

# Reliable Household Demographic Classification

**Javier Carreno**

JAVIER.CARRENO.2023@LIVE.RHUL.AC.UK

**Khuong An Nguyen**

KHUONG.NGUYEN@RHUL.AC.UK

**Zhiyuan Luo**

ZHIYUAN.LUO@RHUL.AC.UK

*Department of Computer Science, Royal Holloway, University of London*

*Surrey TW20 0EX, United Kingdom*

**Andrew Fish**

ANDREW.FISH@LIVERPOOL.AC.UK

*Department of Computer Science, University of Liverpool*

*Liverpool L69 3BX, United Kingdom*

**Editor:** Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

## Abstract

We propose the Hybrid Calibration Score (HCS), a new nonconformity measure for inductive conformal prediction. HCS combines instance-level scoring with global model calibration via Expected Calibration Error. On a real-world demographic classification task, HCS achieves 99% coverage with smaller prediction sets ( $\text{APS} = 1.55$ ) and higher decisiveness ( $\text{OneC} = 55.19\%$ ) than standard measures, while preserving formal coverage guarantees.

**Keywords:** Conformal Prediction, Nonconformity Measure, Audience Modelling

## 1. Introduction

Accurate confidence estimation is essential for deploying machine learning in risk-sensitive settings. Conformal Prediction (CP) offers a principled way to generate prediction sets with formal coverage guarantees under minimal assumptions (Vovk et al., 2005). Its effectiveness hinges on the choice of nonconformity measure (NCM), which scores how atypical a prediction is (Papadopoulos et al., 2002; Aleksandrova and Chertov, 2021).

We propose the **Hybrid Calibration Score (HCS)**, a model-agnostic NCM that combines proper scoring rules (*Brier Score* and *Log Loss*) with a global Expected Calibration Error (*ECE*) element. This ensures both per-instance accuracy and global probabilistic trustworthiness, integrating seamlessly into any CP pipeline.

## 2. Hybrid Calibration Score (HCS)

Given a model’s output distribution  $p(y|x)$ , HCS is defined as:

$$a(x, y) = \alpha \cdot \text{Brier}(x, y) + \beta \cdot \text{LogLoss}(x, y) + \gamma \cdot \text{ECE}, \quad (1)$$

where  $\alpha + \beta + \gamma = 1$ , and weights are tuned using Bayesian optimisation.

The *Expected Calibration Error (ECE)* (Guo et al., 2017) quantifies the mismatch between predicted confidence and empirical accuracy across  $n$  confidence bins:

$$\text{ECE} = \sum_{i=1}^n \frac{|B_i|}{N} \cdot |\text{acc}(B_i) - \text{conf}(B_i)| \quad (2)$$

where  $|B_i|$  is the number of samples in bin  $i$ ,  $N$  is the total number of samples, and  $\text{acc}(B_i), \text{conf}(B_i)$  are the accuracy and mean confidence in bin  $B_i$ . ECE acts as a global regulariser, complementing instance-wise terms by enforcing population-level calibration.

### 3. Empirical Evaluation

We evaluate HCS on a real-world household classification task using 19,386 TV viewing instances across six categories (Carreno et al., 2024), with a Random Forest baseline. Data is split 60/30/10 (train/calibration/test) using stratified sampling. Experiments are implemented in Python with `scikit-learn`; a custom wrapper inspired by `mapie` enables flexible use of non-standard nonconformity measures. Results are averaged over five random seeds. Metrics follow Kato et al. (2023).

- **Coverage**: Fraction of prediction sets containing the true label.
- **APS**: Average Prediction Set size (lower is better).
- **OneC**: Rate of singleton prediction sets (higher implies greater decisiveness).

Table 1 summarises results at a significance level of  $\alpha = 0.01$ , corresponding to a 99% confidence level.

Table 1: Performance comparison of NCMs at  $\alpha = 0.01$ .

NCM	Coverage (%)		APS		OneC (%)	
	Mean	Std	Mean	Std	Mean	Std
Hinge Loss	99.32	0.08	1.64	0.04	49.23	2.95
Gap	99.24	0.28	2.74	0.06	54.30	1.46
Brier Score	99.21	0.29	2.50	0.08	54.78	1.67
<b>HCS</b>	98.99	0.17	<b>1.55</b>	0.04	<b>55.19</b>	2.71

### Per-Class Analysis

To further assess the performance of the two strongest NCMs, we compare *Hinge Loss* and *HCS* across demographic categories at  $\alpha = 0.01$  (99% confidence). Table 2 reports per-class coverage, APS, and OneC.

Table 2: Per-class comparison of *Hinge Loss* vs. *HCS* at  $\alpha = 0.01$ .

Category	Coverage (%)		APS		OneC (%)	
	Hinge   HCS		Hinge   HCS		Hinge   HCS	
Couple w/ adult children	98.26 (0.46)	97.42 (0.83)	2.26 (0.06)	<b>2.10</b> (0.08)	11.43 (2.56)	<b>19.16</b> (3.97)
Couple w/ teenagers	98.54 (0.55)	98.01 (0.81)	2.25 (0.06)	<b>2.11</b> (0.05)	12.32 (2.59)	<b>19.74</b> (3.58)
Couple w/ young kids	94.29 (3.19)	94.29 (3.19)	2.69 (0.28)	<b>2.60</b> (0.19)	2.86 (3.91)	<b>4.29</b> (3.91)
Only middle-aged adults	99.77 (0.16)	99.63 (0.26)	1.34 (0.04)	<b>1.28</b> (0.04)	66.92 (4.09)	<b>72.42</b> (3.35)
Only young adults	100.00 (0.00)	100.00 (0.00)	1.63 (0.04)	<b>1.53</b> (0.04)	37.30 (4.41)	<b>46.93</b> (4.26)
Seniors	99.34 (0.31)	98.95 (0.45)	1.31 (0.05)	<b>1.28</b> (0.05)	81.16 (3.16)	<b>82.26</b> (2.91)

### 4. Conclusion

HCS delivers a compelling trade-off between reliability and efficiency. By incorporating global calibration into instance-wise scoring, it produces smaller, more decisive prediction sets without compromising coverage. Future work includes extensions to multi-label tasks and per-instance calibration.

## References

- Marharyta Aleksandrova and Oleg Chertov. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. In *Conformal and Probabilistic Prediction and Applications*, pages 151–170. PMLR, 2021.
- Javier Carreno, Khuong An Nguyen, Zhiyuan Luo, and Andrew Fish. Unlocking viewer insights in linear television: A machine learning approach. In *International Conference on Business Informatics Research*, pages 53–67. Springer, 2024.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- Yuko Kato, David MJ Tax, and Marco Loog. A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383, 2023.
- Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.
- Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.