2011 11th IEEE International Conference on Data Mining

Handling Conditional Discrimination

Indrė Žliobaitė
Bournemouth University, UK
izliobaite@bournemouth.ac.uk

Faisal Kamiran
TU Eindhoven, the Netherlands
f.kamiran@tue.nl

Toon Calders
TU Eindhoven, the Netherlands
t.calders@tue.nl

***Abstract*—Historical data used for supervised learning may contain discrimination. We study how to train classifiers on such data, so that they are discrimination free with respect to a given sensitive attribute; e.g., gender. Existing techniques that deal with this problem aim at removing all discrimination and do not take into account that part of the discrimination may be explainable by other attributes, such as, e.g., education level. In this context, we introduce and analyze the issue of *conditional* non-discrimination in classifier design. We show that some of the differences in decisions across the sensitive groups can be explainable and hence tolerable. We observe that in such cases, the existing discrimination aware techniques will introduce a reverse discrimination, which is undesirable as well. Therefore, we develop local techniques for handling conditional discrimination when one of the attributes is considered to be explanatory. Experimental evaluation demonstrates that the new local techniques remove exactly the bad discrimination, allowing differences in decisions as long as they are explainable.**

***Index Terms*—discrimination; classification; independence;**

I. Introduction

Discrimination is a biased treatment towards individuals on the basis of their affiliation to different groups, rather than on individual merit. In many countries, several types of discrimination, such as those based on gender, race, sexual preference, and religion are forbidden by law. When humans make subjective decisions, inevitably individual discrimination cases may occur. Such cases can be brought to court for an in-depth analysis of the circumstances. But not only humans can discriminate. Nowadays more and more decisions in lending, recruitment, grant or study applications are partially being automated based on models fitted on historical data.

Supervised learning uses historical data to infer a relation between an instance and its label. That historical data may contain discrimination; for instance, racial discrimination in the recruitment of job candidates. In such a case classifiers are likely to learn the discriminatory relation present in the historical data and apply it when making predictions. Inappropriately trained models may hence discriminate systematically, which is a lot more harmful than in single cases.

It is in the best interest of the decision makers (e.g. banks, consultancies, universities) to ensure that the classifiers they build are discrimination free even if the historical data is discriminatory. The following case illustrates the legal context and the difficulty of the task. Recently one of the world largest consultancy firms was accused of discrimination against ethnic minorities in a law suit [1]. The firm used existing criminal records to turn down candidates in pre-employment screening. Not the use of criminal records itself was considered problematic. In this data race and criminality was correlated, and the use of criminal records indirectly lead to racial discrimination. Thus, even though the company did not intend to discriminate, the decisions were deemed discriminatory by the court, while having been convicted was deemed to be not relevant for pre-screening purposes. This example shows that discrimination may occur even if the sensitive information is not directly used in the model and that such indirect discrimination is as well forbidden. Many attributes can be used only to the extent that they do not lead to indirect discrimination.

The current solutions to make classifiers discrimination free [2]–[5] aim at removing all discrimination present in the data; the probability of a positive decision by the learned classifier must be equal for all subgroups defined by the sensitive attribute (e.g., male and female). As we observe in this paper, however, such approaches have a significant limitation, as they do not take into account the fact that a part of the differences in the probability of acceptance for the two groups may be objectively explainable by other attributes.

For instance, in the Adult dataset [6], females on average have a lower annual income than males. However, one can observe that females work less hours per week on average; see Table I. Assume the task is to build a classifier to determine a salary, given an individual. The previous works would correct the decision making in such a way that males and females would get on average the same income, say 20 K$, leading to a reverse discrimination as it would result in male employees being assigned a lower salary than female for the same amount of working hours. In many real world cases, if the difference in the decisions can be justified, it is not considered as bad discrimination. Moreover, making the probabilities of acceptance equal for both would lead to favoring the group which is being deprived, in this example females.

TABLE I
Summary statistics of the Adult dataset [6].

	hours per week	annual income (K$)
female	36.4	10.9
male	42.4	30.4
all data	40.4	23.9

In this paper we take a step forward in designing discrimination free classifiers and extend the discrimination problem setting: (1) We argue that only the part of the discrimination which is not explainable by other characteristics should be removed. We analytically quantify how much of the difference

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.72

in the decision making across the sensitive groups is objectively explainable. We refer to the discrimination-aware classification under this condition as *conditional discrimination-aware classification*. (2) With our analytical results we advance two existing discrimination handling techniques to the conditional discrimination-aware setting; i.e., for removing the unexplainable (bad) discrimination when one of the attributes is considered to be explanatory for the discrimination. Our techniques are based on pre-processing data before training a classifier so that only the discrimination that is not explainable is removed. These new techniques are called *local massaging* and *local preferential sampling*. (3) In the experimental evaluation we demonstrate that the new techniques remove exactly the bad discrimination, allowing the differences in decisions to be present as long as they are explainable. (4) Finally, we demonstrate how our theory and techniques apply to cases with more than one explanatory attribute.

II. Related Work

Related work in non discriminatory decision making falls into three categories. Firstly, studies in social sciences (e.g. [7]) discuss the processes and implications of discrimination in decision making. Such studies are beyond the scope of this paper, in which we concentrate of the data mining perspective.

The second group of studies concerns finding quantitative evidence of discrimination in decision making at an aggregated level [8]–[13]. These works, however, do not address the problem of how to design non discriminatory models for future decision making when historical data may be discriminatory.

Finally, several papers [2]–[5] study from data mining perspective how to build non discriminatory classifiers, when the historical data contains discrimination. The fundamental difference from our present study is in defining of what is considered to be non discriminatory. The previous works require the acceptance probabilities to be equal across the sensitive groups. It means that if 10% of male applicants is accepted, also 10% of female applicants should be accepted. The previous works solve the problem by introducing a reverse discrimination either in the training data [2], [5] or pushing constraints into the trained classifiers [3], [4]. These works do not consider any difference in the decisions to be explainable, and thus tend to overshoot in removing discrimination so that males become discriminated, as the Adult data in Table I illustrates. We are not aware of any study formulating or addressing this problem of conditional non discrimination from a data mining perspective.

III. Formal Setting

Formally, the setting of conditional discrimination-aware classification is defined as follows. Let X be an instance in p dimensional space, let $y \in \{+, -\}$ be its label. The task is to learn a classifier $\mathcal{L} : X \rightarrow y$. In addition to X, let $s \in \{f, m\}$ be a sensitive attribute. It is forbidden by law to make decisions based on the sensitive attribute, e.g., gender.

A. Discrimination model

In relation to experimental findings in social sciences reported in [7] we assume that discrimination happens in the following way. The historical data originates from human decision making, which can be considered as a classifier $\mathcal{L}$. That classifier consists of three main parts:

1) a function from attributes to a qualification score $r = G(X)$, where X does not include the sensitive attribute;
2) a discrimination bias function $B(s) = \begin{cases} b & \text{if } s = m \\ -b & \text{if } s = f \end{cases}$;
3) the final decision function $y = \mathcal{L}\left(G(X) + B(s)\right)$.

According to this model a decision is made in the following way. First the qualifications of a candidate are evaluated based on attributes in X and a preliminary score is obtained $r = G(X)$. The qualifications are evaluated objectively. Then the discrimination bias is introduced by looking at the gender of a candidate and either adding or subtracting a fixed bias from the qualification score, to obtain $r^* = G(X) + B(s) = r \pm b$. The final decision is made by $\mathcal{L}(r^*)$. Decision making can have two major forms: *online* and *offline*. With the offline decision the candidates are ranked based on their scores r^*, and n candidates that have the highest scores are accepted. With the online decision an acceptance threshold θ is set, the incoming candidates that have the score $r^* > \theta$ are accepted.

This discrimination model has two important implications. First, the decision bias is more likely to affect the individuals that are close to the decision boundary according to their score r. If an individual is far from the decision boundary, adding or subtracting the discriminatory bias b does not change the final decision. This observation is consistent with experimental findings how discrimination happens in practice [7].

Second, traditional classifiers try to learn r^*, whereas discrimination aware classification also involves decomposing r^* into $G(X)$ and $B(s)$ and reverting the influence of $B(s)$. There may be attributes within X, however, that contribute to $G(X)$, but at the same time are correlated with the sensitive attribute s, and through s, with $B(s)$. When observing the decisions it would seem due to correlation that the decision is using s. Previous works have been very conservative in assuming that all the correlation between r^* and s is due to the discrimination bias $B(s)$. In this paper we refine this viewpoint.

B. Discrimination in classification

Even though discrimination happened in the historical data, new classifiers are required not to use the sensitive information in the decision making. Removing the sensitive attribute s from the input space would not help, if some of the attributes in X are not independent from s, that is $P(X|m) \neq P(X|f) \neq P(X)$. For instance,a postal code may be strongly related with the race. If it is not allowed to use race in the decision making, discriminatory decisions still can be made by using postal code. That is indirect discrimination, known as the *redlining*.

To get rid of such discriminatory relations among attributes, one would also need to remove the attributes that are correlated

with s. It is not a good solution if these attributes carry the objective information about the class label, as the predictions will become less accurate. For instance, a postal code in addition to the racial information may carry information about real estate prices in the neighborhood, which is objectively informative for loan decisions. The aim is to use the objective information, but not the sensitive information of such attributes.

The explanatory attribute is the attribute e (among X) that is (cor)related with the sensitive attribute s, and at the same time gives some objective information about the label y. Both relations can be measured in data, for instance, as the information gain about s given e, and about y given e. Our reasoning is built upon only one explanatory attribute. Nevertheless, this setting does not delimit taking into account multiple explanatory attributes if they are grouped into a singe representation, as we will demonstrate in Section VII.

In general there is no objective truth which attribute is more reasonable to use as the explanation for discrimination. For instance, when gender is the sensitive attribute, some attributes, such as relationships (*wife* or *husband*) may not be a good explanation, as semantically they are closely related to gender, while different working hours may be an appropriate reason to have different monthly salaries. What is discriminatory and what is legal to use as an explanation depends on the law and goals of the anti-discrimination policies. Thus, the interpretation of the attributes needs to be fixed externally by law or domain experts. When non-discrimination needs to be enforced, the law sets the constraints, while we build the techniques to incorporate those constraints into classification.

This study is built upon and valid with the assumptions:

1) the sensitive and explanatory attributes are nominated externally by law or a domain expert;
2) the explanatory attribute is *not independent* from the sensitive attribute and at the same time gives objective information about the class label;
3) the bad discrimination contained in the historical data is due to direct discrimination based on the sensitive attribute. It means no redlining (hidden discrimination) in the historical data; however, *redlining* may be introduced as a result of training a classifier on this data.

This study is *not* restricted to one *explanatory* attribute, while it is restricted to one binary *sensitive* attribute.

C. Measuring discrimination in classification

In the existing discrimination-aware classification the discrimination is considered to be present if the probabilities of acceptance for the favored community (denote m) and the deprived community (denote f) were not equal, i.e., $P(y = +|s = m) \neq P(y = +|s = f)$. Discrimination is measured as the difference between the two probabilities

$$D_{all} = P(y = +|s = m) - P(y = +|s = f). \quad (1)$$

In the previous works all the difference in acceptance between the two groups was considered undesirable. In this study, however, we argue that some of the difference may be objectively explainable by the explanatory attribute. Thus we can describe

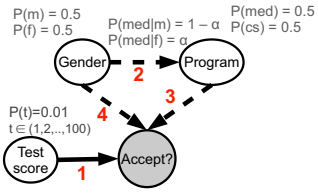

Fig. 1. Toy example with fixed probability distributions.

the difference in the probabilities as a sum of the explainable and bad discrimination

$$D_{all} = D_{expl} + D_{bad}. \quad (2)$$

In this study we are interested to remove and thus measure D_{bad}, which from Eq. (2) is

$$D_{bad} = D_{all} - D_{expl}. \quad (3)$$

For that we need to find an expression for D_{expl}.

IV. Explainable and Bad Discrimination

For analyzing the difference between the explainable and bad discrimination consider a toy model about admission to a fictitious university[1]. Gender is the sensitive attribute; male (m) and female (f) are the sensitive groups, against which discrimination may occur. There are two programs: medicine (med) and computer science (cs) with potentially different acceptance standards. Program is considered to be the explanatory attribute, thus the differences in acceptance rates that can be attributed to different application rates into the programs between male and female are acceptable. All applicants take a test for which their score is recorded (T). The acceptance (+) decision is made personally for each candidate during the final interview. Figure 1 shows the setting.

There are four relations between variables in this example. Relation (1) shows that the final decision whether to accept partially depends on the test score. Notice that the test scores are assumed to be independent from gender or program. Relation (3) shows that the probability of acceptance depends on the program. For example, the competition to medicine may be higher, thus less applicants are accepted in total. Relation (2) shows that the choice of program depends on gender. For instance, the larger part of the female candidates may apply to medicine, while more males apply to computer science. Relation (4) shows that acceptance also depends on gender, which is a bias in the decision making that is clearly a case of bad discrimination. The presence bad, explanatory or both discriminations in the data will depend on the relations (2),(3) and (4), as we will see in the following two examples.

A. How Much Discrimination is Explainable?

With this toy model we develop several scenarios to investigate different combinations of bad and explainable discrimination. **Example 1** demonstrates that all the discrimination

[1]This model does not express our belief how admission procedures happen. We use it for the purpose of illustration only.

may be explainable. Suppose there are 2 000 applicants, 1 000 males and 1 000 females. Each program receives the same number of applicants, but medicine is more popular among females, $P(med|f) = 0.8$. Assume that medicine is more competitive, $P(+|med) < P(+|cs)$. Within each program males and females are treated equally, as described in Table II. However, the aggregated scores indicate that 36% of males were accepted, but only 24% of females. The difference is explained by the fact that more females applied to the more competitive program. Thus, there is no bad discrimination.

TABLE II
EXAMPLE 1: NO BAD DISCRIMINATION.

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	20%	20%	40%	40%
accepted (+)	160	40	80	320

A similar case is reported in the Berkely study [14]. Examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. Overall 44% of males and 35% of female applicants are admitted, thus it seems that there is 9% discrimination (D_{all}) towards female applicants. However, the examination of pooled data w.r.t. different departments, shows that there is a small but statistically significant bias in favor of females. It means that the overall low admission rate for females is explainable by their tendency to apply to graduate departments that were more difficult for applicants of either gender to enter. This case concludes that there was no discrimination.

Example 2 presents a case with both explainable and bad discrimination. Suppose a similar situation to Example 1 occurs, but the decision making is biased in favor of males, $P(+|m, e_i) > P(+|f, e_i)$, where e_i is a program, as presented in Table III. The decisions result in different aggregated acceptance rates for the programs: medicine 17% and computer science 43%. It appears that in total 19% of females and 41% of males are accepted. Our goal is to determine, which part of this difference is explainable by program, and which part is due to bad discrimination.

TABLE III
EXAMPLE 2: BAD DISCRIMINATION IS PRESENT.

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	15%	25%	35%	45%
accepted (+)	120	50	70	360

First, we need to settle what would have been the correct acceptance rates $P^\star(+|med)$ and $P^\star(+|cs)$ within each program, if males and females would have been treated equally. Then we can find which part of the difference between the genders is explainable, and treat the remaining part as bad discrimination that needs to be removed. Finding the correct acceptance rates, however, is challenging, as there is no unique way to do it. Would all the acceptance rate have been as for males now, all as for females, or some average of the two?

To find the correct acceptance rates we refer to the discrimination model given in Section III-A. Under this model, it is reasonable to assume that roughly the same fraction of males benefit from the bias (those that are at most d below the acceptance threshold), as there are females that have a disadvantage due to the bias (those that are at most d above the threshold), as within the programs males and females are assumed to be equally capable. Under this assumption we need to take the average of the acceptance probability of males and females, resulting in $P^\star(+|med) = 20\%$ for medicine and $P^\star(+|med) = 40\%$ for computer science. Alternatively, if we fix the number of positive labels in the groups to the number observed in the discriminatory data, we would get $170/1000 = 17\%$ acceptance for medicine and $440/1000 = 44\%$ for computer science. Following the rationale of the discrimination model, however, these numbers are skewed and would result in programs more popular among females to be perceived as being more selective, leading to *redlining*. This way, when decisions are automated the discrimination would transfer from gender to program; a program with lots of females would receive an overall lower acceptance.

Thus we assume that the acceptance thresholds would have been fixed as the average of the historical acceptance thresholds for males and females. This choice is motivated by the scenario where the candidates come continuously, and that any candidate that is sufficiently qualified would get a position, or salary level, or a loan. Hence, there is no resource constraint and the number of positive outputs only depends on which instances qualify. An alternative scenario would be to assume that all the applications are collected together at a deadline. Then the candidates are ranked and a fixed number of the best candidates are offered a position. Whether to keep the number of accepted individuals fixed or to keep the acceptance threshold fixed depends on the application domain. For instance, in case of scholarships, job application, university acceptance fixing the number of persons may be more reasonable, since the applicants come in batch at the deadline. In case of deciding to grant a credit or what salary level to apply, fixing the threshold makes more sense (accept all individuals that pass qualification requirements), since the individuals come one by one. We argue that the choice of acceptance scenario is situation dependent and hence not part of the design of non-discrimination techniques.

Table IV illustrates calculation of the explainable part for the discrimination towards females, as presented in Example 2. We find the correct acceptance rate within each program as the average of male and female acceptance. Thus, $D_{expl} = 36\% - 24\% = 12\%$. From the original data $D_{all} = 41\% - 19\% = 22\%$. Thus, from Eq.(3) we get $D_{bad} = D_{all} - D_{expl} = 22\% - 12\% = 10\%$ the data has 10% of bad discrimination.

Formally, the explainable discrimination is the difference between acceptance of males and females

$$P^\star(+|e_i) := \frac{P(+|e_i, m) + P(+|e_i, f)}{2}, \qquad (4)$$

TABLE IV
CALCULATING THE EXPLAINABLE DIFFERENCE.

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate (Example 2)	15%	25%	35%	45%
corrected acceptance rate	20%		40%	
accepted explainable	160	40	80	320

if every individual with a fixed value of the explanatory attribute value e_i would have the same chance to be accepted[2], independently of the gender:

$$\begin{aligned} D_{expl} &= \sum_{i=1}^{k} P(e_i|m)P^\star(+|e_i) - \sum_{i=1}^{k} P(e_i|f)P^\star(+|e_i) \\ &= \sum_{i=1}^{k} \left(P(e_i|m) - P(e_i|f)\right) P^\star(+|e_i), \end{aligned}$$

where $e \in \{e_1, \ldots, e_k\}$, $P(e_i|m)$ and $P(e_i|f)$ are observed from data, and $P_c^\star(+|e_i)$ is calculated as in Eq.(4). The bad discrimination can thus be computed as the difference between D_{all} (Eq. (1)) and D_{expl}:

$$\begin{aligned} D_{bad} &= P(+|m) - P(+|f) \\ &- \sum_{i=1}^{k} \left(P(e_i|m) - P(e_i|f)\right) P^\star(+|e_i). \end{aligned} \tag{5}$$

B. Illustration of the Redlining Effect

Now that we formalized what is bad and explainable discrimination, our next step is to analyze under what circumstances a trained classifier risks to capture bad discrimination.

For our analysis we use synthetic data that is generated based on our toy model introduced in Figure 1. We generate 10 000 male and 10 000 female instances. The (integer) test scores $T \in [1, 100]$ are assigned uniformly for any individual. In every experiment all probabilities in the Belief network (given in Figure 1) are fixed, except for the probabilities $P(e_i|s)$: for $\alpha \in [0, 1]$, we generate data with: $P(med|f) = \alpha$, $P(cs|f) = 1 - \alpha$, $P(med|m) = 1 - \alpha$, and $P(cs|m) = \alpha$. In this way we can study the influence of the strength of the relationship between gender and program on the discrimination, while the total number of people applying for medicine (and computer science respectively) remains the same. For interpretation reasons denote $\beta = P(med|f) - P(cs|f) = \alpha - (1 - \alpha) = 2\alpha - 1$, then $\beta \in [-1, 1]$ can be interpreted as correlation between the gender and the program. The closer $|\beta|$ is to 1, the stronger the dependency between the explainable and sensitive attribute becomes; $\beta = 0$ means that the gender and the program are independent. Hence, the closer β will be to 0, the less explainable discrimination there will be.

Following the discrimination model introduced in Section III-A we assign the label to an individual in the toy dataset as

$$y = \delta\left[\left(t + a(-1)^{\delta[med]} + b(-1)^{\delta[f]}\right) > 70\right], \tag{6}$$

[2]Short notation of probabilities: $P(+|e_i)$ means $P(y = +|e = e_i)$.

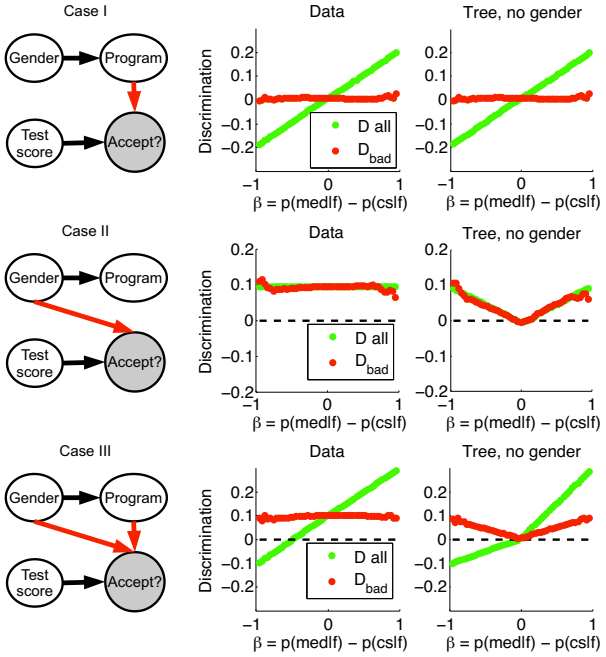

Fig. 2. Interactions between explainable and bad discrimination.

where $\delta[.]$ is a function that outputs 1 if its argument is true and 0 otherwise, t is the test score assigned to an individual, a is the effect to acceptance decisions due to program and b is the effect to the acceptance due to gender discrimination bias.

We report three cases with different acceptance decisions determined from Eq. (6) under discrimination scenarios. The scenarios are summarized in Table V. In Case I acceptance depends only on the program choice and the test, thus all the discrimination is explainable. In Case II both programs have the same acceptance thresholds, but the acceptance decision depends on gender, thus all the discrimination is bad. Case III is a combination of bad and explainable discrimination, the acceptance depends on the test, the program and the gender.

TABLE V
THREE DISCRIMINATION SCENARIOS FOR ANALYSIS.

	$P(t)$	a	b	$P(med\|f)$
Case I, only explainable	0.01	10	0	α
Case II, only bad	0.01	0	5	α
Case III, explainable and bad	0.01	10	5	α

Figure 2 presents the discrimination in function of $\beta = P(med|f) - P(cs|f)$. The left plots show the discriminations D_{all} and D_{bad} in the testing data with the original labels. The right plots show the resulting discriminations with the predicted labels by a decision tree. A decision tree is trained on the data from which gender has been removed, the training data includes only the program and the test score. We analyze the interaction between D_{all} and D_{bad}.

Case I illustrates the situation from Example 1, where all the difference in acceptance is explainable by program. The results indicate no bad discrimination neither in the data, nor

in the trained classifier. The difference in acceptance, that we observe as D_{all}, depends on the relation between gender and program, it is all explainable and thus can be tolerable.

Case II illustrates an opposite situation, where all the discrimination is bad. Therefore, we observe that D_{all} and D_{bad} in the plots overlap. In this case the program and the label are not directly related. When the gender attribute is removed, the learned decision tree captures the discriminatory decisions indirectly through program. This way the *redlining effect* appears, which is strong when gender and program are strongly dependent. If program and gender are independent ($\beta = 0$: $P(med|f) = P(med|m) = P(med) = 0.5$), then no redlining is observed ($D_{bad} = 0$). Notice that in this extreme case the classifier can be easily made discrimination free by removing both gender and program from the input space, without losing any useful information.

In Case III, which corresponds to Example 2, the explainable and the bad discrimination act together. Some of the difference in acceptance appears due to bad discrimination, while some is explainable by the program choice and thus can be tolerated. The learned decision tree shows the same bad discrimination (D_{bad}) as in Case II. However, the probabilities of acceptance for males and females are different in Case II and Case III. D_{all} in Case III becomes negative for $\beta < 0$. We can see that if very few females apply to medicine ($P(med|f)$ is close to zero), which is more competitive program, then $D_{all} < 0$ indicates that females are favored, while in fact they are deprived, as 10% of bad discrimination is present ($D_{bad} \neq 0$). This case illustrates the Simpson's paradox [15], in which a relation present in different groups is reversed when the groups are combined. Thus, to assess the true bad discrimination we need to be able to measure D_{bad}, and we propose the methodology to measure it in this study.

To sum, the experiments demonstrate the following effects:

- removing the sensitive attribute does not remove discrimination if the sensitive attribute is (cor)related with other attributes (Cases II and III);
- if an input attribute is (cor)related with the sensitive attribute *and* the label, and is nominated as explanatory, not all the difference in acceptance is bad and removing all the difference would result in the reverse discrimination;
- Case III demonstrates that there is a need for advanced training strategies to remove discrimination, and at the same time to preserve the objective information that could be captured by one and the same variable.

V. How to Remove Bad Discrimination when Training a Classifier?

As we observed in the synthetic examples, a naive approach to remove the sensitive attribute before training will not work if any other attribute is (cor)related with the sensitive attribute. Removing the explanatory attribute would help to remove bad discrimination, but the accuracy will suffer, as the explanatory attribute at the same time bears the objective information about the label. For instance, in our example the program objectively explains the difference in decisions as acceptance rates differ for different programs. Thus in real life scenarios more involved strategies to remove discrimination are required.

In order to ensure that the built classifier is discrimination free, one needs to control both

1) $P_c(+|e_i, m) = P_c(+|e_i, f)$, where P_c is the probability assigned by the classifier, and
2) $P_c(+|e_i) = P^\star(+|e_i)$, where $P^\star(+|e_i)$ is defined in Eq. (4). This means that the prediction is consistent with the original distribution of the data.

As discussed before, the first condition in isolation is insufficient due to the *redlining effect*. A classifier that only takes this condition into account would underestimate the positive class probability of a group in which females are over-represented.

We distinguish two main strategies that could make classifiers free from bad discrimination. The first strategy is to remove the relation between the sensitive attribute and the class label from the training data, which is the source of the bad discrimination (relation (4) in Figure 1). Note that removing the relation is not the same as removing the sensitive attribute itself, it means making $P(+|med, f) = P(+|med, m) = P^\star(+|med)$. We can achieve that, for instance, by modifying the original labels of the training data.

The alternative strategy is to split the data into smaller groups based on the explanatory attribute. That would remove the relation between the sensitive and the explanatory attributes (relation (2) in Figure 1). Then individual classifiers can be trained for each group. This strategy would also require to correct the training labels in each groups, otherwise the *redlining effect* will manifest. In addition, it would significantly reduce the data available for training a classifier, which may result in much lower accuracy than the global model. Thus, in this study we adopt the first type of strategy.

In this work we propose two new techniques local massaging and local preferential sampling that modify the labels in the historical data so that the historical data satisfies the following conditional non-discrimination constraints: $P'(+|e_i, f) = P'(+|e_i, m) = P'^\star(+|e_i)$ and $P^\star(+|e_i)$ is fixed so that no *redlining* is introduced (P' denotes the probability in the modified data). First we need to fix the desired probabilities of acceptance $P^\star(+|e_i)$, which would have been correct. We set $P^\star(+|e_i)$ to be the average of male and female acceptance rates, Eq. (4), as motivated in Section IV-A. After finding $P^\star(+|e_i)$ for all $e_i \in dom(e)$, the remaining part is to change the labels of the training data so that $P'(+|e_i, f) = P'(+|e_i, m) = P^\star(+|e_i)$. We anticipate that the classifiers trained on the modified data, which does not contain bad discrimination, will produce outputs that would satisfy $P_c(+|e_i, f) = P_c(+|e_i, m) = P^\star(+|e_i)$ (P_c denotes the probability in the outputs of a classifier). The role of the proposed techniques is using our theory on conditional non-discrimination (Section IV) to decide which instances in the historical data need to be modified and in what way.

A. Local Massaging

The local massaging for every partition in the training data induced by the explanatory attribute will modify the values of

labels until both $P'(+|m, e_i)$ and $P'(+|f, e_i)$ become equal to $P^\star(+|e_i)$. The discrimination model in Section III-A implies that discrimination is more likely to affect the objects that are closer to the decision boundary. To this end, massaging identifies the instances that are close to the decision boundary and changes the values of their labels to the opposite. For that purpose individuals need to be ordered according to their probability of acceptance. To be able to order we need to convert the original binary labels (accept or reject) to real valued probabilities of acceptance. For that we learn an internal ranker (a classifier that outputs the posterior probabilities).

Suppose females have been discriminated as in our university admission model and the discrimination is reflected in the historical data. The local massaging will identify a number of females that were almost accepted, and make their labels positive, and identify a number of males that were very likely, but have not been rejected, and make their labels negative.

This technique is related to the massaging proposed in [4], while, given the new theory, now it can handle the explainable discrimination. Algorithm 1 gives the pseudo-code.

Algorithm 1: Local massaging

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \to y^{(i)}$;
  rank males using $\mathcal{H}_i$;
  `relabel DELTA` (male) males that are the closest to the decision boundary from $+$ to $-$ (Algorithm 4);
  rank females using $\mathcal{H}_i$;
  `relabel DELTA` (female) females that are the closest to the decision boundary from $-$ to $+$
end

B. Local Preferential Sampling

The preferential sampling technique does not modify the training instances or labels, instead it modifies the composition of the training set. It deletes and duplicates training instances so that the labels of new training set contain no discrimination and satisfy the criteria $P'(+|m, e_i) = P'(+|f, e_i) = P^\star(+|e_i)$. Following the discrimination model where the discrimination is more likely to affect the objects that are closer to the decision boundary, the preferential sampling deletes the 'wrong' instances that are close to the decision boundary and duplicates the instances that are 'right' and close to the boundary. To select the instances they are ordered according to their probability of acceptance using a ranker learned on each group in the same way as in the local massaging.

In the university example the local preferential sampling will delete a number of males that were almost rejected and duplicate the males that were almost accepted. It will also delete a number of females that were almost accepted and duplicate the females that were almost rejected.

Algorithm 2: Local preferential sampling

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: resampled dataset (a list of instances)

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (see Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \to y^{(i)}$;
  `rank` males using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (male) (see Algorithm 4) males $+$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (male) males $-$ that are the closest to the decision boundary;
  `rank` females using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (female) females $-$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (female) females $+$ that are the closest to the decision boundary;
end

Algorithm 3: subroutine PARTITION$(\mathbf{X}, \mathbf{e})$

find all unique values of e: $\{e_1, e_2, \ldots, e_k\}$;
for $i = 1$ **to** k **do**
  make a group $X^{(i)} = \{X : e = e_i\}$;
end

This technique is related to the preferential sampling [5], while, given the new theory, now it can handle the explainable discrimination. Algorithm 2 gives the pseudo-code.

VI. Experimental Evaluation

We evaluate the performance of the proposed local discrimination handling techniques in line with their global counterparts. The objective is to minimize the absolute value of the *bad* discrimination while keeping the accuracy as high as possible. It is important not to overshoot and end up with a reverse discrimination. The goals of our experiments are:

1) to present a motivation for conditional discrimination-aware classification research,
2) to explore how well the proposed techniques remove bad discrimination as compared to the existing techniques for global non-discrimination, and
3) to analyze the effects of removing discrimination on the final classification accuracy.

We explore the performance of the methods that aim to remove the relation between the sensitive attribute and the label. We

Algorithm 4: subroutine DELTA(gender)

return $G_i|p(+|e_i, \text{gender}) - p^\star(+|e_i)|$,
where $p^\star(+|e_i)$ comes from (Eq. (4)),
G_i is the number of gender people in $X^{(i)}$;

test the local massaging and the local preferential sampling[3].

A. Data

We use two real datasets. In the **Adult** dataset [6], the task is to classify individuals into *high* and *low* income classes. We use a uniform sample of 15 696 instances, which are described by 13 attributes (we discretize the 6 numeric attributes) and a class label. Gender is the sensitive attribute, income is the label. We repeat our experiments several times, where any of the other attributes in turn is selected as explanatory. Figure 3 (left) shows the discrimination in the dataset. The horizontal axis denotes the index of the explanatory attribute.

In the Adult dataset a number of attributes are weakly related with gender (such as workclass, education, occupation, race, capital loss, native country). Therefore, nominating any of those attributes as explanatory would not explain much of the discrimination. For instance, we know from biology that race and gender are independent. Thus, race cannot explain the discrimination on gender; that discrimination is either bad or it is due to some other attributes. Indeed, we observe from the plot that all the discrimination is bad, when treating race (attribute #7) as explanatory.

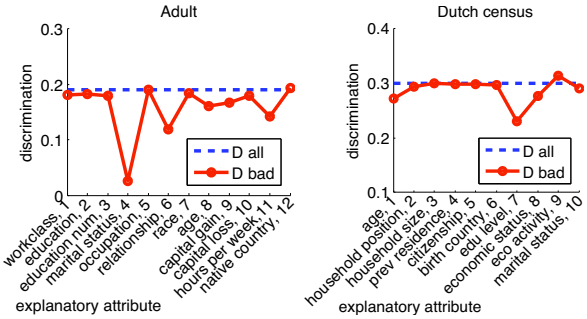

Fig. 3. Discrimination in the datasets.

On the other hand, we observe that the relationship (attribute #6) explains a lot of D_{all}. Whether relationship is an acceptable argument to justify differences in income is for lawyers to determine. Judging subjectively, the values of this attribute '*wife* and *husband* clearly capture the gender information. From a data mining perspective, if we treat it as acceptable, a large part of the discrimination gets explained.

Age, and working hours per week are other examples of explanatory attributes. They justify some of the discrimination. Intuitively, these reasons are perfectly valid for having different income, so it makes sense to treat them as explanatory.

Another dataset that we use is the **Dutch Census of 2001** [16], that represents aggregated groups of inhabitants of the Netherlands. We formulate a binary classification task to classify the individuals into *high income* and *low income* professions, using occupation as the class label. Individuals are described by 11 categorical attributes. After removing the records of under-aged people, several professions in the middle level and people with unknown professions our dataset consists of 60 420 instances. Gender is treated as the sensitive attribute.

Figure 3 (right) presents the discrimination contained in this data. The difference between the all and the bad discrimination is much less than in the Adult data. Here many attributes are not that strongly correlated with gender. Simply removing the sensitive attribute should therefore perform reasonably well. Nevertheless, education level, age and economic activity present cases for conditional non-discrimination, thus we explore this dataset in our experiments.

[3]All datasets and the code of all implementations of our experiments are available at https://sites.google.com/site/conditionaldiscrimination/.

B. Motivation Experiments

To give a motivation for our new approach we demonstrate that the existing techniques do not solve the conditional non discrimination problem.

1) Removing the Sensitive Attribute: First we test a baseline approach, which removes the sensitive attribute from the training data. We learn a decision tree with the J48 classifier (Weka implementation) on all the data except the gender attribute, treated as sensitive. Figure 4 shows the resulting discriminations, when the learned tree is evaluated using 10-fold cross validation. We can clearly observe the *redlining effect*, especially in the Adult data; even though the sensitive attribute is removed, the bad discrimination still manifests.

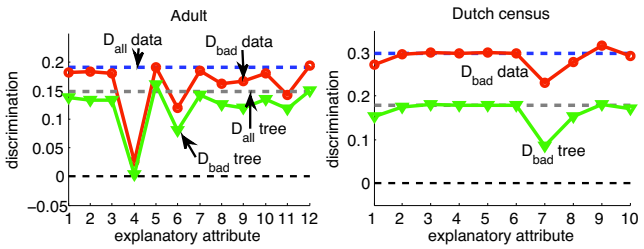

Fig. 4. Removing the sensitive attribute.

2) Global Techniques: Next we investigate to what extent the two existing global techniques [2], [5] remove bad discrimination. Global massaging modifies the labels of the training data to make the probabilities of acceptance equal for the two sensitive groups. Global preferential sampling, resamples the training data so that non-discrimination constraints for the label distribution are satisfied. Both methods aim at making D_{all} equal to 0, which is not the same as removing D_{bad} and will actually reverse the discrimination, as can be seen from Figure 5. The global techniques do not take into account, that the distributions of the sensitive groups may differ and thus some of the differences in probabilities are explainable.

As expected, the massaging and the preferential sampling techniques work well for removing all discrimination, e.g. for the Adult data after massaging $D_{all} = 0$. But, if we treat *marital status* as the explanatory attribute, these results introduce a reverse bad discrimination. The same, but on a smaller scale, holds for several other explanatory attributes, e.g. *hours per week* and *age*. For the Dutch Census data, both techniques overshoot if conditioned on *education level*.

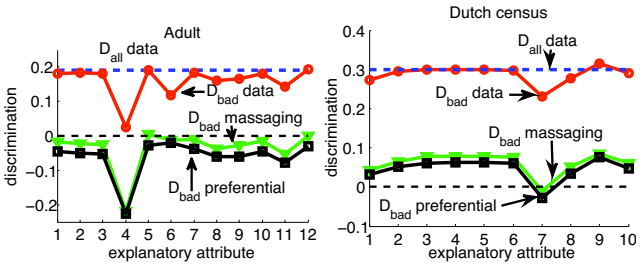

Fig. 5. Discrimination with the *global* techniques.

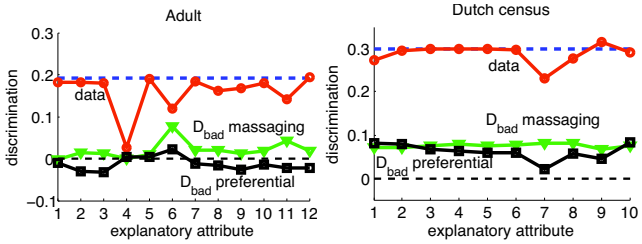

Fig. 7. Discrimination with *the local* techniques.

These results confirm that a reverse bad discrimination is introduced when global discrimination handling techniques are applied raising the necessity for local methods.

3) When are the Local Techniques Essential?: The existing techniques fail the most when the difference between D_{all} and D_{bad} in the data is large. For instance, Figure 3 shows sharp negative peaks when *marital status* or *relationship* act as the explanatory attributes in the Adult data. In such cases, the need for the special techniques that can handle conditional discrimination is essential.

A large difference between D_{all} and D_{bad} implies that a large part of the difference in the decisions is due to the explanatory attribute. We quantify the dependencies between class on the one hand, and sensitive and explanatory attributes on the other hand by the following information gains:

$$G(y, e_i) = H(y) - H(y|e_i), \text{ and}$$
$$G(s, e_i) = H(s) - H(s|e_i).$$

$H(.)$ denotes entropy, s the sensitive attribute, y the label and e_i the explanatory attribute. The information gains for the Adult and the Dutch census datasets are plotted in Figure 6. The figure confirms the intuition that the stronger the relation

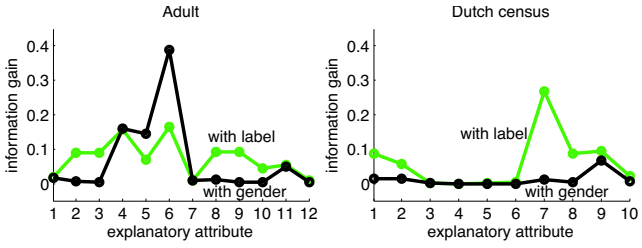

Fig. 6. Relations between sensitive, explanatory attributes and labels.

with the explanatory attribute (higher information gain) the larger the share of the total discrimination that is explainable. Recall Figure 3 for the discriminations.

C. Non-discrimination Using Local Techniques

Let us analyze how the proposed local techniques handle discrimination. We expect them to remove exactly the bad discrimination and nothing more. We test the performance with decision trees (J48) via 10-fold cross validation.

Figure 7 shows the resulting discrimination after applying the local massaging and the local preferential sampling. Both local techniques perform well on the Adult data. Bad discrimination is reduced to nearly zero, except for relationship as explanatory attribute when massaging is applied to the Adult dataset. Our techniques do not produce the reverse discrimination as, e.g., global massaging does.

The proposed solutions do not perform that well with the Dutch census data, as the sensitive attribute is not very strongly correlated with any other attribute in the dataset, because the local techniques are primarily designed to handle high correlations with the sensitive attribute that induce *redlining*.

Note that when the base classifier can also serve as an accurate ranker, there is a *simpler* local approach to employ our conditional discrimination theory and measure. Different rankers can be learned for males and females and used directly for classification by setting the thresholds to $p^\star(+|e_i)$ as in Eq. (4). In the specific case of J48 as a ranker the results are worse than the results of the generic techniques, thus left out. This happens as J48 is not inherited a smooth ranker.

D. Accuracy with the Local Techniques

When classifiers become discrimination free, they may lose some accuracy, as measured on the historical data. We analyze the resulting accuracies after applying the local massaging and the local preferential sampling. Figure 8 presents the testing accuracy of a decision tree (J48) when the original historical data with all the attributes is used for training, and the accuracy after our local techniques have been applied. The accuracy of

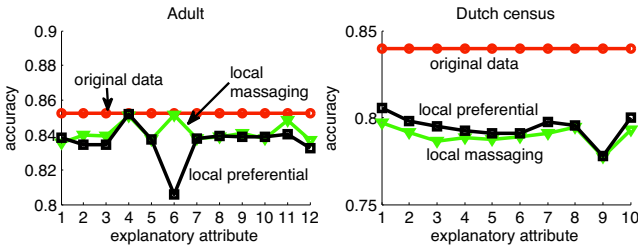

Fig. 8. Accuracy with *the local* techniques.

the local techniques decreases as the evaluation is carried out on the original data that contains discrimination. Nevertheless, the absolute accuracy remains high; it drops by 5% at most. Our experiments demonstrate that the local massaging and the local preferential sampling classify future data with reasonable accuracy and maintain low discrimination.

VII. Handling Multiple Explanatory Attributes

In this section we demonstrate how our theory and techniques extend for handling multiple explanatory attributes that may be required to be taken into account together (e.g. working hours and experience in determining a salary).

We create a new attribute that describes a group to which a person belongs and treat that attribute as explanatory when applying the theory and techniques proposed in this study. A straightforward way to do that is to create a separate group for every unique combination of the values of explanatory attributes. For instance, group 1: long hours and no experience, group 2: long hours and large experience, group 3: average hours and no experience, etc. In reality, however, this approach is not applicable, since with growing numbers of explanatory attributes, it becomes increasingly less likely that two instances will agree on all of them. This is a problem, since if we treat every instance as unique, then there we observe no discrimination, as there is nothing to compare an instance with. Thus we need to form large enough groups to have a pool for comparison within each group. In order not to introduce the *redlining* the grouping *procedure* needs to be independent from the sensitive attribute and the label. The resulting groups themselves are expected to be correlated with the sensitive attribute and the label, as the explanatory attributes are. The main intuition behind grouping is to monitor that individuals that are similar to each other in terms of explanatory attributes (fall into one group) are treated in a similar way in decision making regardless of the gender.

In this study we give an illustration of a grouping approach, while optimal grouping strategies are out of the scope of the present paper and is the subject of further investigation. We report the results of the following experiment on the Adult dataset. In order to form the groups we run the k-means clustering on the input data. To prevent the grouping *procedure* to be influenced by the sensitive attribute indirectly we omit from the clustering input space the attributes that are exceptionally highly correlated with the gender; we omit gender itself, relationship, marital status, occupation and income.

We compare the bad discrimination D_{bad} in the outputs of a decision tree (J48) trained on the original data and on the data that has been preprocessed using the global and our local techniques (massaging and preferential sampling), discussed in Section V. We test the performance via 10-fold cross validation. Figure 9 presents the resulting bad discrimination and accuracies. We observe that the global techniques overshoot and introduce the reverse discrimination, while our local techniques remove exactly the bad discrimination and they preserve reasonable prediction accuracy.

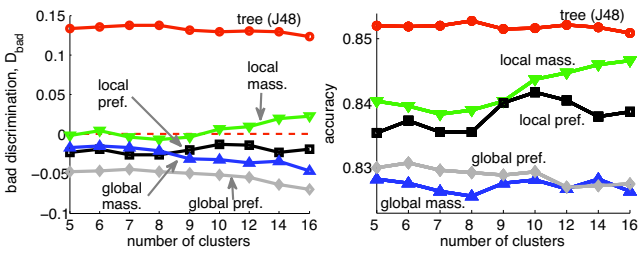

Fig. 9. Discrimination and accuracy with multiple explanatory attributes.

VIII. Conclusion

In this paper we extended the discrimination-aware classification paradigm to the presence of explanatory attributes that are correlated with the sensitive attribute. In such a case, as we demonstrated, not all discrimination can be considered bad and the existing techniques tend to overshoot and start a reverse discrimination. Therefore, we introduced a new way of measuring discrimination, by explicitly splitting it up into explainable and bad discrimination. Local alternatives of the massaging and preferential sampling were introduced and experimentally evaluated. The experiments demonstrated the effectiveness of the new local techniques, especially in cases when the sensitive attribute is highly correlated with the explanatory attribute.

Acknowledgment

F.Kamiran thanks to Higher Education Commission of Pakistan (HEC) for their financial support for this research.

References

[1] T. Ahearn. (2010) Discrimination lawsuit shows importance of employer policy on the use of criminal records during background checks. Online http://www.esrcheck.com/wordpress/2010/04/12/.

[2] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *IEEE ICDM Workshop on Domain Driven Data Mining (DDDM)*, 2009, pp. 13–18.

[3] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[4] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *Proc. of IEEE ICDM Int. Conf. on Data Mining (ICDM)*, 2010, pp. 869–874.

[5] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," in *Proc. of the 19th Ann. Machine Learning Conf. of Belgium and the Netherlands (BENELEARN)*, 2010, pp. 1–6.

[6] A. Asuncion and D. Newman. (2007) UCI machine learning repository. Online http://archive.ics.uci.edu/ml/.

[7] M. Hart, "Subjective decisionmaking and unconscious discrimination," *Alabama Law Review*, vol. 56, p. 741, 2005.

[8] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2008, pp. 560–568.

[9] ——, "Measuring discrimination in socially-sensitive decision records," in *Proc. of SIAM Int. Conf. on Data Mining (SDM)*, 2009, pp. 581–592.

[10] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Trans. Knowl. Discov. Data*, vol. 4, pp. 9:1–9:40, 2010.

[11] A. S. Goldberger, "Reverse regression and salary discrimination," *Journal of Human Resources*, vol. 19, no. 3, pp. 293–318, 1984.

[12] A. H. Munnell, G. M. B. Tootell, L. E. Browne, and J. McEneaney, "Mortgage lending in boston: Interpreting hmda data," Federal Reserve Bank of Boston, Working paper 92-7, 1992.

[13] S. Ross and J. Yinger, *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. MIT, 2002.

[14] P. Bickel, E. Hammel, and J. O'Connell, "Sex bias in graduate admissions: Data from Berkeley," *Science*, vol. 187 (4175), pp. 398–404, 1975.

[15] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society*, vol. 13, pp. 238–241, 1951.

[16] Dutch Central Bureau for Statistics, "Volkstelling," 2001.