PROMIS: A Post-Processing Framework for Mitigating Spatial Bias

Dimitris Kyriakopoulos
dkyriakopoulos@athenarc.gr
Athena Research Center
Athens, Greece

Dimitris Sacharidis
dimitris.sacharidis@ulb.be
Université Libre de Bruxelles
Brussels, Belgium

Giorgos Giannopoulos
giann@athenarc.gr
Athena Research Center
Athens, Greece

Dimitris Gunopulos
dg@di.uoa.gr
National and Kapodistrian University
of Athens
Athens, Greece

Theodore Dalamagas
dalamag@athenarc.gr
Athena Research Center
Athens, Greece

Abstract

The rapid integration of machine learning (ML) into critical decision-making systems has heightened concerns over fairness, particularly regarding spatial biases often tied to sensitive socioeconomic factors. In response, we propose a model-agnostic post-processing method for spatial bias mitigation that operates without access to the original training data. Our approach formulates an optimization problem that minimizes a fairness measure robust to gerrymandering, subject to a constraint specifying the allowable deviation from the original model's performance ensuring spatial fairness while preserving accuracy. This measure has a 0–1 scale, offering an intuitive way to quantify spatial bias. Comprehensive evaluations on real-world datasets show that our framework effectively reduces spatial bias and achieves fairer outcomes with minimal performance loss, outperforming other state-of-the-art post-processing methods. This work advances spatial fairness methodologies, offering practitioners an efficient, interpretable, and adaptable post-processing solution to mitigate location-based discrimination in ML applications.

CCS Concepts

• **Computing methodologies → Machine learning**; • **Information systems → Information systems applications**.

Keywords

Spatial Fairness, Post-Processing, Bias Mitigation.

ACM Reference Format:
Dimitris Kyriakopoulos, Dimitris Sacharidis, Giorgos Giannopoulos, Dimitris Gunopulos, and Theodore Dalamagas. 2025. PROMIS: A Post-Processing Framework for Mitigating Spatial Bias. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3748636.3762760

SIGSPATIAL '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3762760

1 Introduction

The rapid expansion of machine learning (ML) into real-world decision-making has heightened concerns about fairness and bias. As ML models influence critical areas such as employment, credit scoring, and criminal justice, preventing them from reinforcing discrimination has become essential. Algorithmic decision-making can disadvantage protected groups if it inherits historical biases from training data [3]. Fairness in ML is defined in various ways, but all definitions require that a model's performance remains unaffected by protected attributes such as age, gender, or race [12]. Among the most well-known fairness criteria are statistical parity and equal opportunity: the former ensures that positive outcomes occur at equal rates across groups, while the latter requires equal true positive rates.

Spatial fairness specifically addresses scenarios where *location* serves as a protected attribute. Because location often correlates with socioeconomic factors, racial demographics, and historical segregation policies [14], it can raise concerns about location-based discrimination. Location also plays a pivotal role in a range of real-world contexts, from city crime to broader tasks involving spatial data selection and summarization [8]. Model predictions may systematically favor certain locations over others, potentially causing unjust outcomes for particular regions. For instance, spatial bias in crime prediction models may lead to disproportionate over-policing of certain neighborhoods, while in loan approval systems, it can result in systematic denial of credit to residents of specific geographic areas, reinforcing historical inequalities. Despite these social implications, spatial fairness is often overlooked.

Unlike categorical attributes such as gender, location is continuous. Standard approaches to handle continuous protected attributes, such as age or income, typically involve discretizing the domain into predefined groups and comparing algorithmic outcomes across these groups. This approach, however, is not appropriate for spatial fairness. Introducing adhoc boundaries or defining partitionings in a territory is susceptible to *gerrymandering*, where outcomes can be manipulated to mask or exaggerate fairness violations [19].

To address this challenge, [20] introduces a spatial fairness definition that measures the expected deviation of model performance among partitions across all possible partitionings of a territory. As observed in [15], however, this definition does not function as expected when there exist sparse regions with few observations in

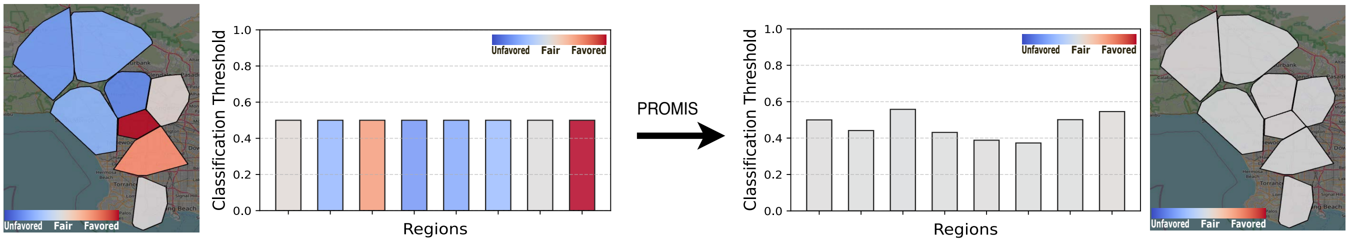


Figure 1: Overview of the PROMIS framework for spatial bias mitigation. Audit regions are colored to indicate bias—blue for unfavored, red for favored, grey for fair treatment—based on true positive rates (TPRs). Left: The initial classifier applies a global decision threshold of 0.5 across all regions, resulting in visible disparities. Right: PROMIS assigns region-specific thresholds, leading to more equitable TPRs. Bar plots reflect this shift, with regions appearing in grey tones after applying PROMIS, illustrating the mitigation of spatial bias.

them. Instead, [15] defines spatial fairness based on the likelihood of the observations coming from a spatially fair model. Although promising, this definition is only used to audit a model for spatial fairness and not to mitigate the spatial bias of an unfair model. The work in [2] proposes a post-processing method suitable for the definition in [15], which however is heuristic and only applies to statistical parity notions of spatial fairness.

This work proposes a post-processing optimization framework that mitigates spatial bias while preserving predictive performance. It builds upon the threshold-based adjustment approach for equal opportunity of [9], and the robust spatial fairness definition of [15]. Our approach, termed PROMIS, extends threshold-based fairness adjustments to the spatial domain by formulating an optimization problem that minimizes the spatial bias index (SBI) measuring the expected spatial bias in a region of a territory. SBI being normalized in [0-1], comprises an intuitive measure for understanding and comparing the extent of bias within the examined areas. Unlike heuristic-based correction methods, e.g., [2], our solution leverages mathematical optimization to derive fairness adjustments that are globally optimal, interpretable, and computationally efficient. Moreover, unlike white-box methods, e.g., [20], PROMIS works with any classification model.

Consider a territory of interest, which includes instances with spatial information, e.g., coordinates, zip code. Further, consider a binary classification model, which produces decisions for the instances. Finally, assume a set of *audit regions* within this territory, for which the user wishes to ensure spatial fairness. Spatial fairness is tied to the performance of the classification model and can be interpreted in various ways, which map to different well-known fairness definitions. For example, if spatial fairness is translated to demanding that true positive rate (TPR) of the classifier's decisions is equal across all audit regions, then we have the equivalent of *equal opportunity* [9] for the spatial domain. Alternatively, if we demand that the positive rate (PR) is equal across all audit regions, then we have the equivalent of *statistical parity* [7] for the spatial domain.

To provide a high-level overview of PROMIS, we focus on equal opportunity and employ a simple example in Fig. 1. The color palette in the figure represents spatial bias: blue shades indicate unfavored behavior, red shades signify favored behavior, and grey denotes fairness. On the left side of the figure, the audit regions are shown in varying colors, highlighting areas where the model exhibits favored or unfavored behavior.

PROMIS mitigates detected unfairness by: (a) determining the optimal number of instances to have their decisions flipped, ensuring that all audit regions achieve similar true positive rates (TPR), and (b) assigning region-specific classification thresholds, aligning the classifier's updated predictions with the identified number of flips from the previous step. This approach enables the bias mitigation process to generalise to new instances within the same audit regions. This process is visualized in the two bar charts in Fig. 1, before and after applying PROMIS. Observe that PROMIS results in changing the classification thresholds per region, which were originally all at 0.5. As a result, the observed spatial bias in these regions is minimized.

Contributions. In this paper, we propose a post-processing, model agnostic framework for enhancing spatial fairness through optimization techniques, addressing key limitations in efficiency and effectiveness. Our contributions are:

- **Intuitive Fairness Metric**: We define and use the Spatial Bias Index (SBI)—an intuitive score ranging from 0 (fair) to 1 (unfair)—that quantifies spatial bias across space.
- **Optimization-Based Mitigation of Spatial Bias**: We propose an optimization-driven approach to minimize spatial bias index. Our method efficiently achieves higher-quality fairness solutions than existing techniques under model performance constraints.
- **Handling Overlapping and Non-overlapping Regions**: We provide formulations for both overlapping and non-overlapping spatial regions, enabling more adaptable and accurate solutions for spatial fairness across complex spatial layouts.
- **Comprehensive Evaluation**: We validate our framework on real and semi-synthetic datasets, evaluating its effectiveness in terms of fairness, accuracy, and computational performance. Our experiments demonstrate the robustness of PROMIS in effectively mitigating spatial bias while maintaining high predictive performance, and outperforming existing techniques.

Overview. The remainder of this paper is organized as follows. Section 2 reviews related work on fairness in machine learning, with an emphasis on spatial fairness. Section 3 formalizes the problem and introduces the spatial bias index. Section 4 presents the PROMIS framework, detailing the optimization formulation, the proposed approximation, the threshold adjustment and inference process, and the batch processing mode for statistical parity. Section 5 reports experimental results comparing PROMIS to existing methods on real and synthetic datasets. Section 6 concludes the paper and outlines directions for future work.

2 Related Work

Algorithmic Fairness. A variety of fairness definitions have been established the last years which distinguish between *individual* and *group* fairness [12, 18]. Efforts to mitigate bias in machine learning models are categorized as *pre-processing*, *in-processing*, or *post-processing*, depending on whether they manipulate training data, intervene during model training, or adjust model predictions, respectively [12].

Spatial Fairness. The majority of proposed definitions and methods are for *group fairness*, where groups refer to spatial regions.

[20] presents a group fairness definition that quantifies the average disparity in model performance across multiple partitionings of the space. This approach ensures robustness against the modifiable areal unit problem and gerrymandering. To promote spatial fairness, the authors introduce an in-processing method that continues training the model. At each epoch, a new partitioning is selected, and the learning rate is adjusted to enhance fairness with respect to that partitioning by emphasizing underperforming regions with respect to a task-specific performance metric. Their method is limited to stochastic training processes (e.g., SGD-based training of neural networks), restricting the range of supported classification models. Additionally, it requires access to the model's internal parameters, such as the learning rate and number of epochs, making it a white-box approach and further constraining its applicability.

To reduce the computational cost of this method, [10] proposes grouping similar spatial regions. [4] employs a similar in-processing approach. It employs a meta-learning framework that jointly learns a fair predictor and a "referee" component guiding training-time updates. Similarly, that work presupposes access to the training process and modifies how a model is learned end-to-end so that it can adapt its fairness behavior to new (unseen) locations during training. PROMIS, in contrast, is a post-processing method that operates without requiring access to the model's original training procedure or additional labels for newly encountered locations. Moreover, the fairness criterion minimized by PROMIS is not tied to the prediction task's utility function, and is therefore incompatible with the optimization strategy based on learning rate adjustments for performance degradation. Whereas the in-processing method of [20] adjusts the training procedure to equalize utility-based metrics across regions, PROMIS explicitly minimizes spatial bias, decoupled from the training process.

In another line of work, [15] proposes an audit mechanism based on spatial scan statistics [11], which calculates the unfairness score for a region by considering the maximum likelihoods of two hypothesis: the first assumes equality inside and outside a region while the second states that model's performance deviates inside and outside a region. [2] introduces a heuristic-based post-processing method, SpatialFlip, to enhance fairness according to this likelihood-based definition. It operates in an incremental manner, iteratively identifying and modifying individual predictions to mitigate spatial bias. However, their method only applies for statistical parity. Similar to [15], PROMIS defines fairness using these two hypotheses but does not perform a statistical test. Instead, it calculates a spatial bias index to quantify unfairness in the space. Unlike [2], PROMIS adopts a principled approach based on constrained optimization, making it applicable to both statistical parity and equal opportunity.

[17] proposes a pre-processing method for calibration fairness. It determines a bias-aware KD-tree partition of the data so that an individual's chance of a positive outcome does not depend unfairly on their location. However, this approach only guarantees spatial fairness for a single partitioning of the space. In contrast, PROMIS seeks to ensure fairness for any region in the space.

Another line of work investigates *individual fairness*, which requires that similar individuals are treated similarly by the model. [13] defines individual similarity based on their features, and measures fairness for location-based recommendations, but does not introduce a mechanism to ensure fairness. In contrast, [16] defines similar individuals as those that either have the same distance from a reference point (e.g., a store or service location) or which have a small distance between them. They introduce fair polynomials as a mechanism for smoothing classification outcomes across space.

Essentially, [16] enforces similarity in model predictions for nearby individuals. However, this approach may not be suitable for certain applications—for example, loan decisions should be influenced by non-spatial attributes. In contrast, PROMIS guarantees that model performance remains location-independent, e.g., ensuring a consistent false negative rate for loan applications regardless of geographical context.

Alternative methods for evaluating spatial bias include the use of discrepancy [1, 6].

3 Problem Definition

We present a spatial fairness framework based on the work of [15], which translates the definitions of statistical parity and equal opportunity [9] into the spatial domain. We first discuss the case of equal opportunity, which requires that the *true positive rate* (TPR)—aka sensitivity or recall—is the same across all protected groups. When location is the protected attribute, equal opportunity translates into the following requirement: for any region r the *population* TPR θ_r inside r should be equal to the *population* TPR $\theta_{\overline{r}}$ outside r. The naive approach of assessing spatial fairness by comparing *empirical* TPRs (i.e., ρ_r, $\rho_{\overline{r}}$), similar to [20], may lead to wrong conclusions as argued in [15]. The reason is that, with limited observations—such as in small or sparse regions—the empirical TPR can deviate significantly from the population TPR.

Instead, [15] proposes to compare the likelihoods of two hypotheses. The *fairness hypothesis* $\mathcal{H}^0$ postulates that the population TPR is the same inside and outside any region (inside = outside, or $\theta_r = \theta_{\overline{r}}$ for all r). The *unfairness hypothesis* $\mathcal{H}^1$ postulates that, in general, the population TPRs differ (inside ≠ outside, or $\theta_r \neq \theta_{\overline{r}}$ for a region r). We want to investigate which hypothesis is more likely given with the observed data.

Consider a set of observations in the *territory* of interest, and let n, p denote the number of actual and true positives. Furthermore, for a region r, let n_r, p_r denote the number of actual and true positives inside r. The empirical TPR ρ_r (resp. $\rho_{\overline{r}}$) observed inside (resp. outside) r is the result of a stochastic process that labels each

Table 1: Notation

Symbol	Meaning
r, R	an audit region, a set of audit regions
n	num. of actual positives in the territory
p	num. of true positives in the territory
$n_r, n_{\overline{r}} = n - n_r$	num. of actual positives inside, outside region r
$p_r, p_{\overline{r}} = p - p_r$	num. of true positives inside, outside region r
$\rho_r = \frac{p_r}{n_r}, \rho_{\overline{r}} = \frac{p_{\overline{r}}}{n_{\overline{r}}}$	empirical true positive rate inside, outside region r
$\theta_r, \theta_{\overline{r}}$	population true positive rate inside, outside region r

actual positive point inside (resp. outside) r as true positive with probability equal to the population TPR θ_r (resp. $\theta_{\overline{r}}$).

On the one hand, the likelihood of the fairness hypothesis $\mathcal{H}^0$ for region r is:

$$\theta_r^{p_r}(1-\theta_r)^{n_r-p_r}\ \theta_r^{p_{\overline{r}}}(1-\theta_r)^{n_{\overline{r}}-p_{\overline{r}}} = \theta_r^{p}(1-\theta_r)^{n-p},$$

which is maximized when the population TPR of r equals the empirical TPR for the territory, i.e., $\theta_r = \rho = p/n$. Therefore, the maximum likelihood of $\mathcal{H}^0$ becomes independent of r:

$$L^0 = \rho^p(1-\rho)^{n-p}. \tag{1}$$

On the other hand, the likelihood of the unfairness hypothesis $\mathcal{H}^1$ for r is:

$$\theta_r^{p_r}(1-\theta_r)^{n_r-p_r}\ \theta_{\overline{r}}^{p_{\overline{r}}}(1-\theta_{\overline{r}})^{n_{\overline{r}}-p_{\overline{r}}},$$

which is maximized when the two parameters of $\mathcal{H}^1$ equal their empirical counterparts, i.e., $\theta_r = \rho_r = p_r/n_r$ and $\theta_{\overline{r}} = \rho_{\overline{r}} = p_{\overline{r}}/n_{\overline{r}}$ [11]. Therefore, the maximum likelihood of $\mathcal{H}^1$ for region r is:

$$L_r^1 = \rho_r^{p_r}(1-\rho_r)^{n_r-p_r}\ \rho_{\overline{r}}^{p_{\overline{r}}}(1-\rho_{\overline{r}})^{n_{\overline{r}}-p_{\overline{r}}}. \tag{2}$$

Having defined the max likelihoods, let us make some observations. Given a region r, we observe *absolute spatial fairness* when the max likelihood of the unfairness hypothesis takes its lowest possible value, i.e., $L_r^1 = L^0$. In that case the empirical TPRs inside and outside r are equal to the overall empirical TPR, i.e., $\rho_r = \rho_{\overline{r}} = \rho$. In contrast, we observe *absolute spatial unfairness* when the max likelihood of the unfairness hypothesis takes its highest possible value for a region r. This occurs when all true positives of the territory and no false negatives appear either inside r or outside r, i.e., when $n_r = p_r = p$ or $n_{\overline{r}} = p_{\overline{r}} = p$.

Therefore, the max likelihood L_r^1 of the unfairness hypothesis for any region r ranges between L^0 (absolute fairness) and 1 (absolute unfairness). Taking the logarithm of these likelihoods, we can quantify the unfairness observed in r.

Definition 3.1. The *local spatial unfairness index (SBI)* for region r is:

$$\mathsf{SBI}_r = 1 - \frac{\log L_r^1}{\log L^0}.$$

The value of SBI_r ranges from 0 (absolute fairness) to 1 (absolute unfairness), with lower values being preferable. The overall unfairness across the territory is then quantified as the expected SBI for any region.

Definition 3.2. The *spatial unfairness index (SBI)* for the territory of interest is:

$$\mathsf{SBI} = \mathbb{E}_r\left[\mathsf{SBI}_r\right] = 1 - \mathbb{E}_r\left[\frac{\log L_r^1}{\log L^0}\right].$$

Enumerating all regions of a territory may be prohibitive. Instead, in practice we consider a set R of *audit regions*, and estimate $\mathsf{SBI} = \frac{1}{|R|}\sum_{r\in R}\mathsf{SBI}_r$.

To illustrate the advantage of assessing spatial fairness using likelihoods instead of TPRs directly, consider the following. Recall that the spatial unfairness metric in [20], considers a collection of partitions $\mathcal{P}$, where each partition $P \in \mathcal{P}$ is a set of non-overlapping regions. For each region in each partition, [20] computes the deviation of the model's performance in that region compared to the model's overall performance. Then, it reports the mean deviation across all regions in all partitions. Let $R = \cup_{P\in\mathcal{P}}P$ be the set of all regions. Therefore, assuming performance is measured by TPR, [20] computes the mean deviation of TPR, denoted as MeanDev:

$$\mathsf{MeanDev} = \frac{1}{|R|}\sum_{r\in R}|\rho_r - \rho|.$$

This metric lacks robustness in regions with few observations, where the observed ρ_r may be close to extreme values. Conversely, in regions with a large number of observations, small deviations of ρ_r from the overall TPR may not be adequately captured. Therefore, MeanDev may report unfairness where it does not exist and fail to report unfairness where it exists.

Let us now introduce a toy example, depicted in Table 2, of a territory with $n = 100$ actual, and $p = 60$ true positive observations; this translates into an empirical TPR of $\rho = 0.6$. We will consider several hypothetical regions with varying numbers of observed actual and true positives. For each region r, we will compare its local SBI with MeanDev computed over the partition $\{r, \overline{r}\}$.

For regions r_1 and r_2, we observe absolute fairness as the TPRs inside and outside are equal and coincide with the overall TPR ρ. For both regions, MeanDev agrees with SBI in that there is no unfairness. Conversely, for regions r_3 and r_4, we observe absolute unfairness; r_3 gathers all false negatives of the territory and no true positives, while the converse holds for r_4. For both regions, MeanDev agrees with SBI in that they exhibit the most extreme case of unfairness.

For regions r_5, r_6, r_7 the MeanDev is the same. Note that the true positive rate and the number of observations increases going from r_5 to r_7. Our intuition is that region r_7, with a large number of observations inside ($n_{r_5} = 60$) and outside ($n_{\overline{r}_5} = 40$), and for which there are no false negatives outside ($\rho_{\overline{r}_7} = 1$), should give a stronger indication of unfairness compared to r_5 that contains fewer observations $n_{r_7} = 10$ and has no false positives inside ($\rho_{\overline{r}_5} = 0$). It is this intuition that SBI captures, assigning a higher unfairness index to r_7 than to r_5.

Table 2: Spatial unfairness metrics for seven regions in a toy-example territory with $n = 100$ actual positives and $p = 60$ true positives ($\rho = 0.6$).

	n_r	p_r	ρ_r	$\rho_{\overline{r}}$	MeanDev	SBI	ASBI
r_1	20	12	0.6	0.6	0	0	0
r_2	50	30	0.6	0.6	0	0	0
r_3	40	0	0	1	0.5	1	4.90
r_4	60	60	1	0	0.5	1	4.90
r_5	10	0	0	0.67	0.33	0.15	2.00
r_6	40	8	0.2	0.87	0.33	0.35	3.27
r_7	60	20	0.33	1	0.33	0.43	3.27

A final note concerns the definition of statistical parity into the spatial domain. Recall that it requires the *positive rate* (PR) to remain constant across protected groups. In the spatial domain, this means that the population PR inside and outside any region should be the same. This notion of spatial fairness can be captured by the same framework and equations simply by substituting the meaning of the symbols: n denotes the number of observations, p the number of positives, and ρ the positive rate.

4 The PROMIS Method

Section 4.1 introduces PROMIS, which is detailed in Sections 4.2 and 4.3. Then Section 4.4 describes inference using the adjusted model, and Section 4.5 presents a special case for statistical parity.

4.1 Overview

PROMIS is a post-processing, model-agnostic framework for spatial fairness on binary classification outputs. It supports two widely adopted fairness definitions: statistical parity and equal opportunity.

The framework requires access to the model's predictions and prediction probabilities, and takes as input (1) a set of audit regions R, (2) a set of observations Z, and (3) the model's default decision boundary τ. Each observation is denoted as $z_i = (R_i, prob_i, y_i) \in Z$, where $R_i \in R$ indicates the set of regions to which the i-th observation belongs, $prob_i$ is the probability produced by the model, and y_i is the true class label. The model's prediction is given by $\hat{y}_i = \mathbb{I}[prob_i > \tau]$. Given R, Z and τ PROMIS outputs for each region $r \in R$ an adjusted classification threshold $\hat{\tau}_r$ and a tie-breaking probability π_r. These values are then used at inference time to assign spatially fair predictions to new instances. While we detail the case of equal opportunity, note that the same PROMIS formulation can be directly applied to statistical parity by replacing the true positive rate (TPR) with the positive rate (PR) throughout.

The goal of PROMIS is to find a small adjustment (i.e., the derived $\hat{\tau}_r, \pi_r$) to the model so as to minimize the spatial bias without hurting its predictive performance. PROMIS operates in two steps. First, it solves a constrained optimization problem to minimize the SBI for the observed instances. Then, PROMIS generalizes this solution to apply for new instances. Since the SBI metric is a complex, non-linear function, we introduce an alternative, simpler linear function, which is easier to optimize.

4.2 Step 1: Constrained Optimization

In the first step of PROMIS, the goal is to derive for each audit region $r \in R$ an adjustment Δp_r to its observed true positive rate so as to mitigate spatial bias while not hurting predictive accuracy. Based on the set of regions R, we distinguish the case where the audit regions do not overlap, which we call *no-overlap*, from the opposite, which we call *with-overlap*.

In the *no-overlap* case, we define the following optimization problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{|R|}\sum_{r\in R}\mathsf{SBI}'_r && \text{(O1a)}\\ &\text{subject to} && \rho-\epsilon \le \rho' \le \rho+\epsilon, && \text{(C1)}\\ &&& \sum_{r\in R}|\Delta p_r| \le B, && \text{(C2a)}\\ &&& -p_r \le \Delta p_r \le n_r - p_r \ \text{ for all } r\in R, && \text{(C3a)} \end{aligned}$$

which finds for each region r an adjustment $\Delta p_r = p'_r - p_r$ to the number of observed true positives p_r so that the resulting SBI$'$ is minimized (objective O1a), under the constraint that the resulting TPR $\rho' = (\sum_{r\in R} p'_r)/n$ does not deviate much from the observed TPR (constraint C1), that the total adjustment does not exceed a maximum allowed budget B (constraint C2a), and that the adjustments are possible (constraint C3a); note that we denote quantities after adjustment with a prime.

For the *with-overlap* case, we introduce for each observation $z_i \in Z$ a variable Δp_i that denotes the adjustment to its predicted label: when $\hat{y}_i = 0$, $\Delta p_i = 1$ suggests a flip to the positive class, and conversely when $\hat{y}_i = 1$, $\Delta p_i = -1$ suggests a flip to the negative class. Then, the optimization problem remains the same except the last two constraints that change as follows:

$$\sum_{z_i\in Z}|\Delta p_i| \le B, \qquad \text{(C2b)}$$

$$0 \le \hat{y}_i + \Delta p_i \le 1 \ \text{ for all } z_i \in Z, \qquad \text{(C3b)}$$

which reformulates the budget constraint (C2b), and ensures that the adjustments are possible (constraint C3b).

From this solution, we compute an adjustment per region by summing up the adjustments for the observations inside the region: $\Delta p_r = \sum_{z_i\in r}\Delta p_i$. Therefore, we obtain an output analogous to that in the *no-overlap* case.

Note that in both optimization problems, we chose to let the variables (Δp_r or Δp_i) be continuous to avoid solving hard integer programming problems, and because the adjustments are only derived to guide the changes at the classification thresholds. Still, the optimization problems are hard as the objective SBI is a non-linear function of the variables. Therefore, we introduce an approximation to the spatial fairness index that is linear.

Recall that spatial fairness exists for region r when the TPR inside and outside r is equal. We would like to use the difference of TPRs as a proxy for SBI$_r$ and avoid the problem with having few observations inside or outside r. We leverage the variance normalization idea from [5]. Specifically, the difference of the empirical TPRs, $\rho_r - \rho_{\overline{r}}$, has variance proportional to $\frac{1}{n_r} + \frac{1}{n_{\overline{r}}}$. Therefore, we define the *approximate local SBI* for r, denoted as ASBI$_r$, as:

$$\mathsf{ASBI}_r = \frac{\sqrt{n_r \cdot n_{\overline{r}}}}{\sqrt{n_r + n_{\overline{r}}}} \cdot |\rho_r - \rho_{\overline{r}}|,$$

and naturally define the global *approximate SBI* as

$$\mathsf{ASBI} = \frac{1}{|R|}\sum_{r\in R}\mathsf{ASBI}_r.$$

Table 2 also depicts the approximate SBI next to the actual SBI for the illustrative scenario introduced earlier. Observe that ASBI distinguishes the level of unfairness between r_5 and r_7 similar to SBI, and in contrast to MeanDev. However, ASBI, being only an approximation, cannot discern between the level of unfairness in r_6 and r_7.

Time Complexity. The approximate optimization problem in PROMIS-A is formulated as a linear program with continuous variables and linear constraints. In the *no-overlap* case, the number of variables and constraints is $O(|R|)$. In the *with-overlap* case, we introduce one variable per instance and additional constraints per region, yielding a total complexity of $O(|Z| + |R|)$, typically simplified to $O(|Z|)$ since $|Z| \gg |R|$ in practice. These problems are solvable in polynomial time using standard LP solvers. Empirically, the optimization remains efficient at scale: in our largest experiment (3,000 overlapping regions; 200,000+ instances), PROMIS-A finished in under 28 seconds (including the next step).

To distinguish between the two objectives, we call PROMIS the method based on SBI, and PROMIS-A the one that uses ASBI.

4.3 Step 2: Classification Threshold Adjustment

The next step is to derive adjusted classification thresholds at the regional level based on the outputs Δp_r of the previous step. We aim to find a classification threshold that flips $|\Delta p_r|$ observations, positives to negatives if $\Delta p_r < 0$, or negatives to positives if $\Delta p_r > 0$. In many cases, the optimization may yield non-integer Δp_r. We write $\Delta p_r^{int} = \lfloor |\Delta p_r| \rfloor$ for the integer part, and $\Delta p_r^{frac} = |\Delta p_r| - \Delta p_r^{int}$, for the fractional part of Δp_r.

Among all observations inside a region r, we select those with negative prediction and sort them in descending order of their predicted probability for $\Delta p_r > 0$, or we select those with positive prediction and sort them in ascending order of their predicted probability for $\Delta p_r < 0$. In either case, the Δp_r^{int}-th prediction probability in that sorted list is called the *first pivot* probability τ_r^1, and the $(\Delta p_r^{int} + 1)$-th the *second pivot* τ_r^2.

If the pivots differ, we set the classification threshold for the region to a midpoint value along the way from τ_r^1 to τ_r^2 as: $\hat{\tau}_r = \tau_r^1 + \frac{1+\Delta p_r^{frac}}{2} \cdot (\tau_r^2 - \tau_r^1)$.

It is likely, however, that the probability pivots are the same. In that case, we set $\hat{\tau}_r = \tau_r^1$, and also assign a *tie-breaking probability* to that region. At inference time, if an instance has prediction probability equal to τ_r, we predict the positive class with probability equal to the tie-breaking probability $\pi_r = \frac{|\Delta p_r| - (\#\{prob_i > \hat{\tau}_r\} - \#\{prob_i > \tau_r\})}{\#\{prob_i = \tau_r\}}$, where the numerator counts the number of instances with probability equal to $\hat{\tau}_r$ that must be flipped, and the denominator counts the total number of instances with probability equal to $\hat{\tau}_r$, assuming $\Delta p_r > 0$. For $\Delta p_r < 0$, we replace the '>' operator with '<' in the equation. In this way, we guarantee that, in expectation, we flip exactly $|\Delta p_r|$ instances in comparison to the original classifier.

At the end of this step, we have computed the adjusted classification thresholds for each region, and if necessary we also assign a tie-breaking probability.

Time Complexity. To determine adjusted thresholds, we first sort all instances once by their predicted probabilities in $O(|Z| \log |Z|)$ time. For each region, we then compute the pivot(s) for threshold adjustment, which takes $O(1)$ per region. The overall complexity of the threshold adjustment step across all regions is therefore $O(|Z| \log |Z| + |R|)$, typically simplified to $O(|Z| \log |Z|)$.

4.4 Inference

Inference is assigning a class to a new instance x. If we are in the *no-overlaps* setting, x falls inside a single region r. Thus, given the model's prediction probability $prob_x$, this step looks up the derived classification threshold $\hat{\tau}_r$, and possibly applies the tie-braking probability, to assign a class $\hat{y}' = \mathbb{I}[prob_x > \hat{\tau}_r]$ to x.

In the *with-overlaps* setting, x receives a class $\hat{y}'_r$ for each region r it falls in. We apply a simple weighted voting scheme. Each region contributes a vote with a weight w_r that is inversely proportional to its size (the number of observations that fall in it), i.e., $w_r = 1/|z_i \in Z : z_i \in r|$. The idea is to give small regions a strong voice. Therefore, the class of the instance is computed as $\hat{y}' = \mathbb{I}[\sum_{r \in R : x \in r} w_r \cdot \hat{y}'_r > 0.5]$, where $\mathbb{I}[]$ is 1 if its argument is true.

Time Complexity. In the *no-overlap* case, inference is a constant-time lookup, i.e., $O(1)$ per instance. In the *with-overlap* case, voting evaluates $O(|R_x|)$ regional predictions, where $|R_x|$ is the number of regions containing x; in practice $|R_x|$ is small, so inference is efficient.

4.5 Batch Processing for Statistical Parity

PROMIS variants make adjustments to a model so that at inference time its predictions exhibit no spatial bias. In the case of statistical parity, however, it is possible to modify the method to post-process in batch a collection of observations so that they are spatial bias-free, similar to how SpatialFlip [2] operates.

To achieve this, PROMIS approaches only solves the optimization problem, resulting in adjustments Δp_r at the region (for regions with *no-overlap*) or Δp_i at the individual (for the *with-overlap* case) level. Then these adjustments are applied directly on the observations: Δp_r^{int} observations are flipped in each region r plus one individual with probability Δp_r^{frac} for *no-overlap*, or each observation z_i is flipped with probability Δp_i for *with-overlap*.

5 Experiments

This section presents the experimental evaluation of the proposed framework. Section 5.1 presents the evaluation setting, including datasets, evaluation measures and model configurations. In Section 5.2, we compare the PROMIS variants with two state of the art methods on spatial mitigation, *Fairness by Where*, denoted as *FairWhere* [20] and *SpatialFlip* [2]. Section 5.3 presents additional analysis that provides insights on the behavior of our proposed method.

5.1 Evaluation Setting

Our evaluation setting considers a binary classification model that produces decisions on instances that include location information and lie within a territory of interest. We are provided with a set of audit regions, for which we want to ensure spatial fairness. In the subsequent experiments, we perform audits for spatial fairness to evaluate and compare the effectiveness of both PROMIS variants. We next elaborate on the specifics of the setting.

Datasets. Two real-world and a semi-synthetic datasets are considered in our evaluation: CRIME[1], LAR[2] and SYNTH.

CRIME is a real-world, publicly available dataset of crime incidents from Los Angeles (2010–2019), derived from police reports. Each record includes details such as crime type, location, and victim demographics. The top 10 most frequently occurring crime categories in the dataset are identified and labeled as serious crimes (label = 1), while all other crime types are categorized as non-serious (label = 0). After preprocessing to remove records with missing coordinates, the dataset was one-hot encoded for categorical features leading to 49 features in total. We end up with 713,204, with 204,204 of them corresponding to serious crimes (0.29 positive rate). To ensure a balanced representation of serious and non-serious crimes, we apply stratified sampling based on the binary crime label and split the dataset into training (60%), validation (20%), and test (20%) sets.

The second real-world dataset focuses on mortgage loan applications. Combining the loan/application register (LAR) records for Bank of America in 2021, with information from the U.S. Census Bureau's Gazetteer files, we ended up with a LAR dataset consisting of 206,418 applications, of which 127,286 were approved, corresponding to an approval rate of 0.62. This dataset is used to assess performance on statistical parity i.e. minimize the deviation of the chance of being granted a mortgage loan.

Finally, we create a semi-synthetic dataset termed SYNTH by processing the test set of the CRIME dataset. We choose this subset of the whole CRIME dataset, so that we can compare our framework to the SpatialFlip method, which is computationally expensive. The spatial coordinates remain unchanged, while we greedily generate predictions to introduce high spatial bias. To achieve this, we first generate predictions from a binomial distribution with a positive rate of 0.8, ensuring fair predictions across all audit regions. Then, for each of the three types of audit regions, we select non-overlapping region pairs and systematically modify the predictions. Within each pair, we perform an equal number of positive-to-negative and negative-to-positive flips to maintain the overall positive rate while introducing spatial bias. For each selected pair, we apply the maximum possible number of flips, aiming to make predictions in one region entirely positive and in the other entirely negative. The flipping process follows a linear strategy, starting with the pair that allows for the most flips and proceeding in descending order. This ensures that we create a scenario with extreme spatial unfairness, simulating an unfair-by-design world.

Classification models. For the CRIME dataset, we first train a Neural Network classifier on the train set (*DNN*), so as to provide a fair comparison with FairWhere [20], which is restricted to neural network-based prediction models. The model consists of two hidden layers with 100 and 50 neurons, both using ELU activation, batch normalization, and dropout (0.1). Training is performed using the Adam optimizer with a learning rate of 0.0001 and a batch size of 4096. We use F1-score loss and the model is trained for 100 epochs (converged) achieving 48% F1-score. To provide additional analysis of the behavior of both versions of PROMIS, we train an XGBoost classifier (*XGB*), with the binary logistic objective, using log loss as the evaluation metric. For this classifier the accuracy achieved for both validation and test sets was 73%.

Partitioning Strategies. To evaluate model performance under different spatial partitioning scenarios, we create both *no-overlap* and *with-overlap* audit regions. For *no-overlap* regions, we employ k-means clustering based on the spatial coordinates of the instances. The resulting partitioning approach is referred to as *Cluster*. Specifically, for the CRIME dataset, we generate 8 distinct regions, whereas for the LAR dataset, we create 100 such regions. For the creation of *with-overlap* regions, we leverage k-means to determine the initial region centers. Given predefined radii, we generate regions by expanding outward from each center, with the radius defining the maximum distance from the center (i.e., the region's boundary). This process yields nested regions, referred to as *Scan*. In particular, for the CRIME dataset, we start with 10 centers and expand using 4 radii, resulting in a total of 40 overlapping regions. Similarly, for the LAR dataset, we use 100 centers and expand with 30 radii, producing 3,000 overlapping regions. Lastly, following the partitioning strategy of [20], we consider *s1* × *s2* grid partitionings called as *Grid*. We set the maximum values for *s1* and *s2* to 5, resulting in 24 equal-sized partitions (excluding the trivial 1 × 1 partitioning). Since the method in [20] exclusively operates on grid structures, we represent both clusters and scan regions as 1 × *s2* grids, where *s2* corresponds to the total number of regions.

Metrics. In our evaluation, we consider three performance dimensions: (a) bias mitigation, (b) maintaining the effectiveness of the original classifier's output and (c) execution runtimes. The fairness measures used in our evaluation are the SBI score, as defined in Section 3 and the MeanDev measure as derived by [20] (also see Section 3). The impact of the mitigation on the overall behavior of the original classifier is measured either through F1-score, when DNN serves as base model, or accuracy when XGB serves as base model.

Experimental Setup. To compare our method with the state of the art, we consider the in-processing method of [20], termed FairWhere, and the post-processing method of [2], termed SpatialFlip.

To evaluate the mitigation solution, we compared different groups of methods per dataset, base model and fairness notion due to the approaches restrictions. For the mitigation of the DNN model all methods were considered for the Statistical Parity definition, while for the Equal Opportunity definition the SpatialFlip was excluded, since its limited to Statistical Parity. Regarding the LAR dataset only PROMIS configurations and SpatialFlip methods were compared, since FairWhere is restricted to neural network-based prediction models. When XBoost serves as base model and considering the equal opportunity fairness notion, only PROMIS methods are applicable.

For the experiment where the DNN serves as the base model, FairWhere continues training on the training set with the goal of mitigating spatial bias. At the start of fairness training, each sampled partitioning is used for 5 epochs before transitioning to the next, iterating over 120 different samples (equivalent to five full enumerations over all 24 partitioning candidates). Subsequently, each epoch samples a new partitioning, continuing for 720 samples

[1] https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/about_data
[2] https://ffiec.cfpb.gov/data-publication/modified-lar/2021

(equivalent to thirty full enumerations). In total, each partitioning is expected to be used for 50 epochs throughout the training process.

For the PROMIS and PROMIS-A we leverage the Gurobi solver for the optimization part. For the PROMIS we set the NonConvex parameter to -1, and for PROMIS-A to 0. The Threads is set to 0. We set WorkLimit to 300 limiting the work unit parameter (approximately one second of computation per unit on a single thread) to ensure practical runtimes for PROMIS method, while PROMIS-A never reaches that limit. The solver returns the best solution found within the preset work limit , which may not guarantee global optimality. To show the impact of increasing the work limit on the performance of PROMIS, we set it to half an hour and six hours in our experiments on the LAR dataset. For both PROMIS configurations we constraint the positive ratio (for statistical parity) or the true positive ratio (for equal opportunity) to maximum 10% shift wrt the ratio of the predictions of the base model.

For both PROMIS implementations and SpatialFlip method we limit the number of flips allowed by trying several increasing budgets constraint. For the experiment, where DNN serves as base model, this budget is computed by the number of different predictions on the validation set between the base DNN and the mitigated by FairWhere model. For the unfair by design experiment the budget respects to number of different predictions between the unfair by design predictions and the original fair predictions to assess whether the methods can effectively mitigate spatial bias within this budget, restoring fairness close to the originally generated fair predictions. For the rest experiments we set freely budgets.

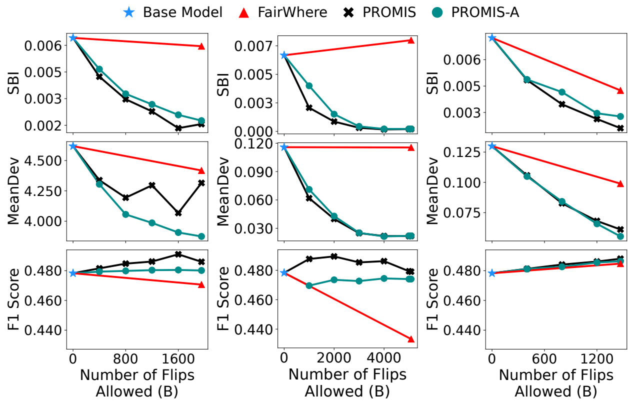
When DNN or XGBoost are used as base models, PROMIS methods apply the thresholds adjustment approach by assessing only the validation set (20%) to mitigate the spatial bias, while for the unfair by design and LAR datasets they mitigate the bias via batch processing by directly flipping predictions based on the optimization solution.

Finally, to audit for spatial fairness 1000 alternative world were created during monte-carlo simulations and the statistical significance level was fixed at 0.005.

All experiments were conducted on a MacBook Air with an M3 chip and 16GB of RAM. Code and data is available online[3].

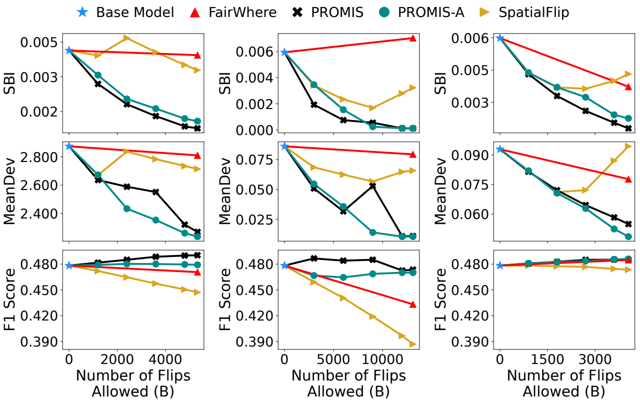
5.2 Comparison with State of the Art

Bias Mitigation for Equal Opportunity. Figure 2 shows a comparison between our approach and FairWhere in mitigating the DNN model for the Equal Opportunity fairness notion, while Figure 3 compares all methods for the Statistical Parity notion. Both PROMIS variants consistently outperform FairWhere in all metrics (SBI, MeanDev, and F1-score) under both fairness notions. Although minimizing MeanDev is not PROMIS's main objective, it still achieves the lowest values for the metric within the maximum budget constraint. PROMIS-A closely follows PROMIS in terms of SBI reduction (Figures 2 and 3). However, in the case of *Scan* under Statistical Parity (Figure 3b), PROMIS-A outperforms PROMIS under one specific budget, because the latter reaches the computational work limit, preventing it from attaining a global optimum.



(a) Grid **(b) Scan** **(c) Cluster**

Figure 2: SBI (Eq. Opp.), MeanDev, F1 vs. number of flips allowed on CRIME-DNN



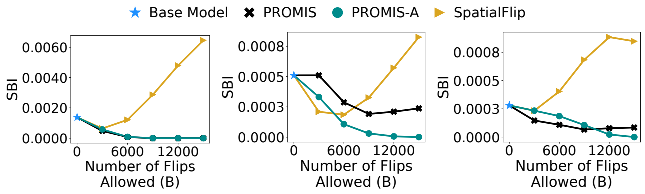
(a) Grid **(b) Scan** **(c) Cluster**

Figure 3: SBI (St. Par.), MeanDev, F1 vs. number of flips allowed on CRIME-DNN

We also observe that while FairWhere improves MeanDev in *Scan* regions, it does not necessarily reduce SBI (Figure 2b and 3b). Conversely, minimizing SBI does not always improve MeanDev, as seen with PROMIS in grid-based audit regions (Figure 2a). Moreover, SpatialFlip proves ineffective at higher budget constraints (Figure 3), often exacerbating its own primary objective (SBI) across all audit region types. Regarding F1-score, our method either improves or roughly maintains upon the base model's performance. In contrast, FairWhere reduces the F1-score for *Grid* and even more drastically for *Scan* regions (Figure 2b and 3b). Meanwhile, SpatialFlip consistently yields the lowest F1-scores among the compared methods (Figure 3).

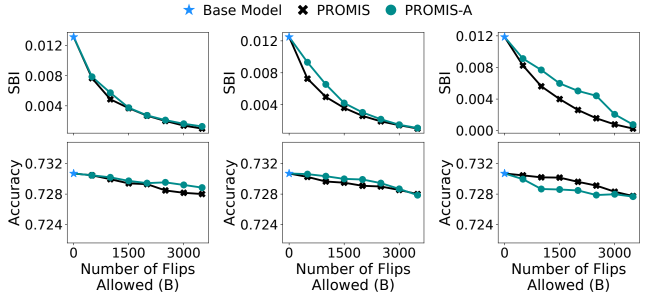
Bias Mitigation for Statistical Parity. Figure 4 shows the results of mitigating the LAR dataset using the two PROMIS methods and SpatialFlip under the Statistical Parity fairness notion. Both variants of PROMIS outperform SpatialFlip, across all region types, early on with respect to flip budget, with the exception of *Scan* regions, where PROMIS start performing better after 7,500 flips. There, PROMIS struggles to find a high-quality solution, likely because this experiment evaluates fairness across 3,000 regions—more than

[3] https://github.com/dimitrskpl/promis-spatial-bias-mitigation



(a) Grid **(b) Scan** **(c) Cluster**

Figure 4: SBI (St. Par.) vs. number of flips allowed on LAR



(a) Grid **(b) Scan** **(c) Cluster**

Figure 5: SBI (Eq. Opp.), Accuracy vs. number of flips allowed on CRIME-XGB

in any other experiment—making it difficult for the solver to identify an optimal solution within the allotted work limit. In contrast, the computationally efficient PROMIS-A provides the best results for *Scan* regions under these conditions. Similarly, at the highest budget constraints for *Cluster* (Figure 4c), PROMIS-A outperforms PROMIS and both of them largely outperform SpatialFlip.

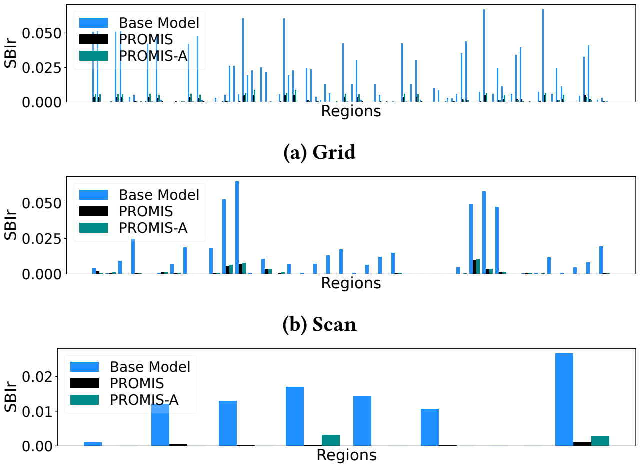
5.3 Analysis of PROMIS Approaches

In this section, we present additional analysis that provides insights on the behavior of our proposed method.

Bias Mitigation with the XGB Model for Equal Opportunity. Figure 5 illustrates the PROMIS methods mitigating the bias on the XGBoost (XGB) model under the Equal Opportunity fairness notion. Both PROMIS and PROMIS-A successfully reduce spatial bias while preserving the original model's performance.

In Figure 6, we observe the SBI_r across individual regions for the base model and for both PROMIS configurations under the maximum budget constraint. The significant reduction in these statistics toward zero represents a balanced model performance across regions. PROMIS-A achieves results comparable to PROMIS, with the latter providing slightly better outcomes in most settings.

Bias Mitigation on Unfair by Design Predictions for Statistical Parity. Figure 8 presents PROMIS strategies' performance in mitigating synthetic biased predictions. Both PROMIS and PROMIS-A successfully remove the introduced bias within the same budget used to create it, as evidenced by the reduction in SBI. Figure 7 provides a more detailed view on how we inserted spatial bias in the SYNTH dataset and the results of the mitigation using the PROMIS-A method. First, a fair world is transformed into an unfair by design world via targeted flips, creating extreme biases in specific regions. Darker red regions in the heatmaps indicate greater favorability (a higher local positive rate relative to the global average), whereas darker blue regions indicate greater unfavorability (a lower local positive rate). Greyish tones suggest more balanced conditions. PROMIS-A effectively reverts these synthetic predictions back to a fair state. Furthermore, the flips executed by PROMIS-A in the final figure precisely counteract those applied to induce bias in the second figure. This is visually represented through color coding: light green denotes a flip from negative to positive, while orange signifies a flip from positive to negative.



(a) Grid

(b) Scan

(c) Cluster

Figure 6: SBI (Eq. Opp.) per region using maximum budget on CRIME-XGB

Runtimes. Our framework is significantly faster than competing methods. Compared to FairWhere, PROMIS approaches are 2× to 7,650× faster, and compared to SpatialFlip, they are 3× to 7,900× faster. Table 3 presents the fit times of all methods across different experimental setups. PROMIS-A is the most computationally efficient method overall, with a maximum recorded runtime of just 31 seconds. Next follows PROMIS with a maximum runtime close to 6 minutes. FairWhere can be computationally intensive, with its longest runtime nearing one hour, whereas SpatialFlip is the most expensive method, taking up to eleven hours for a single experiment.

Table 3: Fit Times (Seconds) Per Experiment

Dataset	Regions	Model	Fair	PROMIS	PROMIS-A	FairWhere	SpatialFlip
Crime	Cluster	DNN	EO	33	1	148	-
Crime	Grid	DNN	EO	367	31	3,556	-
Crime	Scan	DNN	EO	353	8	743	-
Crime	Cluster	DNN	SP	20	0.02	153	158
Crime	Grid	DNN	SP	289	11	3,607	4,102
Crime	Scan	DNN	SP	345	4	673	4,243
LAR	Cluster	-	SP	307	0.05	-	939
LAR	Grid	-	SP	236	16	-	19,306
LAR	Scan	-	SP	240	28	-	37,992
SYNTH	Cluster	-	SP	2	0.02	-	-
SYNTH	Grid	-	SP	263	9	-	-
SYNTH	Scan	-	SP	342	3	-	-
Crime	Cluster	XGB	EO	104	1	-	-
Crime	Grid	XGB	EO	328	31	-	-
Crime	Scan	XGB	EO	341	8	-	-

PROMIS Analysis. Figure 9 illustrates that increasing the work limit for PROMIS generally leads to better solutions, as the solver can explore a larger portion of the search space. Initially, we extended the work limit from 5 to 30 minutes; however, PROMIS still struggled to significantly improve solution quality for *Scan* regions

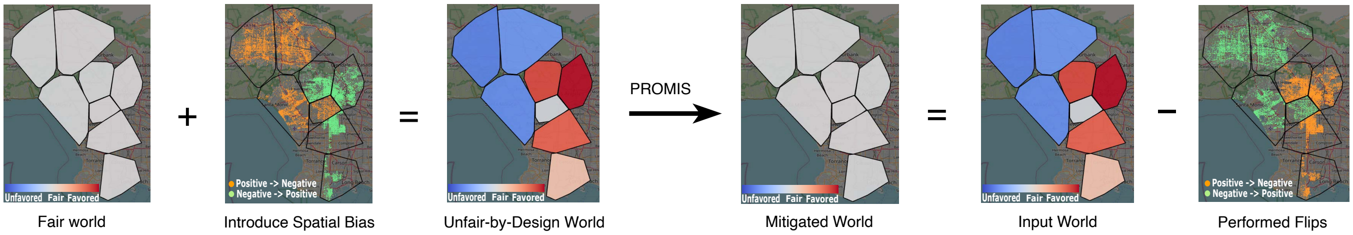
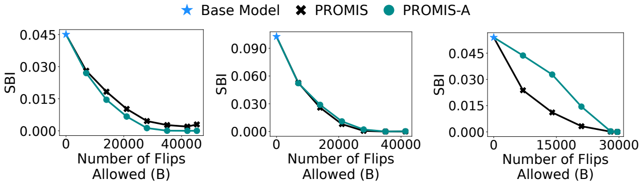
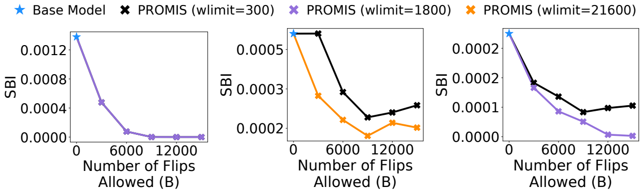


Figure 7: PROMIS-A mitigating an unfair-by-design world. Left: Spatial distribution of outcomes across regions—starting from a fair world, spatial bias is introduced via targeted flips, creating favored (dark red) and unfavored (dark blue) audit regions. Right: PROMIS-A mitigates the bias via relabeling, restoring fairness across regions (grey tones).



(a) Grid **(b) Scan** **(c) Cluster**

Figure 8: SBI (St. Par.) vs. number of flips allowed on SYNTH



(a) Grid **(b) Scan** **(c) Cluster**

Figure 9: SBI (St. Par.) vs. number of flips allowed on LAR with different work limits

when mitigating the LAR dataset under the Statistical Parity fairness notion. Therefore, we further increased the work limit to 6 hours for this setting. The higher work limits yielded substantial improvements in both scan regions (Figure 9b) and clusters (Figure 9c). However, for grid regions (Figure 9a), no notable improvement was observed with the higher work limit. This is because PROMIS already achieved high-quality solutions within the original 5-minute budget. While the SBI does decrease further under the final budget, the reduction is marginal and not visually prominent in the plot.

PROMIS-A Analysis. To demonstrate that PROMIS-A closely approximates PROMIS, we employ a *bias auditing experiment*, using the approach of [15], aiming to detect regions flagged as significantly unfair by either SBI_r or ASBI_r.

In an attempt to show that PROMIS-A closely approximates PROMIS, we replaced SBI_r with ASBI_r. The audit was applied for each combination of model, dataset and fairness notion. To evaluate the quality of the audit results using ASBI_r, we measured accuracy, precision, recall and F1 scores considering as ground truths the audit results of the original approach [15], where for a region 1 means that it was classified as significant and 0 as non-significant.

Table 4 shows the results of our audit experiments. The average accuracy across all experimental settings was 99.7% indicating that PROMIS-A closely replicates PROMIS: for most setups, the significant regions identified by both methods were nearly or entirely the same. In the table, the column labeled "Signif." denotes the number of significant regions detected by the original audit process.

Table 4: Audit results comparing the regions flagged as significantly unfair by PROMIS and its approximation (PROMIS-A), across datasets, audit regions, base models, and fairness notions. The column "Signif." reports the number of significant regions identified using the audit approach of [15].

Dataset	Regions	Model	Fair	Signif.	Acc.	Prec.	Rec.	F1
Crime	Scan	XGB	SP	32	1	1	1	1
Crime	Scan	XGB	EO	31	0.975	1	0.969	0.984
Crime	Cluster	XGB	SP	7	1	1	1	1
Crime	Cluster	XGB	EO	7	1	1	1	1
Crime	Grid	XGB	SP	79	0.996	1	0.988	0.994
Crime	Grid	XGB	EO	77	1	1	1	1
Crime	Scan	DNN	SP	33	1	1	1	1
Crime	Scan	DNN	EO	29	1	1	1	1
Crime	Cluster	DNN	SP	7	1	1	1	1
Crime	Cluster	DNN	EO	7	1	1	1	1
Crime	Grid	DNN	SP	78	1	1	1	1
Crime	Grid	DNN	EO	72	1	1	1	1
LAR	Scan	-	SP	1147	0.997	0.998	0.994	0.996
LAR	Cluster	-	SP	47	0.990	0.978	1	0.989
LAR	Grid	-	SP	108	1	1	1	1
SYNTH	Scan	-	SP	40	1	1	1	1
SYNTH	Cluster	-	SP	8	1	1	1	1
SYNTH	Grid	-	SP	103	0.987	0.970	1	0.985

6 Conclusion

In this paper, we presented PROMIS, a post processing, model agnostic, spatial bias mitigation method for binary classifiers. By building upon a robust, likelihood-based fairness definition and formulating it as optimization problem, PROMIS can effectively enforce spatial fairness constraints without compromising predictive performance or requiring access to training data.

Through extensive experimental evaluation on real-world and synthetic datasets we demonstrated the effectiveness, flexibility, and computational efficiency of PROMIS compared to state-of-the-art methods. Our findings reveal that PROMIS consistently reduces spatial bias as measured by the Spatial Bias Index (SBI), while preserving classification accuracy and outperforming both heuristic and in-processing baselines across different fairness notions.

Our next steps include adapting our method for regression tasks, as well as developing it as a full fledged, user interactive tool.

7 Acknowledgments

This work was partially supported by the EU Horizon program through projects AI-DAPT (GA 101135826), AutoFair (GA 101070568), WiseFood (GA 101181895), and CoDiet (GA 101084642).

References

[1] Deepak Agarwal, Jeff M. Phillips, and Suresh Venkatasubramanian. 2006. The hunting of the bump: on maximizing statistical discrepancy. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*. ACM Press, 1137–1146.

[2] Victor Arroyo and Dimitris Sacharidis. 2023. SpatialFlip: A Postprocessing Method to Improve Spatial Fairness. In *SIGSPATIAL*. doi:10.1145/3589132.3628376

[3] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review'16* (2016).

[4] Weiye Chen, Yiqun Xie, Xiaowei Jia, Erhu He, Han Bao, Bang An, and Xun Zhou. 2024. Referee-Meta-Learning for Fast Adaptation of Locational Fairness. *AAAI* (2024). doi:10.1609/aaai.v38i20.30197

[5] Lionel Cucala. 2014. A Distribution-Free Spatial Scan Statistic for Marked Point Processes. *Spatial Statistics* (2014). doi:10.1016/j.spasta.2014.03.004

[6] David P. Dobkin, Dimitrios Gunopulos, and Wolfgang Maass. 1996. Computing the Maximum Bichromatic Discrepancy with Applications to Computer Graphics and Machine Learning. *J. Comput. Syst. Sci.* 52, 3 (1996), 453–470.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS*. doi:10.1145/2090236.2090255

[8] Georgios J. Fakas and Georgios Kalamatianos. 2023. Proportionality on Spatial Data with Context. *ACM Transactions on Database Systems* (2023). doi:10.1145/3588434

[9] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS'16*. https://dl.acm.org/doi/10.5555/3157382.3157469

[10] Erhu He, Yiqun Xie, Weiye Chen, Sergii Skakun, Han Bao, and Rahul Ghosh. 2024. Learning With Location-Based Fairness: A Statistically-Robust Framework and Acceleration. *IEEE TKDE* (2024). doi:10.1109/TKDE.2024.33714607

[11] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics - Theory and Methods* 26, 6 (1997), 1481–1496.

[12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35. doi:10.1145/3457607

[13] Christopher Riederer and Augustin Chaintreau. 2017. The price of fairness in location based advertising. In *FATREC*.

[14] Richard Rothstein. 2015. The Racial Achievement Gap, Segregated Schools, and Segregated Neighborhoods: A Constitutional Insult. *Race and Social Problems* (2015). doi:10.1007/s12552-014-9134-1

[15] Dimitris Sacharidis, Giorgos Giannopoulos, George Papastefanatos, and Kostas Stefanidis. 2023. Auditing for Spatial Fairness. In *EDBT*. doi:10.48786/EDBT.2023.41

[16] Sina Shaham, Gabriel Ghinita, and Cyrus Shahabi. 2022. Models and Mechanisms for Spatial Data Fairness. *Proceedings of the VLDB Endowment* 16, 2 (2022), 167–179. doi:10.14778/3565816.3565820

[17] Sina Shaham, Gabriel Ghinita, and Cyrus Shahabi. 2024. Fair Spatial Indexing: A paradigm for Group Spatial Fairness. In *EDBT*. https://doi.org/10.48786/edbt.2024.14

[18] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. doi:10.1145/3194770.3194776

[19] David W.S. Wong. 2004. The Modifiable Areal Unit Problem (MAUP). In *WorldMinds: Geographical Perspectives on 100 Problems*, Donald G. Janelle, Barney Warf, and Kathy Hansen (Eds.). Springer, Dordrecht, 571–575. doi:10.1007/978-1-4020-2352-1_93

[20] Yiqun Xie, Erhu He, Xiaowei Jia, Weiye Chen, Sergii Skakun, Han Bao, Zhe Jiang, Rahul Ghosh, and Praveen Ravirathinam. 2022. Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-level Learning Framework. *AAAI* (2022). doi:10.1609/aaai.v36i11.21481