test sample contains 6,782 inputs, of which about 2,776 are boundary and 372 unseen

Dataset	Grouping	Cov.	Size	BND	NEW	Worst class	Worst bin
Nursery	Marginal	0.895	0.960	0.507	0.971	0.257	0.712
	Class	0.894	1.999	0.672	0.929	0.874	0.818
	Bin × class	0.903	3.378	0.795	0.846	0.881	0.892
	Tree-6	0.896	1.012	0.707	0.929	0.467	0.823
	Random maps	0.895	0.969	0.511	0.971	0.258	0.714
	Region	0.897	1.018	0.900	0.983	0.467	0.719
	Region × class	0.896	2.249	0.906	1.000	0.878	0.822
Bank Marketing	Marginal	0.900	1.005	0.833	0.779	0.263	0.748
	Class	0.901	1.222	0.826	0.814	0.896	0.745
	Bin × class	0.904	1.685	0.906	0.840	0.897	0.894
	Tree-6	0.905	1.100	0.869	0.768	0.581	0.827
	Random maps	0.900	1.006	0.833	0.778	0.269	0.749
	Region	0.899	1.060	0.900	0.911	0.431	0.747
	Region × class	0.900	1.268	0.903	0.911	0.897	0.772

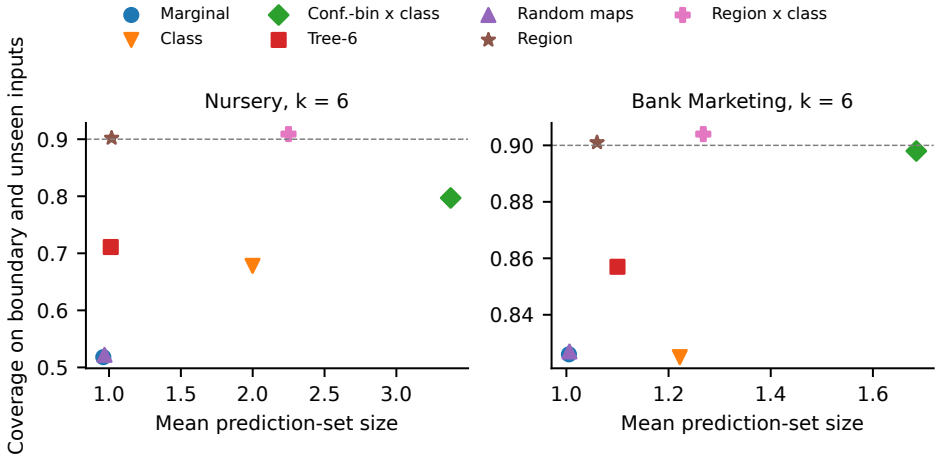


Fig. 3 Coverage on non-positive (boundary and unseen) inputs against mean set size at $k = 6$. The dashed line marks the nominal level 0.90

5.8 Discussion

The experimental results show that rough-set regions are informative on some datasets and uninformative on others. On Nursery and Car, region membership identified failures of marginal conformal prediction that were not captured by the predicted class, the model confidence or the granule size. Both are fully discrete datasets in which a few attributes leave genuinely ambiguous granules. On the remaining datasets, the boundary label largely repeated the information already contained in the confidence, and on Credit-g the additional indicators reduced the predictive performance. Therefore, the general claim that boundary membership predicts conformal failures is not supported by the experiments, whereas the local claim is.

From the perspective of granular computing, these results can be explained by the level of granulation. At coarse granulation, the granules are large and their label conflicts are stable properties of the data, so that boundary membership is informative. As more attributes are added, the granules shrink, the conflicts become rare and an increasing share of new inputs falls outside the training universe. Since unseen

Table 4 Paired differences over all datasets and k (first grouping minus second) with 95% hierarchical bootstrap intervals; set-size differences are relative

Comparison	Non-positive coverage	Mean set size
Region − randomised maps	+0.127 [0.037, 0.260]	+3.1% [1.4%, 5.1%]
Region − Tree-6	+0.024 [−0.004, 0.056]	−4.9% [−7.3%, −2.7%]
Region × class − conf. bin × class	+0.041 [−0.008, 0.123]	−18.3% [−25.4%, −11.1%]

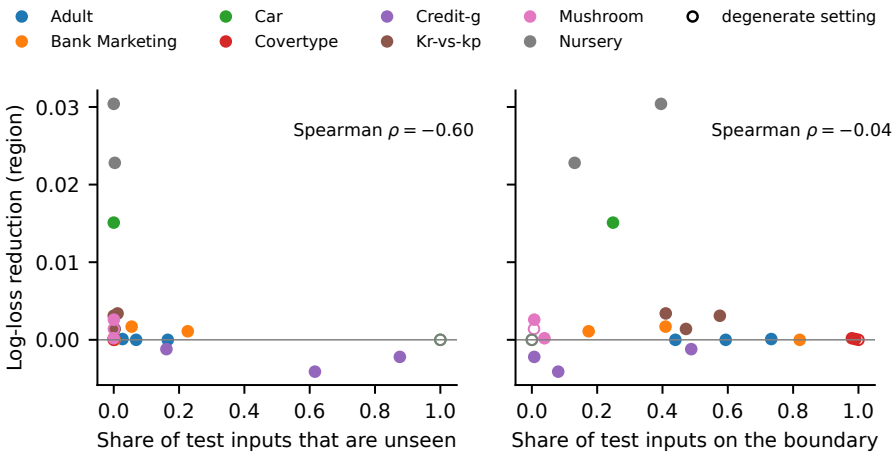


Fig. 4 Log-loss reduction of each of the 23 dataset-by-k settings against the share of unseen (left) and boundary (right) test inputs. Hollow markers indicate degenerate settings

keys carry no label information, the signal fades. At the opposite extreme, when nearly every granule is a boundary granule, as on Covertype, membership no longer distinguishes between inputs. Rough-set regions are therefore most useful at an intermediate level of granulation, which can be identified from the training granules before any calibration is performed.

In practice, the region profile can be inspected before calibration. The results suggest using the regions when the boundary share is moderate and unseen inputs are rare, and avoiding them when nearly all inputs are unseen or nearly all are boundary inputs. Region calibration is an inexpensive way to restore coverage on ambiguous inputs, but it does not repair class imbalance: on Bank Marketing, its worst-class coverage was 0.431. Region × class calibration covers regions and classes with smaller sets than confidence bin × class calibration but covers low-confidence inputs less well. Users who are mainly concerned with low-confidence predictions should prefer confidence bins, whereas users concerned with ambiguous attribute combinations and set size should prefer regions.

This study has several limitations. Only eight tabular datasets, one classifier, one score, one significance level and three levels of granulation were considered. The primary RQ1 interval includes zero. The analysis was revised once after the first full run, and neither the analysis nor the sensitivity filter was registered in advance; the results are therefore exploratory. The randomised maps are not valid Mondrian predictors, and Tree-6 reuses the auxiliary sample to learn its leaves and to fit the miscoverage model. The Tree-6 grouping is not an implementation of the learned-partition methods [7, 8] or the multi-group method [9], so no superiority over these methods is claimed.

5.9 Reproducibility and use of AI tools

The experiments used Python 3.11.16, NumPy 2.4.6, pandas 2.3.3, scikit-learn 1.9.1 and OpenML 0.15.1. The experiment script has SHA-256 digest `148b1f4dcc90822fb1506866f9250fa315450f952c844174968d05ea95a39664`. The data, splits, classifier, groupings and miscoverage model were fixed in the script before its first full run; after that run, only the analysis and logging were revised, and all reported results come from the revised

run. A generative AI assistant (Claude, Anthropic) was used to write and debug the experiment and analysis code, to propose the comparators and the analysis plan, and to draft the manuscript text. The author ran all experiments, checked every reported number against the run outputs, revised the text and takes full responsibility for the content.

6 Conclusion

In this paper, a region-aware framework was proposed to examine whether Pawlak's positive, boundary and unseen regions can serve as a reliability signal for conformal prediction. The regions were used both as predictors of conformal miscoverage and as a training-free Mondrian taxonomy. The experiments on eight datasets showed that, pooled over all settings, the regions added little beyond class, confidence and granule size, with a log-loss reduction of 0.0035 whose interval includes zero. At coarse granulation on discrete datasets, however, they identified failures that standard signals missed, raising the AUC by 0.023 on Nursery. Region calibration restored coverage on ambiguous inputs at a few percent larger sets, and region-by-class calibration provided smaller sets than confidence-bin-by-class calibration at the cost of weaker coverage on low-confidence inputs. The usefulness of the regions depends on the level of granulation and can be assessed from the region profile before calibration.

In future work, the pre-calibration rule suggested here will be tested on further datasets, the framework will be compared with partition-learning and multi-group conditional methods, and the analysis will be extended to adaptive scores, other classifiers and distribution shift.

Statements and Declarations

- **Funding:** No funding was received for conducting this study.
- **Competing interests:** The author has no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate:** Not applicable. The study uses publicly available benchmark data only.
- **Consent for publication:** Not applicable.
- **Data availability:** All datasets analysed in this study are publicly available on OpenML (https://www.openml.org) under the identifiers and versions listed in Table 1.
- **Code availability:** The code, per-run and per-example outputs, and data and split digests are available from the corresponding author on reasonable request.
- **Author contribution:** B.M. conceived the study, ran all experiments, verified the results, and revised and finalised the manuscript. The use of a generative AI assistant is described in the reproducibility subsection of Section 5.

References

[1] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005). https://doi.org/10.1007/b106715

[2] Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023) https://doi.org/10.1561/2200000101

[3] Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA **10**(2), 455–482 (2021) https://doi.org/10.1093/imaiai/iaaa017

[4] Vovk, V.: Conditional validity of inductive conformal predictors. In: Proceedings of the Asian Conference on Machine Learning. PMLR, vol. 25, pp. 475–490 (2012). https://proceedings.mlr.press/v25/vovk12.html

[5] Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11**(5), 341–356 (1982) https://doi.org/10.1007/BF01001956

[6] Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht (1991). https://doi.org/10.1007/978-94-011-3534-4

[7] Martinez Gil, T., *et al.*: Identifying homogeneous and interpretable groups for conformal prediction. In: Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence. Proceedings of Machine Learning Research, vol. 244, pp. 2471–2485 (2024). https://proceedings.mlr.press/v244/martinez-gil24a.html

[8] Kiyani, S., Pappas, G.J., Hassani, H.: Conformal prediction with learned features. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 24749–24769 (2024). https://proceedings.mlr.press/v235/kiyani24a.html

[9] Gibbs, I., Cherian, J.J., Candès, E.J.: Conformal prediction with conditional guarantees. Journal of the Royal Statistical Society Series B: Statistical Methodology **87**(4), 1100–1126 (2025) https://doi.org/10.1093/jrsssb/qkaf008

[10] Jung, C., Noarov, G., Ramalingam, R., Roth, A.: Batch multivalid conformal prediction. In: International Conference on Learning Representations (2023)

[11] Campagner, A., Cabitza, F., Berjano, P., Ciucci, D.: Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches. Information Sciences **579**, 347–367 (2021) https://doi.org/10.1016/j.ins.2021.08.009

[12] Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (2019) https://doi.org/10.1080/01621459.2017.1395341

[13] Romano, Y., Sesia, M., Candès, E.: Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems, vol. 33, pp. 3581–3591 (2020)

[14] Ding, T., Angelopoulos, A.N., Bates, S., Jordan, M.I., Tibshirani, R.J.: Class-conditional conformal prediction with many classes. In: Advances in Neural Information Processing Systems, vol. 36, pp. 64555–64576 (2023). https://doi.org/10.52202/075280-2817

[15] Romano, Y., Barber, R.F., Sabatti, C., Candès, E.: With malice toward none: Assessing uncertainty via equalized coverage. Harvard Data Science Review **2**(2) (2020) https://doi.org/10.1162/99608f92.03f00592

[16] Pawlak, Z., Skowron, A.: Rough membership functions: A tool for reasoning with uncertainty. Banach Center Publications **28**, 135–150 (1993)

[17] Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences **46**(1), 39–59 (1993) https://doi.org/10.1016/0022-0000(93)90048-2

[18] Yao, Y.: Three-way decisions with probabilistic rough sets. Information Sciences **180**(3), 341–353 (2010) https://doi.org/10.1016/j.ins.2009.09.021

[19] Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems **17**(2–3), 191–209 (1990) https://doi.org/10.1080/03081079008935107

[20] Zhu, W., Wang, F.-Y.: On three types of covering-based rough sets. IEEE Transactions on Knowledge and Data Engineering **19**(8), 1131–1144 (2007)

[21] Jensen, R., Shen, Q.: Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering **16**(12), 1457–1471 (2004) https://doi.org/10.1109/TKDE.2004.96

[22] Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. In: Machine Learning: ECML 2002. Lecture Notes in Computer Science, vol. 2430, pp. 381–390. Springer, Berlin, Heidelberg (2002)

[23] Györfi, L., Walk, H.: Nearest neighbor based conformal prediction. Annales de l'ISUP **63**(2–3), 173–190 (2019)

[24] Vanschoren, J., Rijn, J.N., Bischl, B., Torgo, L.: OpenML: Networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2013) https://doi.org/10.1145/2641190.2641198

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)