open proceedings

Legally-Compliant Spatial Fairness Framework: Advancing Beyond Spatial Fairness

Nripsuta Ani Saxena
nsaxena@usc.edu
University of Southern California
Los Angeles, CA, USA

Ronit Mathur
rmathur0725@gmail.com
USC Viterbi SHINE
Los Angeles, CA, USA

Cyrus Shahabi
shahabi@usc.edu
University of Southern California
Los Angeles, CA, USA

ABSTRACT

Although location data is crucial in many decision-making domains, such as mortgages and insurance, identifying spatial bias with respect to legally protected attributes is not straightforward. While the emerging field of spatial fairness examines differences in outcomes based on location, it entirely overlooks legally protected attributes and their correlation with location, meaning the outcomes may differ but may not necessarily violate legal compliance. On the other hand, while the fairness in machine learning (fair-ML) community focuses on legally protected attributes, it has not explored the specific challenges introduced by spatial data. Such challenges, such as the modifiable areal unit problem (MAUP), render extant fair-ML work impractical for spatial fairness. This work extends spatial fairness by incorporating legal compliance from fair-ML, introducing the *legally compliant spatial fairness (LC-spatial fairness)* framework. LC-spatial fairness evaluates ML model outputs for fairness concerning location and legally protected attributes, identifying biased regions that need attention. We introduce a new fairness definition that simultaneously considers both spatial data and legally protected attributes, and demonstrate that it is robust to MAUP. Furthermore, we highlight why previous approaches are not robust to MAUP. Experimental evaluation on real-world data demonstrates the efficacy of our approach, which identifies significantly more spatial unfairness than previous techniques. Furthermore, results highlight the ability of our approach to identify spatial biases that other methods overlook while also revealing areas where competitors found biases that our approach did not, largely because our approach explicitly considers legally protected attributes and unprotected attributes in addition to location, while previous approaches focus on location only.

1 INTRODUCTION

Machine learning (ML) is increasingly used across various fields to facilitate location-based decision-making (LB-DM), where decisions are influenced by an individual's geographic location, such as their address or residential zip code. For instance, ML is being widely adopted in banking and insurance, both industries that heavily rely on location data for decision-making and can have a significant effect on an individual's life, notably affecting financial prospects such as mortgage rates, credit card approvals, and insurance premiums [12, 21, 22]. Such use of ML in location-based decision-making is poised to increase further. Just in the United States alone, over 80% of financial institutions report that they feel confident in using AI- and cloud-based credit-risk decision-making [12], and more than 60% of businesses intend to boost budgets for credit-risk analytics further [12]. However, location can correlate with many personal attributes, such as race, ethnicity, and national origin [34, 43], which are legally protected against discrimination in various decision-making domains [7, 28]. Due to discriminatory historical practices like redlining[1], location's correlation with legally protected characteristics continues to persist after decades, often with harmful consequences. For example, researchers have found that the racial composition of a neighborhood was a stronger predictor of home appraisal values in 2015 than in 1980 [17]. Spatial segregation by race, in particular, has been challenging to remedy [8, 23]. As a result, neglecting to account for location appropriately, combined with the growing use of ML in these domains, risks data-driven models picking up and perpetuating such bias even further. Although the fair-ML literature has noted that ML systems may replicate undesirable trends present in historical data, perpetuating discrimination further, hardly any attention has been paid to unfairness that may creep into ML models due to location's correlation with legally protected attributes.

Meanwhile, the use of location in LB-DM without careful consideration is already leading to (potentially illegal) unintented consequences in many domains. For example, the non-profit ProPublica found that drivers from minority neighborhoods were charged higher car insurance premiums than drivers with similar risk from majority-white neighborhoods [3]. This prompted an investigation by the state of California, leading to the state requiring insurers to adjust rates to remove racial disparities [2]. Other research has also found that riders in neighborhoods with more minority residents and higher poverty levels are significantly associated with higher fares for ride-hailing services [34]. In another instance, Amazon was accused of racism when an analysis of its Prime Same-Day Delivery service in six major U.S. cities by Bloomberg found that Amazon was significantly less likely to offer the service to Prime customers living in predominantly minority neighborhoods [18]. In the most egregious instance, in Boston, the three zip codes constituting the majority-black neighborhood of Roxbury were excluded from Same-Day Delivery. In contrast, zip codes on all sides of it were included. Differences in neighborhoods' household income did not explain all such exclusions.

Although the problem of spatial bias has started receiving attention recently [38, 44, 49], work so far does not consider legally protected attributes at all while assessing spatial fairness. For example, Sacharidis et al. [38] define spatial fairness as the statistical independence of outcomes from location. However, requiring the distribution of a measure (e.g., positive outcomes of a model) to be independent of location ignores the reality that, often, location offers legitimate (and legal) information valuable

 Published in Proceedings of the EDBT/ICDT 2025 Joint Conference, 25th March - 28th March, 2025, ISBN 978-3-89318-099-8 on OpenProceedings.org.

[1] Redlining is the discriminatory practice of systematically denying individuals from specific neighborhoods access to mortgages, insurance, and other financial services based on their race or ethnicity [37, 48]. It was outlawed in 1968 [Title VIII of the 1968 Civil Rights Act].

for decision-making. For example, two similarly-valued houses may not have the same insurance premiums if one house happens to be in a very high fire risk zone while the other is not. In other words, there may often be valid and perfectly legal reasons for the distribution of a measure to vary in different regions. Thus, since location can provide information relevant to decision-making that is perfectly legal to use, simply removing location from decision-making or making systems location-invariant is not a practical solution. Furthermore, since these definitions do not account for legally protected attributes and their correlation with location, they lack practical real-world applications [43] and would not hold up to legal scrutiny. Therefore, it is necessary to consider fairness comprehensively with location and legally protected attributes in tandem.

However, considering location simultaneously with legally protected attributes introduces complications unique to spatial data, such that traditional fair-ML approaches to measure or mitigate bias cannot be directly applied for spatial fairness. The focus of fair-ML research on discrete attributes, combined with the spatial challenges such as the bias introduced by modifiable areal unit problem (MAUP), renders fair-ML work ineffective in handling an attribute that is simultaneously continuous and spatial. Firstly, such work does not consider location at all. Secondly, most fair-ML work focuses on legally protected attributes that are discrete with well-defined values (e.g., gender: male/female; race: black/white) and does not map well to the continuous nature of spatial data. Thirdly, the unique challenges associated with spatial data further complicate the issue. Perhaps the most salient challenge introduced with spatial data is the modifiable areal unit problem (MAUP). MAUP is a type of statistical bias present only in geographical data, occurring when point-based estimates of spatial phenomenon are aggregated into spatial partitions. For a given space subdivided into many partitions, the shape and size of the partitions can affect the outcomes, i.e., they can alter the perceptions of fairness. For example, Figure 1(a) shows a region where positive (green circles) and negative (red circles) outcomes are distributed in space. Depending on how we partition this space can change the perceptions of spatial fairness. Figure 1(b) and (e) appear spatially fair with an even balance of positive and negative outcomes, while the partitions in Figure 1(c) and (d) appear spatially unfair with only one kind of outcome in each partition. Thus, changing the partitioning can lead to a completely different fairness outcome, highlighting how MAUP can pose a problem in assessing spatial fairness in a robust manner. Gerrymandering is a real-world example of MAUP, where the shape of the districts is manipulated to serve a particular political party [26, 30]. Even fair-ML work on proxy discrimination, i.e., when a seemingly innocuous attribute happens to be correlated with a legally protected one, suffers from limitations. For example, one such method is fair-PCA [40, 46], which requires one extra dimension for a discrete protected attribute with two possible values (e.g., gender, with the possible values male or female). But in the case of spatial fairness, there may be numerous protected subgroups (e.g., the various subregions that must be treated as protected in addition to the subgroups of the legally protected attribute), which renders fair-PCA impractical in this setting.

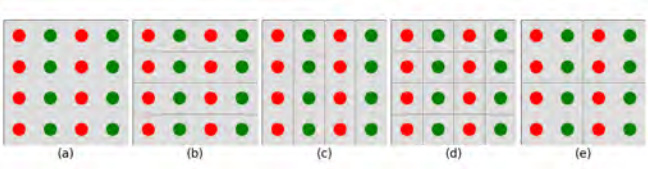

Figure 1: An example illustrating the Modifiable Areal Unit Problem. Green and red denote positive and negative outcomes respectively. The partitioning of the space can significantly change the perceptions of spatial fairness.

To prevent instances of spatial bias, this work presents a novel framework, *Legally Compliant-Spatial Fairness*, or *LC-spatial fairness*, that can holistically assess the outcomes of data-driven models for fairness with respect to location as well as legally protected attributes to ensure that neighborhoods and regions that are different mainly in legally protected attributes are not treated differently. We avoid the pitfalls of spatial fairness work by defining fairness with respect to location and legally protected attributes simultaneously to withstand legal scrutiny. Our definition, formulated for geographic data, can also efficiently handle continuous spaces. Furthermore, we empirically show that our definition is robust to the modifiable areal unit problem (MAUP), enhancing its real-world applicability. Finally, by considering location and legally protected attributes simultaneously, LC-spatial fairness can identify regions where the outcomes of a given model are unfair. Intuitively, we identify pairs of regions in a given space which are similar in non-protected attributes while being different in the protected attribute of interest (such as race). We perform likelihood ratio tests to determine whether the outcomes in such pairs of regions are significantly different, which demonstrates spatial unfairness. Thus, it can be used by regulatory agencies to detect cases of spatial bias, identify which regions deserve more attention, and enforce corrective measures, and by companies to ensure legal compliance. To the best of our knowledge, we formulate the first spatial fairness definition that holistically encompasses location and legally protected attributes. Overall, our main contributions are as follows:

- Motivate and identify the need for defining spatial fairness with respect to location as well as legally protected attributes
- Formalize a spatial fairness definition capable of simultaneously considering location and legally protected attributes holistically
- Intuitively show how our definition of spatial fairness is robust to the modifiable areal unit problem (MAUP)
- Present two use cases to illustrate the utility of our LC-Spatial-Fairness framework: one in the domain of mortgage lending and the other in improving access to healthy food
- Experimental evaluation on real-world datasets show the efficacy of our framework in identifying spatial bias

The rest of this paper is organized as follows. Section 2 reviews related work on fair-ML and spatial fairness. The LC-spatial-fairness framework is covered in Section 3. Use cases for our framework are detailed in 4 along with preliminary results. Section 5 describes the experimental evaluation. We conclude with Section 6.

2 RELATED WORK

Sections 2.1 and 2.2 cover fairness definitions and techniques from fair-ML. Section 2.3 explores the emerging area of spatial fairness.

2.1 Fairness definitions or metrics

The first step towards quantifying the extent of unfairness is defining what it means to be fair. This is not a straightforward task: there is no agreement within the fair-ML community about what it means to be fair in different contexts [28, 41]. The disciplines of philosophy and psychology, from which multiple fair-ML definitions have been inspired, also do not offer a universal definition of fairness despite grappling with this question for a much more extended period of time [28]. This lack of agreement has led to a generally well-accepted consensus that a single definition of fairness would be insufficient across varied contexts [7, 42]. Thus, many fair-ML definitions have been proposed to measure unfairness in diverse computational domains, with over 20 estimated metrics proposed by 2018 already [29]. A typical fair-ML fairness metric measures unfairness concerning legally protected attributes in the outputs of ML models on a particular task either at the level of the individual [11, 24, 28] or at the level of the group [16, 20, 47]. For example, a simple metric might consider true positive rates for different groups (e.g., different racial groups or men and women) for a classification task. More complex metrics have been proposed for different settings and application areas [7]. For example, metrics have been proposed to measure gender bias in machine translation [14] and other NLP tasks [25], fairness definitions for clustering [9], as well as graph machine learning [10] and multi-armed bandits [19, 24] have been proposed. No work so far defines fairness with respect to location and legally protected attributes.

2.2 Fairness techniques

A fairness technique aims to mitigate bias at one of three stages of the ML pipeline: pre-processing, in-processing, and post-processing. Pre-processing methods attempt to "debias" data by transforming datasets before they are used for training an ML model [7]. Such methods may be more general (e.g., focusing on datasets with biased or imbalanced distributions for protected attributes, such as removing bias in embeddings [6]) or geared towards removing specific types of biases from data (e.g., [1] propose a technique to correct for Simpson's Paradox). In contrast, in-processing methods tweak the model itself to remove bias during model training, typically by altering the objective function or adding a fairness constraint [4, 7, 28]. In-processing fairness techniques have been proposed for regression [4] and linear contextual bandits settings [15], among other settings. Finally, post-processing approaches massage a model's outputs so that they are fairer with respect to the given fairness definition [7, 28, 35]. For example, [35] propose a graph smoothing problem that corresponds to Laplacian regularization such that the corrected outputs are aligned with the "treat similar people similarly" definition proposed by [11]. Pre- and post-processing techniques are inherently data-related challenges (i.e., debiasing datasets for use by ML models, or debiasing data that ML models output), rather than a traditional ML challenge. They also have the added advantage of not requiring access to the ML model. Thus, they can still be utilized if access to the model is unavailable and must be treated as a black box or if model retraining would be prohibitively costly.

2.3 Spatial fairness

Most spatial fairness work defines unfairness only with respect to spatial regions or partitions and can be considered a form of group fairness (where every spatial partition is a group). For example, Xie et al. [49] impose a two-dimensional rectangular grid over a given region and consider multiple possible partitionings of the form $s_1 \times s_2$ (where s_1 is the number of rows and s_2 the number of columns). Then, for each partitioning, they measure the variance of performance metric (e.g., accuracy) of a model's outcomes in each partition. Finally, they compute mean variance across all partitionings, with a lower mean variance implying higher fairness. However, as pointed out by [38], their method is apt for assessing the fairness of outcomes that are regularly distributed across space. It does not perform as well when outcomes are distributed irregularly in space. In contrast, [38] formulate a framework to conduct a likelihood ratio test to audit for spatial unfairness. However, they only consider location and the number of positive outcomes (or another measure of outcome). Our approach builds upon the work of Sacharidis et al. and incorporates the notion of non-protected and legally protected features into the assessment of fairness. Furthermore, our method compares two partitions at a time to each other, while the method of Sacharidis et al. compare the number of positive outcomes inside a partition to the number of positive outcomes everywhere outside that partition. The overarching goal of our work is similar to Sacharidis et al. [38] – to detect the presence of unfairness given a spatial dataset and a model's outputs for that data. Unlike our work, however, Sacharidis et al. [38] and Xie et al. [49] do not consider fairness with respect to legally protected attributes when auditing for spatial fairness, which limits the real-world application of their work since bias due to location is illegal only when it concerns location's association with legally protected attributes [43]. They can be considered the state of the art (SOTA) approach in spatial fairness. We compare with them in Section 4.3.

There is also some work exploring individual spatial fairness. Shaham et al. [44] explore individual spatial fairness by adapting the fair-ML fairness definition proposed by Dwork et al. to location. For a given set of locations $L = l_1, l_2, ..., l_m$, where $l_i \in \mathbb{R}^k$ and an output set A, a randomized mapping $M : L \rightarrow \Delta(A)$ will satisfy individual spatial fairness iff the (D, d)-Lipschitz condition is satisfied for every two locations $l_i, l_j \in L$: $D(M(l_i), M(l_j)) \leq d(l_i, l_j)$. In Dwork et al., the (D, d)-Lipschitz condition must hold over every pair of individuals in the set of individuals V, whereas Shaham et al. require it to hold over pairs of locations in L instead (where each location may be that of a specific individual). The authors then use this to define distance-based spatial fairness and zone-based spatial fairness in the following manner. Distance-based fairness considers the distance of individuals from a reference point of interest, and requires two individuals to be treated similarly if they are at similar distances from the reference point. For example, the reference point could be a health store that wants to display discount offers to nearby individuals. Then, the L would be composed of different individuals' distance to this store, and the outcome of the model may be whether to display an offer to a particular individual or not. The authors argue that a strict boundary risks treating two individuals very close to each other differently if they happen to be on opposite sides of the boundary, which they propose is unfair. Zone-based fairness, on the other hand, adapts their spatial fairness definition for coordinate values rather than distances. The authors propose 'c-fair polynomials,' where a polynomial is fit to the outputs of a model to achieve distance- and zone-based spatial fairness. According to their definition, a polynomial $P(x) : \mathbb{R} \rightarrow \mathbb{R}$ would be c-fair iff the condition $|P(x) - P(y)| \leq c|x - y|$ holds for every pair of points x and y

in its domain, where c acts as the knob to control the trade-off between fairness and utility. In the same vein, they define c-fair polynomials for zone-based spatial fairness where the location dataset L is composed of zones or regions where an individual might be located. Both distance-based and zone-based spatial fairness only consider location while determining fairness, and ignore legally protected attributes. Thus, both definitions have limited applicability in the real world as bias due to location alone is not considered illegal.

These studies pioneered the field of spatial fairness and raised awareness for the need of fairness concerning location. We build upon their research to close the gap between their works and fair-ML research by defining fairness with respect to location and legally protected attributes.

3 LEGALLY COMPLIANT SPATIAL FAIRNESS: A FRAMEWORK

3.1 Our spatial fairness definition

We propose LC-spatial fairness, a comprehensive framework that first defines fairness with respect to location and legally protected attributes. Subsequently, it utilizes this definition to audit the outputs of a given location-based decision-making model for fairness concerning legally protected attributes to identify regions that exhibit unfairness, and are in need of more resources and attention.

We present our definition (Definition 3.3) after introducing some relevant terminology.

Definition 3.1. Legally protected attributes refers to features in a data-driven decision-making model that are safeguarded against discrimination by law across various domains (such as banking and credit, for example). Such attributes are typically immutable personal characteristics, such as sex, race, or national origin.

Definition 3.2. Non-protected attributes refers to features that are relevant to decision-making but are not safeguarded against discrimination by law. Such attributes are typically mutable characteristics that can be changed, such as income.

Definition 3.3. Consider a region R divided into n partitions, $r_1, \ldots, r_n$. Let $\mathbf{F}$ be the set of non-protected feature vectors $\mathbf{f}_i \in \mathbb{R}^m$ for each partition r_i, and let $\mathbf{P}$ be the set of protected attribute vectors $\mathbf{p}_i \in \mathbb{R}^p$ for each partition r_i. Let $O : R \rightarrow \mathbb{R}$ be the outcome function of the model, where O_i is the outcome of region i.

A model's outcomes are fair if and only if for any two partitions r_i and r_j:

1. The non-protected features are similar:

$$\text{Sim}(\mathbf{f}_i, \mathbf{f}_j) \geq \epsilon$$

2. The protected attributes are dissimilar:

$$\text{Diss}(\mathbf{p}_i, \mathbf{p}_j) \geq \delta$$

3. The outcomes are similar:

$$|O(r_i) - O(r_j)| \leq \eta$$

where $\epsilon, \delta, \eta > 0$ are small positive thresholds indicating similarity in non-protected features, dissimilarity in protected attributes, and similarity in outcomes, respectively. $\text{Sim}(\mathbf{f}_i, \mathbf{f}_j)$ and $\text{Diss}(\mathbf{p}_i, \mathbf{p}_j)$ represent a similarity and dissimilarity metric respectively.

A notable strength of the LC-spatial-fairness framework is its flexibility in defining the thresholds of (dis)similarity based on relevance to the specific task at hand. Industries required to protect against discrimination by law, such as banking and credit, would have stricter requirements and likely prefer higher ϵ, δ thresholds and a lower η threshold. Other application settings which merely desire to act more ethically without being required by law may prefer lower ϵ, δ thresholds and a higher η threshold. Moreover, in recognition of the fair-ML literature acknowledging that different (dis)similarity metrics may be required for different contexts, our LC-spatial-fairness framework provides the flexibility to incorporate different(dis)similarity metrics tailored for specific tasks.

3.2 Assessing spatial unfairness: Hypotheses and Likelihood Ratio Test

Next, we incorporate our definition into a hypothesis test to audit a given model's outcomes. We define two hypotheses as follows. Our hypotheses assume two regions r_i and r_j under consideration are similar in the non-protected attributes, F, and dissimilar in the protected attributes, P. The null hypothesis (H_0) posits no spatial unfairness between the two regions, i.e., their outcomes are similar. In contrast, the alternate hypothesis (H_a) posits there is spatial unfairness, and the outcomes in r_i and r_j are significantly different. As we build on the work of Sacharidis et al. [38], we adopt their terminology, start with their equation, and build upon it to get to ours. Although the discussion uses positive rate as the measure of interest, this can be easily adapted to other measures. Furthermore, for consistency, we continue with their setting of a model for determining outcomes for individuals. However, it is generalizable to other settings, such as assessing unfairness in funding resources for different school districts.

As defined in [38], let $\rho = Pr(\hat{Y} = 1)$ denote the overall positive rate of a data-driven model, and $\rho(r_i) = Pr(\hat{Y} = 1 | L \in r_i)$ be the local positive rate of a region r_i. As [38] point out, the positive rate of a model can also be interpreted as the probability of being assigned to the positive class or a Bernoulli trial with ρ as the success probability. To be considered fair in [38], the positive rate in every region $r_1, r_2, ..., r_j$ should follow the same Binomial distribution (i.e., this is their null hypothesis). Their alternative hypothesis states unfairness: a region r_i has a positive rate that follows a Binomial with a different success probability than the Binomial distribution of the positive rate outside r_i. In contrast, rather than comparing the positive rate inside a region to the positive rate outside it, our hypotheses posit the following for any two regions $r_i, r_j \in R$ that are similar in F (non-protected attributes) and dissimilar in P (protected attributes). Our null hypothesis, H_0, assumes no difference between the positive rates of r_i and r_j. Our alternative hypothesis, H_a, posits there is a difference in the positive rates of r_i and r_j.

To understand which hypothesis explains the outcomes better, we derive their maximum likelihoods and compute the likelihood ratio as well as [38]. The likelihood for the null hypothesis for Sacharidis et al. [38] is given by:

$$S_0(R, \rho_0) = \rho_0^{p(r_i)} (1 - \rho_0)^{n(r_i) - p(r_i)} \tag{1}$$

where $n(r_i)$ is the number of individuals in region r_i, while $p(r_i)$ represents the number of individuals with positive outcomes in r_i. The likelihood for their alternative hypothesis is the product of the binomial of a region r_i with the binomial for outside r_i. It is defined as follows:

$$S_a(r_i, \rho_0, \rho_1) = \rho_0^{p(r_i)}(1-\rho_0)^{n(r_i)-p(r_i)} \times \rho_1^{P-p(r_i)}(1-\rho_1)^{N-n(r_i)-(P-p(r_i))} \quad (2)$$

where ρ_0 is the positive rate inside region r_i and ρ_1 is the positive rate outside it. N is the total number of individuals across the entire space R, and P is the total number of positive outcomes in the entire space R.

In contrast, the likelihood for our alternative hypothesis, H_a, has some more terms, which we describe individually before putting them all together. For a region r_i, the likelihood will contain the term:

$$\rho_i^{p(r_i)}(1-\rho_i)^{n(r_i)-p(r_i)} \quad (3)$$

where ρ_i is the success probability for region r_i, $n(r_i)$ is the number of individuals in r_i, and $p(r_i)$ is the number of positive labels in r_i. This will be multiplied by the term:

$$\left(\frac{n_G(r_i)}{n(r_i)}\right)^{n(r_i)} \cdot \left(1-\frac{n_G(r_i)}{n(r_i)}\right)^{n(r_i)-n_G(r_i)} \quad (4)$$

where $n_G(r_i)$ is the number of individuals belonging to protected group G in region r_i. Finally, both these terms would be multiplied by:

$$\left(\frac{n_V(r_i)}{n(r_i)}\right)^{n_V(r_i)} \left(1-\frac{n_V(r_i)}{n(r_i)}\right)^{n(r_i)-n_V(r_i)} \quad (5)$$

where $n_V(r_i)$ denotes the number of individuals in non-protected group V in region r_i.

Thus, the likelihood for a region r_i, L_{r_i}, will be the product of Equations 3, 4, and 5. Therefore, the likelihood of our alternative hypothesis for regions r_i and r_j will be given by the product of the likelihood for the two regions:

$$L_a = L_{r_i} \cdot L_{r_j} \quad (6)$$

We maximize this likelihood for each pair of partitions and then compute a likelihood ratio test to determine which hypothesis explains the model's outcomes better. To determine the significance of the test statistic of the likelihood ratio test we conduct Monte Carlo simulations, similar to how Sacharidis et al. also determine significance. By creating m alternative "worlds" with N total data points, with each data point's outcome determined by a Bernoulli trial with appropriate success probability ρ [38]. The τ statistic for each alternative world is ranked. If the τ statistic for the actual observed data is ranked at position k, then the significance or p-value for the observed data would be $k/(m-1)$. We conclude the outcomes of a pair of partitions, r_i, r_j, are spatially unfair if the test statistic is less than the predetermined level of significance.

3.3 Resistance to the Modifiable Areal Unit Problem (MAUP)

In this section, we discuss the resistance of the LC-spatial-fairness framework to MAUP. A significant difference between our framework and that of previous works such as Sacharidis et al. [38] and Xie et al. [49] is that they compare the measure of interest (such as positive rate/rate of approval or prediction accuracy) within each partition with the measure of interest globally. In the Sacharidis et al. framework [38], for instance, the goal is to identify partitions where the measure of interest in a partition differs from the measure of interest across the entire space. For example, in an experiment with loan approval decisions, their framework identifies partitions where the loan approval rate differs significantly from the overall global approval rate of 0.62. Thus, to change the designation of a given partition from spatially unfair to spatially fair according to this framework, an adversary could redraw the boundaries of the partitions such that the local measure of interest becomes more similar to the global measure.

Consider the following scenario where an adversary could "game" the system. Assume a space R divided into n partitions $r_1, r_2, \ldots, r_n$. Let the positive rate (e.g., rate of approval of loans) be the measure of interest. The positive rate across the entire space R is given to be 70% (i.e., this is the global measure). Intuitively, the framework of Sacharidis et al. designates a partition r_k as spatially unfair when the measure of interest in the partition r_k is different from the global measure of interest. Let there be two adjacent partitions, r_i and r_j, with local positive rates of 90% and 50%, respectively (Figure 2a). The framework in [38] would label these partitions as spatially unfair. A malicious adversary, however, could manipulate the system into labeling both partitions as spatially fair by changing the boundaries of r_i and r_j such that the new partitions r_i' and r_j' (Figure 2b) each have a local positive rate of 70%.

(a) The original partitions, r_i with local positive rate of 90% and r_j with local positive rate of 50%, that are deemed spatially unfair by the mechanism in Sacharidis et al. [38]. The global positive rate is 70%.

(b) An adversarial redrawing of the boundary of partitions such that new r_i' and r_j' each have a local positive rate of 70% and are now deemed spatially fair.

Figure 2: A partitioning where the global positive rate is 70%.

In contrast, our LC-spatial-fairness framework does not compare local rates against a global measure to appraise fairness. Since our framework compares the outcomes of a pair of partitions at a time, where both are similar in unprotected and dissimilar in protected attributes, rather than comparing each partition's

local outcomes to the global outcome rate, we avoid this issue. Let us continue with the example of a given space R divided into n partitions $r_1, r_2, ..., r_n$. To recall, spatial fairness in our framework is defined by:

$$\forall i, j \in \{1, 2, \ldots, n\}, \quad i \neq j : \quad \big(F(r_i) \sim F(r_j)\big) \wedge \big(P(r_i) \nsim P(r_j)\big) \implies O(r_i) \sim O(r_j)$$

where $\sim$ denotes similarity in non-protected attributes and model outcomes, while $\nsim$ denotes dissimilarity in protected attributes.

Let r and r_j be two partitions labeled as unfair by the LC-spatial-fairness framework for being similar in unprotected and dissimilar in protected attributes yet having dissimilar outcomes. Thus:

$$\big(F(r_i) \sim F(r_j)\big) \wedge \big(P(r_i) \nsim P(r_j)\big) \quad \text{but} \quad O(r_i) \nsim O(r_j)$$

Let us assume a malicious adversary redraws the boundaries of r_i and r_j. Let the new partitions be r'_i and r'_j. Four possible cases might occur. First, redrawing the boundaries does not change the makeup of the unprotected and protected attributes of r'_i and r'_j. That is, $F(r'_i) \sim F(r'_j)$, while $P(r'_i) \nsim P(r'_j)$. In this case, r'_i and r'_j would still be compared and deemed unfair by our methodology. Second, redrawing the boundaries could result in a change such that r'_i and r'_j are no longer similar in unprotected attributes, while the protected attributes remain dissimilar. Thus, $F(r'_i) \nsim F(r'_j)$, while there is no change in the makeup of the protected attributes $P(r'_i) \nsim P(r'_j)$. In this case, the two partitions will no longer be compared because the LC-spatial-fairness framework requires them to be similar in unprotected attributes for them to be compared. However, the fairness assessment will not be circumvented because now r'_i and r'_j will be compared to other, different partitions that are now similar to them in unprotected attributes and dissimilar in protected attributes. Thus, unfairness will likely resurface elsewhere in the network, as the adversary cannot isolate partitions from comparison without affecting other comparisons. In the third case, r'_i and r'_j are still similar in unprotected attributes, $F(r'_i) \sim F(r'_j)$, but are now similar in protected attributes, $P(r'_i) \sim P(r'_j)$. Similar to the previous case, while r'_i and r'_j will no longer be compared to each other, they will now be compared to other partitions they were not compared to earlier but are now similar to in unprotected and dissimilar in protected attributes. Finally, the last case would be that the redrawn partitions differ in unprotected attributes while becoming similar in protected attributes: $F(r'_i) \nsim F(r'_j)$ and $P(r'_i) \sim P(r'_j)$. While r'_i and r'_j are no longer compared, they will be compared to other partitions they were not eligible to be compared to before. In sum, any redrawing of boundaries will only result in a fresh set of fairness comparisons.

Thus, the LC-spatial-fairness framework's fundamental nature ensures that redrawing any partition's boundary only shifts comparisons and cannot eliminate fairness checks.

4 USE CASES

We utilize two use cases to demonstrate the utility of our LC-Spatial-Fairness framework: the first, in the domain of mortgage lending, was developed previously and serves as a baseline for comparison, while the second, focused on healthy food accessibility, is newly introduced in this paper. These diverse use cases underscore the broad applicability of our framework across varied settings. The LC-spatial-fairness framework helps identify regions that are treated unfairly in approved loans and with disproportionately heavy access to fast food establishments. This section outlines the use cases and presents experimental results for each of them. We then continue the experimental analysis in Section 5, where we compare with baselines and perform experiments with various partitionings to study our framework's resistance to MAUP.

4.1 Mortgage Lending

Given the growing adoption of artificial intelligence in the financial sector [12], our first use case focuses on the scenario of mortgage applications.

4.1.1 Use Case. Consider a data-driven model for classifying mortgage applications in the United States into binary outcomes (e.g., approval or rejection). Then, R would be the U.S., and the n partitions could be defined in many ways. They may be the different states or counties, or a grid could be superimposed and the cells used as partitions. Let us assume a grid creates n partitions $r_1, r_2, ..., r_n$. The set of non-protected attributes F can be relevant non-protected attributes such as an applicant's income and amount of current debt. In contrast, the protected attribute under consideration, P, would be race (since it is protected against discrimination in the credit industry in the United States by federal law [36, 43]). The outcome O would be the decision regarding the mortgage application (i.e., approval or denial).

Many metrics may be incorporated to assess similarity and dissimilarity. Statistical parity [28] and the Mann-Whitney U test statistic are two potential metrics for similarity, while disparate impact measure from the fair-ML literature [28] or the z-score statistic may be utilized for dissimilarity. We present results for them in Section 5. Since discrimination on the basis of race is outlawed in the credit industry in the United States, the ϵ and δ thresholds should be strict. In Section 5, we present experimental results with publicly available mortgage data and set these thresholds to be 0.001.

Other (dis)similarity metrics may also be employed, such as distance-metric inspired ones suggested in Dwork et al. [11]. However, they must be manually crafted specifically for the application setting and typically require deep subject matter expertise.

4.1.2 Data and Experimental Results. We now present results for the mortgage application use case. We utilize the publicly available Loan Application Register (LAR) dataset[2], which contains an anonymized record of mortgage application decisions by financial institutions in the United States. All financial institutions in the U.S. that issue 200 or more open-end lines of credit or closed-end mortgage loans must report such data every year under the Home Mortgage Disclosure Act [45]. For a given financial institution, the dataset contains the applicant's geographic location at the census tract level and the application's outcome (e.g., whether the loan was approved, denied, approved but not taken out, or application withdrawn, et cetera).

To study spatial unfairness in this context, we utilize income as the unprotected attribute of relevance, while race serves as the legally protected attribute (since the Equal Credit Opportunity Act prohibits discrimination based on race in credit decisions). Thus, after filtering for applications that were either approved or denied, we perform a spatial join with data from the U.S. Census

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

Table 1: Results for the LC-Spatial Fairness framework for the mortgage application use case.

Dataset	Grid dimensions	Number of unfair regions
Bank of America	100×50	493
Wells Fargo	100×50	569
United Wholesale Mortgage	100×50	238
Loan Depot	100×50	899

Bureau from the 2020 Census[3] to obtain the racial composition and income distribution of different census tracts. The outcome of interest is the positive rate, or in other words, the rate of approval of loans.

We perform the aforementioned pre-processing for a mix of different types of institutions that offer mortgages in the U.S. We process datasets for Bank of America and Wells Fargo, two popular banks in the U.S., obtaining 224,145 and 311,375 applications respectively. We also process the LAR datasets for United Wholesale Mortgage, a wholesale mortgage lender, and Loan Depot, a non-bank mortgage lender in the U.S. We obtain 687,772 application records for United Wholesale Mortgage and 225,495 for Loan Depot after all pre-processing is complete. We adopt a high-resolution grid partitioning of 100×50, and utilize the Mann-Whitney U test statistic and the z-score statistic for similarity and dissimilarity respectively, while the thresholds for ϵ and δ are set to 0.001. Table 1 presents the number of unfair regions found by the LC-Spatial Fairness framework. For Bank of America, for example, the framework finds that there are 493 regions which would be considered unfair when compared with another region with similar income yet different racial distribution. United Wholesale Mortgage is the largest mortgage provider in the U.S. [33], originating $108 billion USD in loans in 2023. The much higher number of applications received and approved likely results in the considerably lower number of unfair regions found for it, while institutions that receive fewer applications show more unfairness.

We present experimental analysis for this use case with more metrics and partitionings in Section 5.

4.2 Access to healthy food

We now consider a different kind of use case: an application setting where an industry or agency wishes to act more ethically without necessarily being required by law. We consider such uses of the LC-spatial-fairness framework to be *ethical spatial fairness*.

Agencies such as Food Access Advisory Group and States' Department of Food and Agriculture [32] may wish to analyze the distribution of fast food restaurants across a given space to identify which regions need more grocery stores to combat the problem of food deserts. A food desert is a low-income area in which more than a third of the population lives more than a mile from a grocery store or a supermarket (the distance is 10 miles for rural areas) [31]. A higher concentration of fast food chains in such regions exacerbates the food desert crisis, contributing to decreased access to nutritional food. Thus, government agencies may be interested in identifying regions with an unjustified abundance of fast food chains and choose to offer incentives for more healthy food outlets to open in such areas.

Here, R could be the U.S. or the specific state carrying out the analysis, and the unprotected attribute, F, would be income (since an area must be low-income to qualify as a food desert). The protected attribute, P, would be race since food deserts are typically in minority neighborhoods [31, 32]. The outcome of interest would be the number of fast-food restaurants in each region. If two regions r_i and r_j have similarly low income, but r_i is a minority region, and r_j is not, but r_i has significantly more fast food outlets, then we can conclude that it is spatially unfair.

4.2.1 Data and Experimental Results. For experiments for this use case, we utilize SafeGraph Places data[4] [39]. SafeGraph Places provides detailed information about geographical places, also known as point-of-interest (POI) data. It identifies various information including the main "category" (e.g., restaurant) and "sub-category" (e.g., fast food or limited service restaurant) of each geographical place, location, and brand. We compile information for locations of the top 15 most popular fast food brands in the U.S. [27] and perform a spatial join with census data for income and racial data. After all pre-processing, we are left with 106,091 fast food places nationwide. We can use the same metrics as in Section 4.1; however, since governments will likely not have an unlimited budget to offer incentives, we can have lower thresholds for ϵ and δ. We set these thresholds to be 0.01. For a lower resolution grid partitioning of 20×20, the LC-spatial-fairness framework finds 41 unfair regions, about 10% of the total number of partitions. Each of these 41 regions is an area with significantly more fast food outlets than another area with similar income yet different racial makeup. In other words, it has an unfairly high abundance of unhealthy food options that cannot be explained away by the income of the area.

We present more results for this use case with different partitionings in Section 5.

5 EXPERIMENTAL ANALYSIS

This section details the comparison of the LC-spatial-fairness framework with baselines, as well as experimental evaluation of our framework with different partitionings to assess its resistance to MAUP.

5.1 Comparison to Baselines

Here, we compare the LC-spatial-fairness framework with two baselines. The first baseline is a common technique to assess bias or unfairness in the discipline of fair-ML known as disparate impact [28]. The second baseline is a previous spatial fairness methodology proposed by Sacharidis et al. [38].

5.1.1 Fair-ML baseline. We compare the fairness assessment of the LC-spatial-fairness framework with the disparate impact assessment commonly used to measure unfairness in varied decision-making settings in many fair-ML works [28]. Disparate impact is formally defined as follows.

Definition 5.1. Disparate impact is the ratio of the positive outcome rate for a protected group to the positive outcome rate

[3] https://data.census.gov/table

[4] https://www.deweydata.io/data-partners/safegraph

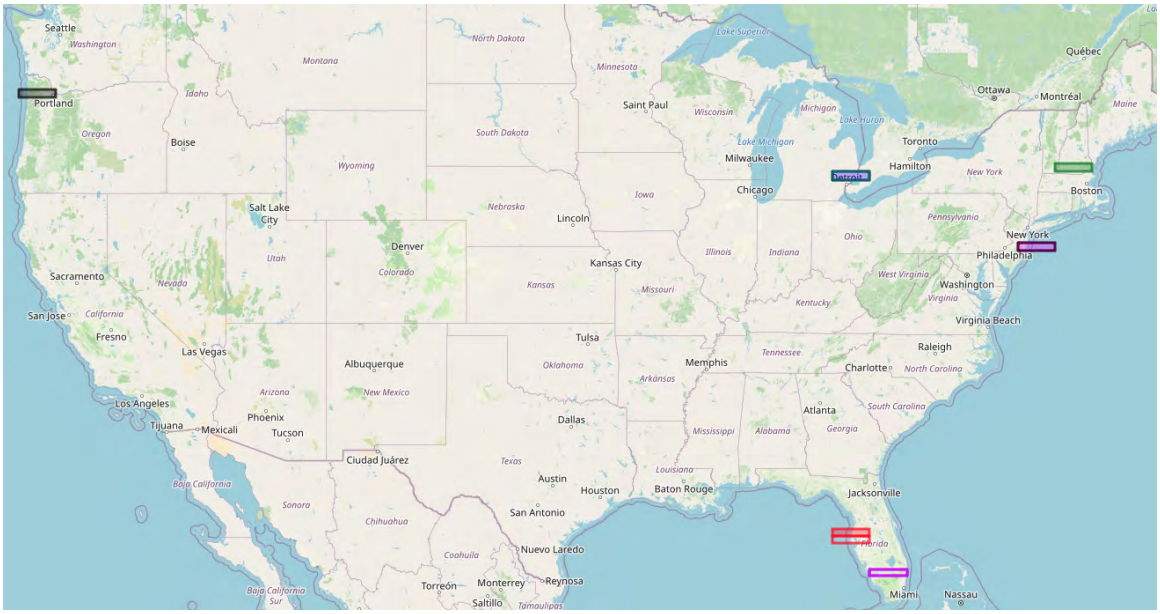

Figure 3: A figure illustrating the 5 most spatially unfair pairs of regions as determined by our method.

for a reference or comparison group. Mathematically, it may be defined as:

$$\text{Disparate Impact} = \frac{P(\text{positive outcome} \mid A = a)}{P(\text{positive outcome} \mid A = b)}$$

where $P(\text{positive outcome} \mid A = a)$ denotes the probability of receiving a positive or desirable outcome (e.g., mortgage approval, job offer in hiring) for individuals in group $A = a$, and $P(\text{positive outcome} \mid A = b)$ is the probability of receiving a positive or desirable outcome for individuals in group $A = b$. $A = a, b$ denote the different groups of the protected attribute (e.g., white/black when the protected group is race).

Like most fair-ML metrics to measure bias in given decisions, disparate impact also only looks at the outcomes and protected features. It does not take other features into account. The threshold of disparate impact that indicates concern is 0.80, modeled on the $p\%$-rule used by the U.S. Equal Employment Opportunity Commission (EEOC) to evaluate bias in hiring [5, 13]. In other words, a disparate impact of less than 0.80 indicates the presence of significant bias. The closer the disparate impact assessment is to 1, the lower the bias. Using this to assess bias in the Bank of America data, we get a disparate impact value of 0.962038. Such a high value indicates the presence of almost no bias. However, this is highly likely to be incorrect since we do not live in a perfect world. Thus, while disparate impact is an efficient tool to assess the presence of bias in many other decision-making scenarios, it is not suitable for spatial settings. Not accounting for the spatial features ignores too much information and leads to faulty assessments of fairness. Next, we shall compare to a baseline that does take spatial features into account, and we see that it does detect the presence of unfairness.

5.1.2 Spatial fairness baseline. We compare the LC-spatial-fairness framework with the previous spatial fairness work closest to ours, Sacharidis et al. [38]. For comparison with Sacharidis et al. [38] we use the LAR dataset for Bank of America for the year 2021. After the pre-processing steps detailed in 4.1 and spatial join with census data, we are left with 224,145 applications.

Like Sacharidis et al. [38], our goal is to audit this dataset for spatial fairness, and the outcome we consider is the positive rate, or in other words, the rate of approval of loans. In contrast to [38], however, we consider other attributes in addition to location. For our methodology, income will be the unprotected attribute of relevance, and race will be the legally protected attribute.

In their study, Sacharidis et al. [38] analyze results from a partitioning of 100×50. The measure of interest is the positive rate or the rate of approval for a mortgage. As a reminder, their approach aims to identify partitions where the positive class (i.e., positive outcomes or mortgage approvals) is assigned differently from the global mean. Sacharidis et al. consider only location (partition) and outcomes. In contrast, we consider income as well as race (the legally protected attribute) in addition to location and outcomes. Therefore, although the methodology of Sacharidis et al. is the closest to our framework, the two techniques still have considerable differences in their fairness assessment.

The technique of Sacharidis et al. [38] identifies 59 statistically significant partitions as spatially unfair. In other words, their framework ascertains that 59 partitions have local positive rates that are statistically significantly different from the global positive rate (which is 0.62).

In contrast, our methodology assesses the presence of significantly more spatial unfairness, and identifies 493 pairs of partitions as spatially unfair. The five pairs with the most spatial unfairness are depicted in Figure 3. The pairs of partitions are color coded, with the partition that is spatially unfair with respect to the other sharing the same color. The spatially unfair partition is colored, while the other is a transparent rectangle. Some partitions are determined to be spatially unfair in comparison to multiple partitions. Intuitively this makes sense, since partitions where there is truly the most unfairness will likely stand out in stark contrast to multiple other partitions which are similar to them in non-protected attributes and dissimilar in protected attributes yet have much better outcomes.

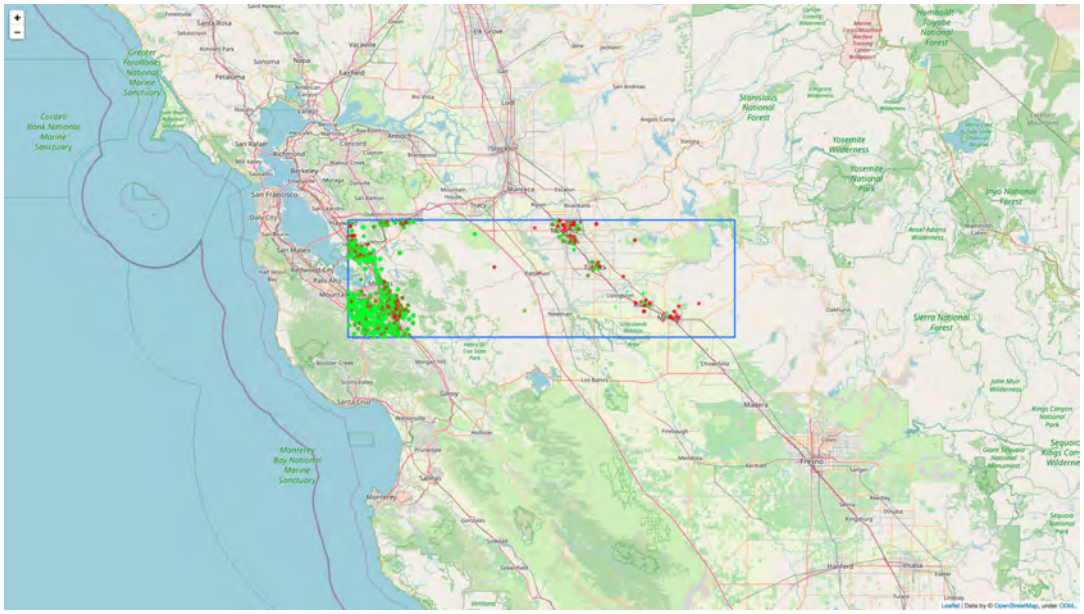

Figure 4: The most spatially unfair region as determined by Sacharidis et al. [38].

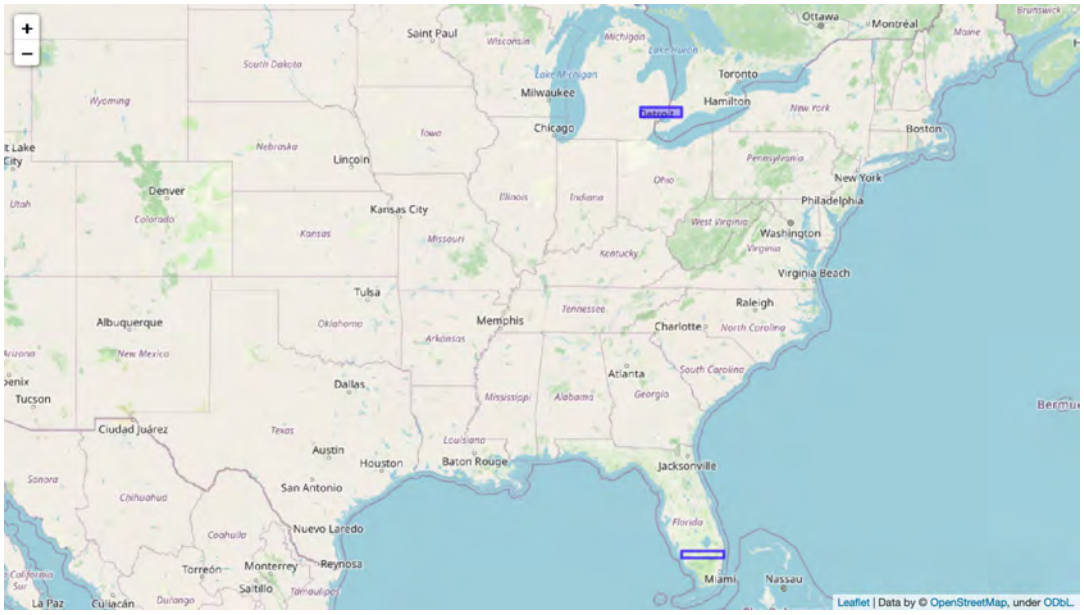

Figure 5: The most spatially unfair region as determined by our methodology.

Next, we analyze the region deemed to be most spatially unfair by each methodology. Figure 4 depicts the partition the method of Sacharidis et al. [38] determines to be the most spatially unfair. It covers a region in Northern California with a positive rate of 84% in comparison to the global positive rate of 62%. This region, however, happens to contain parts of the San Francisco Bay Area, encompassing neighborhoods such as Sunnyvale, CA and parts of Mountain View, CA which has household incomes significantly higher than the national average. Therefore, it is not surprising that the region's rate of approval of mortgages is also significantly higher than the global rate of approval. Plausible and legally valid reasons for observed differences in different locations should not be considered an instance of spatial unfairness.

In contrast, the pair of regions determined to exhibit the most statistically significant spatial unfairness is shown in Figure 5. The region in Detroit is spatially unfair with respect to the region in Florida. Both regions are have similar income. The region in Detroit is majority minority, while the region in Florida is majority white. Finally, despite similar income, the region in Detroit has a significantly lower mortgage approval rate than the region in Florida. The ability of our technique to not only identify regions with spatial unfairness but also identify the region(s) with respect to which it is unfair significantly increases its utility in real-world applications.

Sacharidis et al. [38] also find a partition near Detroit to be spatially unfair because it has a positive rate of 0.47 while the

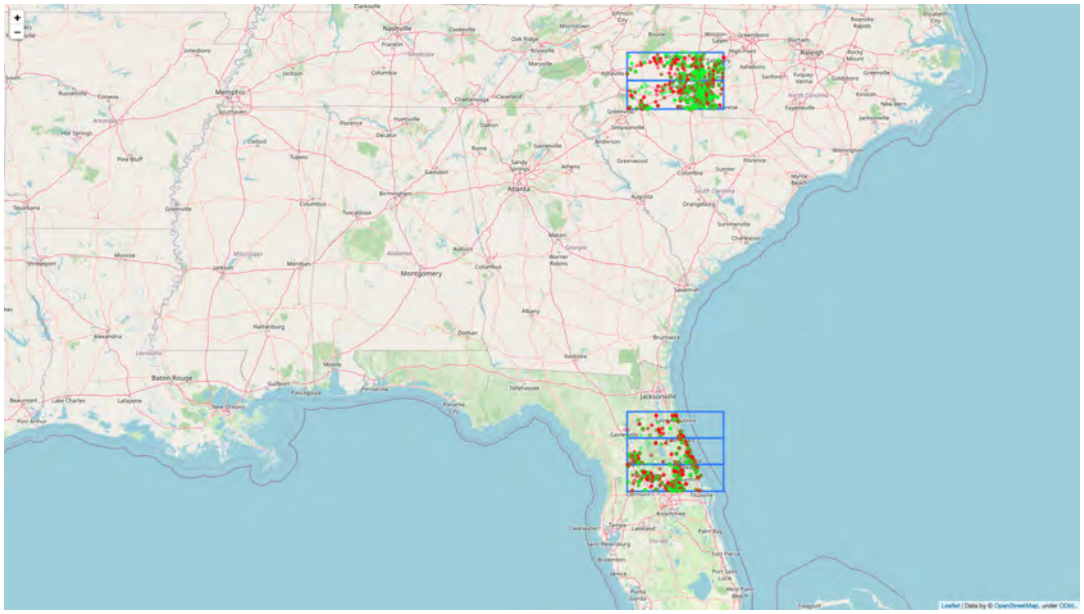

Figure 6: Spatially unfair regions flagged by our method as well as Sacharidis et al. [38].

global positive rate is 0.62. However, the partition they find unfair is separate from the partition deemed unfair by our methodology. The LC-spatial-fairness framework, in contrast, does not find that partition to be spatially unfair.

Finally, we highlight the regions determined to exhibit spatial unfairness by both methodologies. Figure 6 highlights the five partitions labeled as spatially unfair by both techniques. They are all clustered together, with three partitions in the state of Florida. The other two partitions are in the state of North Carolina, near the border with South Carolina.

5.2 Different Partitionings

In this section, we present results for both use cases for different partitioning schema. This is intended to explore how changing the resolution might affect the spatial fairness assessment, and whether a malicious adversary could potentially "game" the framework by changing the grid resolution to change the assessment from spatially unfair to spatially fair. We test for various grid resolutions, ranging from very low resolution of 10×10 to very high resolution of 100×50.

5.2.1 Mortgage Lending. We utilize the Bank of America dataset to analyze the mortgage lending use case with different partitionings. Table 2 presents the results for various grid dimensions, which range from the U.S. being divided into 100 regions to 5,000 regions.

As observed from the table, at very low resolution (the 10×10 partitioning), there is a relatively high number of region pairs deemed to be unfair. This is likely because with such huge region sizes (100 cells is only double the number of states in the country), it is likely that a region may be similar in income but dissimilar in race to many other regions. As the resolution of the partitioning increases, the number of unfair region pairs also increases, but starts to stabilize around the partitioning with size 10×50. On growing the partitioning dimensions further, the number of unfair region pairs found changes but does not vary drastically relative to the number of partitions. Simply changing the partitioning, therefore, is not a reliable way for a malicious actor to "game" the auditing process to change the assessment of the LC-spatial-fairness framework from spatially unfair to spatially fair.

Table 2: Results for the LC-Spatial Fairness framework for the mortgage application use case with Bank of America dataset for different partitionings.

Partitioning	Number of unfair region pairs
10×10	65
10×20	146
10×30	190
20×20	231
10×50	274
20×30	325
20×40	299
50×20	311
40×30	450
30×50	535
40×40	583
90×30	464
70×40	447
90×40	442
80×50	431
90×50	430
100×50	493

5.2.2 Access to healthy food. We now present experimental results for the fast food experiment with various partitionings. Table 3 presents the results for various partitions, with the U.S. divided into grids with 100 regions to 5,000 regions. As shown in the table, very few unfair regions are found at very low resolution (such as 10×10) and at very high resolution (such as 100×50). This is likely because at high resolution, say 100×50, the data gets too sparse. As a reminder, we obtain only 106,091 fast food places across the country after combining outlets of the top 15 most popular fast food chains in the U.S. In contrast, we have

Table 3: Results for the LC-Spatial Fairness framework for the access to healthy food use case with SafeGraph dataset for different partitionings.

Partitioning	Number of unfair region pairs
10×10	7
10×20	22
10×30	42
10×40	53
20×20	41
10×50	51
30×20	73
40×20	103
50×50	18
90×50	13
70×40	14
100×30	15
90×50	13
100×50	5

many times the amount of data for the mortgage lending scenario. The 106,091 fast food outlets spread across 5,000 regions would be incredibly sparse. In the case of very low resolution, the data is so aggregated that most differences are not statically significant. As the resolution of the partitioning increases, the number of unfair regions found usually increases. For partitionings until 50×50, the LC-spatial-fairness framework assesses approximately 10-14% of the total number of regions to be unfair. At the partitioning of size 50×50, the number of unfair regions starts to drop significantly, likely because at this point the resolution becomes too fine. The 106,091 fast food outlets divided between 2500 regions (the 50×50 partition) would mean an average of only 42 fast food outlets per region, which is not significant. For more reasonable partitioning resolutions, the number of unfair regions tends to increase on average as the partition resolution becomes finer; however, the change is not drastic. For instance, from the second-lowest resolution partition (10×20) to a higher resolution of 40×20, the framework identifies approximately 10–14% of regions as unfair. Thus, it would be hard for an adversary to be able to successfully exploit the LC-spatial-fairness framework by simply changing the partitioning.

These results highlight two key observations. The first observation is that changing the resolution does indeed affect the results, but this is expected–similar to the impact of altering the similarity metric. Changing the grid resolution significantly alters the characteristics (both protected and unprotected features) of each region. But the LC-spatial-fairness framework still detects unfair regions with respect to the new resolution. The second observation is that our method is agnostic to MAUP-resistence when it comes to grid resolution. To be MAUP-resistant does not mean finding exactly the same number of unfair regions each time or consistently identifying a specific region as unfair. Rather, it means that the framework can remain robust against intentional manipulation of the partitioning. Specifically, an adversary should not be able to alter the resolution or partitioning in a way that systematically conceals unfairness or falsely indicates fairness. While changes in resolution naturally affect the characteristics of regions and the results, the LC-spatial-fairness framework still reliably identifies unfair regions according to the criteria set for the new resolution.

Table 4: Results for the LC-Spatial Fairness framework with statistical parity as the dissimilarity metric for the mortgage application use case with Bank of America dataset for different partitionings.

Partitioning	Number of unfair region pairs
10×10	69
10×20	150
10×30	174
20×20	290
10×50	316
20×30	281
20×40	350
50×20	784
40×30	553
30×50	532
40×40	539
90×30	417
70×40	644
90×40	837
80×50	674
90×50	684
100×50	740

5.3 Use case with a different metric

To demonstrate the flexibility of our LC-spatial-fairness framework in incorporating various (dis)similarity metrics tailored to different application contexts, we now present results for the mortgage lending scenario using an alternative dissimilarity metric. This is intended to showcase how different similarity metrics can be integrated into our framework for novel application settings. Specifically, we utilize the widely-recognized fair-ML metric of statistical parity to assess the dissimilarity of the protected attribute and conduct the mortgage lending experiment using the Bank of America dataset. The statistical parity metric is defined as follows:

Definition 5.2. Statistical parity evaluates whether a desirable or positive outcome is distributed equitably across different protected groups. In other words, statistical parity examines whether the proportion of individuals receiving a positive outcome is the same across all groups. Mathematically:

$$P(Y = 1 \mid A = a) = P(Y = 1 \mid A = b), \quad \forall a, b \in \text{Protected Groups}$$

In our context, the desirable outcome is the approval of a mortgage application and the protected groups are racial groups. The experimental results are presented in Table 4.

Table 4 shows that incorporating statistical parity as a metric gives somewhat similar results as in Table 2 up until the partition with resolution 20×40. As the partitions get finer, statistical parity leads to an assessment of greater unfairness.

6 CONCLUSION

This work introduces the LC-spatial-fairness framework to assess for legally-compliant spatial fairness and identify regions which exhibit significant unfairness. By considering location in consideration with relevant non-protected attributes and legally protected attributes, we bridge the gap between traditional fair-ML approaches and previous spatial fairness work in a manner that would withstand legal scrutiny. Government agencies and

non-profits can utilize our framework to identify instances of spatial unfairness, while companies can make use of it to ensure they do not contribute to it unintentionally.

While the LC-spatial-fairness framework proposed in this paper represents a significant step forward for assessing spatial fairness, it is not without limitations. The first limitation reflects a broader challenge inherent to the discipline of fairness in artificial intelligence: there is no ground truth for what it means to be "fair." Thus, any framework to measure fairness would require the definition of a metric to measure "similarity" of individuals or groups to assess whether their outcomes are "fair enough." Since there is no universally agreed-upon definition of fairness, measuring it always involves choosing a metric that tries to capture the concept. As no single metric is perfect, selecting the right one for a specific task requires careful consideration. A second, and closely related, limitation is that defining an appropriate (dis)similarity metric for a specific application context may demand substantial subject matter expertise, adding complexity to the implementation of the framework. An independent regulatory body could be tasked with deciding on the appropriate fairness metric for each industry (for example, for mortgage loans in the credit industry), perhaps in tandem with consultations with subject matter experts as needed.

ACKNOWLEDGMENTS

This research has been funded in part by NSF grants CNS-2125530 and 2427150 (PI: Horn). Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the sponsors such as the NSF.

REFERENCES

[1] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Can you trust the trend? discovering simpson's paradoxes in social data. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 19–27.

[2] Julia Angwin and Jeff Larson. 2017. California requires auto insurers to adjust rates after CR, ProPublica investigation. https://www.consumerreports.org/consumer-protection/california-requires-auto-insurers-adjust-rates-after-cr-propublica-investigation/

[3] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2017. Minority neighborhoods pay higher car insurance premiums than white areas with the same risk. *ProPublica, April* 5 (2017), 2017.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[5] Dan Biddle. 2017. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[7] Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *Comput. Surveys* 56, 7 (2024), 1–38.

[8] Raj Chetty, Nathaniel Hendren, and Lawrence F Katz. 2016. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review* 106, 4 (2016), 855–902.

[9] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. 2021. An overview of fairness in clustering. *IEEE Access* 9 (2021), 130698–130720.

[10] Yushun Dong, Oyku Deniz Kose, Yanning Shen, and Jundong Li. 2023. Fairness in Graph Machine Learning: Recent Advances and Future Prospectives. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5794–5795.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[12] Experian. 2021. Global Insights Report September-October 2021. *Experian* (2021). https://www.experian.com/blogs/global-insights/wp-content/uploads/2021/11/GIRw4_21_Final.pdf

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[14] Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).

[15] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. 2018. Online learning with an unknown fairness metric. *Advances in neural information processing systems* 31 (2018).

[16] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[17] Junia Howell and Elizabeth Korver-Glenn. 2021. The increasing effect of neighborhood racial composition on housing values, 1980–2015. *Social Problems* 68, 4 (2021), 1051–1071.

[18] David Ingold and Spencer Soper. 2016. Amazon doesn't consider the race of its customers. should it? https://www.bloomberg.com/graphics/2016-amazon-same-day/#:~:text=The%20most%20striking%20gap%20in,on%20all%20sides%20are%20eligible.

[19] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems* 29 (2016).

[20] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40–40.

[21] Julie Lee. 2023. The future of AI in lending. *Experian Insights* (Jan 2023). https://www.experian.com/blogs/insights/future-ai-lending/

[22] TurnKey Lender. 2020. How machine learning is used in the lending industry. *TurnKey Lender* (Feb 2020). https://www.turnkey-lender.com/blog/how-machine-learning-is-used-in-the-lending-industry/

[23] Huiping Li, Harrison Campbell, and Steven Fernandez. 2013. Residential segregation, spatial mismatch and economic growth across US metropolitan areas. *Urban Studies* 50, 13 (2013), 2642–2660.

[24] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).

[25] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday* (2020), 189–202.

[26] David Manley. 2021. Scale, aggregation, and the modifiable areal unit problem. In *Handbook of regional science*. Springer, 1711–1725.

[27] Erin McDowell. 2024. The 15 biggest fast-food chains in the US, ranked. https://www.businessinsider.com/biggest-fast-food-chains-in-the-us-ranked-2024-7

[28] 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.

[29] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. conf. fairness accountability transp., new york, usa*, Vol. 1170. 3.

[30] Jonathan K Nelson and Cynthia A Brewer. 2017. Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem. *Cartography and Geographic Information Science* 44, 1 (2017), 35–50.

[31] U.S. Department of Agriculture. [n. d.]. Food Access. https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/

[32] California Department of Food and Agriculture. 2012. Improving Food Access in California. https://www.cdfa.ca.gov/exec/public_affairs/pdf/ImprovingFoodAccessInCalifornia.pdf

[33] Jeff Ostrowski. 2024. 10 largest mortgage lenders in the U.S. https://www.bankrate.com/mortgages/largest-mortgage-lenders/#:~:text=ThereâĂŹs%20a%20new%20No.,most%20active%20home%2Dloan%20lenders.

[34] Akshat Pandey and Aylin Caliskan. 2021. Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 822–833.

[35] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems* 34 (2021), 25944–25955.

[36] Richard Rothstein. 2015. The racial achievement gap, segregated schools, and segregated neighborhoods: A constitutional insult. *Race and social problems* 7 (2015), 21–30.

[37] Richard Rothstein. 2017. *The color of law: A forgotten history of how our government segregated America*. Liveright Publishing.

[38] Dimitris Sacharidis, Giorgos Giannopoulos, George Papastefanatos, and Kostas Stefanidis. 2023. Auditing for Spatial Fairness. *Proceedings 26th International Conference on Extending Database Technology, EDBT 2023* (2023).

[39] Safegraph. 2024. Safegraph places documentation. https://docs.safegraph.com/docs/places#section-pattern

[40] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair pca: One extra dimension. *Advances in neural information processing systems* 31 (2018).

[41] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[42] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2020. How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence* 283 (2020), 103238.

[43] Nripsuta Ani Saxena, Wenbin Zhang, and Cyrus Shahabi. 2024. Spatial Fairness: The Case for its Importance, Limitations of Existing Work, and Guidelines for Future Research. *arXiv preprint arXiv:2403.14040* (2024).

[44] Sina Shaham, Gabriel Ghinita, and Cyrus Shahabi. 2022. Models and mechanisms for spatial data fairness. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 16. NIH Public Access, 167.

[45] Wolters Kluwer Compliance Solutions. 2023. Understanding HMDA reporting: A comprehensive guide for lenders. https://www.wolterskluwer.com/en/expert-insights/understanding-hmda-reporting-a-comprehensive-guide-for-lenders#:~:text=HMDA%20reporting%20entities,submit%20their%20reports%20to%20HMDA.

[46] Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie H Morgenstern, and Santosh Vempala. 2019. Multi-criteria dimensionality reduction with applications to fairness. *Advances in neural information processing systems* 32 (2019).

[47] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

[48] Louis Lee Woods. 2012. The Federal Home Loan Bank Board, redlining, and the national proliferation of racial lending discrimination, 1921–1950. *Journal of Urban History* 38, 6 (2012), 1036–1059.

[49] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12208–12216.