JMLR: Workshop and Conference Proceedings 1:1–13, 2021 Algorithmic Fairness through the Lens of Causality and Robustness

Detecting Bias in the Presence of Spatial Autocorrelation

Subhabrata Majumdar* *Applied ML Research, Splunk* SMAJUMDAR@SPLUNK.COM
Cheryl Flynn *Data Science and AI Research, AT&T* CFLYNN@ATT.COM
Ritwik Mitra* RIT.STAT@GMAIL.COM

Abstract

In spite of considerable practical importance, current algorithmic fairness literature lacks technical methods to account for underlying geographic dependency while evaluating or mitigating bias issues for spatial data. We initiate the study of bias in spatial applications in this paper, taking the first step towards formalizing this line of quantitative methods. Bias in spatial data applications often gets confounded by underlying spatial autocorrelation. We propose hypothesis testing methodology to detect the presence and strength of this effect, then account for it by using a spatial filtering-based approach—in order to enable application of existing bias detection metrics. We evaluate our proposed methodology through numerical experiments on real and synthetic datasets, demonstrating that in the presence of several types of confounding effects due to the underlying spatial structure our testing methods perform well in maintaining low type-II errors and nominal type-I errors.
Keywords: ML fairness, spatial fairness, spatial equity, spatial filtering, hypothesis testing

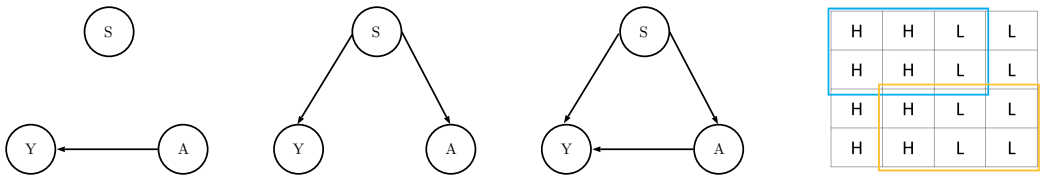
1. Introduction

In recent literature, fairness concerns have been studied in machine learning (ML)-based criminal justice systems, credit scoring, and facial recognition among others. In comparison, ML fairness in problems involving *spatial data* has been studied to a relatively lesser extent. Idiosyncrasies associated with spatial data such as spatial autocorrelation can introduce an underlying dependence structure in the data. As a result, existing fairness definitions and techniques designed for independent data may no longer apply. In this paper, we present a first attempt towards addressing this gap by formalizing a quantitative approach for detecting bias in spatial data applications.

Deployment locations of services such as bike-share locations, cell towers, and user-stops for location-based gaming may need to be constrained by minimum and maximum distances between adjacent pairs or groups of locations. Potential reasons for imposing such constraints may range from an overall goal for optimal coverage and accessibility, to limitations of the underlying technology (e.g. 5G cellular networks). Such service locations, especially when deployed through for-profit corporations, may be selected using ML in a manner that attempts to maximize engagement with or subscription to the service while satisfying any necessary spatial constraints. To assess the bias in such applications, one might consider the association between selected locations and their corresponding demographics. However, for historic and societal reasons, distribution of different demographic groups is also usually spatially autocorrelated (Williams and Emamdjomeh, 2018). These underlying spatial patterns in the data can result in a confounding effect that can invalidate standard bias detection procedures.

* Work done while at Data Science and AI Research, AT&T.

© 2021 S. Majumdar, C. Flynn & R. Mitra.



(a) Confounding effects of underlying spatial structure (b) Simple Tile Example

Figure 1: In presence of spatial autocorrelation, associations between the outcome feature Y and sensitive feature A get confounded with potential associations with S (panel a). A simple tile example demonstrates that this effect translates to increased likelihood of detecting bias in presence of spatial autocorrelation (panel b).

Consider the simple tile setup in Figure 1(b). Each tile is assigned a value of High (H) or Low (L) to represent a demographic attribute such as income or percent minority. Suppose we randomly select 6 tiles and compute the disparate impact (DI) bias metric, where

$$f = \frac{|Selected \cap H|/|H|}{|Selected \cap L|/|L|}, \quad DI = \max(f, 1/f).$$

Assuming $DI > 1.2$ indicates a biased selection process, let us compare the likelihood of detecting bias in two scenarios. In scenario 1, we impose no spatial constraints. Then the number of H and L tiles follows a hypergeometric distribution and the probability distribution of DI can be computed exactly, where bias is detected 60.84% of the time. In scenario 2, we only select 2×3 rectangles. There choices can have either four or two H tiles (blue or yellow border in Fig. 1(b), respectively), so that $DI = 2$ always, and bias is detected with 100% probability. Thus, the combination of the presence of spatial autocorrelation in the demographic attribute and the selected tiles alters the distribution of the bias metric.

Formalizing the above notion, in the presence of a spatial dependency pattern S between partitions of a geographic area, the estimation of any association or cause-effect relationship between a target variable, Y, and a sensitive attribute, A, may be confounded by S. As demonstrated in the causal diagram in Fig. 1(a), there are multiple subcases of the above scenario, which may result in overestimation of the degree of association between Y and A through any spurious correlation either or both of them might have with S. From left to right, these diagrams capture three cases: (1) direct relationship between Y and A, (2) indirect relationship between Y and A through S, and (3) direct relationship between Y and A with additional association through S. Bias detection methods should account for these relationships to avoid a potentially erroneous conclusion.

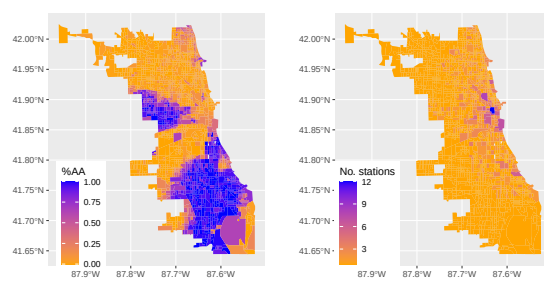


Figure 2: Spatial distribution of bikeshare stations in Chicago (left) is largely complementary to that of the percentage of African Americans (right) at census tract level.

Our contributions We provide a framework to evaluate bias concerns in spatial data, specifically in scenarios where confounding effects of spatial autocorrelation influence the outcomes of interest, sensitive attribute values, or both. Under this setup, we aim to evaluate a number of scenarios regarding the effect of an underlying spatial structure: (a) the detection of a common spatial effect, (b) determining whether these effects are equivalent in magnitude, and (c) testing for association between Y and A while adjusting for this spatial effect. We propose methods that provide answers to them (Sec. 2), and evaluate these methods in a number of data settings, including the Chicago bikeshare data in Fig. 2.

Related work Spatial fairness is an important problem with a myriad of applications that affect the populace in direct and tangible ways. Spatial distribution of resources impact the accessibility (or lack there of) of crucial private and public goods, such as housing, transportation, schools, and healthcare. Historically, systemic, intentional and even unintentional biases have affected distribution of these resources. The oft-noted illegal practice of 'redlining' where credit for housing is often denied to minorities can stem from banks focusing on broad information, such as the neighborhood of the applicant, which ends up disadvantaging those from poorer locations (Corbett-Davies and Goel, 2018). Fairness concerns in bikeshare problems is well-documented. Ogilvie and Goodman (2012) noted inequities in who were accessing London's Barclays Cycle Hire (BCH) scheme—most users of the system tended to be male and from privileged background; see also Goodman and Cheshire (2014); NACTO (2015) and Fig. 2. Availability and access to primary care facilities are essential to the overall well-being of a local community (Guagliardo, 2004). Spatial inequality in this context has also gotten attention in literature. For example, Jin et al. (2015) studied the accessibility and found clear distinctions between rural/poorer and urban/richer areas in Deqing county in China. More recently, evidence has surfaced across multiple geographic areas in the United States regarding racial disparities in COVID-19 vaccine enrollment and distribution attributed to spatial factors (Heath, 2021; Marin, 2021). As a final example, Colley et al. (2017) found that the popular augmented reality game Pokémon GO disadvantaged urban locations with a higher percentage of minority population, as compared to predominantly white neighborhoods.

In spite of the very real needs, to the best of our knowledge there are no known methods to detect and mitigate spatial fairness issues. Till now, most algorithmic fairness methods (Mehrabi et al., 2019) deal with obtaining *point estimates* of any association between the outcome Y and sensitive attribute A, or mitigating this association while estimating Y, and not the *significance* of such association. While some very recent papers do propose hypothesis tests for evaluating fairness (Black et al., 2020; DiCiccio et al., 2020), they assume the data samples analyzed to be independent. As we have seen in the motivating example, the presence of an underlying dependency pattern can result in such tests conflating this dependence with the presence of actual demographic bias.

2. Methodology

Consider n location units $\{l_i\}_{i=1}^n$, the spatial dependency between which is quantified by a weight matrix $\mathbf{W}$. For location unit i, we have non-sensitive input features $\boldsymbol{X}_i \in \mathbb{R}^p$– giving rise to data matrix $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n \times p_1}$, a (discrete or continuous) output feature $\boldsymbol{y} \in \mathbb{R}^n$ indicating the allocation or presence of a resource at location i, and $\mathbf{A} =$

$(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)^T \in \mathbb{R}^{n \times p_2}$ as the data on sensitive input features. The realized value of a random vector is denoted by lowercase; e.g. $\boldsymbol{X}_i$, takes the value $\boldsymbol{x}_i$. For simplicity, we assume a single sensitive attribute A, i.e. $p_2 = 1$, hereafter.

Under the above setup, we aim to answer three questions: (Q1) Are both Y and A significantly associated with the common weight matrix $\mathbf{W}$? (Q2) Given that they are, are their strength of associations with $\mathbf{W}$ similar? (Q3) Can we adjust for spatial association to measure the true degree of association between Y and A?

2.1. Testing for common autocorrelation

Moran's I (Moran, 1950) is a well-known statistic to test for global dependence in spatial data. Given realizations $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ of a continuous random variable Z and a weight matrix $\mathbf{W} = ((w_{ij}))$, Moran's I is defined as

$$I = \frac{n}{\bar{w}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}; \quad \bar{w} = \frac{1}{2} \sum_{i,j} (w_{ij} + w_{ji}), \bar{y} = \frac{1}{n} \sum_i y_i.$$

Under the null hypothesis (Appendix A.1) that no spatial autocorrelation exists, a centered and scaled version of I is asymptotically standard Gaussian (Shapiro and Hubert, 1979; O'Neil and Redner, 1993). Recently, Lee and Ogburn (2020) generalized the Moran's I for realizations of a categorical random variable, and proposed an analogous test statistic. When observations are binary, standardized versions of Moran's I and this statistic above are equivalent, and asymptotically standard Gaussian (Lee and Ogburn, 2020). Asymptotic Gaussian tail-bounds or permutation tests can be used to obtain p-values of either test statistic.

Assume that the standardized Moran's I calculated using the size-n sample $\boldsymbol{y}$ (or $\boldsymbol{A}$) and $\mathbf{W}$ follow the (non-asymptotic) location family distribution $F(\cdot, \mu_y)$ (or $F(\cdot, \mu_a)$) with mean parameter μ_y (or μ_a). Then, to answer the question (Q1) above, we test the following:

$$H^1_{0y} : \mu_y = 0 \text{ vs. } H^1_{ay} : \mu_y \neq 0 \quad \text{and} \quad H^1_{0j} : \mu_a = 0 \text{ vs. } H^1_{aa} : \mu_a \neq 0.$$

Specifically, we are interested in knowing whether the scenario when both the alternate hypotheses hold, i.e. both y and A exhibit significant autocorrelation. Thus we test for the combined null $H^1_0 = H^1_{0y} \bigcup H^1_{0a}$ against the combined alternative $H^1_a = H^1_{ay} \bigcap H^1_{aa}$. To answer question (Q2), we test $H^2_0 = H^{21}_0 \bigcup H^{22}_0$ against $H^2_a = H^{21}_a \bigcap H^{22}_a$, where

$$H^{21}_0 : \mu_y - \mu_a \geq \delta \text{ vs. } H^{21}_a : \mu_y - \mu_a < \delta \quad \text{and} \quad H^{22}_0 : \mu_y - \mu_a \leq -\delta \text{ vs. } H^{22}_a : \mu_y - \mu_a > -\delta,$$

for some fixed $\delta > 0$, and with the additional assumption that $\mu_y \neq 0, \mu_a \neq 0$. Guidance for useful values of δ may be determined by target thresholds of bias metrics (Appendix A.2), such as the disparate impact thresholds of $(0.8, 1.2)$ based on the EEOC four-fifths rule (EEOC, 1979). We take $\delta = 1$ in our experiments to account for small differences that may occur under null between the two standardized Moran's I statistics due to sampling noise.

Working under the above setup–known as the Intersection-Union (IU) principle (Berger, 1982; Berger and Hsu, 1996)–we combine test statistics for each of the pairs of sub-hypotheses above using the concept of *Pivotal Parametric Products*:

Definition 1 (SenGupta (2007)) *For the problem of testing for a (possibly vector-valued) parameter θ using a union null $H_0 \equiv \bigcup_{s=1}^{S} H_{0s}$, a Pivotal Parametric Product (P3) is defined as any function $\eta \equiv g(\theta)$ such that H_0 holds if $\eta = 0$.*

Testing for $H_0 : \eta = 0$ can be done using an (unbiased or consistent) estimator of η. In our context, we define the P3 functions to test for H_0^1 vs. H_a^1 and H_0^2 vs. H_a^2 as $P_1 = \min(|\mu_y|, |\mu_a|)$; $P_2 = \left|(\mu_y - \mu_a)^2 - \delta^2\right|$. Assuming that the individual Moran's I test statistics are called I_y and I_a, respectively for y and A, the corresponding test statistics will be: $T_1 = \min(|I_y|, |I_a|)$; $T_2 = \left|(I_y - I_a)^2 - \delta^2\right|$. Denoting the lower-$p$ ($p \in [0, 1]$) tail of the probability distribution of a statistic T by T_p, rejection regions at level α for the above tests are characterized by large values of T_1 and small values of T_2, specifically by the sets $\{T_1 \geq T_{1,1-\alpha}\}$ and $\{T_2 \leq T_{2,\alpha}\}$, respectively.

We use permutation test (Appendix B.1) to obtain p-values for above tests. We construct the null distributions for each test statistic by calculating them over randomly permuted samples, then compute the tail probabilities of the statistics above with respect to the respective distributions.

2.2. ESF adjustment

Eigenvector Spatial Filtering (ESF) (Dray et al., 2006; Griffith, 2004a; Griffith et al., 2019) is a popular method to remove underlying spatial dependencies from spatial observations. It fits a linear regression on that feature, with eigenvectors of $\mathbf{W}$ and a number of covariates as input features. Residuals from this regression can be used as dependency-free analogues of the original observations for further analysis. Formally, this *spatial autoregression* model assumes $\boldsymbol{y} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$; $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n), \sigma > 0$, where $\rho \in [-1, 1]$ denotes spatial autocorrelation, and $\boldsymbol{\beta} \sim \mathbb{R}^p$ quantifies the (linear) effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$. Given that the spatial dependency in $\boldsymbol{Y}$ is entirely due to the top k few eigenvectors of $\boldsymbol{W}$, Consider now the spectral decomposition $\mathbf{W} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T$, where $\mathbf{E} \in \mathbb{R}^{n\times n}$ is orthogonal and $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ is diagonal. Using straightforward algebra (Griffith et al., 2019), this model simplifies to

$$\boldsymbol{y} = \rho\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T\boldsymbol{y} + \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \simeq \rho\mathbf{E}_k\boldsymbol{\gamma}_k + \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\zeta}, \tag{1}$$

where $\boldsymbol{\gamma}_k$ estimates the effect of the top k eigenvectors, and $\boldsymbol{\zeta} \sim N(0, \sigma_1^2\mathbf{I}_n), \sigma_1 > 0$. The least square estimates of $\boldsymbol{\gamma}_k, \boldsymbol{\beta}$ can be plugged in to obtain a 'sanitized', decorrelated version of $\boldsymbol{y}$: $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \rho\mathbf{E}_k\widehat{\boldsymbol{\gamma}}_k + \mathbf{X}\widehat{\boldsymbol{\beta}}$.

To address (Q3) above, we use ESF to learn features that capture the spatial structure in Y and A, then use these features to adjust the bias detection procedure. Depending on whether the feature of interest is continuous or discrete, we use a squared error or logistic loss–with ℓ_1 penalization to automatically choose the best subset of eigenvectors (Appendix B.2). To choose the optimal tuning parameter λ, we use Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

3. Experiments

3.1. Synthetic data: testing for common autocorrelation

We take the centroid locations of $n = 2080$ census block groups in the city of Chicago, construct a weight matrix $\mathbf{W}$ using the inverse of distances between tracts i, j as the weight

$w_{ij}, 1 \leq i, j \leq n$, and generate samples of the random variables $(Y, \boldsymbol{X}, A)$. When required, we use 1000 permutations to generate approximate null distributions, and compute all empirical rejection rates using 1000 Monte Carlo runs. We evaluate the performance of the IU-combination tests in three different scenarios.

Association by autocorrelation We first focus on the case where both the response Y and the sensitive attribute A are independently correlated with $\mathbf{W}$. We start with standard Gaussian errors $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim N(0, \mathbf{I}_n)$, and given a value of spatial autocorrelation ρ, generate spatially lagged features: $\boldsymbol{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}_1, \boldsymbol{a} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}_2$.

We repeat the experiment for $\rho = 0, 0.1, \ldots, 0.9$, and different sample sizes, by either randomly choosing 1000 and 1500 census tracts or using all 2080 census tracts, to quantify how the degree of spatial autocorrelation and sample size affect the power (i.e. Type-II error) of the testing procedure.

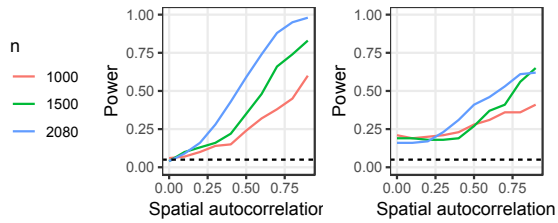


Figure 3: Powers for T_1 (left) and T_2 (right) under association by correlation. Dotted lines indicate size $\alpha = 0.05$.

We summarize the results in Fig. 3. As ρ increases, the test T_1 is able to detect the existence of a non-zero ρ with higher power, and also maintains nominal size, as the value of T_1 is exactly 0.05 for $\rho = 0$. For T_2, note that rejection implies inferring the equality of ρ across $\boldsymbol{y}$ and $\boldsymbol{a}$. Thus Fig. 3 indicates that our permutation test is also able to infer this with higher power as ρ grows larger. Note that for T_2 the alternate hypothesis implies inferring the equality of spatial autocorrelation. This is the case here, so the test rightly maintains rejection rates above 0.05 for all values of ρ.

Association and autocorrelation In this situation Y and A are associated independently, and one or both are spatially autocorrelated. To represent such scenario, we generate synthetic data using the two settings, starting with standard Gaussian errors $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2$ as before:

$$\boldsymbol{a}_0 = \boldsymbol{\epsilon}_2, \boldsymbol{y} = 5\boldsymbol{a}_0 + \boldsymbol{\epsilon}_1, \boldsymbol{a} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{a}_0, \tag{2}$$

$$\text{and} \quad \boldsymbol{a} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}_2, \boldsymbol{y} = 5\boldsymbol{a} + \boldsymbol{\epsilon}_1. \tag{3}$$

In the first case only the sensitive attribute is spatially autocorrelated, while in the second situation both Y and A are.

We summarize the results in Fig. 4. The statistic T_1 geared towards finding evidence of a common level of spatial autocorrelation (rightly) maintains nominal size across different values of ρ in the first case above, and increasingly becomes more powerful for higher ρ in the second case. The statistic T_2 maintains nominal rejection rates for moderate to high ρ in the first case, giving evidence that there is a mismatch of autocorrelation magnitudes for Y and A. In the second case, it however maintains higher rejection rates.

Proxy autocorrelation We now consider the third scenario, of Y being associated with *non-sensitive* attributes, i.e. components of $\boldsymbol{X}$, but not A. However X and A are spatially autocorrelated with the same $\mathbf{W}$: $\boldsymbol{x} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}_2, \boldsymbol{y} = 5\boldsymbol{x} + \boldsymbol{\epsilon}_1; \boldsymbol{a} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}_2$. The results for this setting (not shown) are very similar to Fig. 3. This is because setting 3 can

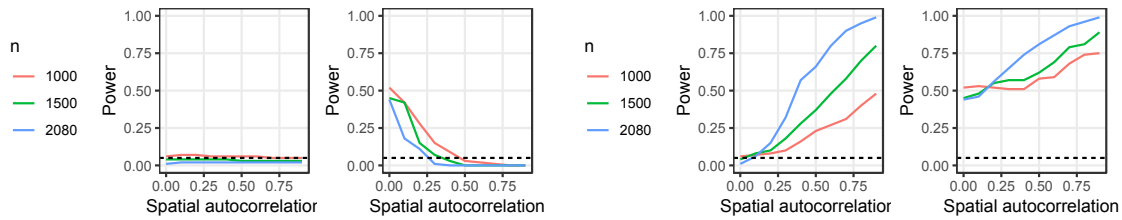
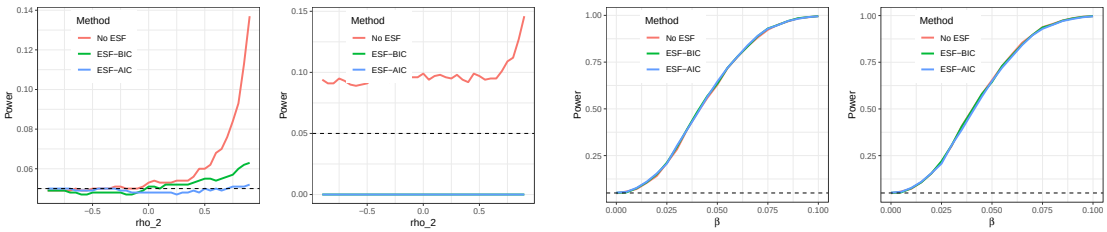


Figure 4: Powers for T_1 (left) and T_2 (right) under association and correlation. Top and bottom rows correspond to settings (2) and (3), respectively.



(a) Association by autocorrelation (b) Association and autocorrelation

Figure 5: In each subfigure, rejection rates are given for the ESF-adjusted tests for (left panels) continuous or (right panels) discrete Y, A.

be seen as the noisy version of setting 1, where we were essentially testing for a common source of autocorrelation between $\boldsymbol{x}$ and $\boldsymbol{a}$, as compared to $\boldsymbol{y}$ and $\boldsymbol{a}$ here.

3.2. Synthetic data: adjusting for autocorrelation

We consider two different values of the underlying spatial autocorrelation, and evaluate the ESF-based testing methods for discrete and continuous attributes across above settings.

Association by autocorrelation For the continuous setup, we consider the model $\boldsymbol{y} = (\mathbf{I} - \rho_1 \mathbf{W})^{-1}\boldsymbol{\epsilon}_1, \boldsymbol{a} = (\mathbf{I} - \rho_2 \mathbf{W})^{-1}\boldsymbol{\epsilon}_2$, and compare error rates of asymptotic tests using the correlation coefficient before and after ESF-adjustment. For the discrete setup, we take $\boldsymbol{y} \leftarrow \mathbb{I}(\boldsymbol{y} > 0), \boldsymbol{a} \leftarrow \mathbb{I}(\boldsymbol{a} > 0)$, and compare using the disparate impact metric. In this case we obtain the null distribution using a permutation procedure. In both cases, we consider the range of values $\rho_1 = 0.9, \rho_2 = -0.9, -0.85, \ldots, 0.85, 0.9$. We summarize the results in Fig. 5(a). Without ESF correction, rejection rates stay above the nominal 0.05 level across all values of ρ_2, but come down below nominal level after correction. The results due to proxy autocorrelation are largely similar to this setting, so we omit those results.

Association and autocorrelation In this setup, we consider the case when both Y, A are discrete. We consider the two subcases:

$$\boldsymbol{a}_0 = \boldsymbol{\epsilon}_2, \boldsymbol{a} = \mathbb{I}((\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{a}_0 > 0), \boldsymbol{y} = \mathbb{I}(\beta\boldsymbol{a}_0 + \boldsymbol{\epsilon}_1 > 0),$$
$$\text{and} \quad \boldsymbol{a} = \mathbb{I}((\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}_2 > 0), \boldsymbol{y} = \mathbb{I}(\beta\boldsymbol{a} + \boldsymbol{\epsilon}_1 > 0).$$

$\mathbf{W}$	$\mathbb{I}$(no. stations ≥ 1)	%AA	Med. inc.	$\mathbb{I}$(%AA ≥ 30)	$\mathbb{I}$(Med. inc. $\geq$50k)
Exp, $V = 1$	96.06 (0)	594.88 (0)	211.94 (0)	529.95 (0)	197.46 (0)
Exp, $V = 0.1$	107.2 (0)	516.73 (0)	234.19 (0)	460.11 (0)	195.15 (0)
Exp, $V = 0.01$	61.01 (0)	213.45 (0)	170.67 (0)	192.4 (0)	123.44 (0)
Exp, $V = 0.001$	32.41 (0.001)	34.93 (0)	34.95 (0)	34.69 (0)	34.11 (0)
1-hop neighbor	2.28 (0.011)	10.22 (0)	5.43 (0)	9.2 (0)	4.85 (0)
2-hop neighbor	1.87 (0.032)	12.67 (0)	8.59 (0)	11.66 (0)	6.11 (0)

Table 1: Standardized Moran's I statistics for the Chicago bikeshare data, with permutation test p-values in parentheses.

We present the results in Fig. 5(b), calculated on a range of values for β, and fixing $\rho = 0.9$—with outputs for the first case on the left and second case on the right. It is evident from the results that the rejection rates remain the same pre- and post-filtering as expected.

3.3. Chicago Bikeshare data

The Chicago bikeshare data (Chicago Data Portal, 2021) contains latest information on the location and other neighborhood-related attributes of all bikeshare stations in the city of Chicago. We join this data with the 2018 ACS Census data to obtain two sensitive features at census tract level—percentage of African American population (%AA), and median income. As we had seen in Fig. 2, there is evidence of potential spatial fairness issues in this context. We perform Moran's I-based tests for a number of features to determine their dependence on the weight matrix $\mathbf{W}$. We consider several choice of $\mathbf{W}$: spatial cross-covariance matrices with exponential covariance function having scaling parameter $V = 10^i; i = 0, -1, -2, -3$, and adjacency matrices indicating 1-hop and 2-hop neighbors of block groups. The results in Table 1 show that both sensitive features and the output feature are spatially auto-correlated. Based on the above single-feature Moran's I, the IU combination test statistics T_1 calculated turns out to be significant for all sensitive features, but T_2 does not. This demonstrates the presence of a common spatial factor, but with uneven magnitudes of effects on the bikeshare indicator versus any of the sensitive features.

We perform ESF on three census tract-level features: number of bikeshare stations, %AA, and median income, based on eigenvectors of the inverse-distance weighted $\mathbf{W}$. Table 2 shows the results from pairwise significance tests based on correlations before and after removing spatial effects using ESF. Without applying ESF the tests conclude significant association between each of the two sensitive features and bikeshare station numbers in a census tract. However, spatial auto-correlation completely accounts for this association, and the ESF-adjusted features do not show any significant correlation.

Feature pair	No ESF	ESF-BIC	ESF-AIC
%AA vs. No. of stations	**0.007**	0.27	0.35
Med. Inc. vs. No. of stations	**4.4e-16**	0.999	0.99
%AA vs. Med. Inc.	**5.8e-129**	**2.6e-5**	**1.3e-8**

Table 2: Pairwise p-values from asymptotic correlation tests with and without ESF. Bold indicates significant evidence of rejecting null (at level 0.05).

4. Conclusion

We have shown that the influence of a common spatial factor may have a confounding effect on the estimation and evaluation of any association between a demographically sensitive feature and an outcome of interest. Our proposed hypothesis testing methods provide a first set of quantitative tools to tackle this problem. Spatial equity and accessibility has received attention in the geospatial data analysis literature (Ashik et al., 2020; Cobb, 2020). But this attention has not yet carried over to the ML fairness community. We aim to bridge this gap through our paper, and foster new developments in this very relevant domain.

An immediate line of future research is the conceptualization of methods to *mitigate* detected spatial fairness issues. It is not enough to ensure that the deployment of a spatially-based service has no source of bias apart from the underlying spatial factor. This may perpetuate underlying inequity structures and skew distribution of resources further away from disadvantaged communities. To this end, a bi-level mitigation may be necessary—mitigating both sources of demographic bias in the outcome of a ML-based deployment prioritization model, with *and* without consideration to the underlying spatial structure. Compared to the numerous existing bias mitigation methods for independent data, this imposes a harder constraint on the underlying utility maximization problem.

We note that methodology similar to ours may be applicable to other dependent data problems as well—such as mixed effect models, and analysis of covariate data attached to nodes in networks—contingent on assumptions relevant for the specific class of problems under consideration.

References

R. Arboretti, E. Carrozzo, F. Pesarin, and L. Salmaso. Testing for equivalence: an intersection-union permutation solution. *Stat. Biopharm. Res.*, 10:130–138, 2018.

Fajle Rabbi Ashik, Sadia Alam Mim, and Meher Nigar Neema. Towards vertical spatial equity of urban facilities: An integration of spatial and aspatial accessibility. *Journal of Urban Management*, 9(1):77–92, 2020.

R. L. Berger and J. C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.*, 11:283–319, 1996.

R.L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300, 1982.

E. Black, S. Yeom, and M. Fredrikson. FlipTest: fairness testing via optimal transport. In *FAT*-2020*, pages 111–121, 2020.

Chicago Data Portal. Divvy trips, 2021. URL https://data.cityofchicago.org/Transportation/Divvy-Trips/fg6s-gzvg.

Casey D. Cobb. Geospatial Analysis: A New Window Into Educational Equity, Access, and Opportunity. *Review of Research in Education*, 44(1):97–129, 2020.

Ashley Colley, Jacob Thebault-Spieker, Allen Yilun Lin, Donald Degraen, Benjamin Fischman, Jonna Häkkilä, Kate Kuehl, Valentina Nisi, Nuno Jardim Nunes, Nina Wenig, et al. The geography of pokémon go: beneficial and problematic effects on places and movement. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1179–1192, 2017.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1467–1477, 2020.

Stéphane Dray, Pierre Legendre, and Pedro R. Peres-Neto. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (pcnm). *Ecological Modelling*, 196(3):483 – 493, 2006.

EEOC. Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures. 1979. https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines.

Anna Goodman and James Cheshire. Inequalities in the london bicycle sharing system revisited: impacts of extending the scheme to poorer areas but then doubling prices. *Journal of Transport Geography*, 41:272–279, 2014.

D. Griffith. A spatial filtering specification for the autologistic model. *Environment and Planning A*, 36:1791 – 1811, 2004a.

Daniel A Griffith. A Spatial Filtering Specification for the Autologistic Model. *Environment and Planning A: Economy and Space*, 36(10):1791–1811, 2004b.

Daniel A. Griffith, Yongwan Chun, and Bin Li. *Spatial Regression Analysis Using Eigenvector Spatial Filtering*. Academic Press, 2019.

Mark F Guagliardo. Spatial accessibility of primary care: concepts, methods and challenges. *International journal of health geographics*, 3(1):3, 2004.

S. Heath. COVID-19 Vaccine Poses Geographic Care Access Health Disparities, 2021. URL https://billypenn.com/2021/03/18/philadelphia-vaccine-registry-equity-internet-access-walkup-fema-site/.

Cheng Jin, Jianquan Cheng, Yuqi Lu, Zhenfang Huang, and Fangdong Cao. Spatial inequity in access to healthcare facilities at a county level in a developing country: a case study of deqing county, zhejiang, china. *International journal for equity in health*, 14(1):1–21, 2015.

Y. Lee and E. L. Ogburn. Testing for Network and Spatial Autocorrelation. In *NetSci-X 2020*, pages 91–104, 2020.

M. Marin. Philly vaccine interest registry has built-in disparities, widening the racial divide, 2021. URL https://billypenn.com/2021/03/18/philadelphia-vaccine-registry-equity-internet-access-walkup-fema-site/.

N. Mehrabi, F. Morstatter, N. Saxena, et al. A Survey on Bias and Fairness in Machine Learning, 2019. arXiv:1908.09635.

Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

NACTO. Walkable station spacing is key to successful, equitable bike share. 2015.

Flora Ogilvie and Anna Goodman. Inequalities in usage of a public bicycle sharing scheme: socio-demographic predictors of uptake and usage of the london (uk) cycle hire scheme. *Preventive medicine*, 55(1):40–45, 2012.

K. A. O'Neil and R. A. Redner. Asymptotic distributions of weighted ustatistics of degree 2. *Ann. Probab.*, 21:1159–1169, 1993.

A. SenGupta. P^3 approach to intersection–union testing of hypotheses. *J. Stat. Plan. Inf.*, 137:3753–3766, 2007.

C. P. Shapiro and L. Hubert. Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *Ann. Statist.*, 7:788–794, 1979.

A. Williams and A. Emamdjomeh. America is more diverse than ever—but still segregated, 2018. URL https://www.washingtonpost.com/graphics/2018/national/segregation-us-cities/.

Appendix A. Preliminary concepts

A.1. Hypothesis testing

In statistical hypothesis testing, the null hypothesis (H_0) represents a default assumption on the data generating process, which is tested against the alternative hypothesis (H_a). Given sample datas, a *test statistic* T is calculated, and if the statistic value falls outside an interval, the test rejects H_0 in favor of H_a. Hypothesis tests can be seen as a decision function $\phi(\cdot)$ that takes a dataset $\mathcal{D}$ as input, and given a significance level $1-\alpha, \alpha \in (0,1)$, along with hypotheses H_0, H_a, returns the decision whether to reject H_0 in favor of H_a or not. Two types of errors can happen in this decision problem: (1) Type-I error, or the error of rejecting H_0 based on the data when it is true in reality, and (2) Type-II error, or the error of failing to reject H_0 when it is actually false. A good test would ideally have both errors small. Generally, in hypothesis problems a hard constraint on the upper bound of the type-I error is imposed: $\Pr\left[\phi(\mathcal{D};\alpha) = \text{Reject } H_0 | H_0 \text{ is true}\right] \leq \alpha$, where the probability is over the randomness of the data.

A p-value refers to the probability that T is at least as extreme (as high or low, depending on what H_a is) as t—the value of T for the sample data—given that H_0 is true. When p-value is $\leq \alpha$, we take the decision $\phi(\mathcal{D};\alpha) = \text{Reject } H_0$, else $\phi(\mathcal{D};\alpha) = \text{do not reject } H_0$. To

calculate the p-value, we need access to the null distribution of T. Depending on the test, this is obtained either using known asymptotic results, or approximated using permutation or bootstrap.

A.2. Bias metrics

In algorithmic fairness, there are a number of bias detection metrics, such as disparate impact, demographic parity, and statistical parity (Mehrabi et al., 2019). Many of these were proposed keeping classification problems in mind, work for binary sensitive attribute A, and either discrete labels Y or probability outputs $\widehat{Y}$. For example, the disparate impact (DI) metric in Section 1 calculates the ratio of positive probabilities given different values of A.

When one or more of Y, A is continuous, statistical tests like Kolmogorov-Smirnov (KS) statistic, or t-tests can be used to distinguish between the distributions of the continuous attribute, conditioned on different values of the discrete attribute. When both Y and A are continuous, methods to correlate samples of continuous random variables, such as the Pearson's (linear) correlation coefficient and Spearman's rank correlation can be used.

Appendix B. Implementation details

B.1. Permutation test

The steps are as follows:

- Suppose Π_n is the set of all permutations of $\{1, \ldots, n\}$. For some permutation $\pi \in \Pi_n$, we calculate the Moran's I statistics from permuted samples, say $I_{y,\pi}, I_{j,\pi}$, and using these the P3 test statistics $T_{1,\pi}, T_{2,\pi}$.
- Given some large integer M, we obtain permuted samples of the test statistics under their respective null hypotheses by repeating the above for random permutations $\pi_1, \ldots, \pi_M$:

$$\mathcal{T}_1^0 = \{T_{1,\pi 1}, \ldots, T_{1,\pi M}\}; \mathcal{T}_2^0 = \{T_{2,\pi 1}, \ldots, T_{2,\pi M}\}.$$

- The permutation p-values are empirical tail probabilities of T_1, T_2 corresponding to the respective (empirical) null distributions:

$$\widehat{p}_1 = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}(T_{1,\pi m} \geq T_1), \widehat{p}_2 = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}(T_{2,\pi m} \geq T_2).$$

Note that to generate the joint null distribution of (T_1, T_2), it is important to use the *same* set of permutations π_i to generate null samples $(\mathcal{T}_1^0, \mathcal{T}_2^0)$ above (Arboretti et al., 2018).

B.2. Details of ESF adjustment

When the feature being analyzed is continuous, we obtain its decorrelated version using residuals from its spatial autoregression

$$(\widehat{\boldsymbol{\gamma}}_k, \widehat{\boldsymbol{\beta}}) = \text{argmin}_{\boldsymbol{\gamma}, \boldsymbol{\beta}} \left[\frac{1}{2n} \|\boldsymbol{y} - \rho \mathbf{E} \boldsymbol{\gamma} - \mathbf{X} \boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\gamma}\|_1^2 \right], \quad \widetilde{\boldsymbol{y}} = \boldsymbol{y} - \rho \mathbf{E}_k \widehat{\boldsymbol{\gamma}}_k + \mathbf{X} \widehat{\boldsymbol{\beta}}.$$

When the feature being analyzed is discrete, we use an autologistic model (Griffith, 2004b), which is simply a logistic version of the spatial autoregression model discussed. Thus, we estimate $\boldsymbol{\gamma}, \boldsymbol{\beta}$ from the below model, again using ℓ_1-penalization to select the number of eigenvectors of $\mathbf{E}$:

$$p_{yi} := \Pr(y_i = 1 | \mathbf{E}, \boldsymbol{y}, \mathbf{X}) = [1 + \exp\{-(\rho \mathbf{E}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta})\}]^{-1}; \quad i = 1, \ldots, n.$$

There are now three possible cases: both Y, A are continuous, both are discrete, or one of them is continuous and the other is discrete. Below we describe ESF-adjusted testing strategies for each of these cases. In each case, we denote a generic bias metric computed before and after ESF by $r(\cdot, \cdot)$.

Case I: both continuous Examples of relevant bias metrics in this situation are Pearson's correlation coefficient, Spearman's rank correlation, and Kendall's Tau. In this case, the post-ESF bias metric is computed by simply using the respective residuals, i.e. $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{a}})$ in place of $r(\boldsymbol{y}, \boldsymbol{a})$, and compared against standard thresholds for that metric.

Case II: both discrete Disparate impact, equalized odds, and equality of opportunity are examples of bias metrics applicable when both Y, A are discrete. In our setting, we first estimate the sample-level probabilities using ESF ($i = 1, \ldots, n$):

$$\widehat{p}_{yi} = \left[1 + \exp\{-(\rho \mathbf{E}_{ky}\widehat{\gamma}_{ky} + \mathbf{X}\widehat{\boldsymbol{\beta}}_y)\}\right]^{-1}, \quad \widehat{p}_{ai} = \left[1 + \exp\{-(\rho \mathbf{E}_{ka}\widehat{\gamma}_{ka} + \mathbf{X}\widehat{\boldsymbol{\beta}}_a)\}\right]^{-1}.$$

Following this, we generate $\widehat{y}_i^b \sim \text{Bernoulli}(\widehat{p}_{yi}), \widehat{a}_i^b \sim \text{Bernoulli}(\widehat{p}_{ai})$, and calculate $r(\widehat{\boldsymbol{y}}^b, \widehat{\boldsymbol{a}}^b)$. We repeat this B times, and obtain the collection of metrics $\{r(\widehat{\boldsymbol{y}}^b, \widehat{\boldsymbol{a}}^b) : b = 1, \ldots, B\}$—fixing $B = 1000$ in our experiments. This simulates the empirical null distribution of the bias metric, i.e. under the assumption that the only bias present in the data is due to $\mathbf{W}$. We now simply compare the unadjusted metric against this distribution to obtain an approximate p-value: $\widehat{p} = B^{-1} \sum_{b=1}^{B} \mathbb{I}(r(\widehat{\boldsymbol{y}}^b, \widehat{\boldsymbol{a}}^b) \geq r(\boldsymbol{y}, \boldsymbol{a}))$.

Case III: one discrete, one continuous The KS statistic can be used as bias metric when one of Y, A is discrete. In this situation, we simply estimate the sample class probabilities for the discrete feature (without loss of generality, Y) and approximate the empirical null distribution of bias metrics using multiple simulated copies of the ESF-adjusted $\widetilde{\boldsymbol{y}}$ and the fixed, continuous fitted values $\widehat{\boldsymbol{a}}$: $\{r(\widehat{\boldsymbol{y}}^b, \widehat{\boldsymbol{a}}) : b = 1, \ldots, B\}$. The p-value is simply the tail probability at $r(\boldsymbol{y}, \boldsymbol{a})$ with respect to this empirical distribution.