Measuring Fairness under Unawareness via Quantification (Extended Abstract)

Alessandro Fabris[1,2], Andrea Esuli[3], Alejandro Moreo[3], Fabrizio Sebastiani[3]

[1]Dipartimento di Ingegneria dell'Informazione
Università di Padova
35131 Padova, Italy
E-mail: fabrisal@dei.unipd.it

[2]Max Planck Institute for Security and Privacy
44799 Bochum, DE
E-mail: alessandro.fabris@mpi-sp.org

[3]Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
E-mail: {alejandro.moreo,fabrizio.sebastiani}@isti.cnr.it

Abstract. Models trained by means of supervised learning are increasingly deployed in high-stakes domains, and, when their predictions inform decisions about people, they inevitably impact (positively or negatively) on their lives. As a consequence, those in charge of developing these models must carefully evaluate their impact on different groups of people and ensure that sensitive demographic attributes, such as race or sex, do not result in unfair treatment for members of specific groups. For doing this, awareness of demographic attributes on the part of those evaluating model impacts is fundamental. Unfortunately, the collection of these attributes is often in conflict with industry practices and legislation on data minimization and privacy. For this reason, it may be hard to measure the *group fairness* of trained models, even from within the companies developing them. In this work, we tackle the problem of measuring group fairness under unawareness of sensitive attributes, by using techniques from *quantification*. We identify five important factors that complicate the estimation of fairness under unawareness and formalize them into five different experimental protocols under which we assess the effectiveness of different estimators of group fairness. We also consider the problem of potential model misuse to infer sensitive attributes at an individual level, and demonstrate that quantification is suitable for decoupling the (desirable) objective of measuring group fairness from the (undesirable) objective of inferring sensitive attributes of individuals.

1 Introduction

The widespread adoption of automated decision-making in high-stakes systems has brought about an increased attention to the underlying algorithms and to

their effects across sensitive groups. Typically, sensitive groups are subpopulations determined by social and demographic factors, such as race and sex. The unfair treatment of such demographic groups is ruled out by anti-discrimination laws and studied by a growing community of algorithmic fairness researchers. Important works in this space have addressed problems that may arise in the judicial system, in healthcare, in job search, and in computer vision, just to name a few domains that may be impacted. A common trait of these works is a careful definition and measurement of *group fairness*, typically viewed in terms of differences in quantities of interest, such as the acceptance rate, recall, or accuracy, across the salient subpopulations. According to popular definitions of fairness, large such differences correspond to low fairness on the part of the algorithms.

Unfortunately, sensitive demographic data, such as the race and sex of users, is often hard to obtain, for various reasons. There are several barriers to demographic data procurement which make measurement of fairness non-trivial even for the company that is developing and deploying a model. Legislation plays a major role in this, forbidding the collection of sensitive attributes in some domains. Even in the absence of explicit prohibition, privacy-by-design standards and a data minimization ethos push companies in the direction of avoiding the collection of sensitive attributes from their customers. Similarly, the prospect of negative media coverage is a clear concern, so companies often err on the side of caution and inaction. For these reasons, in a recent survey of industry practitioners, a majority of respondents stated that the availability of tools supporting fairness auditing without access to individual-level demographics would be very useful. In other words, the problem of *measuring algorithmic fairness under unawareness of sensitive attributes* is pressing, and requires ad-hoc solutions.

In the algorithmic fairness literature, much work has been done to propose techniques directly aimed at improving the fairness of a model (Donini et al., 2018; Hashimoto et al., 2018; He et al., 2020; Zafar et al., 2017). Comparably little attention, though, has been devoted to the problem of reliably measuring fairness. This represents an important and rather overlooked preliminary step to enforcing fairness and making algorithms more equitable across groups. More recent works have studied non-ideal conditions, such as noisy or missing group labels (Awasthi et al., 2020) and non-iid samples (Singh et al., 2021), showing that naïve fairness-enhancing algorithms may actually make a model *less* fair (Mehrotra and Celis, 2021).

In this work, we tackle the problem of measuring algorithmic fairness under unawareness of sensitive attributes, by using techniques from *quantification* (Esuli et al., 2023). Estimating, rather than the class labels of individual data points, the class prevalence values for sets (usually referred to as "samples") of such data points, is precisely the goal of practitioners looking to measure fairness under unawareness of sensitive attributes. When auditing an algorithm for group fairness, the aim is not the development of a model that is accurate for individual predictions (i.e., classification), which may be misused to infer people's demographics, such as a user's race, and may thus lead to the inappropriate and non-consensual utilization of this information. Rather, the central interest of

fairness audits is the reliable estimation of group-level quantities (i.e., quantification), such as the prevalence of women among the instances to which a certain class has been assigned by the model.

We consider several methods that have been proposed in the quantification literature and assess their suitability for estimating the fairness of a classifier under unawareness of sensitive attributes. More precisely, we adapt quantification approaches to measure a classifier's *demographic disparity* (Barocas et al., 2019), defined as the difference in acceptance rate across relevant subpopulations. Overall, we make the following contributions:

- **Five experimental protocols for five major challenges**. Drawing from the algorithmic fairness literature, we identify five important factors for the problem of estimating fairness under unawareness of sensitive attributes. These factors are based on challenges encountered in real-world applications, including the non-stationarity of processes generating the data, and the variable cardinality of the available samples. For each factor, we define and formalize a precise experimental protocol, through which we compare the performance of quantifiers (i.e., group-level prevalence estimators) generated by six different quantification methods (Sections 4.3-4.7).
- **Adaptation and ablation study**. We demonstrate a simple procedure to adapt and integrate quantification approaches into a wider machine learning pipeline with minimal orchestration effort. We prove the importance of each component through an ablation study (Section 4.8).
- **Quantifying without classifying.** We consider the problem of potential model misuse to maliciously infer demographic characteristics at an individual level, which represents a concern for methods based on proxy attributes. Proxy methods are estimators of sensitive attributes which exploit the correlation between available attributes (e.g., ZIP code) and the sensitive attributes (e.g., race) in order to infer the values of the latter. Through a set of experiments, we demonstrate two methods that yield precise estimates of demographic disparity but poor classification performance, thus decoupling the objectives of group-level prevalence estimation and individual-level class label prediction (Section 4.9).

It is worth noting some intrinsic limitations of fairness measures and proxy methods which are also applicable to this work. In essence, proxy methods exploit co-occurrence of membership in a group and display of a given trait, potentially learning, encoding and reinforcing stereotypical associations. Even when labels for sensitive attributes are available, they are not all equivalent. Self-reported labels are preferable to avoid external assignment (i.e., inference of sensitive attributes), which may be harmful. More in general, approaches that define sensitive attributes as rigid and fixed categories are limited since they impose a taxonomy onto people, erasing the needs and experiences of those who do not fit the envisioned categories. While acknowledging these limitations, we hope our work will help highlight, investigate and mitigate unfavourable outcomes for disadvantaged groups brought about by automated decision-making systems.

The outline of this work is the following. Section 2 presents the notation employed throughout this manuscript. Section 3 shows how these approaches can be adapted and integrated to measure demographic disparity. Section 4 discusses our experiments; we omit the actual results for reasons of space, and report them in the extended version of this paper (Fabris et al., 2023). Section 5 contains concluding remarks, discussing limitations and avenues for future work.

2 Notation

In this paper, we use the following notation. By $\mathbf{x}$ we indicate a data item drawn from a domain $\mathcal{X}$, encoding a set of non-sensitive attributes (i.e., features) taken by classifiers and quantifiers as an input. We use $\mathcal{S}$ to denote the domain of a sensitive attribute, binarily encoded to $\mathcal{S} = \{0, 1\}$ for ease of exposition, and by s a value that $\mathcal{S}$ may take. By y we indicate a class taking values on a binary domain $\mathcal{Y} = \{\ominus, \oplus\}$, representing the target of a prediction task.[1]

Symbol σ denotes a *sample*, i.e., a non-empty set of data points drawn from $\mathcal{X}$. By $p_\sigma(s)$ we indicate the true prevalence of attribute s in sample σ, while by $\hat{p}^q_\sigma(s)$ we indicate the estimate of this prevalence obtained by means of quantifier q, which we define as a function $q : 2^{\mathcal{X}} \to [0, 1]$. Since $0 \leq p_\sigma(s) \leq 1$ and $0 \leq \hat{p}^q_\sigma(s) \leq 1$ for all $s \in \mathcal{S}$, and since $\sum_{s\in\mathcal{S}} p_\sigma(s) = \sum_{s\in\mathcal{S}} \hat{p}^q_\sigma(s) = 1$, the $p_\sigma(s)$'s and the $\hat{p}^q_\sigma(s)$'s form two probability distributions across $\mathcal{S}$.

We also introduce random variables $X, S, Y, \hat{Y}$ which denote, respectively, data points from $\mathcal{X}$, their sensitive attributes, true labels, and predicted labels. By $\Pr(V = v)$ we indicate, as usual, the probability that random variable V takes value v, which we shorten as $\Pr(v)$ when V is clear from context. By $h : \mathcal{X} \to \mathcal{Y}$ we indicate a binary classifier that assigns classes in $\mathcal{Y}$ to data points; by $k : \mathcal{X} \to \mathcal{S}$ we instead indicate a binary classifier that assigns sensitive attributes $\mathcal{S}$ to data points (e.g., that predicts if a certain data item $\mathbf{x}$ is "female"). It is worth re-emphasizing that both h and k only use non-sensitive attributes from $\mathcal{X}$ as input variables. For ease of use, we will interchangeably write $h(\mathbf{x}) = y$ or $h_y(\mathbf{x}) = 1$, and $k(\mathbf{x}) = s$ or $k_s(\mathbf{x}) = 1$.

We consider three separate datasets, following the workflow of a realistic machine learning pipeline.

- A *training set* $\mathcal{D}_1$ for h, $\mathcal{D}_1 = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, typically of large cardinality. Given the inherent difficulties in demographic data procurement, we expect this dataset to contain no explicit information on the sensitive attributes $\mathcal{S}$.

[1] In this paper we assume the existence of a single binary sensitive attribute $\mathcal{S}$; however, there is no loss of generality in this, since everything we say can straightforwardly be extended to the case in which multiple sensitive attributes are present at the same time. Moreover, we focus on the case in which the classifier that we want to audit is a binary one, but the definitions and techniques we employ can be straightforwardly extended to a multiclass setting.

- A small *auxiliary set* $\mathcal{D}_2 = \{(\mathbf{x}_i, s_i) \mid \mathbf{x}_i \in \mathcal{X}, s_i \in \mathcal{S}\}$, employed to learn quantifiers for the sensitive attribute. This dataset may originate from a targeted effort, such as interviews, surveys sent to customers asking for voluntary disclosure of sensitive attributes, or other optional means to share demographic information. Alternatively it could derive from data acquisitions carried out for other purposes. Both $\mathcal{D}_1$ and $\mathcal{D}_2$ are in the development domain of our machine learning pipeline.
- A *deployment set* $\mathcal{D}_3 = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathcal{X}\}$ which emulates the production domain for classifier h, whose demographic parity we aim to measure. Alternatively, acting proactively rather than reactively, $\mathcal{D}_3$ could also be a held-out test set available at a company for pre-deployment audits. From the perspective of the estimation task at hand, i.e. estimating the demographic disparity of h, $\mathcal{D}_2$ represents the quantifiers' training set, while $\mathcal{D}_3$ is their test set.

3 Using quantification to measure fairness under unawareness of sensitive attributes

We adapt the above quantification approaches for estimating a classifier's fairness. We define classifier fairness in terms of *demographic parity* (also called *statistical parity* (Dwork et al., 2012) or *independence* (Barocas et al., 2019)), and, in particular, of a flavour of demographic parity based on the distribution of sensitive attribute $\mathcal{S}$ conditional on the prediction of the classifer, as proposed in (Wachter et al., 2020). We call our estimand the *demographic disparity* of classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$ for attribute value s, and define it as

$$\Delta(s) = \Pr(S = s|\hat{Y} = \ominus) - \Pr(S = s|\hat{Y} = \oplus) \tag{1}$$

or, more concisely,

$$\Delta(s) = \Pr(s|\ominus) - \Pr(s|\oplus) \tag{2}$$

It is worth reemphasizing that the sensitive attribute $\mathcal{S}$ does *not* belong to the set of attributes $\mathcal{X}$ which generate the feature space on which classifier h operates (in other words, when training h we are *unaware* of $\mathcal{S}$). Demographic disparity measures whether the prevalence of the sensitive attribute in the group assigned to the positive class is the same as in the group assigned to the negative class; a value $\Delta(s) = 0$ indicates maximum fairness, while values of $\Delta(s) = -1$ or $\Delta(s) = +1$ indicate minimum fairness, with the sign of $\Delta(s)$ indicating whether, for $S = s$, the classifier is biased towards the $\oplus$ class or the $\ominus$ class, respectively.

Example 1. Assume that $\mathcal{S}$ stands for "sex", s for "female", and that the classifier is in charge of recommending loan applicants for acceptance, classifying them as "grant" ($\oplus$) or "deny" ($\ominus$). For simplicity, let us assume the outcome of the classifier to directly translate into a decision without human supervision. The bank might want to check that the fraction of females out of the total number of loan recipients is approximately the same as the fraction of females out of the total number of applicants who are denied the loan. In other words, the bank

might want $\Delta(s)$ to be close to 0. Of course, if the bank is aware of the sex of each applicant, this constraint is very easy to check and, potentially, enforce. If the bank is unaware of applicants' sex, as we assume here, the problem is not trivial, and this is where our techniques come in.

In estimating the demographic disparity of h, our focus is on the deployment set where h is supporting the decision-making process. To highlight this fact, we rewrite Equation 2 by making the dependence of $\Delta(s)$ on $\mathcal{D}_3$ explicit, i.e.,

$$\Delta(s) = p_{\mathcal{D}_3^{\ominus}}(s) - p_{\mathcal{D}_3^{\oplus}}(s) \tag{3}$$

where we define

$$\begin{aligned} \mathcal{D}_3^{\oplus} &= \{\mathbf{x} \in \mathcal{D}_3 \mid h(\mathbf{x}) = \oplus\} \\ \mathcal{D}_3^{\ominus} &= \{\mathbf{x} \in \mathcal{D}_3 \mid h(\mathbf{x}) = \ominus\} \end{aligned} \tag{4}$$

and where we make explicit the fact that, if a value s that attribute $\mathcal{S}$ can take is viewed as a class, the probabilities $\Pr(s|\ominus)$ and $\Pr(s|\oplus)$ of Equation 2 may be seen as the prevalence values of class s in the two samples $\mathcal{D}_3^{\oplus}$ and $\mathcal{D}_3^{\ominus}$. In other words, measuring demographic disparity is reduced to estimating the prevalence values of class s in the two samples $\mathcal{D}_3^{\oplus}$ and $\mathcal{D}_3^{\ominus}$, i.e., *it can be framed as a task of quantification*.

This approach can be easily integrated into existing machine learning pipelines with little orchestration effort. Below, we summarize the workflow we envision:

1. A classifier $h : \mathcal{X} \to \mathcal{Y}$ is trained (under unawareness of sensitive attribute $\mathcal{S}$) on $\mathcal{D}_1$ and ready for production. The assumption that, at this stage, we are unaware of sensitive attribute $\mathcal{S}$ is due to the inherent difficulties in demographic data procurement already mentioned in Section 1.
2. Classifier h naturally imposes a partition of the auxiliary set $\mathcal{D}_2$ into $\mathcal{D}_2^{\ominus} = \{(\mathbf{x}_i, s_i) \in \mathcal{D}_2 \mid h(\mathbf{x}) = \ominus\}$ and $\mathcal{D}_2^{\oplus} = \{(\mathbf{x}_i, s_i) \in \mathcal{D}_2 \mid h(\mathbf{x}) = \oplus\}$. These two disjoint datasets act as the training sets for the two quantifiers $q_{\ominus}$ and $q_{\oplus}$. Quantifier $q_{\ominus}$ (or its dual $q_{\oplus}$) is trained on $\mathcal{D}_2^{\ominus}$ (resp., $\mathcal{D}_2^{\oplus}$) to estimate the prevalence of data points where $S = s$ among the data points labelled with $\ominus$ (resp., $\oplus$).
3. Classifier h also imposes a partition of the deployment set $\mathcal{D}_3$ into $\mathcal{D}_3^{\ominus} = \{\mathbf{x} \in \mathcal{D}_3 \mid h(\mathbf{x}) = \ominus\}$ and $\mathcal{D}_3^{\oplus} = \{\mathbf{x} \in \mathcal{D}_3 \mid h(\mathbf{x}) = \oplus\}$. Quantifiers $q_{\ominus}$ and $q_{\oplus}$ trained in Step 2 are applied to these datasets to obtain an estimate of the prevalence of s in $\mathcal{D}_3^{\ominus}$ and $\mathcal{D}_3^{\oplus}$. The demographic disparity of h, defined in Equation 1, can thus be estimated as

$$\hat{\Delta}(s) = \hat{p}^{q_{\ominus}}_{\mathcal{D}_3^{\ominus}}(s) - \hat{p}^{q_{\oplus}}_{\mathcal{D}_3^{\oplus}}(s) \tag{5}$$

   where, as we recall from Section 2, $\hat{p}^{q}_{\sigma}(s)$ denotes the prevalence of attribute value s in set σ as estimated via quantification method q.

This quantification-based way of tackling demographic disparity is suited for a non-invasive auditing procedure, since it allows unawareness of the sensitive

Table 1: Summary of experimental protocols.

Protocol name	Variable	Section
`sample-prev-`$\mathcal{D}_1$	joint distribution of (S, Y) in $\mathcal{D}_1$, via sampling	§ 4.3
`flip-prev-`$\mathcal{D}_1$	joint distribution of (S, Y) in $\mathcal{D}_1$, via label flipping	§ 4.4
`sample-size-`$\mathcal{D}_2$	size of $\mathcal{D}_2$, via sampling	§ 4.5
`sample-prev-`$\mathcal{D}_2$	joint distribution of $(S, \hat{Y})$ in $\mathcal{D}_2$, via sampling	§ 4.6
`sample-prev-`$\mathcal{D}_3$	joint distribution of $(S, \hat{Y})$ in $\mathcal{D}_3$, via sampling	§ 4.7

attribute $\mathcal{S}$ in the set $\mathcal{D}_1$ used for training the classifier h to be audited and in the set $\mathcal{D}_3$ on which this classifier is going to be deployed; it only requires the availability of an auxiliary data set $\mathcal{D}_2$ where attribute $\mathcal{S}$ is present. Dataset $\mathcal{D}_2$ may originate from a targeted effort, such as interviews, surveys sent to customers asking for voluntary disclosure of sensitive attributes, or other optional means to share demographic information. Alternatively it could derive from data acquisitions carried out for other purposes.

Additionally, we note that this approach is extremely suitable to situations in which the prevalence of attribute value s in $\mathcal{D}_2$ is possibly very different from the prevalence of s in the test set $\mathcal{D}_3$ (a situation that certainly characterizes many operational environments) since the best quantification approaches are robust by construction to distribution drift, as we will show in the next section.

4 Experiments

4.1 General setup

In this section we describe an evaluation of different estimators of demographic disparity. We propose five experimental protocols (Sections 4.3–4.7) summarized in Table 1. Each protocol focuses on a single factor of import for the estimation problem, varying the size and mutual shift of the training, auxiliary, and deployment set. Protocol names are in the form `action-characteristic-dataset`, as they act on datasets ($\mathcal{D}_1$, $\mathcal{D}_2$ or $\mathcal{D}_3$) modifying their characteristics (size or class prevalence) through one of two actions (sampling or label flipping). We investigate the effect of each factor on the performance of six estimators of demographic disparity, keeping the remaining factors constant.

Under each experimental protocol, the size or the prevalence of a given dataset is carefully varied based on the protocol definition. For every protocol, we perform an extensive empirical evaluation as follows:

- We compare the performance of each estimation technique on three datasets (Adult, COMPAS, and Credit Card Default). The datasets and respective preprocessing are described in detail in Section 4.2.
- We split a given dataset into $\mathcal{D}_A, \mathcal{D}_B, \mathcal{D}_C$, three stratified subsets of identical size and same distribution over (S, Y). Five such random splits are

performed. To test each estimator under the same conditions, these splits are the same for every method.
- For each split, we permute the role of the stratified subsets $\mathcal{D}_A, \mathcal{D}_B, \mathcal{D}_C$, so that each subset alternatively serves as the training ($\mathcal{D}_1$), auxiliary ($\mathcal{D}_2$), or deployment set ($\mathcal{D}_3$). All (six) such permutations are tested.
- Whenever an experimental protocol requires sampling from a subset, for instance when artificially altering a class prevalence value, we perform 10 different samplings. To perform extensive experiments at a reasonable computational cost, every time an experimental protocol requires changing a dataset $\mathcal{D}$ into a shifted version $\breve{\mathcal{D}}$, we also reduce its cardinality to $|\breve{\mathcal{D}}| = 500$. Further details and implications of this choice on each experimental protocol are provided in the context of the protocol's setup.
- Different learning approaches can be used to train the sensitive attribute classifier k underlying each quantification method. We test Logistic Regression (LR) and Support Vector Machines (SVM).[2] Sections 4.3–4.7 report results of quantification algorithms wrapped around a LR classifier. Analogous results obtain for SVMs are reported in (Fabris et al., 2023).
- The classifier h, whose demographic disparity we aim to estimate, is LR trained with balanced class weights (i.e., loss weights inversely proportional to class frequencies).
- To measure the effect of a given factor on the performance of different quantifiers, we report the signed estimation error, derived from Equations 3 and 5 as follows:

$$\begin{aligned} e &= \hat{\Delta}(s) - \Delta(s) \\ &= \left[\hat{p}^{q^{\ominus}}_{\mathcal{D}_3^{\ominus}}(s) - \hat{p}^{q^{\oplus}}_{\mathcal{D}_3^{\oplus}}(s)\right] - \left[p_{\mathcal{D}_3^{\ominus}}(s) - p_{\mathcal{D}_3^{\oplus}}(s)\right] \end{aligned} \tag{6}$$

  We summarize the experiments by reporting the Mean Absolute Error (MAE) and Mean Squared Error (MSE), where the mean of errors e_i is computed over multiple experiments E.

Overall, our experiments consist of over 700,000 separate estimates of demographic disparity.[3] The actual results of our experiments are omitted from this paper for reasons of space; for these results we refer the reader to the extended version of this paper (Fabris et al., 2023).

The remainder of this section is organized as follows. Section 4.2 presents the chosen datasets and the applied preprocessing. Sections 4.3–4.7 motivate and detail the experimental protocols, reporting the performance of different demographic disparity estimators. Section 4.8 describes an ablation study, aimed at investigating the benefits of training and maintaining multiple class-specific

[2] Some among the quantification methods we test in this study require the classifier to output posterior probabilities (as is the case for LR). If a classifier natively outputs classification scores that are not probabilities (as is the case for SVM), the former can be converted into the latter via "probability calibration".

[3] Code available at `https://github.com/alessandro-fabris/ql4facct`.

Table 2: Dataset statistics after preprocessing.

Dataset	#data items	#non-sensitive features	sensitive attribute	$S=1$	$Pr(S=1)$	target variable	$Y=\oplus$	$Pr(Y=1)$
Adult	45,222	84	sex	Male	0.675	income	>50K	0.248
COMPAS	5,278	6	race	Caucasian	0.398	recidivist	no	0.498
CreditCard	30,000	81	sex	Male	0.396	default	no	0.779

quantifiers rather than a single one. Finally, Section 4.9 shows that good estimators of demographic disparity are not necessarily good at classifying the sensitive attribute at an individual level, so that reliable fairness auditing may be decoupled from this undesirable misuse of the same models.

4.2 Datasets

We perform our experiments on three datasets. We choose Adult and COMPAS, two standard datasets in the algorithmic fairness community, and Credit Card Default (hereafter: CreditCard), which serves as a representative use case for a bank performing a fairness audit of a prediction tool used internally. A summary of these datasets and related statistics is reported in Table 2. See the extended version of this paper (Fabris et al., 2023) for more details on these datasets.

4.3 Protocol `sample-prev-`$\mathcal{D}_1$

In the first experimental protocol, we evaluate the impact of shifts in the training set $\mathcal{D}_1$, by drawing different subsets $\breve{\mathcal{D}}_1$ as we vary $\Pr(Y=S)$.[4] More specifically, we vary $\Pr(Y=S)$ between 0 and 1 with a step of 0.1. In other words, we sample at random from $\mathcal{D}_1$ a proportion p of instances $(\mathbf{x}_i, s_i, y_i)$ such that $Y=S$ and a proportion $(1-p)$ such that $Y \neq S$, with $p \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. It is worth noting that we defined $\mathcal{D}_1$, in Section 2, as a training set involving $(\mathcal{X}, \mathcal{Y})$. Here we exploit our knowledge of $\mathcal{S}$ to control the dataset shift between training and test conditions, emulating a biased data collection procedure. Once a training set has been selected, the classifier h is learnt exclusively from non-sensitive attributes $\mathcal{X}$, completely disregarding the sensitive attribute $\mathcal{S}$. We choose a limited cardinality $|\breve{\mathcal{D}}_1| = 500$, which lets us perform multiple repetitions at reasonable computational costs, as outlined in Section 4.1. While this may impact the quality of the classifier h, this aspect is not the central focus of the present work.

This experimental protocol aligns with biased data collection procedures, sometimes referred to as *censored data*. Indeed, it is common for the ground truth variable to represent a mere proxy for the actual quantity of interest, with non-trivial sampling effects between the two. For instance, the validity of

[4] Although Y and S take values from different domains, by $Y=S$ we mean $(Y=\oplus \wedge S=1) \vee (Y=\ominus \wedge S=0)$, i.e. a situation where positive outcomes are associated with group $S=1$ and negative outcomes with group $S=0$.

arrest data as a proxy for offence has been brought into question. Indeed, in this domain, different sources of sampling bias may be in action, such as uneven allocation of police resources across jurisdictions and neighbourhoods and lower levels of cooperation in populations who feel oppressed by law enforcement.

By varying $\Pr(Y = S)$ we are imposing a spurious correlation between Y and S, which may be picked up by the classifier h. In extreme situations, such as when $\Pr(Y = S) \simeq 1$, a classifier h may end up confounding the concepts behind S and Y. In turn, we expect this to unevenly impact the acceptance rates for the two demographic groups, effectively changing the demographic disparity of h, i.e., our estimand $\Delta(s)$.

4.4 Protocol `flip-prev-`$\mathcal{D}_1$

Certain biases in the training set resulting from domain-specific practices, such as the use of arrest as a substitute for offence, may be modelled as either a selection bias or a label bias distorting the ground truth variable Y. With this experimental protocol, we impose the latter bias by actively flipping some ground truth labels Y in $\mathcal{D}_1$ based on their sensitive attribute. Similarly to `sample-prev-`$\mathcal{D}_1$, this protocol achieves a given association between the target Y and sensitive variable S in the training set $\mathcal{D}_1$. However, instead of sampling, it does so by flipping the Y label of some data points. More specifically, we impose $\Pr(Y = \ominus | S = 0) = \Pr(Y = \oplus | S = 1) = p$ and let p take values across eleven evenly spaced values between 0 and 1. For every value of p, we firstly sample a random subset $\breve{\mathcal{D}}_1$ of the training set with cardinality 500. Next, we actively flip some Y labels in both demographic groups, until both $\Pr(Y = \ominus | S = 0)$ and $\Pr(Y = \oplus | S = 1)$ reach a desired value of $p \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. Finally, we train a classifier h on the attributes $\mathcal{X}$ and modified ground truth Y of $\breve{\mathcal{D}}_1$.

This experimental protocol is compatible with settings where the training data captures a distorted ground truth due to systematic biases and group-dependent annotation accuracy. As an example, the quality of medical diagnoses can depend on race, sex and socio-economical status. Moreover, health care expenditures have been used as a proxy to train an algorithm deployed nationwide in the US to estimate patients' health care needs, resulting in systematic underestimation of the needs of black patients. In the hiring domain, employer response rates to resumes have been found to vary with the perceived ethnic origin of an applicant's name. Finally, the gender gap in mathematical performance, while negligible in elementary school, has been found to increase with age, possibly due to gender stereotypes arising in this domain from an early age and to the prescriptive nature of these stereotypes. These are all examples where the "ground truth" associated with a dataset is distorted to the disadvantage of a sensitive demographic group.

Similarly to Section 4.3, we expect this experimental protocol to bring about sizeable variations in the demographic disparity of classifier h due to the strong correlation we are imposing between S and Y via label flipping.

4.5 Protocol `sample-size-`$\mathcal{D}_2$

A further factor of interest for the estimation problem is the size of the auxiliary set $\mathcal{D}_2$, whose influence is studied in this experimental protocol. Our goal is to understand how low we can go in the small data regime, before degrading the performance of different estimation techniques. We consider subsets $\breve{\mathcal{D}}_2$ of the auxiliary set, sampling instances uniformly without replacement from it. We let cardinality $|\breve{\mathcal{D}}_2|$ take five values that are evenly spaced on a log scale, between a minimum sample size $|\breve{\mathcal{D}}_2|$=1,000 and a maximum size $|\breve{\mathcal{D}}_2| = |\mathcal{D}_2|$. In other words, we let the cardinality of the auxiliary set take five different values between 1,000 and $|\mathcal{D}_2|$ in a geometric progression. As described in Section 4.1, for each cardinality of the auxiliary set we wish to test, we perform ten samplings over five splits and six permutations, for a total of 300 repetitions per approach per dataset.

This protocol is justified by the well-documented difficulties in demographic data procurement for industry practitioners, which vary depending on domain, company, and other factors of disparate nature. Furthermore, the collection of sensitive attributes in the US is highly industry-dependent, ranging from mandatory to forbidden, depending on the fragmented regulation applicable in each domain. Finally, high quality auxiliary sets may be obtained through optional surveys, for which response rates are highly dependent on trust, and can be improved by making the intended use for the data clearer.

For these reasons, the cardinality of the auxiliary set $\mathcal{D}_2$ is an interesting variable in the context of fairness audits. The estimation methods we consider have peculiar data requirements, with diverse purposes (e.g., estimation of true positive rates – *tpr*) and approaches. For this reason, interesting patterns should emerge from this protocol. We expect key trends for the estimation error to vary monotonically with cardinality $|\breve{\mathcal{D}}_2|$, which is why we let it vary according to a geometric progression.

4.6 Protocol `sample-prev-`$\mathcal{D}_2$

The auxiliary set $\mathcal{D}_2$ can also display significant dataset shifts with respect to the the sets $\mathcal{D}_1$ and $\mathcal{D}_3$ available during training or at deployment. With this experimental protocol, we assess the estimation error under shifts which affect either $\mathcal{D}_2^{\ominus}$ or $\mathcal{D}_2^{\oplus}$, i.e., the subsets of $\mathcal{D}_2$ labelled positively or negatively by classifier h. We consider two experimental sub-protocols, describing variations in the prevalence of sensitive variable S in either subset. More specifically, we let $\Pr(s|\ominus)$ (or its dual $\Pr(s|\oplus)$) take 9 evenly spaced values between 0.1 and 0.9. We avoid extreme values of 0 and 1 which would make either demographic group $S = 0$ or $S = 1$ absent from the training set of one quantifier. To exemplify, in sub-protocol `sample-prev-`$\mathcal{D}_2^{\ominus}$ we let the prevalence $\Pr(s|\ominus)$ in $\breve{\mathcal{D}}_2^{\ominus}$ take values in $\{0.1, 0.2 \ldots, 0.8, 0.9\}$, while the remaining subset $\breve{\mathcal{D}}_2^{\oplus}$ remains at is natural prevalence $\Pr(s|\oplus)$.[5] For each repetition, we set $|\breve{\mathcal{D}}_2^{\ominus}| = |\breve{\mathcal{D}}_2^{\oplus}| = 500$. This makes

[5] The natural prevalence is matched allowing for small fluctuations due to subsampling.

for a challenging quantification setting and allows for fast training of multiple quantifiers across many repetitions.

This protocol captures issues of representativeness in demographic data, e.g., due to non-uniform response rates across subpopulations. Given the importance of trust for the provision of one's sensitive attributes, in some domains this practice is considered akin to a *data donation*. Individuals from groups that historically had worse quality or lower acceptance rates for a service can be hesitant to disclose their membership to said group, fearing it may be used against them as grounds for rejection or discrimination. This may be especially true for individuals who perceive to be at high risk of rejection, bringing about complex selection biases, jointly dependent on S and Y, or S and $\hat{Y}$ if individuals have some knowledge of the classification procedure. For example, health care providers are advised to collect information about patients' race to monitor the quality of services across subpopulations. In a field study, 28% of patients reported discomfort about disclosure of their own race to a clerk, with black patients significantly less comfortable than white patients on average.

This is the first protocol we describe where quantifiers are trained on subsets $\breve{\mathcal{D}}_2^{\ominus}$, $\breve{\mathcal{D}}_2^{\oplus}$ that have a different prevalence for the sensitive variable S with respect to their counterparts $\mathcal{D}_3^{\ominus}$, $\mathcal{D}_3^{\oplus}$ in the deployment set. More specifically, with this protocol, we vary the joint distribution of $(S, \hat{Y})$ to directly influence the demographic disparity of the classifier h on the auxiliary set $\mathcal{D}_2$, and move it away from the value $\Delta(s)$ of the same measure computed on the deployment set $\mathcal{D}_3$. This is a fundamental evaluation protocol as it makes our estimand different across $\mathcal{D}_2$ (or, more precisely, its modified version $\breve{\mathcal{D}}_2$) and $\mathcal{D}_3$, which is typically expected in practice. If this were not the case, a practitioner could simply resort to an explicit computation of demographic disparity on the auxiliary set $\mathcal{D}_2$ and deem it representative of any deployment condition. Given this reasoning, we borrow this protocol from the quantification literature to cause sizeable variations in the demographic disparity of h across $\mathcal{D}_2$ and $\mathcal{D}_3$, which act as the training and test set to different quantifiers. We expect these variations to bring about clear trends in the estimation error of demographic parity for the approaches considered in this work.

4.7 Protocol `sample-prev-`$\mathcal{D}_3$

This is essentially the counterpart of protocol `sample-prev-`$\mathcal{D}_2$ (Section 4.6), focusing on shifts in the test set $\mathcal{D}_3$. Similarly, we consider two sub-protocols that model changes in the prevalence of a sensitive variable S in the test subset of either positively or negatively predicted instances, called $\mathcal{D}_3^{\ominus}$ and $\mathcal{D}_3^{\oplus}$. More in detail, we let $\Pr(s|\ominus)$ (or its dual $\Pr(s|\oplus)$) in $\breve{\mathcal{D}}_3$ take eleven evenly spaced values between 0 and 1. For example, under sub-protocol `sample-prev-`$\mathcal{D}_3^{\ominus}$, we vary the prevalence of sensitive attribute S in $\breve{\mathcal{D}}_3^{\ominus}$, so that $\Pr(s|\ominus) \in \{0.0, 0.1 \ldots, 0.9, 1.0\}$, while keeping the prevalence in $\breve{\mathcal{D}}_3^{\oplus}$ fixed. Contrary to protocol `sample-prev-`$\mathcal{D}_2$, here we also allow for extreme prevalence values of 0 and 1 for the sensitive attribute S, as this does not invalidate the quantifiers' training. For both sub-

protocols, in each repetition we sample subsets of the test set $\mathcal{D}_3$ such that $|\breve{\mathcal{D}}_3^{\ominus}| = |\breve{\mathcal{D}}_3^{\oplus}| = 500$.

This protocol accounts for the inevitable evolution of phenomena, especially those related to human behaviour. Indeed, it is common in real-world scenarios for data generation processes to be non-stationary and change across training and test, due e.g., to seasonality or any sort of unmodelled novelty and difference in populations. Given most work on algorithmic fairness focuses on decisions or predictions about people, and given the unavoidable role of change in human lives, values, and behaviour, the above considerations about non-stationarity seem particularly relevant in this context. For instance, data available from one population is often repurposed to train algorithms that will be deployed on a different population, requiring ad-hoc fair learning approaches and evoking the *portability trap* of fair machine learning. Moreover, agents may be responsive to novel technology in their social context and adapt their behaviour accordingly, causing *ripple effects* and *feedback loops*. Furthermore, as a concrete (although spurious) example of a shift in a popular fairness dataset, the repeated offense rate for black and white defendants in the COMPAS datasetincreases sharply between 2013 and 2014. As a final example, personalized pricing constitutes an increasingly possible practice with non-trivial fairness concerns and inevitable shifts due to changing habits and environments.

In the quantification literature, this is the most common evaluation protocol. Similarly to `sample-prev-`$\mathcal{D}_2$, it imposes shifts in the estimand between the training and testing conditions of a quantifier, represented by the auxiliary set $\mathcal{D}_2$ and the deployment set $\mathcal{D}_3$, respectively. Through this protocol, we expect to find similar patterns to those highlighted in Section 4.6, with the roles of the auxiliary set $\mathcal{D}_2$ and test set $\mathcal{D}_3$ now switched. Under this protocol, $\mathcal{D}_3$ has a smaller cardinality and variable prevalence (and is referred to as $\breve{\mathcal{D}}_3$ for this reason), while $\mathcal{D}_2$ is left to its original cardinality and prevalence of sensitive attribute S.

4.8 Ablation study

In Sections 4.3–4.7 we tested six approaches to estimate demographic disparity. For each approach, we exploited multiple quantifiers for the sensitive attribute $\mathcal{S}$, namely one for each class in the codomain of classifier h. In the binary setting adopted in this work, where $\mathcal{Y} = \{\ominus, \oplus\}$, we trained two quantifiers. One quantifier was trained on the set of positively-classified instances of the auxiliary set $\mathcal{D}_2^{\oplus} = \{(\mathbf{x}_i, s_i) \in \mathcal{D}_2 \mid h(\mathbf{x}) = \oplus\}$ and deployed to quantify the prevalence of sensitive instances (such that $S = s$) within the deployment subset $\mathcal{D}_3^{\oplus}$. The remaining quantifier was trained on $\mathcal{D}_2^{\ominus}$ and deployed on $\mathcal{D}_3^{\ominus}$.

Training and maintaining multiple quantifiers is more expensive and cumbersome than having a single one. Firstly, quantifiers that depend on the classification outcome $\hat{y} = h(\mathbf{x})$ require re-training every time h is modified, e.g., due to a model update being rolled out. Secondly, the cost of maintenance is multiplied by the number of classes $|\mathcal{Y}|$ that are possible for the outcome variable. To ensure these downsides are compensated by performance improvements, we

perform an ablation study evaluating the performance of different estimators of demographic disparity supported by a single quantifier.

In this section, we concentrate on three estimation approaches, namely CC, SLD and PACC. CC is chosen as the naïve baseline adopted by practitioners unaware of ad-hoc approaches for prevalence estimation. SLD and PACC are among the best performers in Sections 4.3–4.7, displaying low bias or variance across all protocols. We compare their performance under two settings. In the first setting, adopted thus far, two separate quantifiers $q_{\ominus}$ and $q_{\oplus}$ are trained on $\mathcal{D}_2^{\ominus}$, $\mathcal{D}_2^{\oplus}$ and deployed on $\mathcal{D}_3^{\ominus}$, $\mathcal{D}_3^{\oplus}$, respectively. In the second setting, we train a single quantifier q on $\mathcal{D}_2$ and deploy it separately on $\mathcal{D}_3^{\ominus}$ and $\mathcal{D}_3^{\oplus}$ to estimate $\hat{\Delta}(s)$ via Equation 5, specialized so that $q_{\ominus}$ and $q_{\oplus}$ are the same quantifier.

4.9 Quantifying without classifying

The motivating use case for this work are internal audits of group fairness, to characterize a model and its potential to harm sensitive categories of users. We envision this as an important first step to empower practitioners in arguing for resources and, more broadly, advocate for deeper understanding and careful evaluation of models. Unfortunately, developing a tool to infer demographic information, even if motivated by careful intentions and good faith, leaves open the possibility for misuse, especially at an individual level. Once a predictive tool, also capable of instance-level classification, is available, it will be tempting for some actors to exploit it precisely for this purpose.

For example, the *Bayesian Improved Surname Geocoding* (BISG) method is intended to estimate population-level disparities in health care. However, it was also used to identify individuals potentially eligible for settlements related to discriminatory practices by auto lending companies. Automatic inference of sensitive attributes of individuals is problematic for several reasons. Such procedure exploits the co-occurrence of membership in a group and display of a given trait, running the risk of learning, encoding and reinforcing stereotypical associations. While also true of group-level estimates, this practice is particularly troublesome at an individual level, where it is likely to cause harms for people who do not fit the norm, resulting, for instance, in misgendering and the associated negative effects. Even when "accurate", the mere act of externally assigning sensitive labels can be problematic. For example, gender assignment may be forceful and lead to psychological harm for individuals.

We here aim to demonstrate that it is possible to decouple the objective of (group-level) quantification of sensitive attributes from that of (individual-level) classification. For protocols in previous sections, we compute the accuracy and F_1 score (defined below) of the sensitive attribute classifier k underlying the tested quantifiers, comparing it against their estimation error for class prevalence of the sensitive attribute S (Equation 6).

5 Discussion and conclusion

Measuring the differential impact of models on groups of individuals is important to understand their effects in the real world and their tendency to encode and reinforce divisions and privilege across sensitive attributes. Unfortunately, in practice, demographic attributes are often unavailable. In this work we have taken the perspective of responsible practitioners, interested in internal fairness audits of production models. We have tackled the problem of measuring group fairness under unawareness of sensitive attributes by applying approaches from the quantification learning literature that are specifically designed for group-level estimation rather than individual-level classification; this is convenient, since practitioners who try to measure fairness under unawareness are precisely interested in group-level estimates.

We have studied the problem of estimating a classifier's demographic disparity at deployment under unawareness of sensitive attributes, with access to a disjoint auxiliary set of data for which demographic information is available. Drawing from the algorithmic fairness literature, we have identified five factors of import for this problem, associating each of them with a formal evaluation protocol. These factors mirror challenges in real-world applications, including dataset shift and variable cardinality for auxiliary datasets comprising demographic information. We have tested five quantification methods under every protocol, comparing them against the naïve Classify-and-Count (CC) method, which represents the default approach for practitioners unaware of quantification. Each quantification approach was shown to outperform CC under most combinations of 5 protocols, 3 datasets, and 2 underlying learners. Moreover, we have shown a simple approach to integrate quantification methods into existing machine learning pipelines with little orchestration effort, and demonstrated the importance of each component through an ablation study.

Finally, we have considered the problem of model misuse to infer demographic characteristics at an individual level, which represents a concern when developing models to measure group fairness via proxy attributes. Through a dedicated set of experiments, we have shown that it is possible to obtain precise estimates of demographic disparity from methods that have poor classification performance. This is a positive result for decoupling these two objectives, which should help deter from the exploitation of these models for individual-level inference.

Acknowledgments

The work by Alessandro Fabris was supported by MIUR (Italian Minister for Education) under the "Departments of Excellence" initiative (Law 232/2016). The work by Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani has been supported by the SoBigData++ project, funded by the European Commission (Grant 871042) under the H2020 Programme INFRAIA-2019-1, and by the AI4Media project, funded by the European Commission (Grant 951911) under the H2020 Programme ICT-48-2020. The authors' opinions do not necessarily reflect those of the European Commission.

Bibliography

Awasthi P, Kleindessner M, Morgenstern J (2020) Equalized odds postprocessing under imperfect group information. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020), Virtual Event, pp 1770–1780

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. fairmlbook.org, URL http://www.fairmlbook.org

Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical risk minimization under fairness constraints. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, CA, pp 2791–2801

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012), Cambridge, US, pp 214–226, DOI 10.1145/2090236.2090255

Esuli A, Fabris A, Moreo A, Sebastiani F (2023) Learning to quantify. Springer Nature, Cham, CH

Fabris A, Esuli A, Moreo A, Sebastiani F (2023) Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. Journal of Artificial Intelligence Research 76:1117–1180, DOI 10.1613/jair.1.14033

Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness without demographics in repeated loss minimization. In: Proceedings of the 35th International Conference on Machine Learning (ICML 2018), Stockholm, SE, pp 1929–1938

He Y, Burghardt K, Lerman K (2020) A geometric solution to fair representations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2020), New York, US, pp 279–285, DOI 10.1145/3375627.3375864

Mehrotra A, Celis LE (2021) Mitigating bias in set selection with noisy protected attributes. In: Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021), Toronto, CA, pp 237–248, DOI 10.1145/3442188.3445887

Singh H, Singh R, Mhasawade V, Chunara R (2021) Fairness violations and mitigation under covariate shift. In: Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021), Toronto, CA, pp 3–13, DOI 10.1145/3442188.3445865

Wachter S, Mittelstadt B, Russell C (2020) Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law and Security Review 41, DOI 10.1016/j.clsr.2021.105567, article 105567

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: Mechanisms for fair classification. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), Fort Lauderdale, US, pp 962–970