SpatialFlip: A Postprocessing Method to Improve Spatial Fairness

Victor Arroyo
Université Libre de Bruxelles
 Brussels, Belgium
victor.arroyo.olea@gmail.com

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

ABSTRACT
In this study, we examine how to improve the Spatial Fairness of a classifier by selecting suitable modifications, flips, to its output. When Machine Learning algorithms are evaluated for Spatial Fairness and found to be unfair, it is important to make corrections to prevent the development of products and services that discriminate against individuals based on their location. We compared five different strategies for improving the Spatial Fairness of a classifier by computing observations’ features. Among these strategies, an ensemble approach proved to be the most effective by modifying points among the recommendations made by simpler strategies.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence.

KEYWORDS
Responsible Data Science, Spatial Fairness

1 INTRODUCTION
We consider a post-processing fairness mitigation approach. We first recall the definition of spatial fairness from [3], and then present the problem and our post-processing method.

2.1 Spatial Fairness
We consider a binary classifier that predicts a binary outcome (e.g., accept loan) for individuals (e.g., loan applicants), and each individual is associated with a location. We say that the classifier is spatially fair if its performance is independent of the individual’s location. This implies that if we consider any region in the space, the model performance inside the region is the same as that outside.

Hereafter, we study the positive rate of the model, i.e., the proportion of positive outcomes (e.g., accepted loans). The work in [3] proposes an auditing approach to investigate if a model is spatially fair. It scans a set of spatial regions, and for each region it determines which among two hypotheses is more likely: the null hypothesis of fairness which says inside = outside, and the alternate of unfairness which says inside ≠ outside. Specifically, it computes the likelihood ratio (LR) of these hypothesis.

2.2 Problem Definition
Our goal is to improve the spatial fairness of a model. Thus we need an unfairness score to quantitatively compare how close one model is to being spatially fair. We propose to use the Mean Likelihood Ratio (MLR) computed over a set of regions, which has an intuitive probabilistic interpretation: the expected fairness deviation (in terms of likelihood ratio) of any region in space.

We consider a post-processing fairness mitigation approach. We observe the outcomes of the ML model, and our goal is to change some of them to improve spatial fairness. The changes we can make are flips of the model’s outcomes for specific instances, e.g., grant the loan when it is denied and vice versa. Note there is a trade-off between a model’s performance and fairness since we may improve spatial fairness but also degrade model performance. Therefore, we operate on a budget, expressed as the maximum number of flips we are allowed to make.
When assigning priority scores to outcomes, we consider all regions.

We propose five strategies to select which outcomes to flip. Each of them seeks to assign a priority score to each model outcome that quantifies how much the spatial fairness will improve by flipping this outcome. Then, for a budget of $n$ flips we should select the outcomes with the top-$n$ priorities.

**NumRegions.** To improve spatial fairness, some regions may require outcome reversals (e.g., more loan approvals), while others may require the opposite (e.g., more loan rejections). We assign a flip direction, either positive or negative, to each region. Since regions can overlap, an outcome may belong to multiple regions. When assigning priority scores to outcomes, we consider all regions they belong to. A region is considered covering if it contains the outcome. The NumRegions strategy prioritizes outcomes that affect the most regions. Specifically, the strategy assigns each outcome a score based on the absolute difference between the number of covering regions of one flip direction versus the other.

**AggFlips.** For a given region, we define the required flips as the number of flips necessary so that it becomes fair. The intuition behind the AggFlips strategy is that we should make flips that can quickly fix the fairness of as many regions as possible. Hence, the strategy assigns to each outcome a score that sums up over all covering regions the opposite of their required flips and scales that with the NumRegions score.

**AggLR.** The AggLR strategy aggregates a score for each covering region of an outcome based on the regions’ likelihood ratio (LR). The intuition is that we should flip outcomes that reside in regions that are highly unfair.

**Ensemble.** This is an ensemble strategy that combines the three basic strategies. It operates iteratively, selecting the best flip among the top-1 recommended from the basic strategies.

**EnsembleLA.** This is also an ensemble strategy with a lookahead (LA) functionality. Specifically, at each iteration it considers all recommended flips (not just the top-1) and among all of them selects the best.

### 3 RESULTS AND DISCUSSION

In this work, we will be using as a binary classifier a Mortgage Loan Applications Register (LAR) submitted by financial institutions in the USA in 2021\(^1\). We will take the features Action Taken, i.e., whether it was accepted or denied, and Census Tract, a code that is translated to a geographic coordinate reference. The preprocessed dataset contains 206,418 applications with an overall 0.62 positive rate. After auditing the classifier, the data points and the detected unfair regions are distributed as Figure 1 shows.

The budget of flips $n$ was set to 400. The contrast of performance of the strategies in terms of MLR drops is shown in Figure 2, where it can be observed that the EnsembleLA and AggLR approaches outperformed the rest of the strategies. By EnsembleLA’s definition, it was expected to obtain the best performance among the heuristics. Concerning the AggLR approach performance, the score associated with the recommended points gives preference to observations belonging to regions with higher LR value, i.e., more spatially unfair partitions. Since the MLR is computed as an arithmetic mean, this metric is highly sensitive to outliers; if we start focusing on improving the most unfair regions, it is expected that this method will produce a more significant impact on the MLR score. On the other hand, AggFlips promotes points that belong to regions with lower LR, i.e., less unfair partitions; then, the computation of MLR is negatively affected since lower values are no longer part of the equation, producing an increase of Spatial Fairness estimation. This reflects the implications of the selected unfairness score. The NumRegions approach demonstrated to not be convenient by itself in recommending points, but in supporting the recommendations made by the AggFlips and AggLR strategies. Lastly, the Ensemble approach does not perform well since it follows local improvements, which tend to concern high-AggFlips-scored observations and, consequently, poor performance.

### REFERENCES


\(1\) https://ffiec.cfpb.gov/data-publication/modified-lar/2021