SHORT-PAPER

SpatialFlip: A Postprocessing Method to Improve Spatial Fairness

VICTOR ARROYO, Université Libre de Bruxelles, Brussels, BRU, Belgium

DIMITRIS SACHARIDIS, Université Libre de Bruxelles, Brussels, BRU, Belgium

Published: 22 December 2023

Citation in BibTeX format

SIGSPATIAL '23: 31st ACM International Conference on Advances in Geographic Information Systems
November 13 - 16, 2023
Hamburg, Germany

Conference Sponsors:
SIGSPATIAL

SIGSPATIAL '23: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (November 2023)
https://doi.org/10.1145/3589132.3628376
ISBN: 9798400701689

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

ACM Reference Format:
Victor Arroyo and Dimitris Sacharidis. 2023. SpatialFlip: A Postprocessing Method to Improve Spatial Fairness. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3589132.3628376

1 INTRODUCTION

Transparency, privacy, and bias are some of the difficulties that have brought doubts to society about Machine Learning (ML) and its applications. In particular, for ML classification models, it is interesting to study its *algorithmic fairness*, i.e., the lack of discrimination (bias) against individuals with specific attribute values. These attributes are called *protected*, and we can typically find gender, religion, or race among them [2]. When the protected attribute is location, we call *Spatial Fairness* [3–5], the notion that algorithms should not discriminate against individuals or groups based on spatial attributes, such as their birthplace or home address.

Awareness of Algorithmic Spatial Fairness is crucial since data scientists may introduce historical discrimination patterns by using location as input for their ML algorithms. While location by itself might not be considered to be a protected attribute, it may be highly correlated to ethnicity or race origin, since, occasionally, people from the same ethnic group gather spatially as communities [1]. There can be found in literature frameworks to define and assess Spatial Fairness in binary classification models, which audit its outcomes and, if it is spatially unfair, can also detect where the discrimination is made [3]. The present work uses this Spatial Fairness framework to develop strategies for correcting a binary classifier. Note that algorithmic fairness for ML models is typically achieved in one of three ways: *pre-processing*, where bias is eliminated from the training data; *in-processing*, where the learning process is altered to become fairness-aware; and *post-processing*, where the outcomes of the model are changed to ensure fairness.

2 METHOD

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.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11…$15.00
https://doi.org/10.1145/3589132.3628376

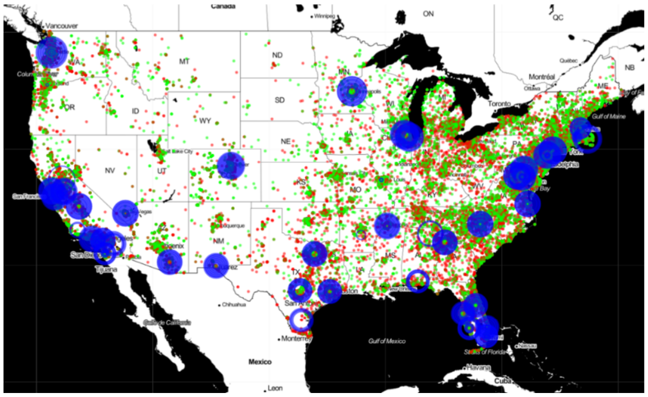

Figure 1: LAR dataset with all its data entries labeled by color; Mortgage Loan Applications are green if accepted and red if denied. Blue circles represent spatially unfair regions

2.3 SpatialFlip Strategies

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.

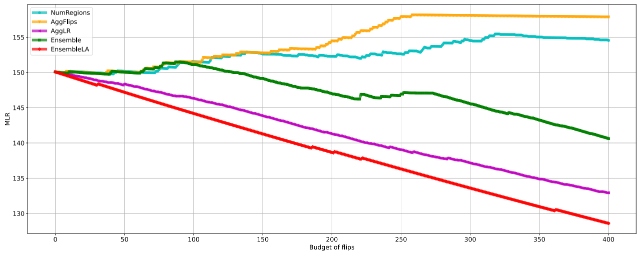

Figure 2: Strategies comparison in terms of *Mean Likelihood Ratio* evolution with a budget of n=400 flips.

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] Esther Havekes, Michael Bader, and Maria Krysan. 2016. Realizing Racial and Ethnic Neighborhood Preferences? Exploring the Mismatches Between What People Want, Where They Search, and Where They Live. *Population research and policy review* 35 (01 2016), 101–126. https://doi.org/10.1007/s11113-015-9369-6

[2] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in Rankings and Recommendations: An Overview. *CoRR* abs/2104.05994 (2021). arXiv:2104.05994 https://arxiv.org/abs/2104.05994

[3] Dimitris Sacharidis, Giorgos Giannopoulos, George Papastefanatos, and Kostas Stefanidis. 2023. Auditing for Spatial Fairness. In *EDBT*. OpenProceedings.org, 485–491. https://doi.org/10.48786/edbt.2023.41

[4] Sina Shaham, Gabriel Ghinita, and Cyrus Shahabi. 2022. Models and Mechanisms for Spatial Data Fairness. *Proc. VLDB Endow.* 16, 2 (oct 2022), 167–179. https://doi.org/10.14778/3565816.3565820

[5] 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. In *AAAI*. AAAI Press, 12208–12216.

[1] https://ffiec.cfpb.gov/data-publication/modified-lar/2021