Predicting Unobserved Multi-Class Sensitive Attributes: Enhancing Calibration with Nested Dichotomies for Fairness

Ana María Patrón Piñerez[a], **Agathe Fernandes Machado**[b], Arthur Charpentier[b], and Ewen Gallic[c]

[a]Universidad de los Andes, Bogotá, Colombia, [b]Université du Québec à Montréal (fernandes_machado.agathe@courrier.uqam.ca), [c]AMSE, Aix-Marseille Université

Motivations

Colorado legislation SB21-169: insurance companies in Colorado are required to assess their big data systems to prevent **unfair discrimination** based on **protected characteristics** such as **race**.

Uncollected sensitive attributes: "*What we can't measure, we can't understand*" [1]. **Race** is often infrequently or incompletely collected by insurers [8].

Bayesian methods for predicting race have emerged, using surname, first name, and geolocation data from an aggregate source, the USA Census data. The **race dummy predictions** can then be used to assess discrimination in insurance premiums or rates.

➤ Bayesian Improved Surname Geocoding (**BISG**): [6] uses surname (S) and geocoding (G) to predict **race** (R) by assuming: $G \perp\!\!\!\perp S \mid R$. For a new individual with $G = g$ and $S = s$,

$$\forall r \in \mathcal{R},\ \mathbb{P}(R = r|G = g, S = s) = \frac{\mathbb{P}(R = r|S = s)\,\mathbb{P}(G = g|R = r)}{\sum_{r' \in R} \mathbb{P}(R = r'|S = s)\,\mathbb{P}(G = g|R = r')},$$

where $\mathbb{P}(R|S = s)$ and $\mathbb{P}(G = g|R)$ are collected from Census data.

The prediction problem is ***imbalanced***.

	White (W)	Hispanic (H)	Black (B)	Other (O)	Asian (A)
Proportion	0.57	0.17	0.11	0.10	0.05
Accuracy for BISG	0.96	0.83	0.54	0.14	0.57

Multi-class calibration: The SOA recommends that the predictive model for race to be calibrated (or auto-calibrated) [3].

Our contribution: using Nested Dichotomies to decompose an imbalanced multi-class classification task into a sequence of balanced binary probabilistic classifiers.

Comparison of multi-classification models and nested dichotomies.

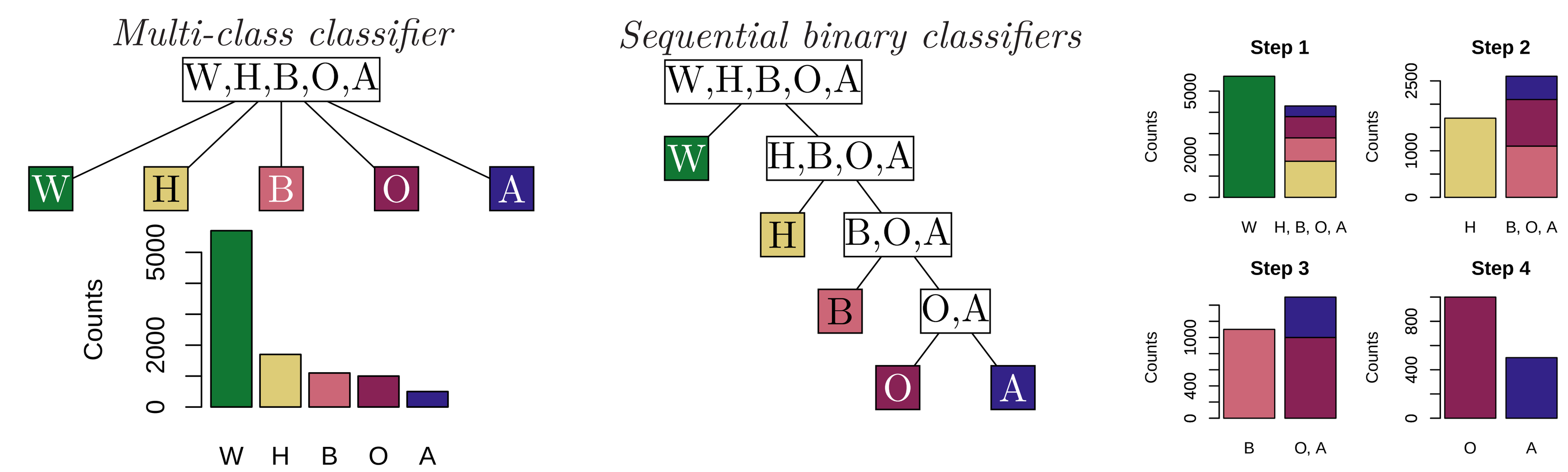
Approach	Strengths	Weaknesses
Multi-class classifier	• Computationally efficient: requires training a single model	• Imbalanced predictive task • Requires multi-class calibration
Nested dichotomies	• Balanced predictive tasks • Requires binary calibration	• Computationally inefficient: requires training multiple models • Accumulates prediction error

Nested Dichotomies

Nested dichotomies decompose a multi-class problem into several binary problems in a **tree-like structure**. The objective of achieving **balanced outcomes** precludes the use of multiple trees in the nested dichotomy algorithm [7], as maintaining balance across the binary problems directly impacts the order of classification tasks.

One-versus-others nested dichotomies

In the following, we examine a structure that progresses **from the most frequent class to the least frequent** in the Census data; the tree structure is known *a priori*.



➤ At the third node, we can estimate using **one binary model**: $\mathbb{P}(R = \text{B}|G, S, R \notin \{\text{W,H}\})$.

➤ The multi-class probability estimates are obtained by **multiplying** the conditional probability estimates **along the tree structure**. For example, $\mathbb{P}(R = \text{B}|G, S)$ is calculated as:

$$(1 - \mathbb{P}(R = \text{W}|G, S)) \cdot (1 - \mathbb{P}(R = \text{H}|G, S, R \neq \text{W})) \cdot \mathbb{P}(R = \text{B}|G, S, R \notin \{\text{W,H}\}).$$

Assessing Calibration

In a **multi-class prediction setting**, where $Y \in \mathcal{Y} = [K]$. A model $h \in \mathcal{H}$ is strongly calibrated (or auto-calibrated) when [12]

$$\mathbb{P}(Y = k \mid h(\mathbf{X})) = h(\mathbf{X})_k, \quad \forall k \in [K].$$

where $h(\mathbf{X}) = (h(\mathbf{X})_1, \ldots, h(\mathbf{X})_K)$. This definition can be weakeaned to:

Marginal calibration [13]: $\mathbb{P}(Y = k \mid h(\mathbf{X})_k) = h(\mathbf{X})_k \quad \forall k \in [K]$, corresponding to a **binary calibration** problem. Here, we aim to improve the marginal calibration of **minority classes**.

Indeed, for a binary response variable Y, a model h is calibrated when [11]

$$\mathbb{P}[Y = 1 \mid h(\boldsymbol{X}) = p] = \mathbb{E}[Y \mid h(\boldsymbol{X}) = p] = p, \forall p \in [0, 1].$$

To assess marginal calibration of a model h, we face K binary scenarios. The literature suggest using **graphical techniques** and **metrics**, typically beginning with the estimation of a **calibration curve**

$$\mathrm{g}: \begin{cases} [0,1] \to [0,1] \\ p \mapsto \mathrm{g}(p) := \mathbb{E}[Y \mid \hat{s}(\mathbf{X}) = p] \end{cases}$$, the identity function, $\mathrm{g}(p) = p$ for a calibrated model.

➤ **Visualization tools** We estimate g for each class $k \in [K]$ by fitting a **local regression** model of degree 0 on $(\hat{s}(\mathbf{x}_i), y_i)$ (to compute the local mean of observed events in the vicinity of predicted scores for each class) using the `locfit` package in R [10].

➤ **Calibration metrics** The **Brier Score** [4], often used to assess a model's calibration, is a proper scoring rule:
$$\mathrm{BS} = n^{-1} \textstyle\sum_{i=1}^{n} \left(\hat{s}(\mathbf{x}_i) - y_i\right)^2.$$

Austin and Steyerberg [2] introduced the **Integrated Calibration Index**, a metric derived from the calibration curve estimated using local regression techniques, $\hat{\mathrm{g}}$. As a "pure" calibration metric, its empirical version writes:

$$\mathrm{ICI} = n^{-1} \textstyle\sum_{i=1}^{n} \left|\hat{s}(\mathbf{x}_i) - \hat{\mathrm{g}}(\hat{s}(\mathbf{x}_i))\right|.$$

Correcting Miscalibration

To address **multi-class miscalibration**, binary calibration techniques have been extended to multi-class settings, such as temperature scaling derived from Platt scaling. However, these methods often rely on **binary decomposition** [9, 5], focusing on the highest predicted score and neglecting minority classes, where models tend to be underconfident.

We apply **binary post-calibration** techniques at each node of the nested dichotomy model, where the classification task is binary and **the problems are balanced**.

➤ **Post-calibration technique** The **local regression** method serves a dual role, functioning both as a **visualization tool** and as a **post-calibration approach**. Therefore, we transform the predicted scores at each node of the nested dichotomy by $\hat{\mathrm{g}}(\hat{s}(\mathbf{x}_i))$ where $\hat{\mathrm{g}}$ is a fitted local regression of degree 0.

References

[1] Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. *ACM FAcct*.

[2] Austin, P. C. and Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38(21):4051–4065.

[3] Baeder, L., Erica, B., Brinkmann, P., Long, J., Stracke, C., Togba-Doya, K., Usan, G., Weaver, N., and Woldeyes, M. (2024). Statistical methods for imputing race and ethnicity. *Society of Actuaries*.

[4] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

[5] Coz, A. L., Herbin, S., and Adjed, F. (2024). Confidence calibration of classifiers with many classes. *NeurIPS*.

[6] Elliott, M. N., Morrison, P. A., Fremont, A. M., McCaffrey, D. F., Pantoja, P. M., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83.

[7] Frank, E. and Kramer, S. (2004). Ensembles of nested dichotomies for multi-class problems. *ACM ICML*.

[8] Haley, J. M., Dubay, L., Garrett, B., Caraveo, C. A., Schuman, I., Johnson, K., Hammersla, J., Klein, J., Bhatt, J., Rabinowitz, D., et al. (2022). Collection of race and ethnicity data for use by health plans to advance health equity. *Working Paper*.

[9] Johansson, U., Löfström, T., and Boström, H. (2021). Calibrating multi-class models. *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, pages 111–130.

[10] Loader, C. (1999). *Fitting with LOCFIT*, chapter 3, pages 45–58. Springer New York, New York, NY.

[11] Schervish, M. J. (1989). A General Method for Comparing Probability Assessors. *The Annals of Statistics*, 17:1856–1879.

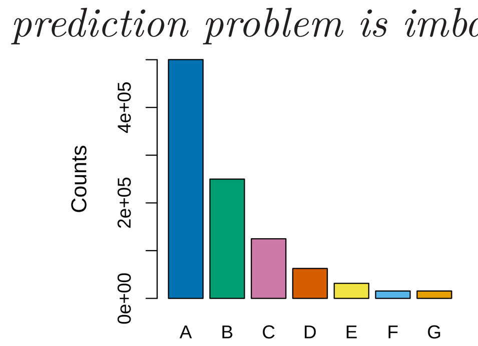
[12] Widmann, D., Lindsten, F., and Zachariah, D. (2019). Calibration tests in multi-class classification: A unifying framework. *NeurIPS*.

[13] Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *ACM SIGKDD*.

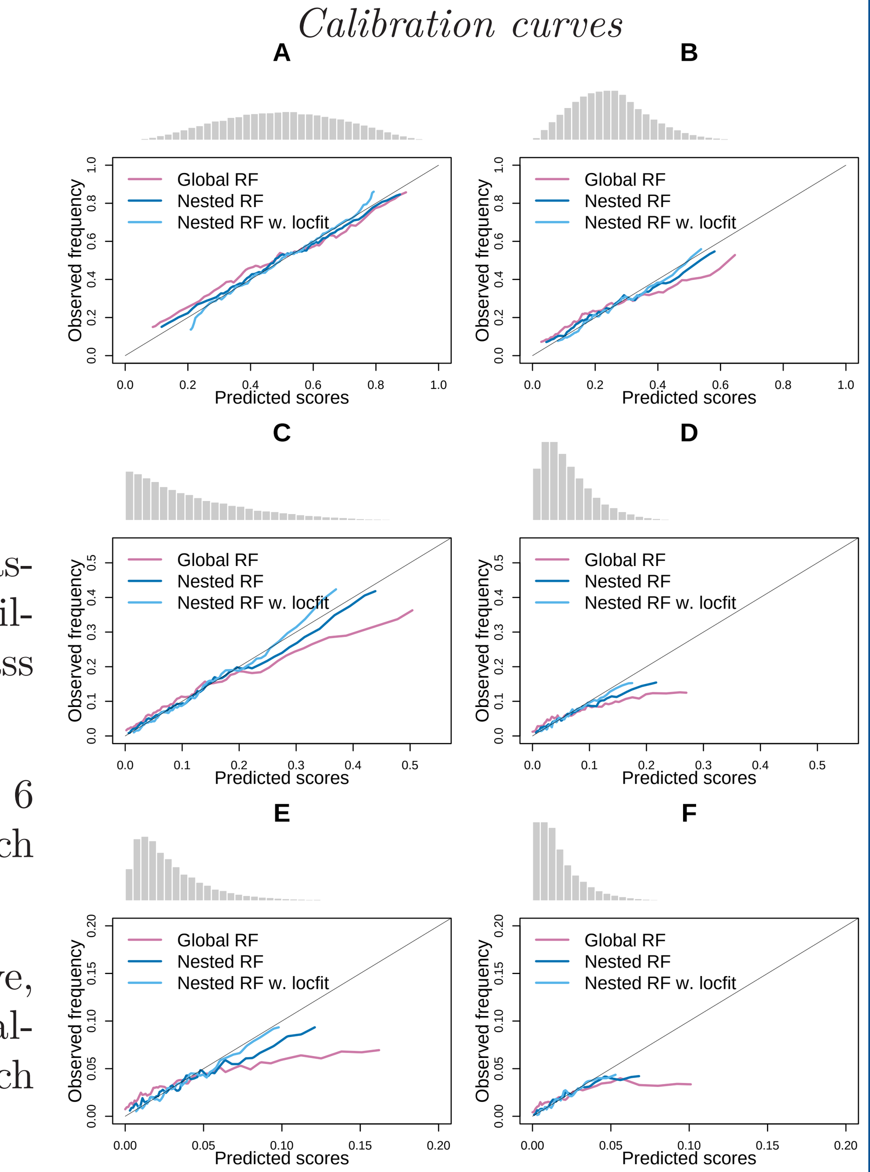
Application on Synthetic Data

We simulate $Y \in \{\text{A, B, C, D, E, F, G}\}$ ($K = 7$) from a multinomial distribution, with probabilities defined based on interactions of degree 2 and powers of $X_1, X_2, X_3, X_4 \sim$ Beta.

The prediction problem is imbalanced.



Calibration curves



We propose to compare three models:

1. A **Random Forest** (RF) multi-class classifier that directly estimates the probability of an individual belonging to a class $k \in [K]$.
2. A **nested dichotomy** with $K - 1 = 6$ nodes, where a binary RF is fitted at each node.
3. The same **nested dichotomy** as above, but **with local regression** applied to calibrate the predicted binary scores at each node for recalibration.

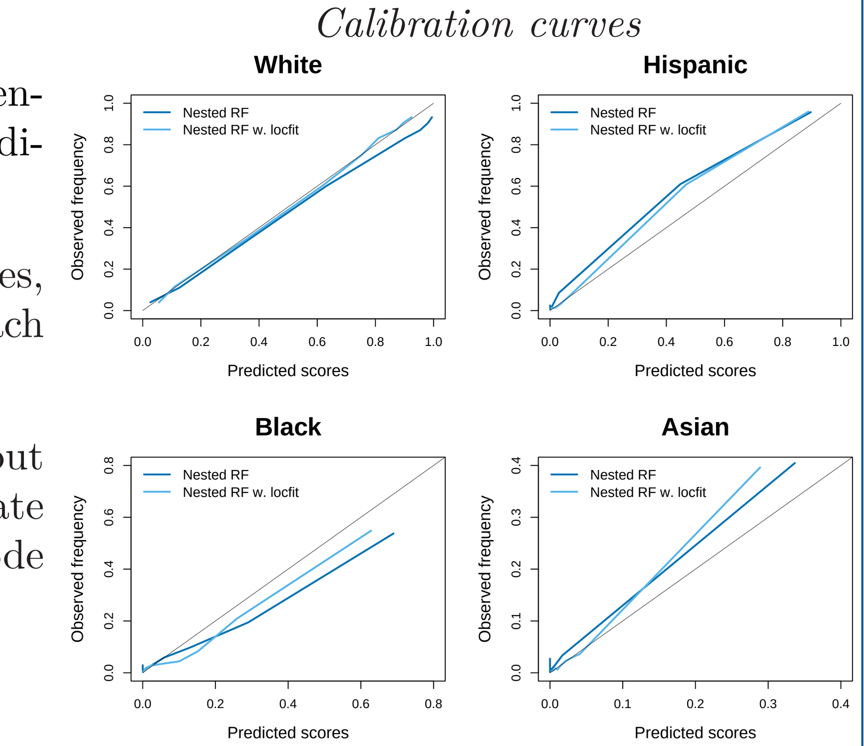
Calibration and performance metrics

	BS			ICI			F1-score		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
A	0.21	0.22	0.21	0.04	0.02	0.01	0.69	0.69	0.70
B	0.18	0.17	0.17	0.04	0.02	0.01	0.37	0.37	0.35
C	0.10	0.10	0.10	0.02	0.01	0.01	0.31	0.30	0.27
D	0.05	0.05	0.05	0.02	0.01	0.01	0.14	0.11	0.07
E	0.03	0.03	0.03	0.02	0.005	0.004	0.10	0.07	0.02
F	0.01	0.01	0.01	0.01	0.004	0.004	0.11	0.04	0.00
G	0.01	0.01	0.01	0.01	0.003	0.003	0.08	0.04	0.00

Application on Real Data: Predicting Ethnicity

➤ We return to the BISG algorithm with $K = 5$ classes to predict race probabilities.

Calibration curves



Here, we compare two models since, with Census Data frequencies, multi-class and nested dichotomy BISG give identical results:

1. A **nested dichotomy** with $K - 1 = 4$ nodes, where a binary BISG is calculated at each node.
2. The same **nested dichotomy** as above, but with **local regression** applied to calibrate the predicted binary frequencies at each node for recalibration.

Calibration and performance metrics

	BS		ICI		F1-score	
	(1)	(2)	(1)	(2)	(1)	(2)
White	0.57	0.50	0.63	0.59	0.86	0.86
Hispanic	0.22	0.22	0.25	0.24	0.85	0.85
Black	0.25	0.24	0.27	0.27	0.52	0.52
Asian	0.07	0.06	0.06	0.05	0.69	0.68
Other	0.58	0.56	0.58	0.56	0.10	0.12