# Affinity Clustering Framework for Data Debiasing Using Pairwise Distribution Discrepancy

Siamak Ghodsi[1,2], Eirini Ntoutsi[3]

[1]L3S Research Center, Leibniz Universität Hannover, Germany
[2]Freie Universität Berlin, Dept. of Mathematics and Computer Science, Berlin, Germany
[3]Research Institute CODE, Bundeswehr University Munich, Germany

Abstract

Group imbalance usually caused by insufficient or unrepresentative data collection procedures, is among the main reasons for the emergence of representation bias in datasets. Representation bias can exist with respect to different groups of one or more protected attributes and might lead to prejudicial and discriminatory outcomes toward certain groups of individuals; in case if a learning model is trained on such biased data. In this paper, we propose *MASC* a data augmentation approach based on affinity clustering of existing data in similar datasets. An arbitrary target dataset utilizes protected group instances of other neighboring datasets that locate in the same cluster, in order to balance out the cardinality of its non-protected and protected groups. To form clusters where datasets can share instances for protected-group augmentation, an affinity clustering pipeline is developed based on an affinity matrix. The formation of the affinity matrix relies on computing the discrepancy of distributions between each pair of datasets and translating these discrepancies into a symmetric pairwise similarity matrix. Furthermore, a non-parametric spectral clustering is applied to the affinity matrix and the corresponding datasets are categorized into an optimal number of clusters automatically. We perform a step-by-step experiment as a demo of our method to both show the procedure of the proposed data augmentation method and also to evaluate and discuss its performance. In addition, a comparison to other data augmentation methods before and after the augmentations are provided as well as model evaluation performance analysis of each of the competitors compared to our method. In our experiments, bias is measured in a non-binary protected attribute setup w.r.t. *racial* groups distribution for two separate minority groups in comparison with the majority group before and after debiasing. Empirical results imply that our method of augmenting dataset biases using real (genuine) data from similar contexts can effectively debias the target datasets comparably to existing data augmentation strategies.

**Keywords**

Distribution Shift, Affinity Clustering, Bias & Fairness, Maximum Mean Discrepancy, Data Debiasing, Data augmentation

## 1. Introduction

Recent years have brought extraordinary advances in the field of Artificial Intelligence (AI) such that now AI-based technologies replace humans at many critical decision points, such as who will get a loan [1] and who will get hired for a job [2]. There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored [3], and can take into account orders of magnitude with more factors than people can. However, like people, data-driven algorithms are vulnerable to biases that render their decisions "unfair". In automated decision-making, fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired protected attributes such as 'race' or 'gender'. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people.

*EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland*
ghodsi@l3s.de (S. Ghodsi); eirini.ntoutsi@unibw.de (E. Ntoutsi)
https://siamakghodsi.github.io/ (S. Ghodsi)
0000-0002-3306-4233 (S. Ghodsi); 0000-0001-5729-1003 (E. Ntoutsi)

One of the leading causes of unfair automated decisions in many real-world scenarios is due to unrepresentative, insufficient, or biased data fed to the learning algorithm [4]. Consequently, such biases can lead to certain discriminatory and prejudicial decisions harming sensitive groups e.g. racial/gender minorities in practice. To overcome this issue, in this paper, we propose a mechanism for **M**inority **A**ugmentation of biased datasets coming from separate but similar sources of data (described in the same feature space) through a **S**pectral **C**lustering scheme (**MASC**). The method proposes a way to augment underrepresented minority groups of an arbitrary task (hereafter we use the terms dataset and task interchangeably) by increasing their instances from a subset of contextually similar datasets that belong to the same cluster. Our proposed method performs an affinity clustering based on distribution discrepancy (that is used as a distance measure) among tasks to group similar tasks into a pre-defined number of clusters. Within each cluster, any member dataset can use instances shared by neighboring (mutually most similar) tasks to augment their underrepresented groups as compensation for group cardinality difference (that leads to representation and imbalance bias) according to a protected attribute.

Our main contributions can be summarized as follows:

- A new data augmentation framework for data debiasing towards statistical balancing between non-protected and protected group(s) based on most similar neighbors.
- Utilizing distribution shift metrics to quantify pairwise discrepancy between different datasets/ joint distributions
- A spectral clustering framework to group similar datasets based on the discrepancy between the joint distribution of these datastes.
- Clustering into an optimal number of clusters using a graph theoretic heuristic, known as "Eigen-gap or Spectral-gap" to avoid parameter selection and thus avoid any additional bias in the pipeline.

The rest of the paper is organized as follows: Preliminaries, related works, and motivation are presented in Section 2. In Section 3, we present the proposed data augmentation pipeline. Experimental evaluation results are provided in Section 4, including an intuitive example of applying the proposed method. Finally, Section 5 concludes this work, discusses its limitations and points out to future directions.

## 2. Preliminaries and Related Works

In this section, the necessary theoretical background and a brief literature review of these necessary notions are discussed.

### 2.1. Distribution Shifts

Distribution shift [5] is a broad topic studying how test data can differ from training data and how would such differences affect model performance. There are several possible causes for dataset shift, out of which two are deemed to be the most important reasons: Sample selection bias and non-stationary environments according to [6]. The motivation for referring to notions of data distribution shift in our paper is to utilize measures of distribution shift that provide practical tools as well as rich theoretical backgrounds that enable us to quantify similarity (and/or distance) among pairs of datasets that we will later use for data debiasing. Next, we look at formal definitions and different types of distribution shifts.

If we consider a dataset $(X, Y)$ to be a set of independent and identically distributed (*i.i.d.*) set of instances drawn randomly from an unknown continuous probability density function, then a classification problem is defined by a joint distribution $P(X, Y)$ of features a.k.a. covariates $X$

and target variables $Y$ [6]. According to the Bayesian Decision Theory [7], a classification can be described either by the prior probabilities of the classes $P(Y)$ and the class conditional probability density functions $P(X|Y)$ for all classes $Y = 1, \ldots, c$ where $c$ is the number of classes or by the covariate probabilities $P(X)$ and conditional probability density functions $P(Y|X)$. Thus the joint distribution $P(X, Y)$ can be decomposed in both following forms:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}, \quad P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

where $P(X) = \sum_{Y=1}^{c} P(Y)P(X|Y)$ and similarly $P(Y) = \sum_{X=1}^{c} P(X)P(Y|X)$ in $P(Y|X)$ and $P(X|Y)$ classification problems respectively. The two forms of problem formulation will be formalized as $Y \rightarrow X$ and $X \rightarrow Y$ (pronounced as $Y$ given $X$) respectively in the rest of this paper.

The literature on distribution shift detection and adaptive learning domain indicates that there are three types of distribution shifts [8, 9]:

1. **Covariate shift** appears only in $X \rightarrow Y$ problems when the probability of input features $P(X)$ changes, but the decision boundary defining the relationship between covariates and target labels $P(Y|X)$ remains the same. In other words, the distribution of the input changes, but the conditional probability of a label given an input remains the same. These shifts are known as "virtual shifts".
2. **Prior probability or target shift** appears only in $Y \rightarrow X$ problems when the probability of target labels $P(Y)$ changes but $P(X|Y)$ remains the same. For example, consider the case when the output distribution changes but for a given output, the input distribution stays the same.
3. **Concept drift** basically can appear in both types of problems namely in problems of type $X \rightarrow Y$ where the probability of $P(Y|X)$ changes between train and test data or in $Y \rightarrow X$ problems where $P(X|Y)$ changes. Concept drift happens when the input distribution remains the same between the two datasets but the conditional distribution of the output given an input changes. In other words, the decision boundary defining the relationship between covariates and labels changes.

## 2.2. Quantifying Distribution Shift

We are interested in measuring distribution discrepancy between two datasets $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, and $\boldsymbol{Z} \in \mathbb{R}^{m \times d}$ defined over the same feature space of $d$ features and having an arbitrary size of input samples. Discrepancy between two datasets can be due to differences in their feature and label distributions, so the conditional probability of labels given an input can remain the same. Since our goal is to develop a method to cluster similar distributions to enable us to augment a test dataset and also for the sake of maintaining the generality of the problem, we assume we don't have access to target labels and thus do not use prior probability information for shift quantification. As a result, we only use covariate distribution similarities. Moreover, we assume the similarity of the attribute space between different tasks; meaning that all the datasets have the same number of feature with the same range of values.

One of the most used measures for quantifying pairwise distribution differences is the Kullback-Leibler (KL for short) distance [10]. KL has nice theoretical properties, but it is not considered a metric as it is not symmetric ($KL(P, Q) \neq KL(Q, P)$ if $P, Q$ are probability distributions of covariate sets $\boldsymbol{X}$ and $\boldsymbol{Z}$ respectively) and it does not satisfy the triangle inequality [8]. A modified version of KL-divergence which belongs to a symmetrized sub-category of KL-divergence is the Jensen-Shannon divergence [11] which is a metric but still, as with all measures based on KL-divergence, it is sensitive to the sample size and requires both datasets to have the same cardinality.

Another well-known metric for measuring the distance between two distributions is the Maximum Mean Discrepancy (MMD for short) [12]. It is a multi-variate non-parametric statistic calculating the maximum deviation in the expectation of a function evaluated on each of the random variables, taken over a reproducing kernel Hilbert space (RKHS). MMD can equivalently be written as the L2-norm of the difference between distribution mean feature embeddings in the RKHS. In contrast to KL-Divergence, MMD is not sensitive to the number of instances and can be highly scalable to any arbitrary number of instances for each of the distributions depending on the kernel function employed for its calculation.

The MMD between two data distributions $\boldsymbol{X} \sim P$ and $\boldsymbol{Z} \sim Q$ is given by:

$$MMD\left(P,Q\right) = \parallel \mu_P - \mu_Q \parallel^2_{\mathscr{H}} \tag{1}$$

Where $\mu_P$ is the kernel mean of $\boldsymbol{X}$ estimated using $\mu_P\left(\phi(X)\right) = \frac{1}{n}\sum_{i=1}\phi(x_i)$ and similarly $\mu_Q$ is kernel mean of $\boldsymbol{Z}$ assuming $\phi : \boldsymbol{X} \rightarrow \mathscr{H}$ to be a feature map embedding $\boldsymbol{X}$ to the embedding Hilbert space $\mathscr{H}$. Then Eq. 1 can be substituted as:

$$MMD\left(P,Q\right) = \left\| \frac{1}{n}\sum_{i=1}\phi(x_i) - \frac{1}{m}\sum_{i=1}\phi(z_i) \right\|^2_{\mathscr{H}} \tag{2}$$

The inner product (indicated by $\langle \bullet \rangle$ ) of feature means of $\boldsymbol{X} \sim P$ and $\boldsymbol{Z} \sim Q$ can be written in terms of the kernel function such that:

$$\left\langle \mu_P\left(\phi\ (\boldsymbol{X})\right), \mu_Q\left(\phi\ (\boldsymbol{Z})\right) \right\rangle_{\mathscr{H}} = E_{P,Q}\left[ \left\langle \phi\left(\boldsymbol{X}\right), \phi\left(\boldsymbol{Z}\right) \right\rangle_{\mathscr{H}} \right] = E_{P,Q}\left[\ k\left(\boldsymbol{X},\boldsymbol{Z}\right)\ \right] \tag{3}$$

Substituting Eq. 3 into Eq. 1 we can rewrite it such that:

$$MMD\left(P,Q\right) = E_P\left[\ k\left(\boldsymbol{X},\boldsymbol{X}\right)\ \right] - 2E_{P,Q}\left[\ k\left(\boldsymbol{X},\boldsymbol{Z}\right)\ \right] + E_Q\left[\ k\left(\boldsymbol{Z},\boldsymbol{Z}\right)\ \right] \tag{4}$$

Finally, expanding the Eq. 4, the two sample MMD-test can be calculated by:

$$\begin{aligned} MMD\left(\boldsymbol{X},\boldsymbol{Z}\right) = \frac{1}{n(n-1)}\sum_i\sum_{j\neq i} k(x_i,x_j) - 2\frac{1}{n.m}\sum_i\sum_j k(x_i,z_j) \\ + \frac{1}{m(m-1)}\sum_i\sum_{j\neq i} k(z_i,z_j) \end{aligned} \tag{5}$$

In [12] it is suggested to use linear statistic if the datasets are sufficiently large. Since our sample sizes are large enough, we use the linear kernel for MMD calculations in Eq 5 in our experiments. To avoid scale differences it is a good practice to normalize the values in the [0,1] range.

## 3. Proposed Method

In this section, the detailed procedure of the proposed ***MASC*** method is described in 4 steps. A procedural overview of each of these steps of the proposed data augmentation method is provided in Algorithm 1. Before approaching these steps with details, an overall overview of the process is discussed in the following.

Assume having $r$ number of biased datasets $\boldsymbol{D}_{all} = \left\{ \boldsymbol{X}_1 \cup \boldsymbol{X}_2 \cup \ldots \cup \boldsymbol{X}_r \right\}$ according to $r$ different tasks. For instance, these datasets could each belong to different branches of a franchised hypermarket or be from civil registration offices in different cities (or states) and many other similar cases. Nevertheless, Our final goal is to find a clustering of these datasets based on a similarity score

**Algorithm 1:** Procedure of the proposed debiasing method **MASC**

**Input:** $D_{all}$ — Set of all tasks;

$b$ — Index of target task $X_b$; ▷ $D_{all} = \bigcup_{i=1}^{r} X_i, \;\; b \in \{1, \dots, r\} \Rightarrow X_b \subset D_{all}$

**Output:** $\widehat{X}$ — Minority-augmented (debiased) target task

**for** $i \leftarrow 0$ **to** $r$ **do** ▷ $r = |D_{all}|$

  **for** $j \leftarrow 0$ **to** $i$ **do**

    $W(i,j) \leftarrow MMD(X_i, X_j)$ ; ▷ `Pairwise MMD:` Eq. 5

    **if** $i \neq j$ **then**

      $A(w_i, w_j) \leftarrow exp\left(-\gamma \parallel w_i - w_j \parallel^2\right)$ ; ▷ `Gaussian Kernel:` Eq. 6

    **else**

      $A(w_i, w_j) \leftarrow 0$ ; ▷ `Zeros on diagonals`

    **end**

  **end**

**end**

$L \leftarrow D - A$: where $D_{i,i} \leftarrow \sum_{j=1}^{r} A_{i,j}$; ▷ Graph Laplacian: Sec 3.2

$L \leftarrow U \Sigma V^T$; ▷ SVD: Eq 7

$e \leftarrow \left[\lambda_2 - \lambda_1, \lambda_3 - \lambda_2, \dots, \lambda_l - \lambda_{l-1}\right]$; ▷ Eigengap vector: Eq 8

$k$ (optimal number of clusters): apply eigen-gap technique on $e$, $k$ is the index of largest gap;

$U \in \mathbb{R}^{r \times k} \leftarrow \begin{bmatrix} \vdots & & \vdots \\ u_1, & \dots, & u_k \\ \vdots & & \vdots \end{bmatrix} \overset{\wedge}{=} \min_{1 \dots k} \lambda_k$; ▷ Top $k$ eigenvectors: Eq. 9

$D_{all} \leftarrow Kmeans(U)$; ▷ $= Kmeans(U) = \{C_1 \cup \dots \cup C_k\}, \quad k \leq r$: Eq. 10

$C_c \leftarrow \{X_1 \cup \dots \cup X_t\}, \quad t \leq r \quad \& \quad c \in \{1, \dots, k\}$; ▷ Where $|C_c| = \sum_{i=1}^{p} N_i = N$: Eq. 11

**for** $i \leftarrow 1$ **to** $p$ **do**

  $|X_{S_i}| \leftarrow n_i$; ▷ Subgroup cardinality of $X = X_b$, $\sum_{i=1}^{p} n_i = n$: Eq. 12

**end**

$n_l \leftarrow max(n_1, \dots, n_p)$; ▷ Cardinality of majority group

**if** $N_g > n_l$ **then**

  $\widehat{X} \leftarrow X \cup \bigcup_{j=1}^{(l_g - n_g)} C_{S_g}(j)$; ▷ Augment $X \subseteq C_c$ in the c-th cluster

**else**

  $\widehat{X} \leftarrow C_{S_g}$;

**end**

such that in each cluster, tasks can share their instances. This way, an arbitrary dataset $X_b \subset D_{all}$ that is biased by over-representing a majority[1] group w.r.t. a protected attribute, can borrow instances of minority group(s) from its neighboring tasks and construct an augmented unbiased training set. For the clustering procedure, a spectral clustering algorithm is utilized that can identify the optimal number of clusters automatically based on an Eigen-gap or Spectral-gap heuristic introduced in [13]. In order to perform the clustering step, initially we need to construct an *affinity* matrix from the pairwise distances that we obtain by *MMD* metric.

[1](*majority, non-protected*), and also (*minority*, and *protected*) groups will be used interchangeably in this paper.

### 3.1. Affinity Matrix Computation

The first step in the proposed method is to compute pairwise distance (or discrepancy) between each pair of datasets $\boldsymbol{X}_i \subset \boldsymbol{D}_{all}$ and $\boldsymbol{X}_j \subset \boldsymbol{D}_{all}$ using Eq. 5. The distances are then transformed into a symmetric matrix of pairwise distances $\boldsymbol{W} \in \mathbb{R}^{r \times r}$ such that the diagonal of the matrix is all zeros. An intuitive way to convert a pairwise distance matrix into an "Affinity" matrix is by applying a *Radial Basis Function a.k.a. Gaussian Kernel* [13, 14]:

$$\boldsymbol{A}\left(w_i, w_j\right) = \begin{cases} exp\left(-\gamma \left\| w_i - w_j \right\|^2\right), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $w_i$ and $w_j$ are two entries of the distance matrix $\boldsymbol{W}$. Eq. 6 results in a weighted undirected symmetric affinity matrix $\boldsymbol{A}$ with zero diagonal elements with weights being Gaussian functions of the pairwise distances.

### 3.2. The Optimal Number of Clusters k

In order to perform spectral clustering on the affinity matrix, we need to calculate the unnormalized graph Laplacian [13] $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ where $D_{i,i} = \sum_{j=1}^{r} A_{i,j}$ is the diagonal degree matrix of the affinity matrix $\boldsymbol{A}$. Graph Laplacian is key to spectral clustering; its eigenvalues and eigenvectors reveal many properties about the structure of a graph.

According to the "Perturbation Theory", an optimal number of clusters $k$ for a dataset can be given through the eigengap identification of eigenvalues of the graph Laplacian, which is the largest difference between eigenvalues [15, 16]. Thus, computing the eigenvalues of the Laplacian matrix and finding it's biggest gap can discover the optimal number of clusters. This way, one can avoid the difficult and tricky decision of the cluster number parameter. Thus, similar to the instructions in *step 5* of [17] we perform a "*Singular Value Decomposition (SVD)*" to calculate the eigenvalues of the Laplacian matrix $\boldsymbol{L}$:

$$\boldsymbol{L} = \boldsymbol{U} \Sigma \boldsymbol{V}^T \tag{7}$$

where $\boldsymbol{U}, \boldsymbol{V}$ are unitary matrices called left and right singular matrices, respectively containing eigenvectors corresponding to eigenvalues in $\Sigma$. Next, we create an eigengap vector $e$ using the eigenvalues from $\Sigma$ in Eq. 7 as follows:

$$e = \left[\lambda_2 - \lambda_1, \lambda_3 - \lambda_2, \ldots, \lambda_l - \lambda_{l-1}\right] \tag{8}$$

where $\lambda_k$, is the *k-th* sorted eigenvalue in ascending order. Note that, if $\left(\lambda_k - \lambda_{k-1}\right)$ implies the largest difference i.e. eigengap according to Eq. 8, then index $k$ is the optimal number of clusters.

### 3.3. Spectral Clustering

After obtaining $k$, the desired number of clusters in Section 3.2, there is one more step to finally be able to partition the affinity matrix. In this step, we find the top k eigenvectors $u_1, \ldots, u_k$ according to the top k smallest eigenvalues of the Laplacian, stack them as columns of a new matrix $\boldsymbol{U} \in \mathbb{R}^{r \times k}$ such that:

$$\boldsymbol{U} = \begin{bmatrix} \vdots & & \vdots \\ u_1, & \ldots, & u_k \\ \vdots & & \vdots \end{bmatrix} \hat{=} \min_{1 \ldots k} \lambda_k \tag{9}$$

where $\hat{=}$ stands for the term "Corresponding to". Then, a k-means clustering [18] is performed on the rows of matrix $\boldsymbol{U}$ which is equivalent to a clustering of the r datasets:

$$\boldsymbol{C} = Kmeans(U) = \{\boldsymbol{C}_1 \cup \boldsymbol{C}_2 \cup ... \cup \boldsymbol{C}_k\} \quad \text{and} \quad k \leq r \quad \boldsymbol{C} \equiv \boldsymbol{D}_{all} \tag{10}$$

and ≡ sign represents the equivalence of its operands. Note that, in practice, spectral clustering is often followed by another clustering algorithm such as k-means to finalize the clustering task. The main property of spectral clustering is to transform the representations of the data points of $\boldsymbol{X}_b$ into the indicator space in which the cluster characteristics become more prominent and passes much more processed/meaningful information to the next step clustering algorithm.

### 3.4. Data Augmentation Within Clusters

Now that the set of input tasks/datasets $\boldsymbol{D}_{all}$ is clustered into $k$ partitions according to Eq. 10, the data augmentation process for minority group(s) can be fulfilled. If cluster $c$ consists of $t$ datasets:

$$\boldsymbol{C}_c = \{\boldsymbol{X}_1 \cup ... \cup \boldsymbol{X}_t\} \quad \text{where} \quad t \leq r \quad \& \quad c \in \{1, ..., k\} \tag{11}$$

Initially, we create a pool of instances in the cluster $\boldsymbol{C}_c$ by collecting all the instances from each dataset belonging to the cluster. The number of instances in this cluster $|\boldsymbol{C}_c| = N$, (where $|\bullet|$ denotes cardinality) can be written as a sum of the number of instances belonging to each of the $p$ protected groups $\sum_{i=1}^{p} N_i = N$. Given a protected attribute $S = \{S_1, ..., S_p\}$ with $p$ groups and knowing $|\boldsymbol{X}| = n$, the augmentation process for task $\boldsymbol{X} \subset \boldsymbol{D}_{all}$ is a very straightforward process based on protected groups cardinality. We calculate the cardinality of each group corresponding to the number of instances belonging to that group such that:

$$\left|\boldsymbol{X}_{S_i}\right| = n_i \quad \text{for i} \in \{1, ..., p\} \tag{12}$$

where $\sum_{i=1}^{p} n_i = n$. Next, we identify the biggest group and indicate it as the majority/non-protected group through a procedure like $max(n_1, ..., n_p) = n_l$. Ideally, the intention would be to balance every minority subgroup $g$ to have a cardinality as big as the majority group $l$, so that $|\widehat{\boldsymbol{X}}_{S_g}| = n_l$. Thus, every protected group needs to be augmented by a difference of $n_l - n_g$. However, it is only the case if the pool of shared protected group instances includes this number of instances otherwise we augment by as many instances as there exist in the shared pool. Thus, the augmented version of dataset $\boldsymbol{X}$ has the following number of instances depending on the number of shared instances:

$$\widehat{\boldsymbol{X}} \leftarrow \begin{cases} \boldsymbol{X} \cup \bigcup_{j=1}^{(n_l - n_g)} \boldsymbol{C}_{S_g}(j) & \text{if } N_g > n_l \\ \boldsymbol{C}_{S_g} & \text{otherwise} \end{cases} \tag{13}$$

where $\boldsymbol{C}_{S_g} = \{\boldsymbol{C}_c \mid S = S_g\}$. Note that, $\boldsymbol{C}_c$ and $\boldsymbol{X}_b$ are substituted by $\boldsymbol{C}$ and $\boldsymbol{X}$ respectively, to avoid syntax complication.

## 4. Experimental Results

In order to evaluate the effectiveness of the proposed *MASC* method, in this section the conducted experimental results on a number of real-world datasets are analyzed. The organization of this section is as follows: First, details of the datasets employed are presented. Next, the evaluation measures used in the experiments and also methods used for comparisons are described. Finally, the experimental results and discussions on them are provided.

**Table 1**
Statistics for five chosen states including their Racial group imbalance ratio, Abbreviations denoted as (Abbr), and Class Ratio for positive (Pos) and Negative (Neg) classes.

| | | | | Group Distribution Ratio | | | Class Ratio |
|---|---|---|---|---|---|---|---|
| **States** | **Abbr** | **Samples** | **Cleaned** | **White** | **Black** | **Other** | **(Pos\|Neg)** |
| Colorado | CO | 57,142 | 32,264 | **87.98%** | 2.56% | 9.46% | 1:1.26 |
| North-Dakota | ND | 7,960 | 4,455 | **92.03%** | 1.35% | 6.76% | 1:2.2 |
| Maryland | MD | 60,237 | 32,865 | **64.59%** | 22.4% | 13.01% | 1.05:1 |
| Mississippi | MS | 29,217 | 13,159 | **66.43%** | 29.92% | 3.66% | 1:2.51 |
| Montana | MT | 10,649 | 5,547 | **92.03%** | 0.36% | 7.61% | 1:2.09 |

## 4.1. Datasets

To evaluate *MASC*'s performance in addressing group imbalance and representation bias, we used the recently released US Census datasets [19], which comprise a reconstruction of the popular Adult dataset [20]. These datasets provide a suitable benchmark with 52 datasets representing different states, effectively capturing the problem of group imbalance between states with varying numbers of instances but similar feature spaces.

The datasets [19] include census information on demographics, economics, and working status of US citizens. Spanning over 20 years, they allow research on temporal and spatial distribution shifts and incorporate various sources of statistical bias. As already mentioned in Section 2.2, in this study we assume that the conditional probability of labels given specific inputs remains constant. Therefore, we focus on the latest release, specifically the year *2019* (till the date of submission), and examine the spatial context to explore the connection between covariate shifts and bias.

The feature space consists of 286 features, with only 10 deemed relevant [19]. The target variable, *Income Value*, is transformed into a binary vector to predict whether an individual earns an income of more than 50k: $Income \in \{\leq 50K, > 50K\}$, the positive class being $" > 50K"$. We selected *"Race"* as the protected attribute due to the challenge it poses compared to gender or age, given the highly imbalanced distribution of racial groups across states. The *"Race"* attribute has 9 categories, but due to a very small representation of seven of these categories which usually comprise less than 1% of the instances in the dataset, we aggregate them to a bigger group called *"Other"*. Thus, the categories in our experiments are aggregated into 3 groups: *White, Black, and Other*. Categorical features are transformed into numerical features and all the features are normalized by standard scaling using their mean ($\mu$) and standard deviation ($\sigma$) values such that each $z = (x - \mu)/\sigma$ is a standard representation of its $x$ and lies within the range $[0, 1]$.

Refer to Table 1 for detailed information on the filtered (cleaned) datasets, including racial distribution, class imbalance ratio, name abbreviation conventions, and other details. The table summarizes information for 5 out of 51 datasets. The intuition behind this specific selection of states will be addressed in detail in Section 4.4. The datasets exhibit significant racial bias, with the *White* group representing the majority (also referred to as non-protected) in all 5 datasets.

## 4.2. Metrics

In this paper, we adopt five measures in total. We use accuracy [1] for analyzing models predictive performance along with four measures for bias and fairness quantification; *Disparate Impact* [21], *Statistical (or Demographic) Parity* [22], *Equalized Odds* [1], and a new proportionality metric that we introduce, the *Group Distribution Ratio* for quantification of bias on datasets before and after debiasing. The measures that take into account model outcome or in other words, which involve

model training and prediction (e.g. Accuracy and Equalized Odds) are not relevant for the first part of experiments. Given a dataset $\boldsymbol{X} = \{D, S, Y\}$, with $D$ regular features, a protected feature $S$ (i.e. *Race*) and a binary target class, the disparate impact (DI short for) of the given dataset is calculated as follows:

$$DI = \frac{P\big(Y = 1 \,\big|\, S = 0\big)}{P\big(Y = 1 \,\big|\, S = 1\big)} \tag{14}$$

which basically calculates the ratio of the probability of being a member of the protected group having positive outcomes to the probability of the non-protected group with positive outcomes. DI ranges between zero and one $DI \in (0, 2)$ with 1 being the best value i.e. implies there is no bias. $0, 2$ mean maximum bias toward one group or the other respectively.

The measure statistical parity (SP for short) also computes a quite similar value, where it reflects the mentioned change as a difference instead of a ratio:

$$SP = P\big(Y = 1 \,\big|\, S = 0\big) - P\big(Y = 1 \,\big|\, S = 1\big) \tag{15}$$

Since we consider two protected groups of "*Black*" and "*Other*" in our experiments, the results are calculated for each of the measures twice; for each of the two protected groups against the non-protected group of "White". So in our analysis $S \in \{\, 0, 1\, , 2\, \}$. SP takes values in the range $SP \in (-1, 1)$ with 0 as the best possible value implying zero bias.

The measure *Equalized Odds* (Eq.Odds for short) calculates the difference in prediction errors between the protected and non-protected groups for both classes as $|\delta FPR| + |\delta FNR|$ where $\delta FNR$ stands for "*False Negative Rates*" and $\delta FPR$ stands for "*False Positive Rates*" that are also known as *Equal Opportunity* and *Predictive Equality* respectively. The $\delta FNR$ measures the difference of the probability of subjects from both the protected and non-protected groups that belong to the **positive** class to have a negative predictive value and similarly, the $\delta FPR$ calculates the difference of the probability of subjects from both the protected and non-protected groups that belong to the **negative** class to have a positive predictive value. So, the Eq.Odds is formulated as follows:

$$Eq.Odds = \Big|P(\hat{Y} = 0|Y = 1, g = w) - P(\hat{Y} = 0|Y = 1, g = b/o)\Big| + \tag{16}$$

$$\Big|P(\hat{Y} = 1|Y = 0, g = w) - P(\hat{Y} = 1|Y = 0, g = b/o)\Big| \tag{17}$$

where $\hat{Y}$ is the predicted label, Y is the actual label and $g \in G = \{w, b, o\}$ is the protected attribute. The value range for each of $\delta FNR$ and $\delta FPR$ is [0,1], where 0 stands for a classifier satisfying perfectly the measure with no discrimination and 1 stands for maximum discrimination. Thus, Eq.Odds can range between [0,2]. In this study, $w$ is taken as the majority (non-protected) group and $b$ and $o$ are minority (protected) groups.

Finally, we introduce a group-proportional measure: the group distribution ratio (GR for short) essentially calculates group imbalance or the proportion of instances belonging to each of the protected groups or the non-protected group w.r.t. the total number of instances in the dataset. Similar to the definition of protected attribute and its member groups in Section 3.4, the group distribution ratio for a protected group $g$ is obtained as follows:

$$GR_g = P\big(\, X \big|\, S = S_g\big) = \frac{|\, X_{S_g}\,|}{|\, X\,|} = \frac{n_g}{\sum_{i=1}^{p} n_p} \tag{18}$$

where the denominator of the fraction in Eq. 18, is the sum of the cardinality of all subgroups of task $X$ or the total number of its instances $\sum_{i=1}^{p} n_p = n$. Clearly, the cumulative probability of all subgroups $\sum_{i}^{p} P\big(\, X \big|\, S = S_i\,\big)$ is $=\ 1$. Thus, a dataset is group balanced w.r.t. a protected attribute

if Eq. 18 is proportionately equal for each subgroup. In other words, the optimal balance for each subgroup in the dataset is given by $GR^* = 1/(\sum_{i=1}^{p} S_i)$ which implies balanced groups with the same number of instances. As a result, for a protected attribute with two subgroups, the optimal group distribution ratio would be $GR^* = 1/2$ and similarly for a protected attribute with three subgroups $GR^* = 1/3$.

## 4.3. Competitors

In order to compare the results of our *MASC* augmentation method, we compare it with 4 different competitors including the original shape of untouched datasets along with three other strategies. Specifically, we use a variation of *SMOTE* [23] for synthetic minority protected group over-sampling instead of minority class augmenting, and similarly we use a variation of *RUS* [24], as a random group under-sampling. In addition, we also introduce a natural geographical neighborhood augmentation by concatenating datasets within their local clusters of geographical neighbors based on the formal region categorization as in [25]. All the augmentation methods are also analyzed by feeding their outputs to a Logistic Regression classifier (LR for short).

Note that we implement a variation of SMOTE and RUS to over/under-sample based on the protected group distribution of the protected attribute such that: for SMOTE we over-sample both minority groups until their cardinality is as large as the majority group. For the RUS method, we under-sample the majority group and the bigger minority group until they contain as few samples as the smallest minority group.

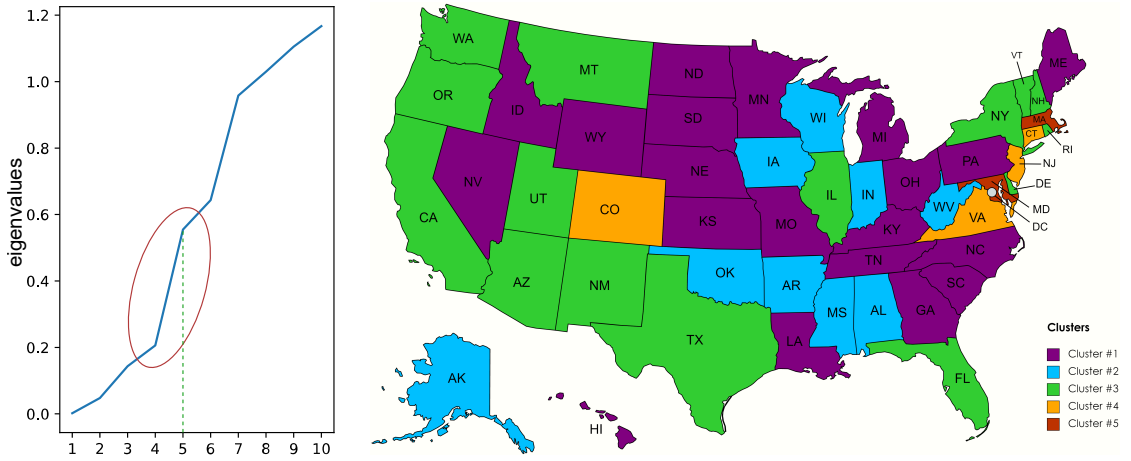
## 4.4. Empirical Results

The forthcoming experiments in this section are conducted in order to compare the initial dataset biases of the original datasets before and after the proposed data augmentation and also in comparison with the three other augmentation strategies, respectively as mentioned in section 4.3 based on the introduced measures in Section 4.2. Following that, we also compare predictive performance and fairness of a LR classifier on the different augmentation strategies to see how would each of the augmentation methods affect model performance[2]. Meanwhile, to get an intuition about the step-wise procedure of the proposed *MASC* method, a demonstration of implementation on the aforementioned US-Census datasets, following the steps in Section 3 is illustrated before discussing performance results.

### 4.4.1. A Demo implementation

Initially, an affinity matrix according to steps 1-9 of Algorithm 1 is generated. Following that, based on instructions in lines 11-12 of Algorithm 1, initially a graph Laplacian and afterward its SVD decomposition are calculated from the affinity matrix, in order to obtain eigenvalues of the Laplacian and find the spectral eigengap as in steps 13-14. According to the spectral graph theory [26], in an ultimately well-shaped problem, one can observe that there exists an ideal case of k completely disconnected components which constitute a block diagonal Laplacian matrix that has k zero eigenvalues and corresponding k eigenvectors of ones. In this extreme case, the (k+1)-th eigenvector which is non-zero, has a strict gap. This gap identifies the optimal number of connected components that can be clustered as highly similar objects. The eigengap heuristic is an advanced guide to avoid parameter selection, although our problem as well as the majority of real-world problems do not produce such a well-formed block diagonal Laplacian. In Figure 1a the first ten eigenvalues and the major eigengap are demonstrated. It depicts that our datasets can be semi-optimally clustered into five categories.

[2]The source code of the proposed *MASC* and the comparisons can be found at: Github/SiamakGhodsi/MASC




(a) The largest eigen-gap

(b) States data clustering shown as a 5-nn graph for legibility

**Figure 1:** In **(a)** the first 10 eigenvalues of the Graph Laplacian are shown. The largest eigengap indicating the optimal number of clusters according to Eq. 8 is identified between eigenvalues $eig_4$ and $eig_5$ implying that 5 clusters can best partition the graph affinity matrix defined in Section 3.1. In **(b)** Clustering result of the graph Laplacian according to the *Unnormalized Spectral Clustering* method introduced in Section 3.3 to the optimal number of 5 clusters obtained in the previous step. States are labeled with 2-letter standard US state abbreviations and are coloured based on the clusters they belong to.

We partition the obtained Affinity matrix into five clusters following instructions in Section 3.3 and accordingly steps 16-17 of Algorithm 26. The clustering is illustrated in Figure 1b. Each color represents a cluster.

Next, according to steps in Section 3.4 and steps 18-26, we augment each of the input datasets using the shared protected-group instances from their neighbors in the same clusters. The results of the minority group(s) augmentation are summarized in Table 2 and group distribution and GR scores are shown in Figure 2b.

### 4.4.2. Results

The *MASC* is applied to all the input datasets and augments each task based on the cluster that they belong to and therefore the neighborhood instances that they share. Since the method indicates 5 clusters as depicted in Figure 1b, for the sake of readability we evaluate the results for 5 of the datasets, each chosen from one of the clusters. Namely, the states are; *Montana* (MT), *Mississippi* (MS), *North-Dakota* (ND), Colorado (CO), and Maryland (MD). The selection of states within each cluster is based on diversity; we chose states from western and more central regions to northern and east-most states that allow comparing population texture of different regions of the US according to [25].

In Figure 2a the GR values according to the distribution of each of the original datasets is shown. Comparing them to distributions in Figure 2b which is the results obtained by our method, the performance of our method in reducing the group differences is inferable. The proposed method borrows similar instances for each minority group from similar states and balances exactly perfectly four of the states and to a very good extent also the Colorado state. In Colorado's case, the number of minority group instances borrowed from other states in the cluster is not enough to equalize the representation of minority groups, but still decreases the worst imbalance in the original dataset in terms of difference between GR-values of Maj-min1 from $87.98\% - 2.56\% \approx 85\%$ to $49.16\% - 26.13\% \approx 23\%$ and reduces the 85% difference to 23%.

Table 2 summarizes the evaluation of the five datasets based on previously introduced measures DI, SP, GR (refer to Section 4.2 for details) comparing the results of original states with

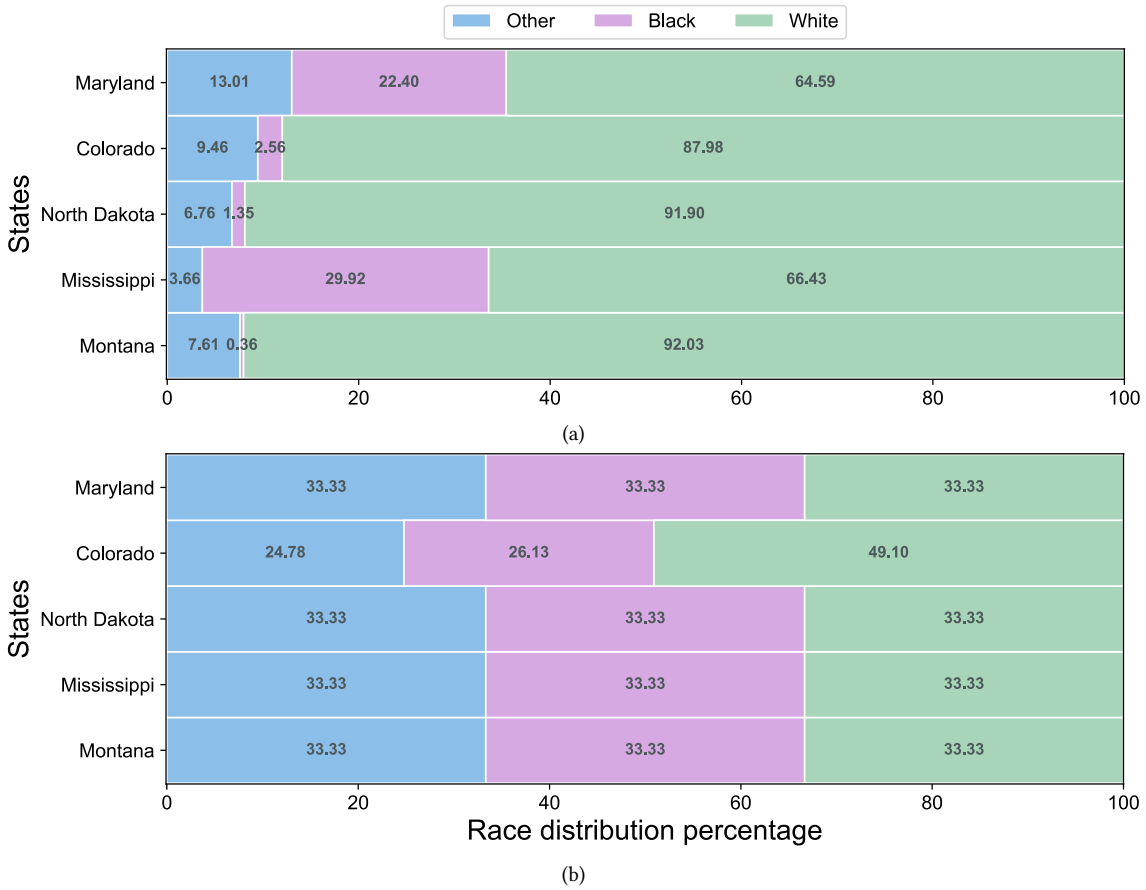



**Figure 2:** Comparison of groups distribution ratio (GR) percentage between the original datasets before data augmentation and their augmented versions after applying *MASC* w.r.t. *Race* as protected attribute. categories are: *{White, Black, Other}* per state.

*MASC*, Geographical-neighborhood grouping, the SMOTE, and the RUS methods. Note that the Geographical-neighborhood augmentation is abbreviated as Geo-nei in the table. The Maj, Min1, and Min2 notations correspond to Majority (*White*) and two minority groups (*Black* and *Other*), respectively. It is empirically shown in the table that the results of the MASC, alleviate group imbalance to a good extent for all the datasets according to the GR column and subsequently achieve good DI and SP rates compared to the original datasets. Moreover, in comparison to Geo-nei, our method performs better for all the states except for the Min1 group in the Colorado state and still achieves much better balances for the protected groups but only the class distribution is slightly worse. Compared to SMOTE and RUS, our method performs comparably well in terms of GR-values but w.r.t. DI and SP metrics, the method reports slightly worse results that is because our method only balances out group distributions and doesn't take into account the distribution of target-class. However, in model performance, our method outperforms the SMOTE and RUS for all the states w.r.t. accuracy and eq.odds metrics that we will see in the followings. Also, there are some technical issues/limitations that may arise using the SMOTE and RUS methods which will be discussed more detail in Section 4.5.

In Figure 3 the performance results of a LR model trained on each of the augmentation methods and tested on the corresponding are illustrated. Note that, there are two legends where the first one represents the three augmentations (including our method MASC) that are based on real (genuine) data and the other represents synthetic augmentation methods. In Figure 3a the Eq.Odds values are shown where we can see the purple bar, representing our method MASC gets

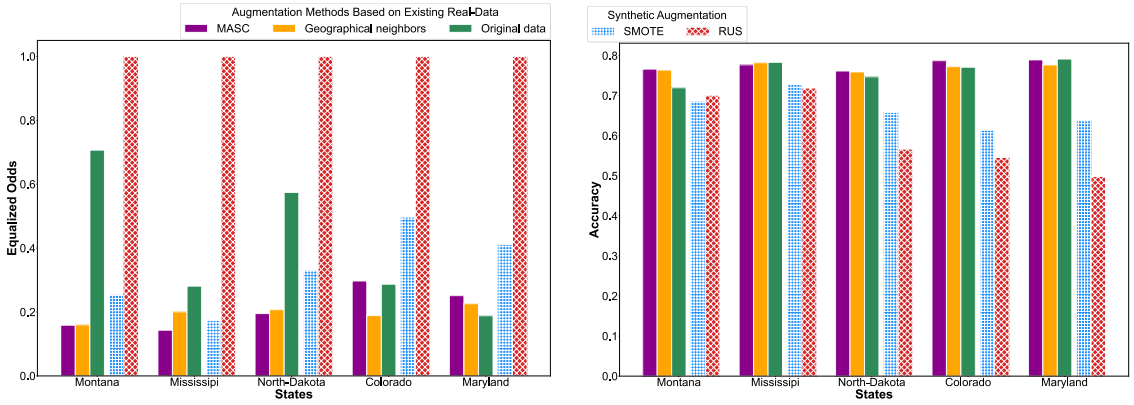
**Table 2**
Evaluation results of the five datasets: the initial stats of the original data in comparison to 4 different augmentation strategies w.r.t. Group Distribution Ratio (GR), Statistical Parity, and Disparate Impact measures. Maj, Min1, and Min2 stand for the Majority (White) and the two Minority (Black, and Other) groups respectively. Boldfaced values imply better results in SP and DI measures. For GR results, the balancedness of the three groups implies better results.

| States | Method | Group Distribution Ratio | | | Statistical Parity (SP) | | Disparate Impact (DI) | |
|---|---|---|---|---|---|---|---|---|
| | | Maj | Min1 | Min2 | Min1 | Min2 | Min1 | Min2 |
| **Montana** | Initial | 92.03% | 0.36% | 7.60% | -0.124 | -0.122 | 0.617 | 0.634 |
| | MASC | 33.33% | 33.33% | 33.33% | -0.054 | 0.093 | 0.855 | 1.285 |
| | Geo-nei | 83.35% | 01.89% | 14.75% | -0.112 | -0.070 | 0.726 | 0.832 |
| | SMOTE | 33.33% | 33.33% | 33.33% | **-0.002** | **-0.015** | **0.991** | **0.954** |
| | RUS | 33.33% | 33.33% | 33.33% | 0.075 | -0.150 | 1.272 | 0.571 |
| **Mississippi** | Initial | 66.42% | 29.91% | 3.65% | -0.172 | -0.056 | 0.487 | 0.803 |
| | MASC | 33.33% | 33.33% | 33.33% | -0.072 | 0.094 | 0.782 | 1.297 |
| | Geo-nei | 80.98% | 14.83% | 04.17% | -0.154 | **-0.062** | 0.559 | 0.810 |
| | SMOTE | 33.33% | 33.33% | 33.33% | -0.092 | -0.006 | 0.705 | **1.022** |
| | RUS | 33.33% | 33.33% | 33.33% | **0.026** | 0.004 | **1.093** | 1.014 |
| **North-Dakota** | Initial | 91.89% | 1.34% | 6.75% | -0.282 | -0.181 | 0.261 | 0.535 |
| | MASC | 33.33% | 33.33% | 33.33% | -0.098 | -0.077 | 0.762 | 1.216 |
| | Geo-nei | 90.74% | 03.60% | 5.65% | -0.146 | -0.112 | 0.598 | 0.691 |
| | SMOTE | 31.31% | 34.34% | 34.34% | **-0.005** | **-0.020** | **0.984** | **0.945** |
| | RUS | 31.31% | 34.34% | 34.34% | -0.058 | -0.058 | 0.820 | 0.820 |
| **Colorado** | Initial | 87.98% | 2.56% | 9.46% | -0.176 | -0.097 | 0.605 | 0.783 |
| | MASC | 49.09% | 26.12% | 24.77% | -0.199 | -0.100 | 0.492 | 0.723 |
| | Geo-nei | 83.02% | 02.87% | 14.10% | -0.125 | -0.128 | 0.668 | 0.672 |
| | SMOTE | 33.33% | 33.33% | 33.33% | -0.007 | -0.016 | 0.983 | 0.963 |
| | RUS | 33.33% | 33.33% | 33.33% | **-0.005** | **-0.012** | **1.012** | **0.971** |
| **Maryland** | Initial | 64.58% | 22.4% | 13.0% | -0.108 | -0.082 | 0.798 | 0.842 |
| | MASC | 33.33% | 33.33% | 33.33% | -0.101 | 0.060 | 0.82 | 0.931 |
| | Geo-nei | 74.82% | 16.15% | 09.02% | -0.147 | -0.061 | 0.64 | 0.871 |
| | SMOTE | 33.33% | 33.33% | 33.33% | -0.060 | 0.006 | 0.886 | **0.988** |
| | RUS | 33.33% | 33.33% | 33.33% | **0.001** | **0.001** | **0.996** | 0.996 |

the best results for three states Montana, Mississippi, and North-Dakota as well as standing in the third best for two other states Colorado, and Maryland. Interesting observation comparing to results in Table 2 where SMOTE and RUS had better DI and SP results, it is observed that in model performance analysis, along with Geographical neighbors our method outperforms the SMOTE and RUS in all the states (Except for Mississippi where Geographical neighbors augmentations stands slighly worse than RUS) for both the metrics, accuracy and eq.odds. In Figure 3b where the accuracy results are compared, again the same situation is observed where MASC outperforms RUS and SMOTE and has the best accuracy in four of the states Montana, North-Dakota, Colorado, and Maryland and also stands in the third best for Mississippi.

## 4.5. Discussion

From an analytical perspective although our method MASC seems to stand statistically comparable to or lower than the SMOTE and RUS in terms of DI and SP in Table 2, but it outperforms both these methods in model performance results reported in Figure 3. The reason for the former is because in our experiments, we implement a version of SMOTE and RUS that statistically




(a) Equalized odds of each method on each of the five states

(b) Accuracy of each method on each of the five states

**Figure 3:** The results of a Logistic Regression model trained on each of the five augmentation methods and tested on the five states. In **(a)** the Eq.Odds values of the five approaches are depicted. In **(b)** the Accuracy of the model w.r.t. each augmentation method for each state.

augment protected groups, but still w.r.t. model performance measures, their augmentation is not comparable to real-world (genuine) data augmentations (e.g. MASC and Geographical neighbors). In that case, in Figure 3a and Figure 3b it is observed that MASC and Geographical neighbors (except for one case) outperform in all cases the two synthetic augmentations and once more we can highlight the importance/difference of real-world (genuine) data augmentation compared to synthetic/generated data.

Moreover, there are ethical and technical issues with SMOTEing and RUSing for protected group imbalance augmentation. Starting with RUS: looking at Figure 2a only 0.36%, *and* 1.35% of the population belong to the minority group1 (Black) of the states Montana and North-Dakota, respectively. For the cleaned dataset it is no more than 20, and 60 instances each. So, with such a small number of instances, it is very unlikely for any learning algorithm to produce reliable predictions while being imposed to test data. This was also observed in Figure 3b where the Eq.Odds results of the RUS method always report one because it basically predicts all the under-sampled data to belong to the majority class. Subsequently, this lack of reliable performance might even get worse in cases where learning parameters are applied to out-of-distribution (OOD) data. An example of OOD is training on the augmented data (in our experiments are 2019 US-Census dataset) and then applying the model for future data, e.g. 2020, and later data of the same state. This is left as an open question to interested readers to test and analyze the results. Furthermore, another question is: what about inter-sectional groups when there is imbalance also w.r.t. more than one attribute; for example how would SMOTE and RUS perform if gender and ethnicity are studied simultaneously? For example, if only exists one instance of coloured-skin females within the 20 samples in Montana dataset, the algorithm will only learn to infer one class label among these group of instances which could lead to highly unreliable and deficient predictions on test data.

In case of SMOTE, it over-samples the minority group of 20 or 60 instances to generate hundreds of times more data. So, these synthetically generated data are only specifically applicable to this application because they need to be very carefully tailored for the application. This may describe the worse performance in Accuracy and Eq.Odds despite balancing the groups in training data perfectly (GR, DI, and SP measures) in the experiments. One of the limitations of SMOTEing is data types. How would it work with categorical data? One has to define a multi-valued vector of features and statistically over-sample the outnumbered categories while they are encoded numerically which results in severe performance deterioration because of much larger search

space. However, our method is easily adaptable to categorical or other data types.

We would like to also highlight once again that in this study we only study the 2019 data so that conditions 2 and 3 of Section 2.1 do not apply to our analysis. In future works, it can be studied also where the distribution of target class (condition 2) or when the decision boundary changes (condition 3) which can happen when analyzing different historical records for each state, e.g. comparing the 2014-2019 data of each state. Also it is worth mentioning that there is a lack of similar datasets especially from the European countries that can be provided for research which can open up space for more studies in this direction.

## 5. Conclusion

In this paper, we propose a spectral clustering-based methodology to tackle data representation and protected-attribute group-imbalance biases. The motivation for developing this pipeline is to utilize contextually similar but separate datasets coming from similar sources, to augment one another in order to provide unbiased or less biased training sets using shared instances from contextually similar neighboring datasets. Our *MASC* approach identifies an optimal number of clusters based on inherent similarities of the input tasks and clusters them according to a robust and scalable MMD two-sampled test. Furthermore, it categorizes similar tasks based on their pairwise distribution discrepancies in a kernel-based affinity space. Experimental results on *New Adult* datasets reveal the promising performance of the proposed *MASC* in dataset debiasing and superior performance in improving predictive and fairness of learning models trained using the augmented training sets obtained by it. Moreover, it is preferable over synthetic data augmentation methods such as SMOTE and RUS since it augments based on genuine (real) existing data in contrary to the synthetic ones which are usually used under many ethical concerns. In future work, we will study the effect of normalized spectral clustering on the size and shape of clusters produced. We also encourage to extend our analysis to temporal aspects of the datasets by assuming change in the conditional probability of outcomes $P(Y|X)$ in $X \rightarrow Y$ problems for each year of the input datasets. Another interesting study would be to compare our method and the Geo-neib with a version of the SMOTE and RUS for multi-class imbalance or regression problems where the targets are multi-class or continuous.

## Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project ''NoBIAS - Artificial Intelligence without Bias''. This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

## References

[1] S. Verma, J. Rubin, Fairness definitions explained, in: FairWare@ICSE, ACM, 2018, pp. 1–7.

[2] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven artificial intelligence systems - an introductory survey, WIREs Data Mining Knowl. Discov. 10 (2020).

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021) 115:1–115:35.

[4] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (2023) 51:1–51:44.
[5] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset shift in machine learning, Mit Press, 2008.
[6] J. G. Moreno-Torres, T. Raeder, R. Alaíz-Rodríguez, N. V. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern Recognit. 45 (2012) 521–530.
[7] R. O. Duda, P. E. Hart, D. G. Stork, Pattern classification, 2nd Edition, Wiley, 2001.
[8] I. Goldenberg, G. I. Webb, Survey of distance measures for quantifying concept drift and shift in numeric data, Knowl. Inf. Syst. 60 (2019) 591–615.
[9] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under concept drift: A review, IEEE Trans. Knowl. Data Eng. 31 (2019) 2346–2363.
[10] Y. Zhang, W. Liu, Z. Chen, K. Li, J. Wang, On the properties of kullback-leibler divergence between gaussians, CoRR abs/2102.05485 (2021).
[11] J. Deasy, N. Simidjievski, P. Lió, Constraining variational inference with geometric jensen-shannon divergence, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 10647–10658.
[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. J. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (2012) 723–773.
[13] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (2007) 395–416.
[14] S. Jayasumana, R. I. Hartley, M. Salzmann, H. Li, M. T. Harandi, Kernel methods on riemannian manifolds with gaussian RBF kernels, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2464–2477.
[15] A. V. Little, M. Maggioni, J. M. Murphy, Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms, J. Mach. Learn. Res. 21 (2020) 6:1–6:66.
[16] Y. Nataliani, M. Yang, Powered gaussian kernel spectral clustering, Neural Comput. Appl. 31 (2019) 557–572.
[17] T. J. Park, K. J. Han, M. Kumar, S. Narayanan, Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap, IEEE Signal Process. Lett. 27 (2020) 381–385.
[18] A. Likas, N. Vlassis, J. J. Verbeek, The global k-means clustering algorithm, Pattern Recognit. 36 (2003) 451–461.
[19] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, in: NeurIPS, 2021, pp. 6478–6490.
[20] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.
[21] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: KDD, ACM, 2015, pp. 259–268.
[22] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, The zoo of fairness metrics in machine learning, CoRR abs/2106.00467 (2021).
[23] A. Fernández, S. García, F. Herrera, N. V. Chawla, SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, J. Artif. Intell. Res. 61 (2018) 863–905.
[24] S. Mishra, Handling imbalanced data: Smote vs. random undersampling, Int. Res. J. Eng. Technol 4 (2017) 317–320.
[25] J. K. Summers, L. M. Smith, L. C. Harwell, J. L. Case, C. M. Wade, K. R. Straub, H. M. Smith, An index of human well-being for the u.s.: A trio approach, Sustainability 6 (2014) 3915–3935. doi:10.3390/su6063915.
[26] G. W. Stewart, J.-g. Sun, Matrix perturbation theory, Academic press, 1990.