UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Heili Aavola

In-Depth Analysis of Miscalibration In Binary Classification

Master's Thesis (15 ECTS)

Supervisor: Mari-Liis Allikivi, MSc

Tartu 2025

In-Depth Analysis of Miscalibration In Binary Classification

Master's Thesis

Heili Aavola

Abstract. Reliable probability estimates from binary classifiers are crucial for decision-making. While standard evaluation metrics provide an overall assessment of calibration quality, a deeper examination of miscalibration patterns can offer further insights into how calibration methods perform. This thesis presents an in-depth analysis of miscalibration patterns for five post-hoc calibration methods: Isotonic Calibration, Logistic Calibration, Beta Calibration, Histogram Binning, and Simplified Venn-Abers. Using a synthetic data framework with five diverse, known true calibration maps, we performed 100 simulation runs for each method-map combination. A suite of five specialized characterization plots was employed to visualize and understand nuanced error profiles, including accuracy, bias, variance, and directional tendencies in misestimation.

The results reveal distinct behavioral characteristics and trade-offs. Parametric methods (Logistic, Beta) exhibited high stability but incurred significant systematic bias when their functional assumptions did not match the true probability landscape. Non-parametric methods (Isotonic, SVA) demonstrated superior adaptability and lower average error but with step-like outputs and slightly higher variance in complex regions. Histogram Binning showed considerable artifacts tied to its fixed-bin structure. The characterization plots successfully highlighted consistent directional biases and other nuanced error patterns not evident from aggregate metrics. This granular understanding reveals the precise behavior of different calibration methods, offering a more nuanced basis for selecting approaches tailored to specific application needs and risk sensitivities, particularly in complex or risk-sensitive contexts, moving beyond single performance scores.

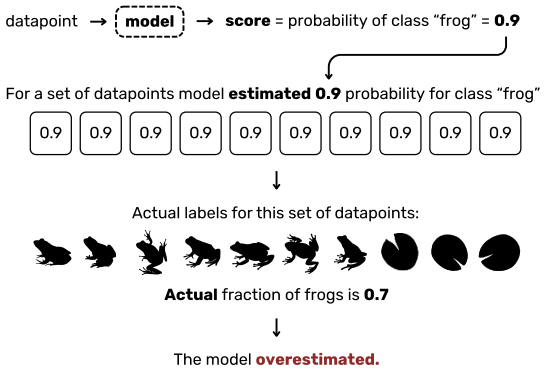
Keywords: Machine learning, calibration, binary classification.

CERCS: P170 Computer science, numerical analysis, systems, control, P176 Artificial intelligence.

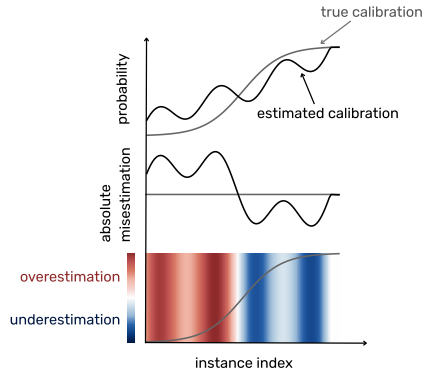
In-Depth Analysis of Miscalibration in Binary Classification

Data Science (MSc), 2025

Calibration definition



In-depth analysis of miscalibration



Author: Heili Aavola
Supervisor: Mari-Liis Allikivi (MSc)

#UniTartuCS

UNIVERSITY OF TARTU
Institute of Computer Science

Kalibreerimisvigade süvaanalüüs binaarses klassifitseerimises

Magistritöö

Heili Aavola

Lühikokkuvõte. Usaldusväärsed tõenäosushinnangud binaarsetelt klassifikaatoritelt on otsustusprotsessides kriitilise tähtsusega, kuid standardsed hindamismõõdikud varjavad sageli kalibreerimisvigade keerukat olemust. Käesolev magistritöö esitab süvaanalüüsi kalibreerimisvigade mustritest viie laialt levinud järelkalibreerimismeetodi puhul: isotooniline kalibreerimine, logistiline kalibreerimine, beeta kalibreerimine, histogrammipõhine kalibreerimine ja lihtsustatud Venn-Abers. Kasutades sünteetiliste andmete raamistikku, mis tugines viiele erinevale teadaolevale tõesele kalibratsioonikaardile ja seeläbi teadaolevatele tegelikele tõenäosustele, viidi läbi 100 simulatsioonikäiku iga meetodi ja kaardi kombinatsiooni kohta. Kasutati viit spetsialiseeritud karakteriseerimisgraafikut, et visualiseerida ja mõista nüansirikkaid veaprofiile, sealhulgas täpsust, nihet, dispersiooni ja suunatud tendentse valehinnangutes.

Tulemused näitavad selgelt eristuvaid käitumuslikke omadusi ja kompromisse. Parameetrilised meetodid (Logistiline, Beeta) näitasid suurt stabiilsust, kuid tekitasid märkimisväärset süstemaatilist nihet, kui nende funktsionaalsed eeldused ei vastanud tegelikule tõenäosusmaastikule. Mitteparameetrilised meetodid (Isotooniline, SVA) demonstreerisid paremat kohanemisvõimet ja väiksemat keskmist viga, kuid tulemuseks olid astmelised väljundid ja veidi suurem dispersioon keerukates piirkondades. Histogrammidel põhinev klassidesse jaotamine tekitas märkimisväärseid artefakte, mis olid seotud selle fikseeritud klassijaotuse struktuuriga. Karakteriseerimisgraafikud tõid edukalt esile järjepidevad suunatud nihked ja muud nüansirikkad veamustrid, mis koondmõõdikutest ei ilmne. Selline detailne arusaam aitab praktikutel teha teadlikumaid valikuid kalibreerimismeetodite osas, mis on kohandatud konkreetsetele rakendusvajadustele ja riskitundlikkusele, minnes kaugemale üksikutest tulemusnäitajatest.

Võtmesõnad: Masinõpe, kalibreerimine, binaarne klassifitseerimine.

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine; P176 Tehisintellekt.

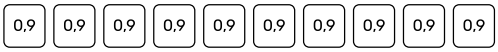
Kalibreerimisvigade süvaanalüüs binaarses klassifitseerimises

Andmeteadus (MSc), 2025

Kalibreerimise definitsioon

andmepunkt → **mudel** → **skoor** = klassi "konn" tõenäosus = **0,9**

Mudel **hindas** selle andmepunktide hulga puhul klassi „konn" tõenäosuseks **0,9**

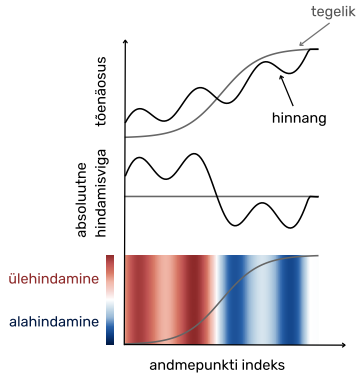


Selle andmepunktide hulga tegelikud märgendid:

Konnade **tegelik** osakaal on **0,7**

Mudel **ülehindas.**

Kalibreerimisvigade süvaanalüüs



Autor: Heili Aavola
Juhen: Mari-Liis Allikivi (MSc)

#UniTartuCS

TARTU ÜLIKOOL
arvutiteaduse instituut

Contents

Introduction

Supervised machine learning models, particularly classifiers, are increasingly pivotal in decision-making processes across diverse domains. A crucial aspect of their utility lies in the confidence scores they produce, which ideally indicate the model's certainty regarding its predictions. For these scores to be truly reliable and guide effective action, especially in high-stakes environments such as medical diagnosis, financial forecasting, or autonomous systems, they must be well-calibrated (Guo et al., 2017; Kull et al., 2017). Calibration, as detailed in Section 1.1, ensures that a predicted probability accurately reflects the actual propensity for the corresponding event to occur, meaning that if a large group of instances are assigned a particular probability p by the model, the event should indeed be observed in approximately a proportion p of those instances.

A variety of methods exist to improve the probabilistic outputs of classifiers, broadly categorized into train-time adjustments and post-hoc calibration techniques applied to pre-trained models. This thesis focuses on several prominent post-hoc methods, including Isotonic Calibration (Section 1.2), Logistic Calibration (Section 1.3), Beta Calibration (Section 1.4), Simplified Venn-Abers Predictors (SVA) (Section 1.5), and Histogram Binning (Section 1.6). While these methods aim to align predicted probabilities with observed frequencies, achieving perfect calibration is practically unattainable with finite real-world data (Allikivi et al., 2024). Consequently, even after applying state-of-the-art calibration techniques, some degree of misestimation—deviations between the calibrated probabilities and the true underlying probabilities—persists. The primary goal of this thesis is to provide a detailed characterization of these residual misestimation patterns, thereby laying foundational work for future investigations into their impact on decision-making.

A significant challenge in evaluating calibration methods in real-world settings is the absence of a known ground truth for the true calibrated probabilities. This makes it difficult to precisely quantify the nuances of misestimation from the single sample of data typically available. To overcome this limitation and enable a rigorous, in-depth analysis, we employ a synthetic data generation process, detailed in Section 2.1.1. Crucially, this involves the creation of *known true calibration maps* (Section 2.1.1), against which the outputs of various calibration methods can be directly compared. This experimental design allows for an exact measurement of misestimation at every point along the score range, facilitating a detailed study of how different calibration methods behave. While current practice often relies on aggregate metrics such as Expected Calibration Error (ECE) or proper scoring rules (e.g., Brier score, log loss) to evaluate the overall quality of calibration, these measures inherently obscure the precise patterns and characteristics of misestimation. This work is motivated by the intuition that a

more granular understanding of these misestimation patterns is essential, particularly for understanding the reliability and potential biases of different methods. Knowing precisely where a calibration method tends to err, and whether these errors manifest as systematic underestimation or overestimation, could inform a preference for one calibrated model over another. Furthermore, understanding the tendency for over- or underestimation can be crucial when the costs associated with different types of errors are asymmetric, a key consideration in frameworks like cautious calibration (Allikivi et al., 2024).

This thesis, therefore, undertakes a comprehensive analysis of the miscalibration profiles exhibited by selected post-hoc calibration methods when applied to simulated data derived from these known true calibration maps. Moving beyond aggregate performance, we aim to show the distinct ways in which these methods deviate from perfect calibration. The primary contribution of this work is to foster a deeper, more intuitive understanding of these misestimation patterns, as presented in the visual "Visualization Methods" (Section 2.3). By systematically applying calibration techniques (Section 2.3) across diverse, controlled true probability landscapes and analyzing the results through a series of diagnostic plots (Sections 2.3.1 through 2.3.5), we seek to highlight the nuances that aggregate metrics often conceal. This improved understanding can better equip practitioners and researchers to select and interpret calibration methods, particularly in contexts where the fine-grained behavior of probability estimates is crucial.

AI-powered tools assisted with language editing and formatting during the preparation of this thesis.

The subsequent chapters will further detail the experimental setup, present a comprehensive visual and qualitative analysis of the results (Chapter 3), and discuss the implications of these findings for the practical application and evaluation of calibration techniques in binary classification.

1 Background

1.1 Binary Classification and Probabilistic Calibration

Binary classification is a fundamental problem in machine learning, where each data instance $X \in \mathcal{X}$ (where $\mathcal{X}$ is the feature space) is associated with a binary outcome $Y \in \{0, 1\}$ (Murphy, 2012). Many modern classifiers do not directly output a class decision (0 or 1) but rather a continuous confidence score. This score forms the basis for making the final decision and often carries a probabilistic interpretation; for example, neural networks frequently use a sigmoid function to output a value between 0 and 1. We define such a scoring model as $s : \mathcal{X} \to \mathbb{R}$, where a higher score $s(X)$ for an instance X indicates greater model confidence that X belongs to the positive class ($Y = 1$).

Ideally, these scores can be transformed into reliable probability estimates. For a given instance X, the classifier (or a subsequent calibration step) aims to predict the probability of X belonging to the positive class. This prediction is formalized as a probability estimate function $\hat{p} : \mathcal{X} \to [0, 1]$. This function maps the features of X (or the score $s(X)$ derived from them) to a value in the unit interval, which should approximate the true conditional probability:

$$\hat{p}(X) \approx P(Y = 1 \mid X).$$

Here, $P(Y = 1 \mid X)$ is the true (but typically unknown in real-world scenarios) probability that instance X belongs to class 1, given its features. This is often referred to as the true posterior probability.

A primary goal for these predicted probabilities $\hat{p}(X)$ is for them to be well-calibrated. Perfect calibration means that the predicted probabilities accurately reflect the actual observed frequencies of the positive class across all prediction levels (Guo et al., 2017; Kull et al., 2017). Formally, a probabilistic classifier (represented by its output $\hat{p}$) is perfectly calibrated if, for any probability value $p_0 \in [0, 1]$ that the model predicts:

$$P(Y = 1 \mid \hat{p}(X) = p_0) = p_0. \tag{1}$$

In simpler terms, this condition means that if we collect all instances X for which our model predicts a specific probability, say $\hat{p}(X) = 0.8$, then approximately 80% of those instances should actually belong to the positive class ($Y = 1$). The probability $P(Y = 1 \mid \hat{p}(X) = p_0)$ in Equation (1) refers to this observable, empirical frequency of positive instances among those that received the prediction p_0. In practice, however, classifiers often exhibit systematic deviations from this ideal, leading to miscalibration (Allikivi et al., 2024).

Miscalibration means that the predicted probabilities do not match the true likelihoods. For a given predicted probability value p' output by the model, this can manifest in two main ways. First, **overestimation** occurs when the model's predicted probability p' is systematically higher than the actual observed frequency of positives for instances given that prediction. This means that if the model predicts $\hat{p}(X) = p'$, the true proportion of positives is lower:

$$P(Y = 1 \mid \hat{p}(X) = p') < p'.$$

Conversely, **underestimation** occurs when the model's predicted probability p' is systematically lower than the actual observed frequency for that prediction level. In this case, if the model predicts $\hat{p}(X) = p'$, the true proportion of positives is higher:

$$P(Y = 1 \mid \hat{p}(X) = p') > p'.$$

It is important to note that a model might overestimate for some predicted values p' (e.g., high confidence predictions) and underestimate for others (e.g., low confidence predictions). Both types of miscalibration can lead to suboptimal decisions in high-stakes contexts such as medical diagnostics, autonomous vehicle control, or financial risk assessment (Zadrozny and Elkan, 2002).

The quality of probabilistic predictions is typically evaluated using strictly proper scoring rules (Gneiting and Raftery, 2007). These mathematical rules are designed such that they are uniquely minimized (lower is better) only when the predicted probabilities $\hat{p}(X)$ perfectly match the true posterior probabilities $P(Y = 1 \mid X)$. Consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of N instances, where each $x_i \in \mathcal{X}$ is an instance (represented by its feature vector) and $y_i \in \{0, 1\}$ is its corresponding true label. Two common strictly proper scoring rules are the log loss (or cross-entropy) and the Brier score (Brier, 1950):

The log loss is defined as:

$$\ell_{\log} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \ln \hat{p}(x_i) + (1 - y_i) \ln \left(1 - \hat{p}(x_i)\right) \right],$$

and the Brier score is given by:

$$\ell_{\text{Brier}} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{p}(x_i)\right)^2.$$

Because these metrics are minimized when $\hat{p}(x_i) = P(Y = 1 \mid X = x_i)$ for each instance x_i, they also incentivize models to produce probabilities that satisfy the calibration

condition in Equation (1) (Gneiting and Raftery, 2007). In practice, another common metric for evaluating calibration, especially for visualization through reliability diagrams, is the Expected Calibration Error (ECE). ECE directly measures average deviations from perfect calibration (Equation (1)) by grouping predictions into bins.

Many machine learning models—particularly complex ones like deep neural networks (Guo et al., 2017) or those based on margin optimization such as Support Vector Machines (Platt, 1999)—produce systematically miscalibrated probabilities. Post-hoc calibration methods address this issue. These methods learn a calibration mapping function, $g : [0, 1] \to [0, 1]$, which takes the raw model outputs (or initial probability estimates $\hat{p}(X)$) and transforms them into new, hopefully better-calibrated, probabilities:

$$\hat{p}_{\text{calibrated}}(X) = g(\hat{p}(X)).$$

For binary classification, to ensure that the calibration process doesn't degrade the model's ability to distinguish between classes (i.e., its ranking of instances), the calibration function g is typically constrained to be monotonically non-decreasing. This preserves the relative confidence ordering between instances while adjusting the absolute probability values to be more accurate (Kull et al., 2017; Zadrozny and Elkan, 2002).

1.2 Isotonic Calibration

Isotonic calibration, a non-parametric method introduced by Zadrozny and Elkan (2002), constructs a piecewise constant (step-like) mapping to transform uncalibrated model outputs (scores or initial probabilities) into calibrated probabilities. The core idea is to apply isotonic regression to a given calibration dataset. This process fits a monotonically non-decreasing function that aims to minimize the mean squared error (MSE) between the function's outputs (the calibrated probabilities) and the empirical frequencies of the positive class observed in that calibration dataset. Given a calibration set of n instances with scores $z_1 \leq z_2 \leq \ldots \leq z_n$ (these z_i are the model's outputs for the instances in the calibration set, sorted) and corresponding true binary labels $y_i \in \{0, 1\}$, isotonic calibration learns a mapping $m : \mathbb{R} \to [0, 1]$. This mapping m is chosen to be the function f that minimizes the sum of squared errors:

$$\sum_{i=1}^{n} (y_i - f(z_i))^2 \tag{2}$$

subject to the constraint that f must be monotonically non-decreasing: if $z_i \leq z_j$, then $f(z_i) \leq f(z_j)$. The monotonicity constraint is crucial as it ensures that the calibration function preserves the original ranking implied by the scores. The solution to this optimization problem, the function m, is a piecewise constant function. It can be

computed efficiently using the Pool Adjacent Violators Algorithm (PAVA) (Zadrozny and Elkan, 2002).

The non-parametric nature of isotonic calibration means it makes minimal assumptions about the true form of the miscalibration. This flexibility allows it to adapt to complex miscalibration patterns, for example, where a model might overestimate probabilities in some score regions and underestimate in others. However, precisely because it directly optimizes its fit on the calibration data with such high flexibility (it can learn a very complex step function), isotonic calibration can be particularly prone to overfitting. This is especially a concern when the calibration dataset is small, as the learned stepped function might capture noise specific to that calibration set rather than reflecting the true underlying calibration trend. An important characteristic, though, is that isotonic calibration does not alter the relative ranking of the original scores; if the underlying model already has good discriminative power (i.e., it ranks positive instances higher than negative ones effectively), this ability is preserved after isotonic calibration.

While powerful in its adaptability, this tendency to overfit on limited data is a key consideration (Allikivi et al., 2024; Kull et al., 2017). Even with sufficient data, the resulting calibration map is inherently step-like. Furthermore, as isotonic calibration optimizes for average calibration performance (by minimizing MSE on the calibration set), it does not inherently guarantee that all its probability estimates will avoid overestimation.

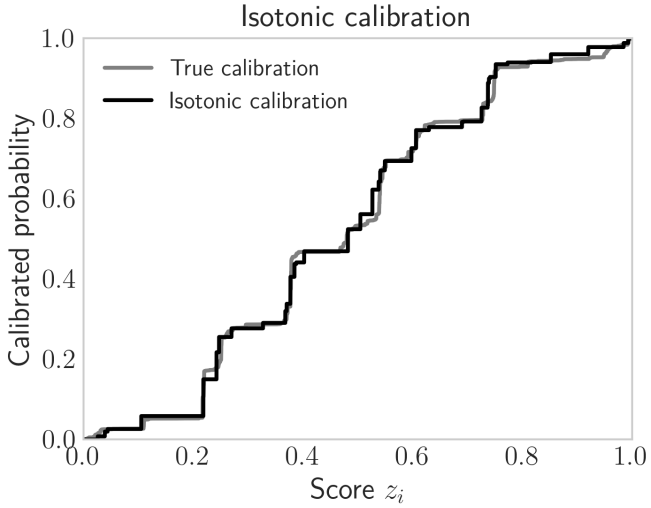


Figure 1: Isotonic calibration: an example of a piecewise constant mapping transforming scores to calibrated probabilities.

1.3 Logistic Calibration (Platt scaling)

Logistic calibration, also widely known as Platt scaling (Platt, 1999), is a parametric post-hoc calibration method. It applies a logistic (sigmoid) transformation to map uncalibrated model scores (denoted as $s(x_j)$ for an instance x_j) to calibrated probabilities. While originally developed for calibrating Support Vector Machines (SVMs), its effectiveness has been demonstrated across many types of classification models (Niculescu-Mizil and Caruana, 2005).

Unlike isotonic calibration, which can learn a calibration map of almost any monotonic shape, logistic calibration assumes a specific, simpler parametric form for the relationship between the model's scores and the true probabilities. Specifically, it posits that this relationship can be well-approximated by a sigmoid function. For a given score $s(x_j)$ corresponding to instance x_j, the calibrated probability $\hat{p}_{\mathrm{L}}(s(x_j))$ is obtained through:

$$\hat{p}_{\mathrm{L}}(s(x_j)) = \frac{1}{1 + \exp(-(a_L s(x_j) + b_L))} \tag{3}$$

Here, a_L and b_L are the calibration parameters that respectively scale and shift the input scores $s(x_j)$ before they are passed through the logistic function. These parameters are learned from a calibration dataset, $\mathcal{D}_{\text{cal}} = \{(x_j, y_j)\}_{j=1}^{n}$, where $y_j \in \{0, 1\}$ are the true binary labels. The parameters a_L and b_L are estimated by finding the values that minimize the negative log-likelihood (NLL) of the observed labels y_j given the calibrated probabilities $\hat{p}_{\mathrm{L}}(s(x_j); a_L, b_L)$ on this dataset:

$$(a_L^*, b_L^*) = \arg\min_{a_L, b_L} \left(-\sum_{j=1}^{n} [y_j \ln \hat{p}_{\mathrm{L}}(s(x_j); a_L, b_L) + (1 - y_j) \ln(1 - \hat{p}_{\mathrm{L}}(s(x_j); a_L, b_L))] \right) \tag{4}$$

The parametric simplicity of logistic calibration offers several advantages. Because it fits a function with only two parameters (a_L and b_L), it generally requires fewer calibration samples than non-parametric methods and is consequently less prone to overfitting on small datasets, especially if the true calibration map is indeed reasonably sigmoid-shaped. The resulting calibration mapping is inherently smooth, avoiding the step-like artifacts characteristic of isotonic calibration. Furthermore, the optimization problem for finding a_L and b_L (minimizing the NLL in Equation (4)) is convex, which guarantees that a unique optimal solution can be found. Logistic calibration tends to generalize well when the underlying relationship between the model's scores and the log-odds of the true probabilities is approximately linear (as this implies a logistic calibration curve).

However, logistic calibration has notable limitations. Its primary assumption of a simple sigmoid relationship may not hold for all models or datasets. When the true calibration function deviates significantly from this logistic form, the method can lead

to systematic calibration errors across parts of the score range (Kull et al., 2017). Kull et al. (2017) have shown that logistic calibration performs well for models that produce approximately symmetrical score distributions for the positive and negative classes. However, when these distributions are skewed or have different variances, logistic calibration may not fully correct complex miscalibration patterns.

In the context of binary neural networks, the relationship between the model's logits (the inputs to the final activation function) and the true probabilities often approximates a logistic function, making Platt scaling (applied to logits) a particularly effective and common choice (Guo et al., 2017). For other model types or more complex neural architectures, this relationship can be more nuanced, potentially requiring more flexible calibration methods. Computationally, logistic calibration is efficient, and this efficiency, combined with its reasonable performance in many scenarios, contributes to its popularity as a baseline method in calibration research.

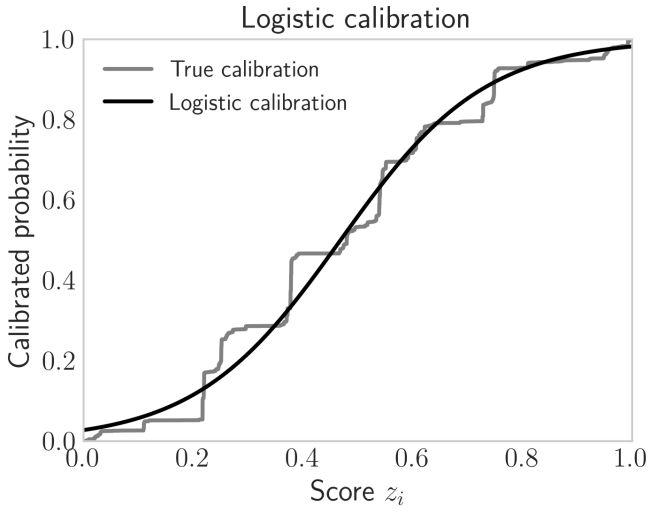


Figure 2: Logistic calibration: an example of a sigmoid fit transforming raw model scores to calibrated probabilities.

1.4 Beta Calibration

Beta calibration, introduced by Kull et al. (2017), extends logistic calibration by offering a more flexible parametric form derived from the beta distribution's properties. This method addresses key limitations of logistic calibration, particularly by better accommodating the asymmetric score distributions that often arise in practice.

The beta calibration function transforms scores $s_k \in (0, 1)$ (original model scores are typically mapped to this open interval if they are not already, for example, by clipping near 0 and 1 and adding a small epsilon to avoid issues with logarithms) using three

parameters (a, b, c). The calibrated probability $\hat{p}_{\text{beta}}(s_k)$ for a score s_k is given by:

$$\hat{p}_{\text{beta}}(s_k) = \frac{1}{1 + \exp(-(c + a\ln(s_k) + b\ln(1 - s_k)))} \tag{5}$$

This three-parameter form creates a highly flexible calibration map capable of handling various patterns of miscalibration, including asymmetric overestimation and underestimation across different probability ranges. The parameters (a, b, c) are estimated by minimizing the negative log-likelihood (NLL) on a calibration dataset $\mathcal{D}_{\text{cal}} = \{(x_k, y_k)\}_{k=1}^{n}$ (where s_k is the score for instance x_k):

$$\begin{aligned}(a^*, b^*, c^*) = \arg\min_{a,b,c} \Bigg\{ -\sum_{k=1}^{n} \Big[& y_k \ln \hat{p}_{\text{beta}}(s_k; a, b, c) \\ & + (1 - y_k) \ln\Big(1 - \hat{p}_{\text{beta}}(s_k; a, b, c)\Big) \Big] \Bigg\}\end{aligned} \tag{6}$$

Despite the increased parameter count compared to logistic calibration, this optimization problem remains convex, ensuring a unique global minimum.

Beta calibration offers several theoretical and practical advantages. It generalizes both logistic calibration (which can be seen as a special case when $a = b$) and linear transformations of log-odds, subsuming them. By more explicitly modeling the score distributions of positive and negative classes (through its link to the beta distribution's properties), it can provide better calibration when these distributions differ significantly in shape. The resulting calibration map maintains smoothness while offering greater flexibility than logistic calibration, yet it typically requires relatively few calibration samples compared to non-parametric methods like isotonic regression.

Kull et al. (2017) demonstrated that beta calibration consistently outperforms logistic calibration across various classifiers and datasets, especially when underlying score distributions are skewed. The method addresses a fundamental limitation in Platt scaling: its inability to effectively handle asymmetric miscalibration patterns where, for example, overestimation occurs in one probability range and underestimation in another. The interpretability of beta calibration parameters can also provide insights: parameter a primarily influences calibration at high scores, while b affects low scores. If $a = b$, the calibration curve tends to be symmetric; if $a \neq b$, it accommodates asymmetry.

Despite its advantages, beta calibration may still underperform compared to non-parametric methods if the true calibration function is highly complex and cannot be well-approximated by the beta calibration family. However, its balance of flexibility, sample efficiency, and theoretical grounding makes it a strong and often preferred parametric method for many practical applications.

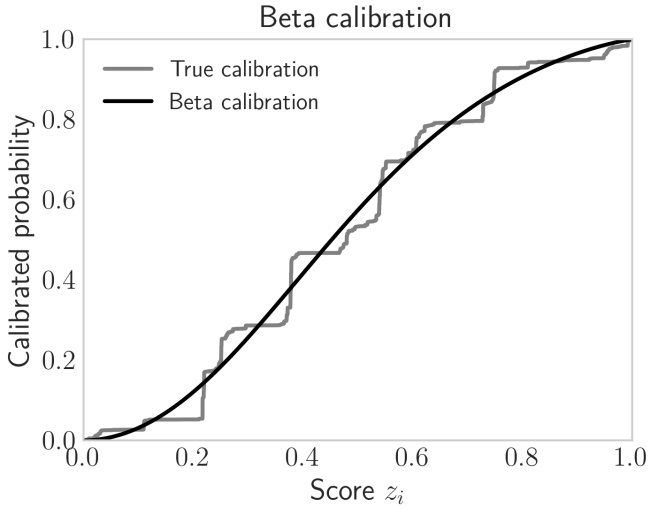


Figure 3: Beta calibration: an example of a flexible three-parameter fit transforming scores to calibrated probabilities.

1.5 Simplified Venn-Abers Predictors (SVA)

The Venn-Abers prediction framework, rooted in conformal prediction theory, offers a way to produce probability predictions with certain theoretical validity guarantees (Shafer and Vovk, 2008; Vovk, Gammerman, et al., 2005). A key property of full Venn-Abers predictors is that one of the multiple probability values they output for a test instance is guaranteed to be well-calibrated in a specific sense (related to average performance over the randomness of data splits). However, obtaining these predictions typically requires retraining or recalibrating the underlying model for each test instance by hypothetically adding it to the calibration set with each possible label, which is computationally very expensive.

Simplified Venn-Abers predictors (SVA), as described by Vovk and Petej (2012), offer a computationally more tractable approximation. While SVA sacrifices some of the strong theoretical guarantees of the full framework, it aims to retain much of its adaptive nature. Given a base scoring model $s(X)$ and a calibration dataset $\mathcal{D}_{\text{cal}} = \{(s_i, y_i)\}_{i=1}^{n}$ (where s_i are the scores and y_i are binary labels), SVA operates as follows for a new test instance x_{test} with score $s(x_{\text{test}})$:

1. **scenario 0 (hypothetical label 0):** Temporarily add $(s(x_{\text{test}}), 0)$ to $\mathcal{D}_{\text{cal}}$. Fit an isotonic regression model, h_0, on this augmented dataset. The SVA lower probability is $p_0(x_{\text{test}}) = h_0(s(x_{\text{test}}))$;
2. **scenario 1 (hypothetical label 1):** Temporarily add $(s(x_{\text{test}}), 1)$ to $\mathcal{D}_{\text{cal}}$. Fit another isotonic regression model, h_1, on this second augmented dataset. The SVA upper probability is $p_1(x_{\text{test}}) = h_1(s(x_{\text{test}}))$.

This process yields two probability estimates, $p_0(x_{\text{test}})$ and $p_1(x_{\text{test}})$, for the test instance. The interval $[\min(p_0, p_1), \max(p_0, p_1)]$ can be seen as an implicit measure of uncertainty for the prediction. For a single point prediction, SVA typically uses the midpoint:

$$\hat{p}_{\text{SVA}}(x_{\text{test}}) = \frac{p_0(x_{\text{test}}) + p_1(x_{\text{test}})}{2}. \tag{7}$$

This midpoint is what we refer to as the SVA output in our experiments.

SVA can adapt to complex miscalibration patterns due to its use of isotonic regression. The lower of the two probabilities, $\min(p_0, p_1)$, can sometimes serve as a conservative estimate, though without the formal guarantees of methods designed specifically for cautious calibration (Allikivi et al., 2024). However, SVA is computationally more intensive than parametric methods, as it requires fitting two isotonic regressions for each test prediction (though not retraining the original scoring model s). With larger calibration sets, the SVA midpoint prediction often behaves similarly to standard isotonic calibration, sometimes acting as a slightly regularized version. Extending SVA to multiclass problems is also not straightforward. Vovk and Petej (2012) demonstrated SVA's potential for good calibration, particularly in challenging scenarios.

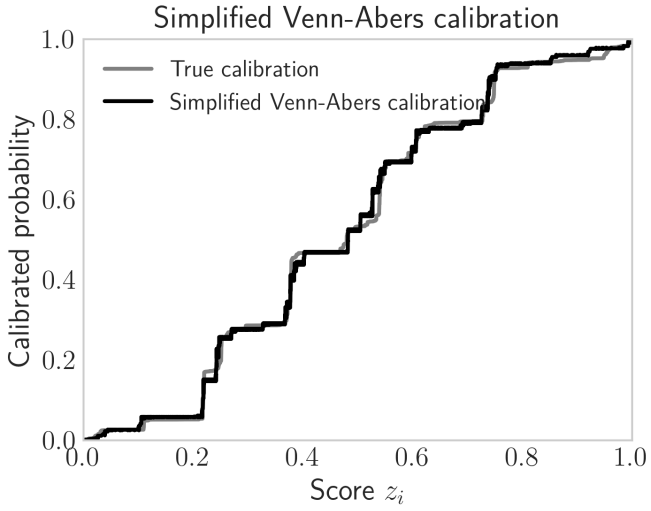


Figure 4: Simplified Venn-Abers predictors: conceptual illustration of generating two probability estimates, p_0 and p_1, based on hypothetical labels for a test instance.

1.6 Histogram Binning

Histogram binning (Zadrozny and Elkan, 2001) is one of the earliest and most intuitive non-parametric approaches to post-hoc calibration. The core idea is to partition the range of the original model's scores (typically assumed to be in or mapped to $[0, 1]$) into a predefined number of discrete bins. A single calibrated probability is then assigned

to all scores falling within the same bin, based on the observed empirical frequency of positive instances in that bin from a calibration dataset.

The calibration procedure generally involves two main steps. First, the score range (e.g., $[0, 1]$) is divided into B contiguous, non-overlapping bins, $I_1, I_2, \ldots, I_B$, such that their union covers the entire score range. For instance, in equal-width binning, each bin I_k would cover a range like $[\frac{k-1}{B}, \frac{k}{B})$. Second, for each bin I_k, the calibrated probability is computed as the fraction of positive instances among all calibration instances whose original scores fall into I_k.

Formally, given a calibration dataset $\mathcal{D}_{\text{cal}} = \{(s_j, y_j)\}_{j=1}^{n}$ (where s_j are the original model scores and $y_j \in \{0, 1\}$ are the true labels), the calibrated probability $\hat{p}_{\text{HB}}(s)$ for any new score s that falls into a particular bin I_k is:

$$\hat{p}_{\text{HB}}(s) = \frac{\sum_{j \text{ s.t. } s_j \in I_k} y_j}{N_k}, \quad \text{if } s \in I_k, \tag{8}$$

where $N_k = |\{j \mid s_j \in I_k\}|$ is the total number of calibration instances whose scores fall into bin I_k. If $N_k = 0$ (an empty bin), a strategy is needed, such as assigning a default probability (e.g., the global mean of labels) or merging with adjacent bins. This bin-wise estimate, $\hat{p}_{\text{HB}}(s)$, represents the empirical frequency of positive instances within bin I_k.

Two common strategies for defining these bins are **equal-width binning**, which divides the score range into B intervals of identical width, and **equal-frequency (or equal-mass) binning**, which aims to create B bins such that each contains approximately the same number of calibration instances.

The conceptual simplicity and ease of implementation are key advantages of histogram binning. Like other post-hoc methods, it can be applied to the outputs of any classification model. It provides a direct empirical estimation of probabilities within each bin without making strong parametric assumptions about the shape of the calibration curve, leading to an intuitive interpretation of the calibration process.

However, histogram binning also has significant limitations. The resulting calibration function is inherently piecewise constant, introducing artificial discontinuities at the bin edges. The choice of the number of bins, B, and the binning strategy itself (e.g., equal-width vs. equal-frequency) are critical hyperparameters that significantly impact calibration quality; too few bins can smooth out important details (underfitting), while too many bins can lead to unreliable probability estimates in sparse bins (overfitting). This creates a trade-off between the granularity of the calibration map and the statistical reliability of the estimates within each bin. An important variant is Bayesian binning (Naeini et al., 2015), which addresses some of these limitations by averaging over multiple binning schemes, weighted by their posterior probabilities, thereby improving

robustness at the cost of increased computational complexity.

Compared to other methods, histogram binning shares similarities with isotonic calibration, as isotonic regression can be viewed as an adaptive form of binning where bin boundaries are optimally placed to minimize squared error while ensuring monotonicity. However, standard histogram binning, unlike isotonic calibration, does not inherently guarantee that the resulting calibration function g will be monotonic. This lack of a monotonicity guarantee can be a drawback if preserving the rank-order of the original scores is crucial, but it could also potentially allow it to fit certain non-monotonic true calibration patterns if they were to exist (though this is generally not desired for score calibration).

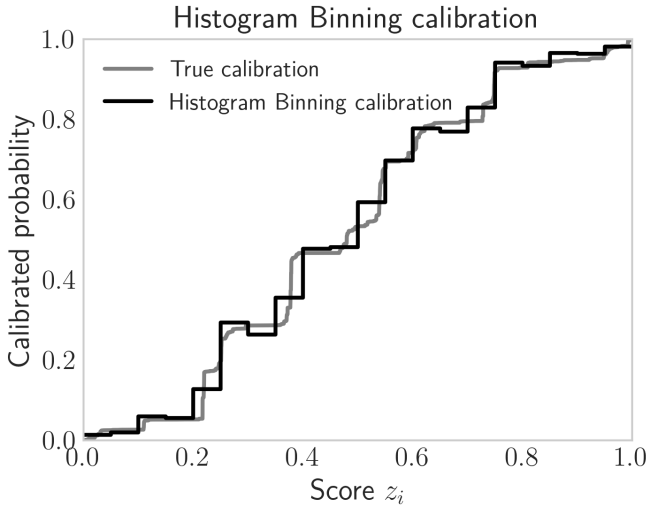


Figure 5: Histogram binning: an example of an empirical frequency-based calibration with discrete bins.

2 Methodology

This chapter details the experimental framework underpinning our in-depth analysis of miscalibration in binary classification. A central tenet of our approach is the utilization of synthetically generated data. This grants us complete control over and precise knowledge of the *true calibration maps*—the actual conditional probabilities $P(Y = 1 \mid z_i)$, where z_i is an uncalibrated model score. Such control is paramount for a granular assessment of how various post-hoc calibration methods perform, particularly in identifying systematic deviations between their *estimated calibration* outputs, $\hat{p}_{\text{calibrated}}(z_i)$, and the known ground truth. This synthetic paradigm circumvents a fundamental limitation of real-world datasets, where true conditional probabilities are typically unknown, thus precluding a similarly detailed error analysis.

Our methodology unfolds in three stages. First, we describe the process of constructing these diverse true calibration maps and the subsequent simulation of observed binary labels from these maps, which form the basis for training and evaluating the calibration methods (Section 2.1). Second, we outline the specifics of our experimental setup (Section 2.2). This includes the selection of five prominent post-hoc calibration techniques—Isotonic Calibration, Logistic Calibration, Beta Calibration, Histogram Binning, and Simplified Venn-Abers Predictors—along with details of their software implementations and parameter choices. To ensure robust findings, each calibration method is applied to data derived from each true calibration map across $N_{\text{runs}} = 100$ independent simulation runs, allowing for an assessment of both average performance and variability. Finally, although detailed in a subsequent section (Section 2.3), we will employ a suite of five specialized visualization techniques designed to dissect the performance of these methods, moving beyond aggregate metrics to reveal nuanced patterns of misestimation, such as consistent over- or underestimation, across the entire score spectrum and under varied true probability landscapes.

2.1 Synthetic Data Generation

The cornerstone of our empirical investigation is a carefully designed synthetic data generation process. This process allows us to establish an unambiguous ground truth against which the outputs of calibration methods can be compared. It consists of two primary steps: the definition of various *true calibration maps* representing different underlying probability landscapes, and the simulation of binary labels based on these maps, which serve as the data for applying calibration techniques.

2.1.1 True Calibration Map Generation

A *true calibration map*, denoted as $c = (c_1, c_2, \ldots, c_N)$, defines the true conditional probability $P(Y = 1|z_i)$ of an instance belonging to the positive class ($Y = 1$) given a specific uncalibrated model output score z_i. In our experiments, each true calibration map consists of $N = 10,000$ discrete probability values. These values correspond to an ordered sequence of N instances, sorted by their model scores z_i, effectively providing a dense sampling of the probability landscape across the model's score range. We assume, in line with Allikivi et al. (2024) and standard expectations for useful classifiers, that this true calibration map c is monotonically non-decreasing with respect to the scores z_i. That is, higher initial scores should, on average, correspond to higher true probabilities of positive class membership.

To comprehensively evaluate the calibration methods, we designed five distinct true calibration map shapes, intended to represent a variety of challenging and realistic miscalibration scenarios. These are depicted in Figure 6. Three of these maps are based on deterministic functional forms: the "Smooth" map uses a power function ($c_i = \text{low} + (\text{high} - \text{low}) \times (x^p/(x^p + (1 - x)^p))$, with $p = 3$) to create a gentle S-shaped curve; the "Plateau" map is a piecewise function featuring three distinct probability plateaus (near 0.02, 0.5, and 0.98) connected by smooth cosine transitions, designed to test adaptability to non-uniform miscalibration; and the "Steep" map employs a sigmoid function ($c_i = \text{low} + (\text{high} - \text{low}) \times (1/(1 + \exp(-x')))$ with a steepness parameter of 50) to model scenarios with abrupt changes in probability. For all these deterministic shapes, low is set to 0.001 and high to 0.999, and x (or x') represents the normalized score index. The specific parameters for the "Plateau" map, such as p1_level = 0.02, p1_start = 0.05, and p1_end = 0.35, are carefully chosen to create significant regions where simple calibration models might struggle.

In addition to these deterministic shapes, we include two true calibration maps generated using recursive algorithm proposed by (Allikivi et al., 2024). This method produces a sequence of N random probabilities that are inherently monotonically non-decreasing, bounded between specified minimum (0.001) and maximum (0.999) values. These maps, named "Random42" and "Random63" after their generation seeds, introduce less structured and more irregular probability landscapes. This diverse set of five true calibration maps serves as the ground truth for all subsequent experiments.

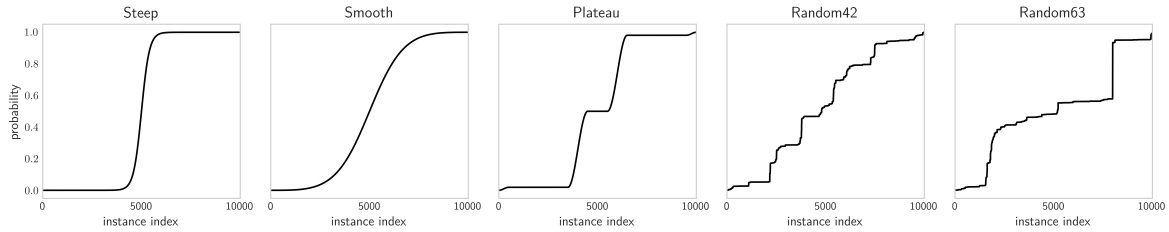


Figure 6: Overview of the five generated true calibration map shapes used in the experiments. The x-axis represents the instance index (sorted by score, effectively the score z_i) from 0 to 9999, and the y-axis shows the true probability $P(Y = 1 \mid z_i)$.

2.1.2 Simulated Label Generation

With a true calibration map $c = (c_1, \ldots, c_N)$ established, the next step is to simulate the binary labels $y_i \in \{0, 1\}$ that a calibration method would observe. For each point i on the true calibration map (corresponding to an uncalibrated score z_i and its true probability c_i), we perform a Bernoulli trial: $y_i \sim \text{Bernoulli}(c_i)$. This generates a sequence of N binary labels, $(y_1, \ldots, y_N)$, which constitutes one realization of an observed "calibration set," $\mathcal{D}_{\text{cal}} = \{(z_i, y_i)\}_{i=1}^{N}$. The uncalibrated scores z_i themselves are represented by a fixed, linearly increasing sequence from approximately 0 to 1, corresponding to the instance indices $0, \ldots, N - 1$.

To rigorously evaluate the stability and average behavior of the calibration methods, this label simulation is repeated $N_{\text{runs}} = 100$ times for each of the five true calibration maps. Each of these 100 runs employs a unique random seed, ensuring that the calibration methods are trained and evaluated on 100 independent sets of observed labels for each underlying true probability landscape. The seed for each run is determined systematically using the formula: seed = (LABEL_SEED_BASE + run_index + hash(map_name)) (mod 2^{32}), where LABEL_SEED_BASE = 2024, run_index ranges from 0 to 99, and map_name identifies the specific true calibration map. This procedure yields a total of 5 maps $\times$ 100 runs = 500 unique label sequences, each comprising $N = 10,000$ labels. These simulated label sequences, paired with the corresponding fixed sequence of uncalibrated scores, form the input for the calibration algorithms.

2.2 Experimental Setup

The experimental setup is designed for a systematic comparison of selected post-hoc calibration methods. We aim to understand how their *estimated calibration* outputs, $\hat{p}_{\text{calibrated}}(z_i)$, deviate from the known *true calibration* c_i across different scenarios. The setup encompasses the choice of calibration techniques, their specific implementations, and the software environment.

2.2.1 Calibration Methods and Implementations

We evaluate five widely recognized post-hoc calibration methods, chosen to represent a range of parametric and non-parametric approaches. These methods learn a calibration mapping function $g : [0, 1] \to [0, 1]$ (or from scores z_i to $[0, 1]$), which transforms the initial uncalibrated scores z_i (or initial probability estimates $\hat{p}(z_i)$ if the scores are already in $[0, 1]$) into *estimated calibration* probabilities $\hat{p}_{\text{calibrated}}(z_i) = g(z_i)$ or $g(\hat{p}(z_i))$.

Isotonic Regression (Isocal) is a non-parametric method that fits a monotonically non-decreasing piecewise constant function to the scores z_i and observed labels y_i (Barlow et al., 1972; Zadrozny and Elkan, 2002). To enhance stability, particularly with noisy labels, we apply Platt smoothing (Platt, 1999) to the binary labels y_i before fitting the isotonic regressor. The implementation from `sklearn.isotonic.IsotonicRegression` is used, with the `out_of_bounds='clip'` option to ensure predicted probabilities remain within the $[0, 1]$ interval.

Logistic Calibration (Logcal), commonly known as Platt Scaling (Platt, 1999), is a parametric method that fits a logistic (sigmoid) function to map scores z_i to probabilities. The form is $\hat{p}_{\text{L}}(z_i) = (1 + \exp(-(a_L z_i + b_L)))^{-1}$, where a_L and b_L are learned parameters. We employ the `sklearn.linear_model.LogisticRegression` implementation. A large value for the regularization parameter C (10^6) and the 'liblinear' solver are used, consistent with common practice for Platt scaling to minimize regularization and enhance numerical stability.

Beta Calibration (Betacal), introduced by Kull et al. (2017), is another parametric method that offers greater flexibility than logistic calibration by fitting a transformation related to the Beta distribution. The functional form is $\hat{p}_{\text{beta}}(z_i) = (1 + \exp(-(c' + a' \ln(z_i) + b' \ln(1 - z_i))))^{-1}$, where z_i are assumed to be initial probabilities in $(0, 1)$, and a', b', c' are learned. This method is particularly adept at handling asymmetric miscalibration. We utilize the `betacal` Python library, specifically with the `parameters="abm"` setting, which incorporates affine transformations (bias and scaling) of the logit-transformed scores alongside the core Beta parameters. Our input scores z_i are already scaled to $[0, 1]$ for this method.

Histogram Binning (Histbin) is a non-parametric technique that divides the score range $[0, 1]$ into a predetermined number of bins, B. It assigns an *estimated calibration* probability to all scores z_i falling within a bin based on the empirical frequency of positive labels y_i in that bin from the calibration data $\mathcal{D}_{\text{cal}}$ (Zadrozny and Elkan, 2001). We implemented this method ourselves. For empty bins, the global mean of labels in the calibration set is assigned. After experimenting with various bin counts, we selected $B = 15$ (`num_bins=15`) for the experiments presented, as this provided a reasonable balance between granularity and estimate stability for $N = 10,000$ data points.

Simplified Venn-Abers (SVA) predictors, derived from conformal prediction theory

(Vovk and Petej, 2012), offer a way to generate interval predictions $[p_0(z_i), p_1(z_i)]$. For a point prediction, SVA typically uses the midpoint of these lower and upper probability bounds. These bounds are derived by fitting two isotonic regression models per test point z_i, under hypothetical assignments of label 0 and label 1 to that point when temporarily added to $\mathcal{D}_{\text{cal}}$. In our experiments, we use this midpoint, $(\hat{p}_{\text{SVA}}(z_i) = (p_0(z_i) + p_1(z_i))/2)$, as the *estimated calibration* probability from SVA.

2.2.2 Implementation Details

The implementations of Isotonic Regression and Logistic Calibration are sourced from the `scikit-learn` library (Pedregosa et al., 2011). Beta Calibration relies on the `betacal` library. The Histogram Binning and Simplified Venn-Abers methods were implemented by us using `NumPy` (Harris et al., 2020) for numerical operations. All experiments are conducted in Python 3.9. Visualization of results, including the diagnostic plots described in Section 2.3, is performed using `Matplotlib`. To ensure transparency and enable replication of our findings, the complete source code for data generation, calibration method application, and analysis is made available on `GitHub`.

2.3 Visualization Methods

To characterize the performance of the selected calibration methods, we generate five distinct types of **characterization plots** for each combination of calibration method and true calibration map. Each plot offers a unique perspective on the accuracy, bias, variance, and consistency of the *estimated calibration* outputs relative to the known *true calibration*. The x-axis in all these plots represents the instance index, which ranges from 0 to $N - 1$ (where $N = 10,000$). Since our true calibration maps and input scores z_i are ordered, this index effectively represents the score z_i increasing from its lowest to highest value. For each instance index (and thus, for each score z_i), our experiments generate $N_{\text{runs}} = 100$ independent *estimated calibration* probabilities, one from each simulation run. These **characterization plots** are designed to provide a much more granular understanding of miscalibration than what aggregate metrics typically offer.

2.3.1 True Calibration and Estimated Calibration Curves

The first **characterization plot**, exemplified in Figure 7, provides a direct visual comparison between the known *true calibration* map and the multiple *estimated calibration* curves produced by a given method across the $N_{\text{runs}} = 100$ simulation runs. For any specific instance index (e.g., index 5000, corresponding to a score z_{5000} notionally in the mid-range of the score distribution), the calibration method yields one hundred different estimated probability values, $\hat{p}_{\text{calibrated}}(z_{5000})$, one from each label set simulation. This

plot visualizes all one hundred of these *estimated calibration* curves. Each individual curve, plotting the estimated probability (y-axis) for every instance index (x-axis), is rendered as a semi-transparent grey line. The *true calibration* map c, which defines the correct probability c_i for each instance index i (score z_i), is overlaid as a prominent solid black line.

This visualization facilitates an initial qualitative assessment of the calibration method's general behavior. It allows for observation of how well, on average, the *estimated calibration* curves follow the trend of the *true calibration*. Any systematic tendency for the bundle of grey lines to lie above or below the black line indicates a propensity for overestimation or underestimation, respectively. The vertical spread (or "tightness") of these grey lines at any given instance index reveals the method's stability or variance; a tight bundle signifies robustness to the stochasticity in the observed labels across different runs, whereas a wide spread indicates higher sensitivity and thus greater variability in the *estimated calibration*. Furthermore, this plot can highlight unusual patterns or artifacts introduced by the calibration method, such as excessive smoothness that fails to capture nuances in the true map, or overly complex step-like behavior that might indicate overfitting to a particular label set. The y-axis represents probability, typically ranging from 0 to 1.

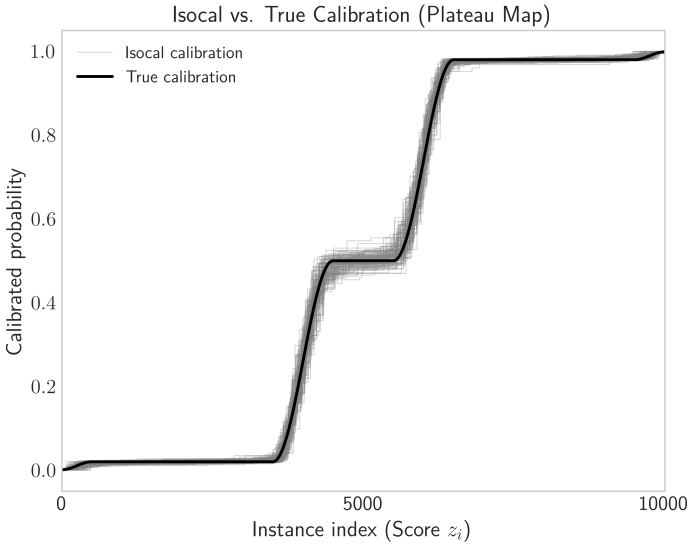


Figure 7: Example: Overlay of True and Estimated Calibration Curves. Grey lines represent individual *estimated calibration* curves from $N_{\text{runs}} = 100$ different simulated label sets; the black line is the known *true calibration* map.

2.3.2 Misestimation Distribution Visualization

The second **characterization plot**, illustrated by the example in Figure 8, directly visualizes the misestimation—the error of each *estimated calibration* curve relative to

the *true calibration* map. Using the same one hundred *estimated calibration* curves from the first plot, we calculate the pointwise difference: $\text{misestimation}_i = \hat{p}_{\text{calibrated}}(z_i) - c_i$, for every instance index i and for each of the 100 runs. For instance, if for index 5000, one run estimated a probability of 0.7 while the true probability c_{5000} was 0.6, the misestimation for that run at that point is $0.7 - 0.6 = +0.1$, signifying an overestimation of 0.1. Conversely, an estimate of 0.55 would result in a misestimation of $0.55 - 0.6 = -0.05$, an underestimation of 0.05.

These difference values are plotted for all instance indices across all runs, forming one hundred "misestimation curves." A horizontal dashed line at $y = 0$ represents perfect calibration (zero misestimation). The y-axis limits for this plot are dynamically set to encompass the typical range of these errors observed for the specific method and true map combination. This plot enables a more focused examination of the errors. It clearly highlights score regions (ranges on the x-axis) where the method tends to produce positive differences (overestimation) or negative differences (underestimation). The vertical distance from the $y = 0$ line quantifies the magnitude of the misestimation for each run. Observing the bundle of these misestimation lines also reveals the consistency of the error profile; if the bundle is tightly packed and consistently above or below zero in a region, it indicates a systematic error pattern, whereas a widely spread bundle crossing zero frequently suggests more random error behavior.

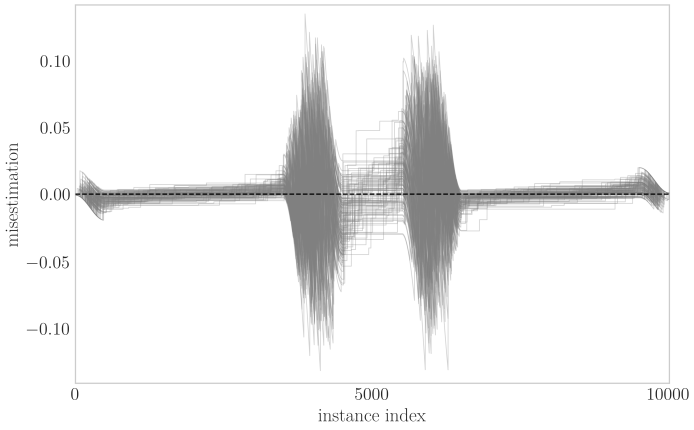


Figure 8: Example: Misestimation Curves. Grey lines show the pointwise difference (misestimation) between each *estimated calibration* curve and the *true calibration* map. The dashed black line at $y = 0$ indicates zero misestimation (perfect calibration).

2.3.3 Visualization of Percentile Envelopes

The third **characterization plot**, exemplified in Figure 9, summarizes the distribution of all one hundred *estimated calibration* curves using percentiles. This provides a robust view of the method's typical performance and its variability, indicating the range wherein 90% of the *estimated calibration* probabilities fall. For any single instance index

i (score z_i), the one hundred estimated probabilities $\hat{p}_{\text{calibrated}}(z_i)$ from the simulation runs are collected and sorted. From this sorted list, the 5th percentile (the value below which 5% of estimates lie), the 50th percentile (the median estimate), and the 95th percentile (the value below which 95% of estimates lie) are determined. This process is repeated for every instance index across the score range.

The plot then displays the area between the 5th and 95th percentile curves as a shaded band, often referred to as a percentile envelope. The median (50th percentile) curve is shown as a solid line (e.g., blue), and the 5th and 95th percentile curves are typically rendered as dashed lines bounding the shaded region. The *true calibration* map (black line) is also included for direct comparison. This statistical summary offers insights into the method's reliability. The median line represents the typical output; its proximity to the *true calibration* map indicates the method's average bias. The width of the shaded 5th-95th percentile band is a key indicator of the method's stability and variance: a narrow band suggests consistent outputs across different noisy label sets, while a wide band implies greater sensitivity to the specific calibration data and thus higher uncertainty in the *estimated calibration*. This envelope effectively defines a range within which 90% of the estimated probabilities are expected to lie for any given score z_i.

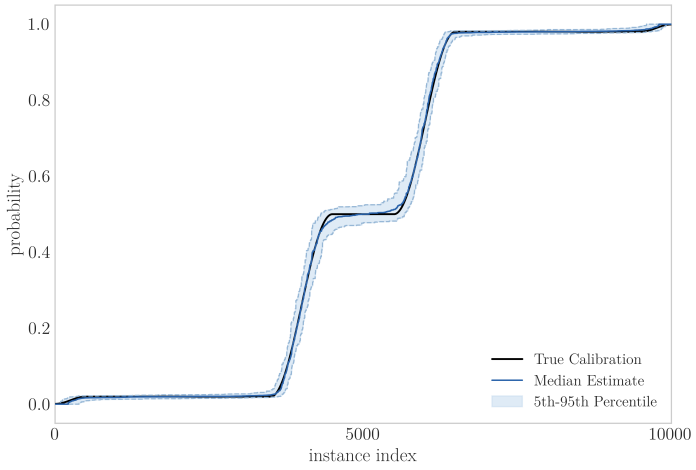


Figure 9: Example: Percentile Envelopes of Estimated Calibration. The shaded area shows the 5th-95th percentile range of all *estimated calibration* curves. The solid blue line is the median estimate, and the black line is the *true calibration* map.

2.3.4 Visualizing Mean Misestimation Size

The fourth **characterization plot**, depicted in the example Figure 10, visualizes the *average* misestimation across all one hundred simulation runs. This information is condensed into a single color-coded horizontal strip. For each instance index i (score z_i), the one hundred misestimation values (calculated as $\hat{p}_{\text{calibrated}}(z_i) - c_i$ for

each run) are averaged. For instance, if individual misestimations at a point were $[+0.1, -0.05, +0.02, \ldots]$, their average might be $+0.015$. This average misestimation is computed for every instance index along the x-axis.

The resulting series of average misestimations is then displayed as a thin horizontal colored strip. A diverging colormap is employed, where, for example, red shades signify a positive average misestimation (indicating an average tendency to overestimate at that score), blue shades denote a negative average misestimation (average underestimation), and white or a neutral color represents an average misestimation close to zero. The intensity of the color reflects the magnitude of this average misestimation. The *true calibration* map (black line) is overlaid on this strip for context, showing the true probability level at which these average errors occur. This plot effectively reveals the systematic average error of the calibration method. By averaging over many runs, random fluctuations due to label noise are smoothed out, exposing persistent biases. If a region of the strip is distinctly colored (e.g., red), it signifies that the method, on average, tends to produce *estimated calibration* probabilities that are higher than the true ones for those scores. The depth of the color indicates the magnitude of this systematic bias. This visualization is crucial for identifying problematic score regions where the calibration method's average output deviates from the *true calibration*. A colorbar provides the scale, mapping color to the numerical value of the average misestimation.

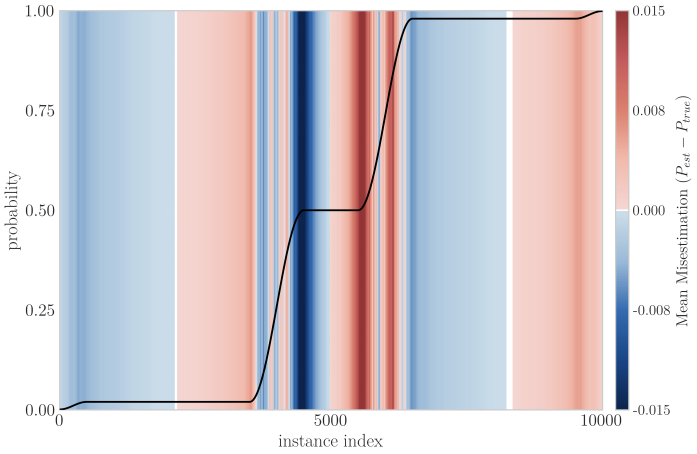


Figure 10: Example: Mean Misestimation Heatmap Strip. Color indicates the average difference between *estimated calibration* probabilities and *true calibration* probabilities (average error). Red typically indicates average overestimation, blue average underestimation. The black line is the *true calibration* map, shown for context.

2.3.5 Visualizing Overestimation Percentage

The fifth **characterization plot**, illustrated by the example in Figure 11, visualizes the dominant tendency of the calibration method to either overestimate or underestimate at each score point, focusing on the *frequency* of such directional errors. This is

achieved by showing the "percentage of times the method overestimates" across all one hundred runs, presented as a color-coded horizontal strip, similar in layout to the mean misestimation plot. For each instance index i (score z_i), the one hundred *estimated calibration* probabilities $\hat{p}_{\text{calibrated}}(z_i)$ are examined relative to the true probability c_i. The number of times these estimates are greater than the true probability, N_{over}, is counted. Similarly, the number of times they are less than the true probability, N_{under}, is counted. Due to the use of continuous probabilities, instances where $\hat{p}_{\text{calibrated}}(z_i) = c_i$ are extremely rare and typically do not affect this counting process as they would not contribute to N_{over} or N_{under}.

The plot then visualizes a metric derived from these counts, typically $N_{\text{over}}/(N_{\text{over}} + N_{\text{under}})$ (if the denominator is non-zero, otherwise undefined or neutral), which is interpreted as an "overestimation percentage." The colorbar legend for this plot ranges from 0% to 100% overestimation. A strong red color (near 100% overestimation) indicates that for that score point, in all or nearly all runs, the *estimated calibration* probability was higher than the *true calibration* probability, showing consistent overestimation. Conversely, a strong blue color (near 0% overestimation) means that in all or nearly all runs, the estimated probability was lower, signifying consistent underestimation (equivalent to nearly 100% underestimation). A white or neutral color (around 50% overestimation) implies a balance, where overestimations and underestimations occurred with roughly equal frequency, suggesting no consistent directional bias for that score point. The color intensity grades between these extremes; for example, 70% overestimation (a reddish hue) implies overestimation in approximately 70 runs and underestimation in 30. The *true calibration* map (black line) is overlaid for context.

This plot is crucial for understanding the consistency of the error's direction. It highlights if a method is systematically skewed towards overestimation or underestimation in certain score regions. This is distinct from the mean misestimation (Plot 4). For instance, a method might have a near-zero mean misestimation if it frequently overestimates by a large amount and equally frequently underestimates by a similar large amount. In such a case, Plot 4 would appear neutral, and Plot 5 would also likely appear neutral (around 50%). However, if the mean misestimation is small (faint color in Plot 4) but Plot 5 shows a strong red, it means the method consistently overestimates, albeit by a small average amount. Such consistent, even if small, overestimation can be particularly problematic in risk-sensitive applications where any tendency to overstate confidence is undesirable, as explored in cautious calibration frameworks (Allikivi et al., 2024). This plot thus provides a valuable tool for analyzing this aspect of calibration behavior. A neutral color in Plot 5 for a given score region suggests that there is no apparent systematic directional preference in the errors for that method and score, which indicates an absence of a consistent skew rather than necessarily poor performance.

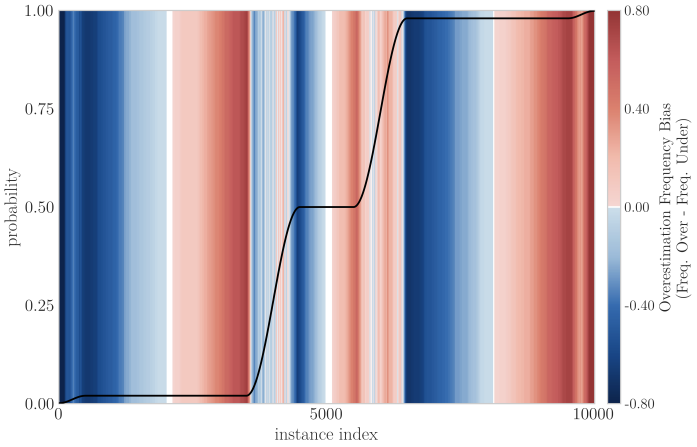


Figure 11: Example: Net Overestimation Frequency Heatmap Strip. Color indicates the percentage of runs where the method overestimated. Strong Red: approaches 100% overestimation. Strong Blue: approaches 0% overestimation (i.e., nearly 100% underestimation). White: around 50% overestimation (balanced directional error). The black line is the *true calibration* map, shown for context.

3 Results and Discussion

This chapter presents the empirical findings derived from the systematic application of five distinct post-hoc calibration methods to datasets generated from five unique *true calibration map* shapes. As detailed in Chapter 2, our experimental design leverages synthetic data to enable a precise and granular analysis of miscalibration patterns, which is often unachievable with real-world data due to the unknown nature of the true conditional probabilities. The calibration methods under investigation are Isotonic Calibration (Isocal), Logistic Calibration (Logcal), Beta Calibration (Betacal), Histogram Binning (Histbin), and the midpoint of Simplified Venn-Abers (SVA) predictors. Each method was applied to $N_{\text{runs}} = 100$ independent sets of simulated binary labels for each of the true map shapes: "Steep," "Smooth," "Plateau," "Random42," and "Random63."

The core of our analysis relies on a suite of five *characterization plots*, previously introduced in Section 2.3. These plots are designed to move beyond aggregate performance metrics and offer a multifaceted view of how the *estimated calibration* produced by each method aligns with, or deviates from, the known *true calibration*. For clarity, these characterization plots are:

1. **true calibration and estimated calibration curves:** an overlay of the true calibration map with 100 estimated calibration curves from the simulation runs;
2. **misestimation distribution:** visualization of the pointwise difference (error) between each estimated curve and the true calibration;
3. **percentile envelopes:** depiction of the 5th, 50th (median), and 95th percentiles of the estimated calibration probabilities across all runs;
4. **mean misestimation size:** a heatmap showing the average misestimation at each point along the score range;
5. **overestimation percentage:** a heatmap indicating the frequency with which a method overestimates the true probability at each point.

In the subsequent sections, we will dissect the performance and characteristic behaviors of each calibration method individually. For each method, a comprehensive 5x5 grid visualization will be presented and discussed. Each such grid displays all five characterization plots (rows) for that specific calibration method applied to each of the five true calibration map shapes (columns). This systematic presentation allows for a detailed understanding of how each method responds to different underlying probability landscapes. Following the individual analyses, we will conduct a comparative discussion, highlighting general trends, the relative strengths and weaknesses of the methods, the impact of true map complexity, and overarching patterns of misestimation. Finally, we will reflect on the broader implications of these detailed findings for the practical selection and interpretation of calibration methods in binary classification tasks.

3.1 Performance of Isotonic Calibration

Isotonic Calibration (Isocal), being a non-parametric method, generally demonstrates strong adaptability across the diverse true calibration map shapes, as illustrated in the comprehensive 5x5 grid of characterization plots presented in Figure 12. The *estimated calibration* curves produced by Isocal are inherently piecewise constant (step-like), a characteristic evident in all visualizations. One of the most notable features is the consistently small magnitude of misestimation, with the mean misestimation strip (row 4 of Figure 12) operating on a scale of approximately ± 0.029, among the smallest observed across all tested methods.

On the **"Steep" map** (column 1 of Figure 12), Isocal's estimated curves (row 1) closely follow the abrupt transition, with the adaptive steps capturing the sharpness well. The individual runs show some variability, and the misestimation (row 2) is concentrated into a tight "spindle" shape directly at the inflection point, indicating errors primarily occur during this rapid change. The 5th-95th percentile envelope (row 3) remains remarkably narrow throughout, even at the transition, signifying high stability across the 100 runs. The mean misestimation strip (row 4) is largely neutral (white), with only very faint, symmetric, alternating regions of pale red (slight average overestimation) and pale blue (slight average underestimation) immediately around the transition. The overestimation percentage strip (row 5) similarly shows a balanced pattern overall, with localized tendencies for slight overestimation just before the true curve's steepest ascent and slight underestimation immediately after.

For the **"Smooth" map** (column 2 of Figure 12), Isocal again provides a faithful approximation (row 1), with its adaptive steps closely tracking the gentle S-shape of the true calibration. The misestimation (row 2) consists of low-amplitude, somewhat oscillatory errors, primarily in the mid-range of scores. The percentile envelope (row 3) is very thin, indicating excellent consistency. The mean misestimation strip (row 4) displays very pale, alternating bands of red and blue, with average errors generally less than 0.01. Correspondingly, the overestimation percentage strip (row 5) is mostly neutral, indicating no large regions of consistent directional bias, though a subtle trend from slight underestimation at lower scores to slight overestimation at higher scores can be discerned.

Isocal's performance on the **"Plateau" map** (column 3 of Figure 12) highlights its adaptive nature. The step-like estimated calibration curves (row 1) align well with the flat plateau regions of the true map. However, misestimations (row 2) are more pronounced at the transitions between plateaus, where Isocal tends to overshoot slightly upon entering a plateau and lag upon exiting. This is reflected in the percentile envelope (row 3), which, while narrow within the plateaus, widens at these transition points. The mean misestimation strip (row 4) clearly shows this pattern: thin red stripes (average

overestimation) at the beginning of plateaus and blue stripes (average underestimation) at the ends. The overestimation percentage strip (row 5) mirrors this, with localized but strong directional biases at these entry/exit points—red as the estimated curve jumps up to meet the plateau and blue as it drops down.

When applied to the **"Random42"** and **"Random63"** maps (columns 4 and 5 of Figure 12), Isocal effectively captures the irregular, staircase-like structures (row 1). The variability between runs, as seen in the misestimation plots (row 2) and the width of the percentile envelope (row 3), is generally low but tends to increase at the sharpest jumps in the true random maps. The mean misestimation strip (row 4) often resembles a barcode, with thin alternating red and blue stripes corresponding to the local steps of the true map, indicating very localized average errors. Similarly, the overestimation percentage strip (row 5) shows narrow vertical bars of alternating directional bias, without large, contiguous regions of systematic over- or underestimation.

A characteristic quirk of Isotonic Calibration, visible especially on flatter regions of the "Plateau" or random maps, can be the appearance of very small "teeth" or minor steps in the estimated calibration curves (row 1). This may suggest a slight tendency to overfit to minor fluctuations in the observed labels, though the overall magnitude of misestimation remains very low (within ± 0.029). The adaptive placement of its steps allows Isocal to conform closely to varied true calibration shapes, making it a highly flexible method.

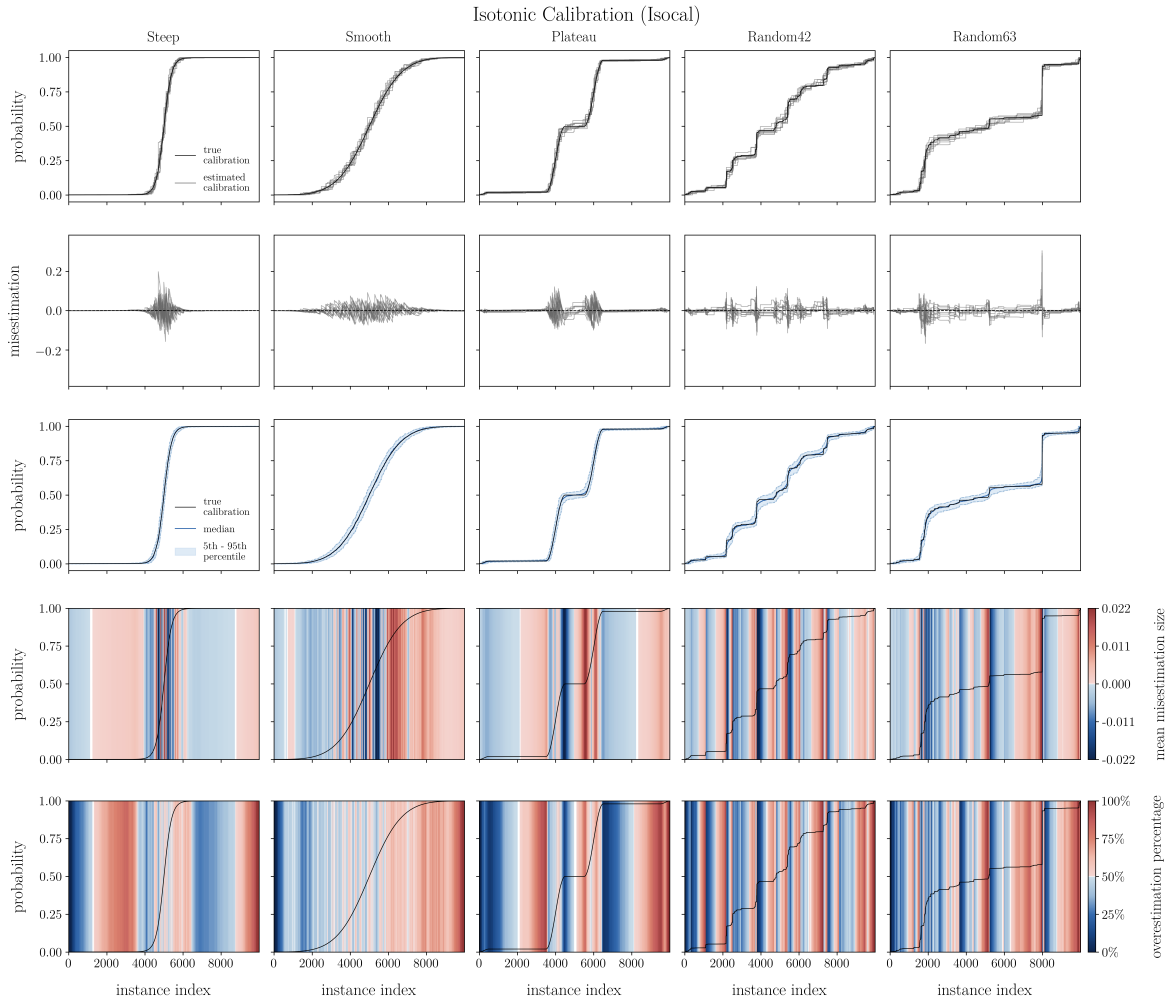


Figure 12: Characterization of Isotonic Calibration. Each column corresponds to a true calibration map shape (Steep, Smooth, Plateau, Random42, Random63), and each row to a characterization plot type (True vs. Estimated, Misestimation Distribution, Percentile Envelopes, Mean Misestimation Size, Overestimation Percentage).

3.2 Performance of Logistic Calibration

Logistic Calibration (Logcal), also known as Platt Scaling, exhibits markedly different behavior compared to the non-parametric methods, as vividly illustrated in its 5x5 grid of characterization plots (Figure 13). Being a parametric method constrained to fit a single sigmoid function, Logcal's *estimated calibration* curves are exceptionally smooth and demonstrate almost no variability across the $N_{\text{runs}} = 100$ simulations; the individual estimated curves (row 1 of Figure 13) are virtually indistinguishable from one another, and the 5th-95th percentile envelope (row 3) is almost invisibly thin for all true map shapes. This extreme stability, however, comes at the cost of adaptability. The method's performance is highly contingent on how closely the *true calibration* map resembles a logistic sigmoid. When there is a mismatch, Logcal can

introduce substantial and systematic misestimations, operating on a much larger error scale (approximately ± 0.221 for the mean misestimation strip in row 4) than the more flexible methods.

For the **"Steep" map** (column 1 of Figure 13), Logcal's inherent smoothness prevents it from capturing the abruptness of the true transition. The estimated calibration curve (row 1) rises much more gradually than the true map. This leads to significant, systematic misestimation (row 2): pronounced underestimation before the true transition point and overestimation after it. The mean misestimation strip (row 4) shows intense blue (underestimation) and red (overestimation) regions around the transition, reflecting the largest average errors observed for this method. The overestimation percentage strip (row 5) is striking, displaying nearly 100% consistent underestimation (deep blue) before the transition and 100% consistent overestimation (deep red) after, with a very sharp boundary between these regions.

The **"Smooth" map** (column 2 of Figure 13), being generally sigmoid-like, offers the best-case scenario for Logcal among the tested shapes. Here, the estimated calibration (row 1) aligns reasonably well with the true map. However, subtle systematic deviations are still present. The misestimation plot (row 2) shows a consistent pattern of slight underestimation at the lower and higher ends of the probability range, and a slight overestimation in the middle. This is mirrored in the mean misestimation strip (row 4) as alternating pale blue and red bands. The overestimation percentage strip (row 5) also clearly delineates these regions of consistent, albeit smaller, directional bias.

When confronted with the **"Plateau" map** (column 3 of Figure 13), Logcal's rigid sigmoid form is entirely unable to reproduce the true map's distinct plateaus (row 1). The estimated sigmoid cuts through the plateaus, leading to large, oscillating systematic errors (row 2) across the entire score range. The mean misestimation strip (row 4) shows intense, broad bands of alternating blue (underestimation on the first and third plateaus) and red (overestimation on the middle plateau and the transitions). The overestimation percentage strip (row 5) indicates near-perfect consistency in these directional biases, with large blocks of either 0% or 100% overestimation.

The performance on the **"Random42"** and **"Random63"** maps (columns 4 and 5 of Figure 13) further underscores Logcal's lack of flexibility. The single smooth sigmoid estimated by Logcal (row 1) only captures the gross global trend of these irregular true maps, completely missing all local steps and variations. This results in substantial and complex patterns of misestimation (row 2) throughout the score range. The mean misestimation strips (row 4) for these random maps exhibit strong, alternating red and blue regions where the true map deviates significantly from the fitted sigmoid. Similarly, the overestimation percentage strips (row 5) show dramatic, large blocks of highly consistent directional bias (either almost pure red or pure blue), highlighting the

model's inability to adapt to non-sigmoid structures.

In summary, Logistic Calibration provides extremely stable (low variance) *estimated calibration* outputs. However, its strong parametric assumption of a sigmoid shape means it can only achieve good calibration if the true underlying relationship is indeed sigmoid. For true calibration maps that deviate from this form, Logcal introduces significant systematic bias, leading to large and consistent misestimations. This behavior is a direct consequence of its limited two-parameter functional form.

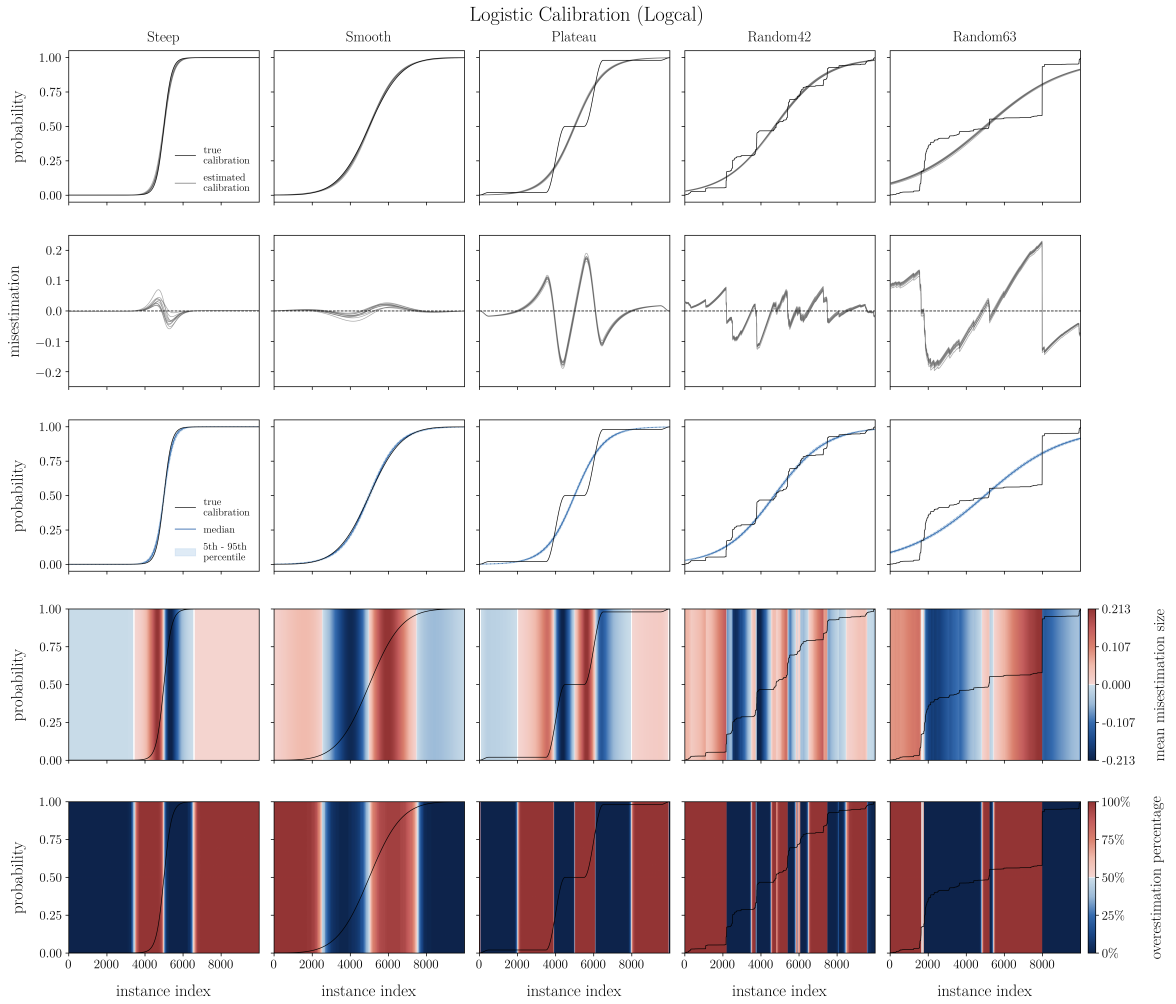


Figure 13: Characterization of Logistic Calibration. Columns and rows are organized as in Figure 12.

3.3 Performance of Beta Calibration

Beta Calibration (Betacal) distinguishes itself as a parametric method offering enhanced flexibility over Logistic Calibration, aiming to model a broader range of smooth, monotonic true calibration maps. Its characteristic performance is detailed in Figure 14. A defining trait of Betacal, shared with Logistic Calibration, is its exceptional stability: the 5th-95th percentile envelopes (row 3 of Figure 14) are consistently very narrow

across all true map shapes, indicating minimal variation in the *estimated calibration* across the 100 simulation runs. However, despite its increased adaptability due to a three-parameter model, Betacal can still incur significant systematic misestimation when the true underlying probability landscape deviates from shapes it can effectively represent. The scale for its mean misestimation (row 4) is approximately ± 0.196, suggesting that average errors, while often less severe than those of Logistic Calibration in complex scenarios, remain substantially larger than those achieved by non-parametric methods like Isotonic Calibration.

When applied to the **"Steep" map** (column 1 of Figure 14), Betacal's estimated curve (row 1) manages a steeper ascent than a standard logistic curve might, yet it visibly smooths the true map's abrupt transition. This smoothing results in a clear pattern of systematic error (row 2): underestimation before the true curve's sharpest rise, followed by overestimation immediately after. These deviations are reflected in the mean misestimation strip (row 4) as distinct blue and red bands. The consistency of this directional error is high, with the overestimation percentage strip (row 5) showing nearly complete underestimation (deep blue) changing sharply to complete overestimation (deep red) around the transition point.

Betacal's strength is most apparent on the **"Smooth" map** (column 2), which aligns well with its parametric capabilities. Here, the estimated calibration curve (row 1) tracks the true sigmoid-like map almost perfectly. Consequently, misestimations (row 2) are minimal, and the mean misestimation strip (row 4) is predominantly white, indicating negligible average error. The overestimation percentage strip (row 5) is also largely neutral, confirming the absence of significant systematic directional bias in this ideal scenario.

The limitations of Betacal's smooth functional form become evident with the **"Plateau" map** (column 3). While its estimated curve (row 1) shows more undulation than Logcal's simple sigmoid in an attempt to follow the plateaus, it cannot reproduce the flat regions or sharp corners. This leads to substantial, wave-like systematic misestimations (row 2), characterized by underestimation on the true plateaus and overestimation during the transitions between them. The mean misestimation strip (row 4) vividly displays this with broad, intensely colored bands of red and blue. The overestimation percentage strip (row 5) underscores the systematic nature of these errors, showing large, contiguous blocks of nearly 100% consistent directional bias.

For the highly irregular **"Random42" and "Random63" maps** (columns 4 and 5), Betacal's inherent smoothness dictates that its estimated curves (row 1) can only capture the general increasing trend, effectively smoothing over all local steps and variations. This results in complex and often large misestimations (row 2) wherever the smooth estimated curve diverges from the jagged true calibration. The mean misestimation

strips (row 4) for these maps show broad, alternating red and blue regions, highlighting significant systematic average errors. Similarly, the overestimation percentage strips (row 5) feature large blocks of highly consistent overestimation or underestimation, indicating that while the specific error patterns are complex, the directional bias in any given region is quite stable across runs.

In summary, Beta Calibration occupies an intermediate position. It offers the high stability of parametric methods but with greater flexibility in shape modeling than Logistic Calibration, allowing it to better fit asymmetric or more nuanced smooth curves. However, it retains a strong smoothness constraint, which leads to significant systematic misestimations when confronted with true calibration maps featuring sharp discontinuities like plateaus or highly irregular, step-like random patterns. While its errors on such complex maps may be less extreme than Logcal's, they are still considerably larger than those from adaptive non-parametric techniques.

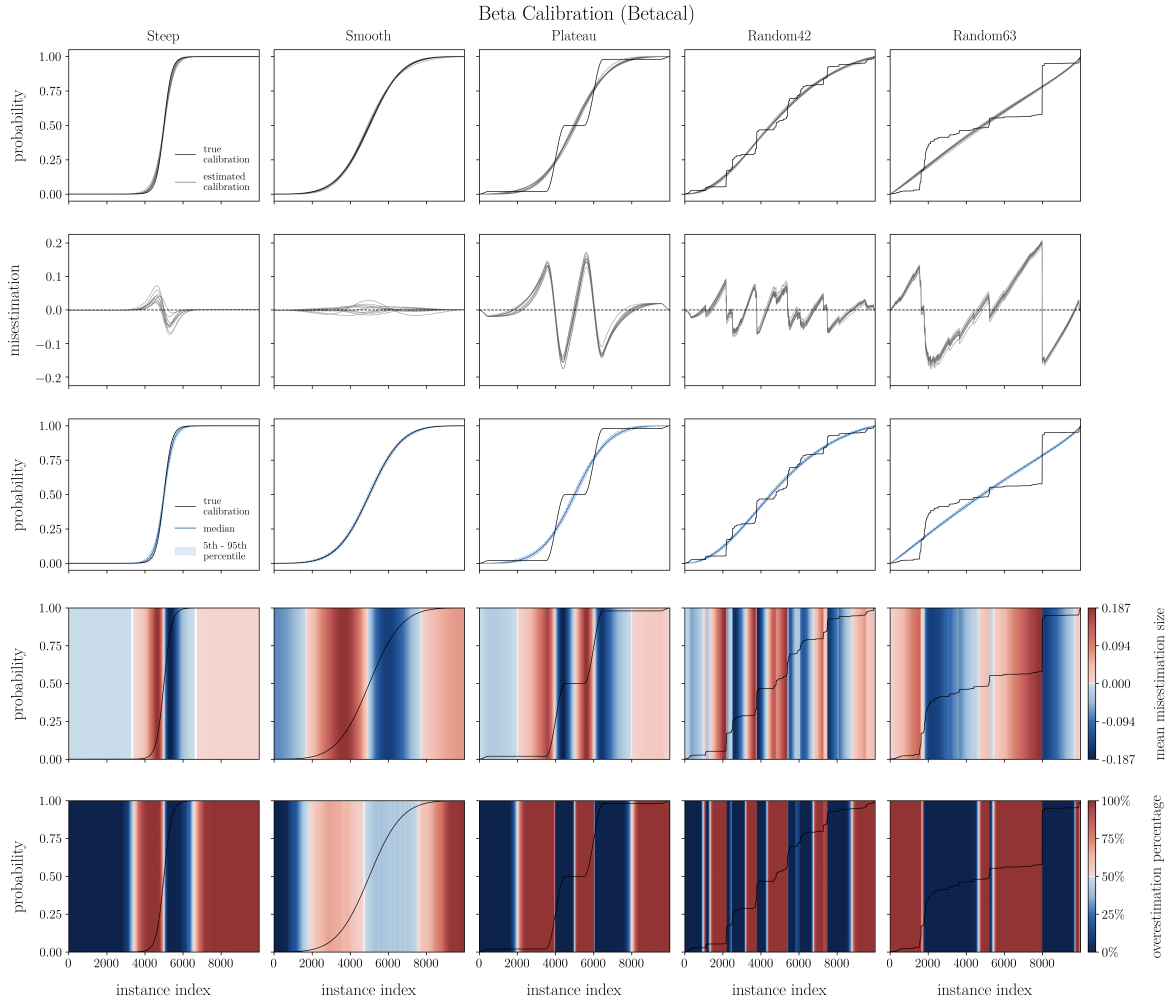


Figure 14: Characterization of Beta Calibration. Columns and rows are organized as in Figure 12.

3.4 Performance of Simplified Venn-Abers (SVA)

The Simplified Venn-Abers (SVA) midpoint, derived from averaging the outputs of two isotonic regressions under hypothetical label assignments, demonstrates a performance profile that closely mirrors that of standard Isotonic Calibration. The comprehensive characterization of SVA is presented in Figure 15. Consistent with its Isotonic Regression underpinnings, SVA produces *estimated calibration* curves that are piecewise constant (step-like), allowing for considerable adaptability to diverse *true calibration map* shapes. A key characteristic, shared with Isotonic Calibration, is the remarkably small magnitude of average misestimation, with the mean misestimation strip (row 4 of Figure 15) operating on a tight scale of approximately ± 0.029.

On the **"Steep" map** (column 1 of Figure 15), SVA's estimated curves (row 1) effectively capture the abrupt transition, with the adaptive steps aligning well with the true map's sharpness. Misestimations (row 2) are primarily concentrated in a narrow "spindle" shape around the inflection point. The 5th-95th percentile envelope (row 3) is very narrow, even through the steep change, indicating high stability across simulation runs. The mean misestimation strip (row 4) is largely neutral, showing only faint, localized, alternating pale red and blue hues immediately around the transition. The overestimation percentage strip (row 5) reveals a consistent directional pattern in this critical region: a tendency for slight overestimation (pale red) just before the steepest ascent, followed by slight underestimation (pale blue) immediately after.

SVA demonstrates excellent performance on the **"Smooth" map** (column 2). The step-like estimated curves (row 1) closely approximate the true S-shape, with the adaptive nature of the steps allowing for a good fit. Misestimations (row 2) are small and scattered, with the percentile envelope (row 3) remaining very thin across the entire score range, signifying high consistency. The mean misestimation strip (row 4) is almost entirely white, reflecting minimal average error. The overestimation percentage strip (row 5) is mostly neutral, though a very subtle, broad trend from slight underestimation at lower scores to slight overestimation at higher scores can be observed, indicating a well-balanced error profile for this map shape.

For the **"Plateau" map** (column 3), SVA's adaptive step functions effectively capture the flat plateau regions and the transitions between them (row 1). Similar to Isotonic Calibration, misestimations (row 2) are most pronounced at the edges of the plateaus, where the estimated curve might slightly lead or lag the true transitions. The percentile envelope (row 3) is narrow within the plateaus and widens modestly at these transition points. The mean misestimation strip (row 4) shows thin red bands (average overestimation) at the start of plateaus and blue bands (average underestimation) at their ends. This is mirrored in the overestimation percentage strip (row 5), which displays localized but strong directional biases at these entry and exit points from the

plateaus.

When applied to the irregular **"Random42" and "Random63" maps** (columns 4 and 5), SVA's flexibility allows its estimated curves (row 1) to adapt well to the complex, staircase-like structures. Misestimations (row 2) are generally small and scattered, though the percentile envelope (row 3) shows some widening in regions with particularly sharp or frequent jumps in the true map. The mean misestimation strip (row 4) often presents a "barcode" appearance, with thin, alternating pale red and blue bands corresponding to the local steps of the true map. The overestimation percentage strip (row 5) also exhibits a complex pattern of localized directional biases, without large contiguous areas of systematic over- or underestimation.

Given that the SVA midpoint calculation is fundamentally based on Isotonic Regression, its performance characteristics are, as expected, very similar. SVA might be "slightly 'shrunk' toward the centre thanks to averaging p_0 and p_1," which could imply a subtle regularization effect. Visually comparing Figure 15 with Figure 12 (for Isotonic Calibration), the differences are indeed very subtle if present. Both methods offer high adaptability, excellent stability for non-parametric approaches, and maintain a very low magnitude of average misestimation. Any potential regularization effect from the SVA midpoint averaging appears to be minimal in these experiments, with both methods performing almost identically.

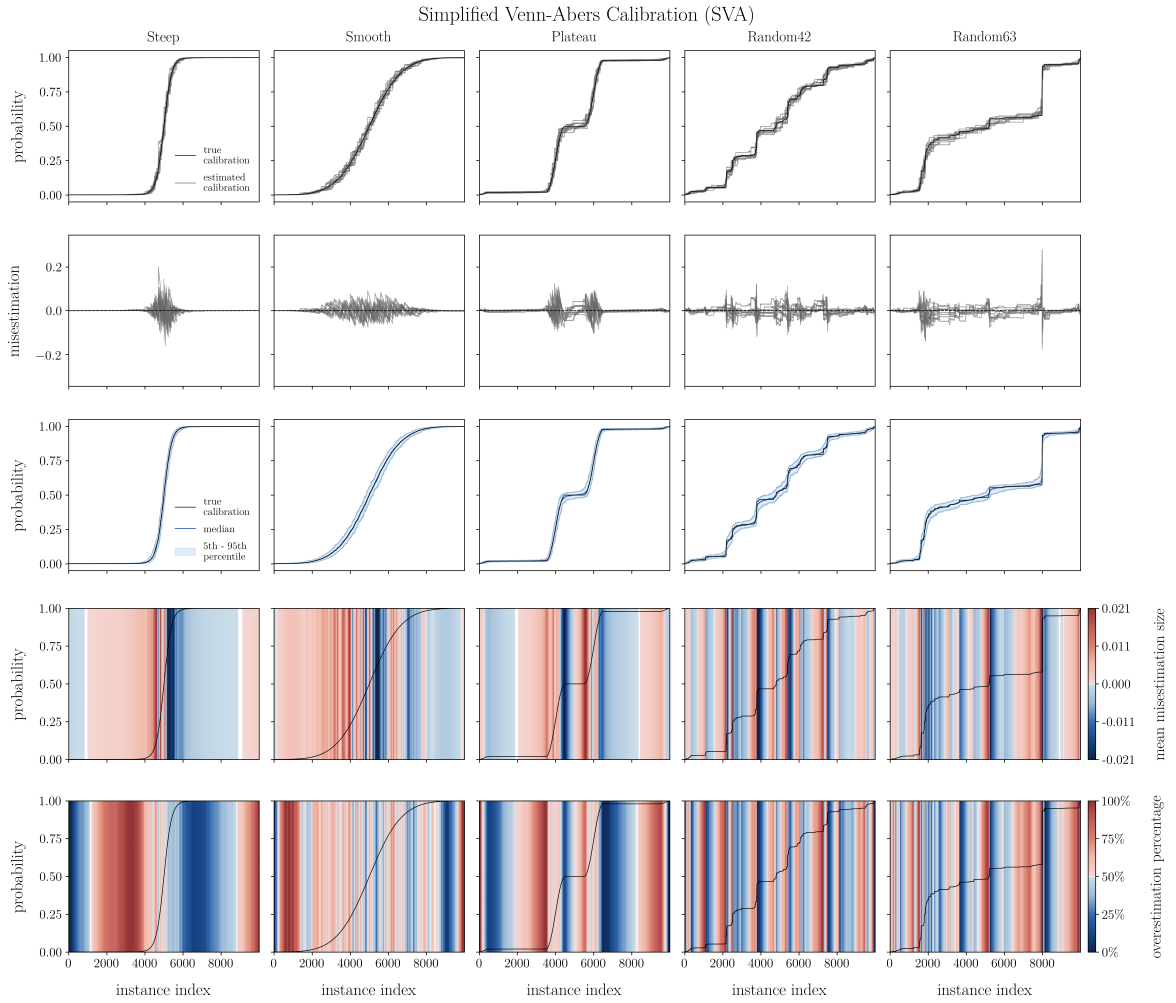


Figure 15: Characterization of Simplified Venn-Abers (SVA) Midpoint. Columns and rows are organized as in Figure 12.

3.5 Performance of Histogram Binning

Histogram Binning (Histbin), configured with 15 equal-width bins for these experiments, produces *estimated calibration* curves with a distinctly coarse, step-like appearance, as shown in Figure 16. This rigid structure, a direct consequence of assigning a single probability value to all scores within each fixed bin, is the most salient visual characteristic of this method across all true map shapes and characterization plots. The method's tendency for misestimation can be significant, with the mean misestimation strip (row 4 of Figure 16) operating on a scale of approximately ± 0.165.

On the **"Steep" map** (column 1 of Figure 16), the fixed bins provide a crude, stepped approximation of the true map's rapid transition (row 1). This coarse discretization results in sharp, spiky misestimations (row 2) around the inflection point, with characteristic "sawtooth" patterns emerging as the true curve passes through the flat estimates within each bin and jumps at bin boundaries. The percentile envelope (row

3) is notably wide at the transition, indicating considerable variability and sensitivity to the specific label set when estimating probabilities in this critical region. The mean misestimation strip (row 4) and the overestimation percentage strip (row 5) both exhibit striking vertical stripes of alternating red and blue, clearly demarcated by the bin edges, indicating sharp changes in average error and consistent directional bias from one side of a bin boundary to the other.

This characteristic striped artifact is even more pronounced on the **"Smooth" map** (column 2). The regular, equal-width bins impose a very obvious staircase structure onto the smooth true calibration (row 1). This leads to a highly regular, repeating sawtooth pattern in the misestimation plot (row 2) across the entire score range. The mean misestimation strip (row 4) shows a "zebra-stripe" pattern of alternating red and blue, perfectly aligned with the bin boundaries. The overestimation percentage strip (row 5) is perhaps the most dramatic illustration of this artifact, displaying an almost perfectly periodic pattern of deep red (consistent overestimation) and deep blue (consistent underestimation) bars, again defined by the fixed bin locations.

When applied to the **"Plateau" map** (column 3), Histogram Binning's performance is heavily influenced by how the fixed bin boundaries align with the true plateaus and transitions. The estimated curves (row 1) create a jagged approximation. If a bin happens to align well with a plateau, the error within that segment might be small, but significant misestimations (row 2) with pronounced spikes and sawtooth patterns occur at transitions and where bin edges cut across plateaus. The mean misestimation (row 4) and overestimation percentage (row 5) strips show intense, alternating colored bars, particularly visible at the true plateau boundaries, reflecting the abrupt changes in error due to the binning.

For the irregular **"Random42" and "Random63" maps** (columns 4 and 5), the fixed 15-bin structure results in a heavily simplified and coarse approximation of the true underlying probability landscapes (row 1). While the general increasing trend might be captured, most local features and steps are lost. Misestimations (row 2) are scattered throughout, with spikes often occurring at bin boundaries that are misaligned with the true map's features. The percentile envelope (row 3) can be quite wide in regions of high true map complexity, reflecting the instability of bin-based estimates. The error strips (rows 4 and 5) show irregular, striped patterns dictated by the bin edges rather than the true map's nuanced structure, often resulting in blocks of consistent directional bias within individual bins.

In summary, Histogram Binning is characterized by its highly artifact-prone *estimated calibration* curves. The fixed nature and predetermined number of bins are its defining limitation, leading to a performance that is strongly dependent on the interplay between bin boundaries and the features of the *true calibration* map. While

simple to understand and implement, it generally exhibits higher variability and larger, more systematic (within-bin) errors compared to adaptive non-parametric methods like Isotonic Calibration or SVA, especially when the true map is complex or does not align well with the binning scheme. The characteristic sawtooth error patterns and striped heatmaps are a direct visual consequence of its underlying mechanism.

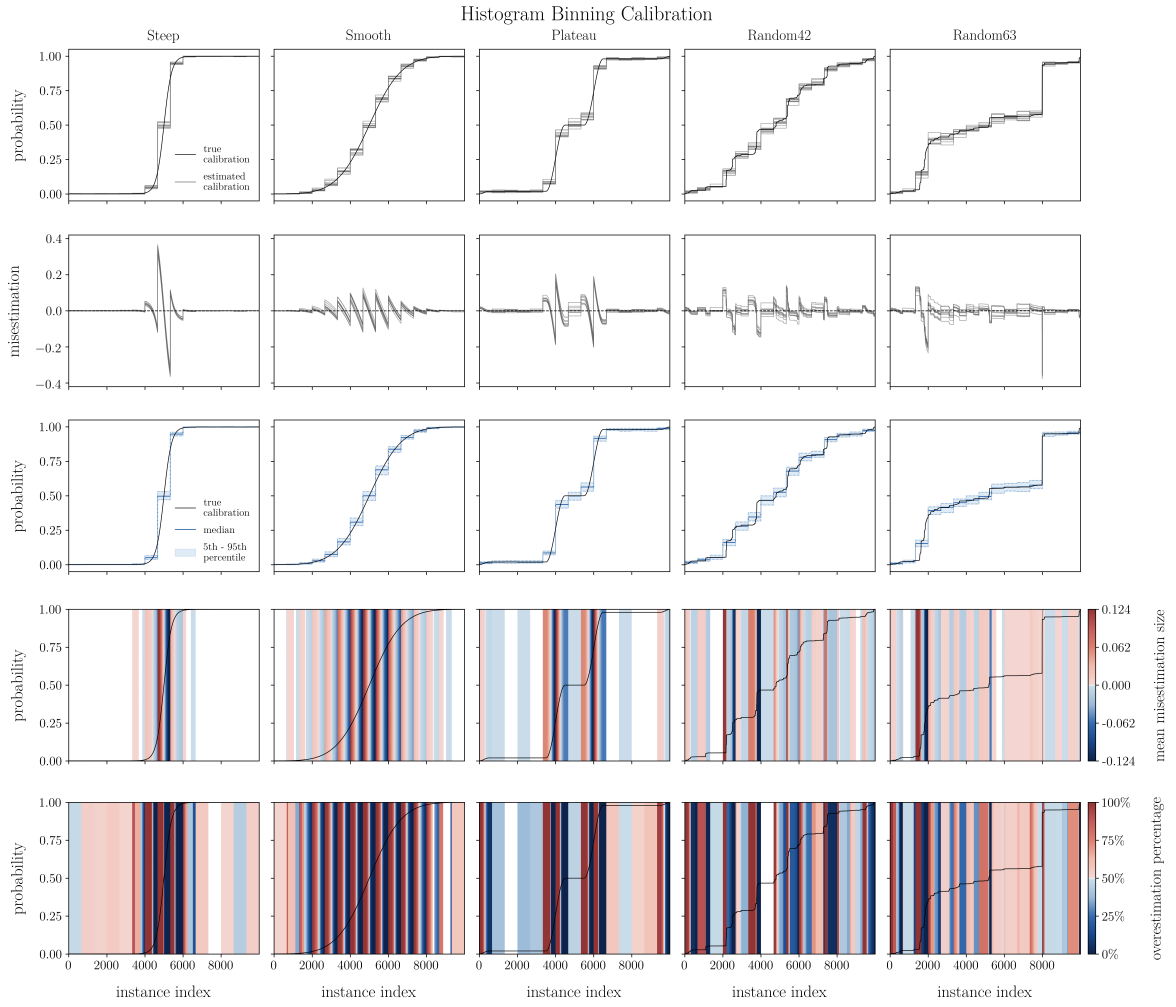


Figure 16: Characterization of Histogram Binning (15 bins). Columns and rows are organized as in Figure 12.

3.6 Comparative Analysis and General Trends

The individual analyses of the five calibration methods reveal distinct behavioral profiles and highlight several overarching trends regarding their adaptability, error characteristics, and stability. This comparative discussion aims to synthesize these observations, providing a clearer understanding of the relative strengths and weaknesses inherent in different calibration approaches. The primary visual references for this comparison are the comprehensive 5x5 characterization plot grids presented for each method (Figures 12 through 16).

Adaptability to true calibration map shapes: a clear distinction emerges in the ability of the methods to adapt to the diverse true calibration map shapes. The non-parametric methods, Isotonic Calibration (Isocal) and Simplified Venn-Abers (SVA) Midpoint, demonstrated the highest degree of adaptability. Their piecewise constant *estimated calibration* curves (row 1 of Figures 12 and 15) were able to closely follow all five true map shapes, including the "Steep" transition, the "Smooth" curve, the distinct "Plateau" regions, and the irregular "Random" patterns. Histogram Binning (Histbin), while also non-parametric, exhibited limited adaptability due to its fixed bin structure (row 1 of Figure 16). While it could approximate general trends, its coarse, pre-determined steps often failed to capture finer details or align well with features not conforming to its bin boundaries.

In stark contrast, the parametric methods, Logistic Calibration (Logcal) and Beta Calibration (Betacal), showed significantly less adaptability. Logcal, constrained by its two-parameter sigmoid function (row 1 of Figure 13), performed well only when the true map was inherently sigmoid-like (e.g., the "Smooth" map). It failed conspicuously to represent plateaus, steep transitions, or random irregularities. Betacal, with its three-parameter model (row 1 of Figure 14), offered more flexibility than Logcal, providing a better fit to the "Steep" map and being able to model some asymmetry. However, it still fundamentally produced smooth curves and could not adequately capture the discontinuities of the "Plateau" or "Random" maps.

Parametric vs. non-parametric trade-offs: the classic trade-off between flexibility and stability (or bias and variance) is evident. Parametric methods (Logcal and Betacal) exhibited extremely high stability across the 100 simulation runs, as seen in their almost invisibly thin 5th-95th percentile envelopes (row 3 of their respective figures). Their *estimated calibration* curves were virtually identical from one run to the next. However, this stability came at the cost of potentially large systematic bias when the true map shape did not conform to their parametric assumptions, leading to substantial mean misestimations (row 4).

Non-parametric methods, particularly Isocal and SVA, achieved much lower systematic bias on average across varied map shapes, due to their ability to adapt. Their percentile envelopes (row 3), while still indicating good stability, were generally wider than those of Logcal and Betacal, especially in regions of high true map complexity or sharp transitions, reflecting a slightly greater sensitivity to the specific label set realization. Histogram Binning showed even greater variability in its percentile envelopes and often produced more pronounced local artifacts.

Magnitude and nature of misestimation: the scale of average misestimation (row 4 of each figure) varied dramatically. Isocal and SVA consistently operated on the smallest error scale (approximately ± 0.029). Beta Calibration showed a larger scale

(approx. ±0.196), and Logistic Calibration exhibited the largest potential for average misestimation (approx. ±0.221) when mismatched with the true map. Histogram Binning's scale (approx. ±0.165) was also considerable. The nature of these errors also differed. Parametric methods, when mismatched, produced broad, smooth regions of misestimation. Non-parametric methods, when they erred, tended to have more localized errors—Isocal/SVA at the edges of their adaptive steps, and Histbin very distinctly at its fixed bin boundaries, often creating "sawtooth" patterns in the misestimation distribution (row 2).

Consistency of directional bias: the overestimation percentage strips (row 5 of each figure) provided crucial insights into the consistency of error direction. Logistic and Beta Calibration, due to their stable functional forms, often displayed large, contiguous blocks of nearly 100% consistent overestimation (deep red) or underestimation (deep blue) when their fitted shape systematically deviated from the true map. This indicates that for a given score region, they would almost always err in the same direction across different label sets. Isocal and SVA showed more nuanced patterns; directional bias was often localized around transitions or steps in the true map. While these methods could exhibit consistent directional error in such local regions, they rarely showed the sweeping, map-wide consistent biases seen with mismatched parametric models. Histogram Binning created very regular, alternating bands of consistent over- and underestimation, dictated entirely by its bin structure.

Characteristic artifacts: each method imprinted characteristic artifacts on its *estimated calibration*. **Logistic calibration:** overly smooth sigmoid curves, failing to capture non-sigmoid features. **Beta calibration:** smoother than Logcal and more flexible, but still unable to model discontinuities; prone to wave-like errors on plateaus. **Isotonic calibration & SVA midpoint:** adaptive, piecewise constant (step-like) functions; potential for small "teeth" or minor wiggles on flat regions. **Histogram binning:** rigid, coarse, equal-width steps; prominent sawtooth error patterns at bin boundaries.

In conclusion, no single method excelled across all true map shapes and all aspects of performance. Non-parametric methods like Isotonic Calibration and SVA offered the best general adaptability and lowest average misestimation across diverse scenarios, while parametric methods like Logistic and Beta Calibration provided superior stability but at the risk of substantial systematic bias if their underlying functional assumptions were violated. Histogram Binning, while simple, introduced significant artifacts tied to its fixed binning scheme. These comparative observations underscore the importance of understanding the underlying characteristics of calibration methods beyond aggregate performance metrics.

3.7 Discussion of Miscalibration Patterns and Implications

The detailed characterization of five prominent post-hoc calibration methods across diverse true probability landscapes, as presented in this chapter, offers several key insights into the nature of miscalibration and the behavior of corrective techniques. This granular analysis, facilitated by our synthetic data framework and the suite of characterization plots, moves significantly beyond what can be gleaned from single aggregate performance metrics like Expected Calibration Error (ECE) or Brier score. While such metrics provide a useful summary of overall calibration quality, they inherently obscure the rich and often complex patterns of misestimation that this study has sought to illuminate.

A primary contribution of this work is the visual and qualitative demonstration of how different calibration methods err. We observed, for instance, the stark trade-off inherent in parametric methods like Logistic and Beta Calibration: their *estimated calibration* functions are exceptionally stable across different realizations of training data, but this stability comes at the cost of potentially large and systematic biases when the true calibration map deviates from their assumed functional forms. Conversely, non-parametric methods like Isotonic Calibration and SVA Midpoint exhibit remarkable adaptability, conforming closely to varied true map shapes and generally maintaining low average misestimation. However, their flexibility can lead to slightly higher variance (i.e., greater sensitivity to the specific calibration dataset) and the introduction of step-like artifacts, which might be undesirable in some applications requiring smooth probability estimates. Histogram Binning, while conceptually simple, clearly demonstrated how a rigid, pre-determined structure can introduce significant and predictable artifacts, with its performance being highly contingent on the alignment of bin boundaries with features in the true calibration map.

The characterization plots, particularly the mean misestimation size (row 4 of Figures 12 through 15) and the overestimation percentage strips (row 5), provide crucial information for risk-sensitive decision-making. The overestimation percentage strip, for example, highlights regions where a method *consistently* overestimates or underestimates the true probabilities, even if the average magnitude of this misestimation (shown in the mean misestimation strip) is small. As discussed in the context of cautious calibration (Allikivi et al., 2024), even small but consistent overestimation can be detrimental in applications where the cost of overconfidence is high (e.g., medical diagnosis, safety-critical systems). Our visualizations allow for the direct identification of such tendencies for each method and true map combination, offering a more nuanced basis for method selection than aggregate error scores alone. For instance, while Isotonic Calibration generally had very low average error, its overestimation percentage plot still revealed localized regions of consistent directional bias around sharp transitions or steps.

These findings have direct implications for practitioners. The choice of a calibration method should not be based solely on its reported aggregate performance on benchmark datasets. Instead, consideration should be given to:

- **the expected complexity of the true calibration map**, if the underlying relationship between scores and true probabilities is believed to be simple and smooth (e.g., roughly sigmoid), a stable parametric method like Beta Calibration (or even Logistic Calibration if symmetry is also expected) might suffice and offer benefits in terms of low variance. However, if complex, non-monotonic, or step-like patterns are anticipated, more flexible non-parametric methods like Isotonic Calibration or SVA are likely to provide more accurate *estimated calibration*;
- **the tolerance for different types of errors**, is systematic bias more or less acceptable than higher variance in the probability estimates? Are there asymmetric costs associated with overestimation versus underestimation for the specific application? The characterization plots provide a means to assess these different error profiles;
- **the desire for smooth vs. step-like probability outputs** some applications might prefer smooth probability transitions, while for others, the step-like nature of Isotonic or Histogram Binning might be acceptable or even align with discrete decision thresholds.

This in-depth analysis underscores that "good calibration" is multifaceted. A method that performs well according to one metric or on one type of true calibration map may exhibit undesirable behaviors on another. The visualization framework presented here offers a powerful tool for researchers and practitioners to gain a deeper, more intuitive understanding of these behaviors, fostering more informed decisions about the application and interpretation of calibration techniques in binary classification. While this study explored five representative true map shapes, the framework itself is general and could be extended to analyze performance on an even wider array of probability landscapes or with different calibration method parametrizations, further enriching our understanding of miscalibration phenomena. Future work could also focus on developing new summary metrics that better capture these nuanced aspects of miscalibration character, moving beyond single scalar values to more descriptive profiles.

Conclusion

The reliability of probability estimates from binary classifiers is essential for effective decision-making across numerous domains. However, assessing the true quality of these probabilities is often challenging, as conventional aggregate metrics can obscure the intricate ways in which calibration methods may deviate from the true underlying likelihoods. This limitation motivated the central goal of this thesis: to conduct an in-depth analysis and detailed characterization of miscalibration patterns for common post-hoc calibration techniques.

To achieve this, we developed an experimental framework centered on synthetic data generation, allowing for precise control and knowledge of five diverse *true calibration map* shapes. Five widely-used methods—Isotonic Calibration, Logistic Calibration, Beta Calibration, Histogram Binning, and the SVA Midpoint—were applied to data simulated from these true maps across 100 independent simulation runs for each scenario. A suite of five specialized *characterization plots* was then employed to visualize and dissect the resulting *estimated calibration* outputs, focusing on patterns of accuracy, bias, variance, and directional error.

Our findings vividly illuminated the distinct behavioral profiles and inherent trade-offs of these methods. Parametric techniques like Logistic and Beta Calibration demonstrated exceptional stability across simulation runs but were prone to substantial systematic bias when their functional assumptions were violated by the true calibration map's shape. In contrast, non-parametric methods, particularly Isotonic Calibration and SVA Midpoint, showcased superior adaptability to varied true probability landscapes and maintained lower average misestimation magnitudes, albeit with characteristic step-like outputs and slightly higher variance in complex regions. Histogram Binning's performance was notably constrained by its rigid, fixed-bin structure, often leading to pronounced artifacts. Crucially, the characterization plots, especially those visualizing mean misestimation size and overestimation percentage, revealed nuanced error profiles—such as consistent directional biases even when average error was small—that would be obscured by single aggregate metrics.

This granular understanding of miscalibration patterns carries significant implications. It equips practitioners with a more profound basis for selecting calibration methods, urging consideration not just of overall error scores but also of the specific types of misestimation a method is prone to and how those align with the risk profile of their application. The insights gained are particularly relevant for risk-sensitive domains where the nature and consistency of calibration errors, especially tendencies towards overconfidence, are critical.

Ultimately, this work underscores that "good calibration" is a multifaceted concept.

The analytical framework and visualization techniques presented herein offer a valuable paradigm for gaining a deeper, more intuitive comprehension of calibration method behaviors. Future research could beneficially extend this approach to a broader array of calibration techniques and true probability scenarios. Furthermore, the development of new summary metrics, inspired by these detailed characterizations and capable of capturing such nuanced aspects of calibration quality beyond single scalar values, remains a promising avenue for advancing the field.

References

Allikivi, Mari-Liis, Joonas Järve, and Meelis Kull (2024). "Cautious Calibration in Binary Classification". In: *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*, pp. 1503–1510.

Barlow, Richard E., David J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. John Wiley & Sons.

Brier, Glen (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1, pp. 1–3.

Gneiting, Tilmann and Adrian E. Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Guo, Chuan, Geoff Pleiss, Yu Sun, and K. Q. Weinberger (2017). "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1321–1330.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Andreas Haldane, Jaime Fernández del Róio, Mark Wiebe, Pearu Peterson, Prokop Hapka, Travis Oliphant, et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362.

Kull, Meelis, Telmo Silva Filho, and Peter Flach (2017). "Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, pp. 623–631.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Naeini, Mahdi Pakdaman, Gregory F. Cooper, and Milos Hauskrecht (2015). "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI'15. AAAI Press, pp. 2901–2907.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting Good Probabilities with Supervised Learning". In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. ACM, pp. 625–632. DOI: 10.1145/1102351.1102430.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher,

Matthieu Perrot, and Édouard Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Platt, John C. (1999). “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. In: *Advances in Large Margin Classifiers*, pp. 61–74.

Shafer, Glenn and Vladimir Vovk (Mar. 2008). “A Tutorial on Conformal Prediction”. In: *Journal of Machine Learning Research* 9, pp. 371–421.

Vovk, Vladimir, Alex Gammerman, and Glenn Shafer (2005). *Algorithmic Learning in a Random World*. New York: Springer.

Vovk, Vladimir and Ivan Petej (2012). “Venn-Abers Predictors”. In: *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pp. 829–838.

Zadrozny, Bianca and Charles Elkan (2001). “Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers”. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML ’01. Morgan Kaufmann Publishers Inc., pp. 609–616.

— (2002). “Transforming classifier scores into accurate multiclass probability estimates”. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699.

I Licence

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Heili Aavola,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis In-Depth Analysis of Miscalibration in Binary Classification, supervised by Mari-Liis Allikivi;
2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;
3. am aware of the fact that the author retains the rights specified in points 1 and 2;
4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Heili Aavola
15/05/2025