Estimation of Uncertainty Bounds on Disparate Treatment when using Proxies for the Protected Attribute

Taha Racicot[†,*], Richard Khoury[†], Christophe Pere[‡]
[†] Departement of Computer Science and Software Engineering, Université Laval, Québec, Canada
[‡] La Capitale Insurance and Financial Services, Québec, Canada

Abstract

This paper proposes a new method for uncovering discrimination in decision making systems with continuous value outcomes when lacking the protected attribute values. We demonstrate our method over race discrimination using name and surname proxies. Also, we use a new method for estimating uncertainty in the disparate treatment evaluation to allow for better judgment when using imprecise proxies for the protected attribute. We carry out tests using synthetic data.

Keywords: Algorithmic discrimination; Causal fairness; Bounds estimation; Disparate treatment; Sensitive attribute proxy

1. Introduction

There is increased public interest today in ensuring that the services provided by companies and governments meet the values of fairness and equality between the different demographic groups of society [1]. Not only is this task complicated by decision-making algorithms which are increasingly difficult to explain [2], but concerns and restrictions on the collection of personal data lead to sensitive attributes (race, religion, sexual orientation, etc.) being deleted or never collected at all, even though they are essential to evaluate discrimination between different groups. Therefore, we need to estimate the disparity of treatment between groups of individuals without knowing who belongs to a protected group. The solution is to use a *proxy* variable that correlates with the protected attribute [3][4].

It is therefore necessary to set up a system for evaluating the disparity of treatment which performs well when: (1) Instances of individuals belonging to the minority group are unknown due to the missing protected attributes; (2) The minority group represents a small proportion of the population; (3) The treatment model or decision-making process is completely inaccessible to the people doing the fairness assessment; and (4) There is uncertainty about the correlation of the proxy with the protected attribute. In addition, given the competitive conditions in which private companies operate, we must also assess the certainty of the estimation of the disparity of treatment, to allow informed decision-making by companies to minimize disparity. Thus our main contributions are (1) Measuring uncertainty in disparate treatment measures when using poorly-correlated attributes like surname and name, and (2) Proposing an easy framework for estimating disparity while using familiar fairness metrics like *demographic disparity*.

This paper is structured as follows. Section 2 will define important metrics from the literature. Section 3 will present our proposed solution, the Bayesian Tarnet. Section 4 will present experiments using our solution, and section 5 will discuss the results obtained.

[*] taha.racicot.1@ulaval.ca

2. Related Work

2.1. Disparity estimation

A number of approaches exist for the evaluation of disparity. They are used, for instance, to measure the effectiveness of a drug or the effect of certain policies [5–7]. Here, the term "treatment" refers to the action taken or the effect induced before observing the result, "outcome" is the effect induced by the treatment, and "disparity" refers to the difference in the outcome between different groups. However, sensitive data about the individuals are often missing and the amount of variables that come into play in the end result is large, making it difficult to accurately calculate the effect belonging to a minority group has on the outcome of the treatment. Many fairness metrics have been used to estimate disparity of outcomes of a treatment on minority groups, both at the group and individual levels [8].

2.1.1. Definitions

Demographic Parity (DP) [9] considers a treatment fair when the probability of a certain outcome conditional to the protected attribute is equal for both protected groups:

$$P(y|t=0) = P(y|t=1), \tag{2.1}$$

where y is the treatment decision or outcome (the treatment) and t is the protected attribute, with $t=1$ representing the majority group and $t=0$ the minority group.

Average Treatment Effect (ATE) measures how the treatment will vary if we modify the protected attribute from $t=1$ to $t=0$. Given n individuals ATE is defined as:

$$ATE = \frac{\sum[y|t=1] - \sum[y|t=0]}{n}. \tag{2.2}$$

While fairness metrics at the group level are popular, they can lead to the Simpson's paradox in which varying the number or granularity of comparison groups can lead to contradictory fairness results and unfair conclusions[10]. Therefore, individual fairness conditions are generally considered more representative.

Individual Treatment Effect (ITE) measures the difference in treatment when we change only the value of the protected attribute t of an individual, without changing their non-protected attributes x. Thus, the ITE $\tau(x)$ of individuals with attributes x is:

$$\tau(x) = \frac{\sum[y|x,t=1] - \sum[y|x,t=0]}{n}, \tag{2.3}$$

However, the ATE and ITE define the disparity as a difference between two average treatments. We can also define it as difference between two probability distributions.

Demographic Disparity (DD) [9] measures the disparity in the probability that a treatment will be chosen for individuals by conditioning on the protected attribute:

$$DD = P(y|t=1) - P(y|t=0). \tag{2.4}$$

When the protected attribute can take more than two values, $t \in 0, ..., t_n$, the metric is computed pairwise on all combinations of values:

$$DD_{ij} = P(y|t=t_i) - P(y|t=t_j). \tag{2.5}$$

2.1.2. Estimation methods

To estimate the ITE, one of the simplest methods is to find several individuals with the same attributes x and separating them into groups for $t=1$ and $t=0$ before calculating $\tau(x)$. One problem with this method is that the set of features x can be very large and its values very precise, which means there will not be enough (or any) individuals with the same values of $x$in the population for this evaluation. Even if many instances of the same

x are available, the minority value of t is rare and some values of x will have few (or no) individuals of that class. Two methods are used to overcome this problem.

Propensity Score (PS) deals with the rare occurrences of x by assigning to each individual a score $PS(t|x)$, which is a discrete value in a predefined range assigned so that if two individuals have similar vectors x they will have similar PS values. This score is used to group similar individuals together and to calculate $\tau(x)$ for the group. This updates equation 2.3 as:

$$\begin{aligned} \tau(x) &= P(y|g_i, t=1) - P(y|g_i, t=0) \\ g_i &= k_{i-1} < PS(t|x) \leq k_i \end{aligned} \tag{2.6}$$

where g_i represents a group of individuals with propensity scores within a range k_{i-1} to k_i.
Regression Adjustment tackles the problem of not having enough individuals with the rare value of the protected attribute by estimating the effect of the protected attribute on the outcome. The idea is to estimate the probability distribution of the result conditioned on the protected attribute and on the other attributes, $P(y|x,t)$.

The problem with these methods is that they attempt to solve the issues of ITE by relaxing the comparison to a group of similar people instead of an individual, which challenges the notion of individual fairness and reintroduces the risk of Simpson's paradox.

Conterfactual Regression (CFR) uses a deep neural network (DNN) to estimate the ITE using observational data [11]. The network learns to generate a similar representation for similar vectors x, and then uses that representation to estimate different outcomes based on the protected attribute value of each individual. This allows the method to simultaneously find similar individuals for comparison and discover the disparity between groups. The Tarnet architecture implementation of CFR is composed of two main parts. The first is the shared representation network which learns a shared representation of similar people from different groups. That network minimizes $G(x_i)$, the distance between the representations of an instance x_i conditional to each possible value of t [12, 13]:

$$G(x_i) = \sum_{j=1}^{n} IPM_G(\{\phi(x_i)\}_{i:t_i \neq j}, \{\phi(x_i)\}_{i:t_i=j}), \tag{2.7}$$

where IPM represents an integral probability metric like the Wasserstein distance[14], ϕ is a function that generates a shared representation of x_i, and n is the number of values the protected attribute can take. The second part of the CFR is composed of multiple branches, each an independent network. Depending on the protected attribute value of the instance, the shared representation of the individual is sent to one specific branch. Each branch thus learns how each protected group is treated independently from the others.

2.1.3. **Uncertainty estimation**

Monte Carlo dropout (MC): Dropout [15] is a popular method of neural network regularization, which works by ignoring random nodes with a certain probability. Normally, dropout is applied at training time, while at testing time all nodes are used. Therefore, the network is deterministic during testing: given a test sample, it will always make the same prediction. For MC dropout [16], the dropout is applied at both training and testing time. This means that the the prediction given a test sample will vary stochastically. MC dropout interprets these stochastic predictions as samples from a probability distribution on the value of the output variable being predicted. We can also use this distribution to estimate the uncertainty on the network prediction [17].

2.2. **Proxy model: Sensitive attribute estimation**

To estimate the disparity between different protected groups, we need to consider the values of the individuals' protected attributes. However, these are typically missing from datasets. The most popular solution is to estimate them using proxies. One of the best methods to do this is the Bayesian Improved Surname Geocoding (BISG) [18] which uses surname and geocoding to estimate the protected attribute with Bayesian statistics. Given a set of n people with g_i their geolocation, i, s_i their surname, and r_i their unknown race:

$$P(r_i|s_i, g_i) = \frac{P(g_i|r_i)P(r_i|s_i)}{\sum_{i=1}^{n} P(g_i|r_i)P(r_i|s_i)} \tag{2.8}$$

One of the main limitations of this method is that it relies on collected data, namely statistics on race given geolocation and name. While the relationship between name and race is easily available, there are many regions for which racial makeup has not been measured, and thus this method is not usable. Consequently, we decided to use a more generic method which relies only on the name to predict the race [19]. This method uses an LSTM that takes in the sequence of character embeddings of a name and predict its ethnicity. It was trained on the first names and last names from the 2010 US Census [20].

2.3. **Important metrics**

Wasserstein Distance Metric (WDM) is used to measure the difference between two probability distributions in the previous metrics. Let P be a distribution associated to a random variable X, such that for a subset of values $A \subseteq \mathbb{R}$, we have $Pr\{X \in A\} = P(A)$. The cumulative distribution function of P is then

$$F_P(x) := Pr\{X \leq x\} = \int_{-\infty}^{x} P(dx), \tag{2.9}$$

where the inverse distribution function of P defined over the interval (0, 1] is

$$F_P^{-1}(u) := inf\{x : F_P(x) = u\}. \tag{2.10}$$

For $1 \leq p \leq \infty$, the p-Wasserstein metric W_p between two probability distributions P and Q over $\mathbb{R}$. is defined using the inverse cumulative distribution function:

$$W_p(P, Q) := \left(\int_0^1 \left| F_P^{-1}(u) - F_Q^{-1}(u) \right|^p du \right)^{1/p}, \tag{2.11}$$

This probability distance metric has many interesting properties [21][22] such as:

- *scale sensitivity*: $W_p(cP, cQ) = |c|^\beta W_p(P, Q)$, where $\beta > 0$ and $c > 0$ are constants;
- *sum invariance*: $W_p(A + P, A + Q) \leq W_p(P, Q)$, where A is another probability distribution.
- *unbiased sample gradients*: the expected gradient of the sample loss equals the gradient of the true loss when using this metric as a loss function during training in gradient descent [21].

Unlike other metrics like the total variance $\frac{1}{2} \int |P(X) - Q(X)|$ which only takes into account the probabilities of the outcomes for each distribution, WDM uses the inverse distribution function, which allows it to consider both the probability and the distance between various outcomes. See [14] for more details.

Relative Standard Deviation (CV) is a metric used to measure the dispersion in the prediction of N outcomes $\hat{y}_i$ relative the mean outcome value μ:

$$CV = \frac{100}{\mu} \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \mu)^2}{N}}. \tag{2.12}$$

Mean Absolute Percentage Error (MAPE) is used to evaluate model performance in a regression task by estimating the relative error between the N predicted outcomes $\hat{y}_i$ and the true outcomes y_i.

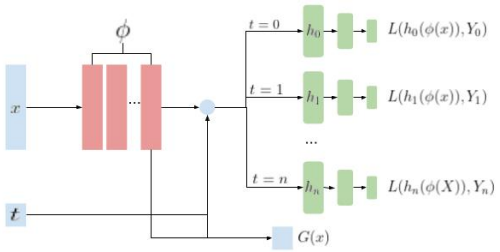
$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_y - \hat{y}_i}{y_i}\right| \tag{2.13}$$

3. Methodology

Our main contribution is providing a method to evaluate disparity along with an uncertainty measure when the protected attributes are estimated by a proxy. We accomplish this by adding uncertainty estimation on the Tarnet architecture and the name proxy model using MC dropout method.

3.1. Bayesian Tarnet architecture

Our architecture, presented in Figure 1, is heavily inspired by the works [12, 13]. The Tarnet architecture was adopted but modified to support more than two values of the protected attribute. Uncertainty estimation was also added to the network as described in [16] by adding dropout between each weight layer h_i. We call our resulting network the Bayesian Tarnet (BT).



(a) Bayesian Tarnet *(b)* Bayesian name proxy

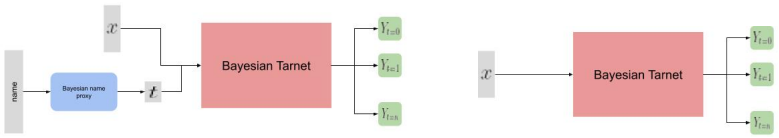
Figure 1. (a) Bayesian Tarnet: Architecture of the disparity estimation model where x denotes the non-protected attribute vector, t denote the protected attribute, h_i represents each branch learning the treatment effect for each protected attribute. (b) Bayesian name proxy: The name proxy model build in MC dropout fashion. r is represents the proxy value and t the predicted protected value.

The first component in Figure 1 is the layer ϕ which learns the shared representation of individuals across the different protected groups. This layer is regularized by an IPM that measures the distance between the representations of an instance x_i with every combination of values of t, as described in equation (2.7). This allows the model to better generalize, especially when the classes t are imbalanced [12]. Next, the output of ϕ is sent into multiple treatment layers h_i, and each layer has a different value of the protected attribute t appended to it from the set of n possible values the attribute can take. Each layer learns the treatment for a single protected attribute value. In our implementation, the representation layer ϕ is composed of 2 dense layers with an input dimension of length of the x feature vector and an output of 20. Each treatment layer h_i is composed of 3 dense layers with an input of 20 and output of 1. The input t is only used to funnel the output $\phi(x)$ to the correct h_i layer. There is a 0.1 dropout between each dense layer of h_i and no dropout between the layers of ϕ. Each hidden dense layer has a *tanh* activation function. L is the loss function.

To uncover disparity as well as uncertainty using a proxy we propose a new way of training this type of model where we combine our BT model with a bayesian proxy model.

In this case, the name proxy model of section 2.2 during training built with dropout in the mc dropout fashion. The proxy model will output the protected attribute prediction t and the BT uses it to make predictions. The bayesian proxy model output will reflect the uncertainty of the proxy and it will be learned by the BT model and taken into account in the uncertainty of the final outcome.

Once the BT model has been trained to differentiate between different treatments depending on the protected attribute value t, we can use it to make predictions on the treatment of a new instance x_i with an unknown t by predicting its treatment for each possible value of t. This allows us to measure the DD for this new individual using equation (2.3). Moreover, since we use MC Dropout, we can also estimate the uncertainty on this outcome.



(a) Training mode *(b)* Testing mode

Figure 2. (a) Training mode: the BT model is used in conjunction with proxy model. (b) Testing mode: the model is used to make a new prediction on an individual without protected attribute information.

3.2. **Lower and upper bound on the disparate treatment**

The computation of DD in equations (2.4) and (2.5) takes the difference between two probability distributions given the protected attribute. Since these distributions are uncertain when using a proxy, we propose **minimum and maximum disparity (MMD)** as a way to estimate bounds over the difference while taking into account the uncertainty over each probability distribution.

Since we use MC dropout during testing, n tests with the same set of input values $x_i \in X$ will lead to a distributions of outcomes $P_1(X), ..., P_n(X)$. Using these distributions, we can calculate a mean distribution $P_\mu(X)$ and a standard deviation distribution $P_\sigma(X)$:

$$P_\mu(X) = \frac{1}{n}\sum_{i=1}^{n} P_i(X),\ P_\sigma(X) = \sqrt{\frac{\sum(P_i(X)-\bar{P}(X))^2}{n}}. \tag{3.1}$$

To derive an upper bound (up) and lower bound (low) for the mean distribution, we simply use $P_\mu(X)$ and $P_\sigma(X)$ accordingly:

$$P_{low}(X) = P_\mu(X) - P_\sigma(X),\ P_{up}(X) = P_\mu(X) + P_\sigma(X). \tag{3.2}$$

Once we have upper and lower bounds for a probability distribution, we can use them to obtain the lower and upper bounds of the difference between two probability distributions with known bounds $P(X) = \{P_{low}(X), P_{up}(X)\}$ and $Q(X) = \{Q_{low}(X), Q_{up}(X)\}$, such as the two distributions to compute DD:

$$\begin{aligned} DD_{low}(X) &= min\{P_{low}(X) - Q_{up}(X), Q_{low}(X) - P_{up}(X)\}, \\ DD_{up}(X) &= max\{P_{low}(X) - Q_{up}(X), Q_{low}(X) - P_{up}(X)\}. \end{aligned} \tag{3.3}$$

where DD_{low} and DD_{up} represent the lower bound and upper bound for the demographic disparity.

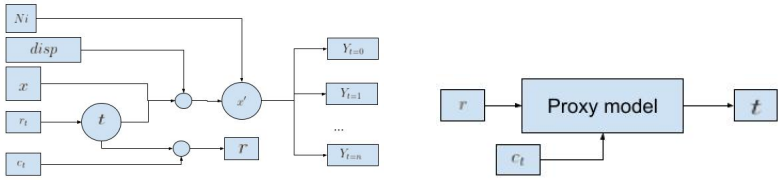
3.3. **Synthetic dataset**

We created a synthetic dataset generator to test our model. This allows us to easily adjust the dataset using hyper-parameters to simulate certain conditions. It notably allows us to

test the reliability of the model in conditions were the proxy e presents low correlation with the protected attribute t, and also in conditions where the protected classes are imbalanced. These conditions happen often in practice as minority groups represent a low percentage of the population and it is difficult to find proxies that are readily-available and highly-correlated with the protected attribute. Consequently, our dataset generation has four hyper-parameters. The first is the percentage of each protected group, denoted $p_{t=i}$, and the second is the relationship between the proxy and the protected attribute, c_t. The proxy value is generated by taking the generated correct protected value and swapping it to another random protected attribute value with rate c_t. For example, a correlation of $c_t = 0.75$ means that the proxy is the protected attribute 25% of the time and is changed the other 75%. The third hyper-parameter *disp* is used to create disparity between the different protected groups. It is a number representing the mean treatment value between each group. For example, $disp = [0, 100, 200]$ means that there are three groups and individuals in each one are assigned treatment values randomly distributed around a mean value of 0, 100, or 200. The final hyper-parameter is a noise parameter Ni, which adds noise in our dataset by randomly switching the feature vectors x of two individuals (but not their protected attributes, proxies, or treatments) with probability Ni. Combining together all these parameters, the synthetic data generation system is illustrated in Figure 3, and a sample of the dataset is given in Table 1.

Table 1. A sample of the generated dataset for the regression task. The columns X_i, $i = 1, ..., k$ represent the features of x where k is the dimension (20 in our case), t is the value of the protected attribute, e is the value of the proxy for the protected attribute and Y_t is the outcome given each of the n values of t.

X_0	X_1	X_2	...	X_k	t	r	$Y_{t=0}$	$Y_{t=1}$	...	$Y_{t=n}$
0.2	-0.1	-0.5	...	-0.3	0	1	100	130	...	180
0.3	0.2	0.4	...	0.7	1	0	40	100	...	300
...	...	...	...	...	...	...	...	...	...	...
-0.1	0.3	0.1	...	0.6	2	1	230	200	...	90



(a) Synthetic data generation *(b)* Synthetic proxy model

Figure 3. In (a), x represents the generated input feature vector. Values for the protected attribute t are generated according to their proportions in r_t and then used with the parameter c_t to generate the proxy values r. This results in an intermediary set of features x' which are used to calculate Y and then swapped randomly according to the parameter Ni. In (b), The synthetic proxy model generates the protected attribute given the proxy value r according to the correlation c_t.

4. **Experiments**

4.1. **Method**

Training was done for 80 epochs with Adam optimizer, a learning rate of 0.001 and a mean absolute error loss. During testing, for each feature vector input to the network, all

protected attribute branches were used simultaneously to estimate the different treatments. Hence, a different outcome was estimated for each person depending on the value of the protected attribute value. The test was repeated 100 times using MC dropout to obtain a different outcome each time. Consequently, an estimation of the disparity along with an uncertainty measure could be computed.

In result Tables 3 and 5, DD_{True} represents the true DD, DD_{Pred} represents the predicted DD by the algorithm, DD_{low} represents the lower bound and DD_{up} represents the upper bound. $MAPE$ is the performance metric, CV is the uncertainty metric, and $\bar{CV}$ represents the mean uncertainty.

In Figures 4 and 5, portion (a) shows the residuals between the correct outcome and the prediction as well as the uncertainties for the prediction outcome of the model for each protected attribute. The residuals correlate with the uncertainty around the prediction and thus demonstrate the performance of the model. Portion (b) shows the lower bound distribution outcome, the higher bound distribution outcome, the estimated distributions and the real distributions for each protected group.

4.2. **Assessing disparity using proxy reliability and protected attribute proportion**

The experiments with the synthetic data were done by varying the percentage of the protected groups r_t and the predictive power of the proxy c_t. We study the performance of our model in rediscovering the correct disparity with an uncertainty measure. Twelve tests were done. Each test tries to measure the performance in a certain situation. These tests and situations are summarized in Table 2. To summarize, we consider cases where the proxy is a poor, average, or excellent predictor of the class (low, medium, or high c_t), where the classes are balanced or imbalanced (high or low r_t), and where the disparity between classes is small or important (low or high $disp$). In all cases we will use three classes, labeled a and b for the minority classes and c for the majority class.

Table 2. Experimental setups with twelve different situations. Each situation has a different set of hyper-parameters representing different proxy correlation, proportions of protected groups and disparity. ↑ means high, ↓ means low and - means mild.

Situation	Description	c_t	r_t	$disp$
			a b c	a b c
1	↓ c_t ↓ r_t ↓ $disp$	0.1	0.05 0.3 0.65	0.0 100.0 200.0
2	↓ c_t ↓ r_t ↑ $disp$	0.1	0.05 0.3 0.65	0.0 500.0 1000.0
3	↓ c_t ↑ r_t ↓ $disp$	0.1	0.3 0.3 0.4	0.0 100.0 200.0
4	↓ c_t ↑ r_t ↑ $disp$	0.1	0.3 0.3 0.4	0.0 500.0 1000.0
5	- c_t ↓ r_t ↓ $disp$	0.5	0.05 0.3 0.65	0.0 100.0 200.0
6	- c_t ↓ r_t ↑ $disp$	0.5	0.05 0.3 0.65	0.0 500.0 1000.0
7	- c_t ↑ r_t ↓ $disp$	0.5	0.3 0.3 0.	0.0 100.0 200.0
8	- c_t ↑ r_t ↑ $disp$	0.5	0.3 0.3 0.4	0.0 500.0 1000.0
9	↑ c_t ↓ r_t ↓ $disp$	1.0	0.05 0.3 0.65	0.0 100.0 200.0
10	↑ c_t ↓ r_t ↑ $disp$	1.0	0.05 0.3 0.65	0.0 500.0 1000.0
11	↑ c_t ↑ r_t ↓ $disp$	1.0	0.3 0.3 0.4	0.0 100.0 200.0
12	↑ c_t ↑ r_t ↑ $disp$	1.0	0.3 0.3 0.4	0.0 500.0 1000.0

4.2.1. **Result**

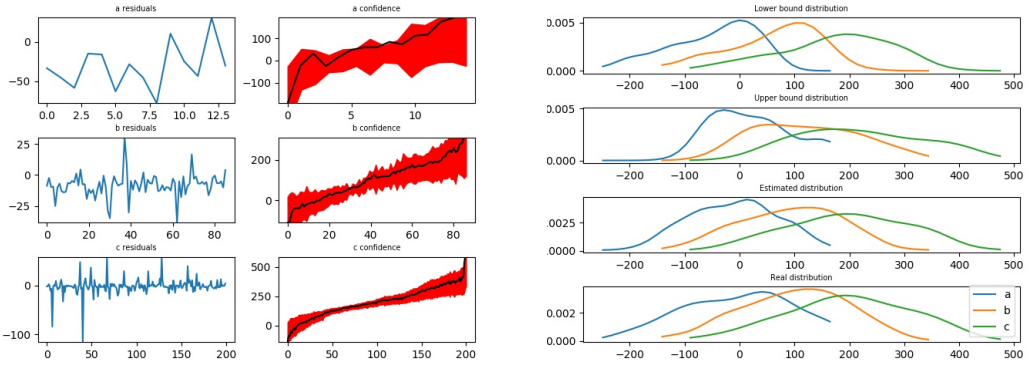
The objective of the first experiment is to test our BT model in different situations with the various combinations of c_t, r_t and $disp$ described in Table 2. The goal is to see the performance in predicting the real disparity DD and estimating lower and upper bounds DD_{low} and DD_{up}. The results presented in Table 3 show that a lower value of the

probability c_t leads to higher values of the $MAPE$ and the CV, which indicate that the model struggles in predicting the true disparity. The difference between DD_{low} and DD_{up} increases also, further indicating the model struggles when using a poorly-correlated proxy. The distribution of classes also affects the uncertainty metrics. For the same values of c_t and $disp$, balanced classes show lower values of $MAPE$ and CV and a smaller difference between DD_{low} and DD_{up}, indicating the system is more confindent in its prediction of DD. The disparity $disp$ has the effect of accentuating the uncertainty about the disparate treatment, which is the result of the prediction. As the disparity gets higher, the values for the outcome are high as well which in return affects the $MAPE$ and the CV in the same way and hurts the performance of the model to identify closer DD_{low} and DD_{up} bounds.

As an example, Figure 4 shows the results of Situation 1, one of the more challenging ones. As explained, the uncertainty on the DD prediction is a lot higher for class a, the much smaller minority class at 0.05% of the dataset, but smaller for class b and very small for class c, the majority class. The predicted outcomes for each of the three classes are also very close to their real values, and well within the lower and upper bounds.

Table 3. Experimental result of the disparity estimation for each situation described in 2

S	DD_{True} a-b a-c b-c	DD_{Pred} a-b a-c b-c	DD_{low} a-b a-c b-c	DD_{up} a-b a-c b-c	$MAPE$ a b c	CV a b c	$\bar{CV}$
1	97.9 196.7 98.7	87.5 184.4 96.9	18.3 79.0 38.5	170.5 290.3 158.4	1.0 0.2 0.1	3.2 0.5 0.4	1.3
2	497.9 996.7 498.7	493.1 989.4 496.4	356.3 741.1 320.6	629.9 1237.8 672.1	1.7 0.0 0.0	5.2 0.1 0.1	1.8
3	97.9 196.7 98.7	96.1 185.4 89.3	56.4 132.8 31.9	143.6 237.9 149.3	0.2 0.3 0.3	0.2 1.3 0.6	0.7
4	497.9 996.7 498.7	455.3 956.9 501.6	375.0 812.2 313.9	535.7 1101.6 689.3	0.3 0.1 0.1	0.3 0.1 0.1	0.2
5	97.9 196.7 98.7	92.4 185.5 93.1	50.0 131.7 47.1	137.7 239.4 142.0	0.4 0.1 0.2	0.5 0.5 0.8	0.6
6	497.9 996.7 498.7	477.8 961.4 483.7	393.6 810.3 345.1	561.9 1112.6 622.2	0.7 0.0 0.0	2.1 0.1 0.1	0.8
7	97.9 196.7 98.7	96.8 195.4 98.5	60.0 149.0 51.6	135.1 241.7 146.6	0.2 0.2 0.2	0.3 0.8 1.0	0.7
8	497.9 996.7 498.7	508.1 967.1 458.9	436.3 837.6 296.6	580.0 1096.6 621.3	0.3 0.1 0.1	0.4 0.1 0.1	0.2
9	97.9 196.7 98.7	101.2 193.6 92.4	55.8 142.0 46.0	148.6 245.3 142.0	0.1 0.1 0.1	0.5 0.8 0.7	0.7
10	497.9 996.7 498.7	505.3 968.6 463.4	428.5 850.8 323.1	582.0 1086.4 603.6	0.4 0.0 0.0	0.7 0.1 0.1	0.3
11	97.9 196.7 98.7	98.9 198.6 99.7	58.2 148.5 53.0	142.3 248.6 149.1	0.1 0.2 0.1	0.3 0.7 0.4	0.5
12	497.9 996.7 498.7	484.9 983.5 498.6	420.3 866.3 354.6	549.6 1100.7 642.6	0.7 0.0 0.0	0.3 0.1 0.1	0.2



(a) Residuals and uncertainties *(b)* Distributions

Figure 4. Performance of the BT model in situation 1 (low c_t, low r_t and low $disp$) as example.

4.3. **Assessing disparity using the name as a proxy**

In this second experiment we test the performance of assessing the disparity using the first name and last name as a proxy for the protected attribute. First, we begin by training the bayesian name proxy model of section 2.2 using the data from the 2010 US Census [20]. To remain coherent with our previous experiment which used three protected groups, we limit the model to predicting three ethnicities, Whites (w, formerly majority class c), Blacks (b, for the extreme minority class a), and Hispanics (h, replacing minority class b). Table 4 shows the performance of the LSTM in predicting each ethnicity based on their first and last names. We replace the protected attribute value in our dataset (situation 3 in Table 2) with names chosen randomly from the 2010 Census dataset according to the ethnicity to which they belong. We also test the performance when there is noise Ni in the data.

4.3.1. **Result**

The results are given in Table 5 for one situation with an increasing level of noise. When the noise in the data is low, CV is also low which indicates that the name acts as a good proxy for the protected attribute. However as Ni gets higher, the uncertainty in the predicted disparity gets larger as well. This means that the performance in predicting narrow bounds for the disparity is dependant on the performance of the proxy model in predicting the correct outcome. However, the increase in uncertainty is not linear to the increase in noise. Looking at $\bar{CV}$ for instance, it does increase linearly for low values of noise, for Ni from 0.0 to 0.4, but then doubling the noise to $Ni = 0.8$ leave $\bar{CV}$ virtually unchanged.

The illustration of Figure 5 is coherent with that of Figure 4. The more rare classes again have higher uncertainty bounds while the majority class w has the thighest uncertainty bounds. Nonetheless, the predicted outcomes are very close to the real outcomes.

Table 4. Performance of the proxy model in predicting the protected attribute value given the last name and first name.

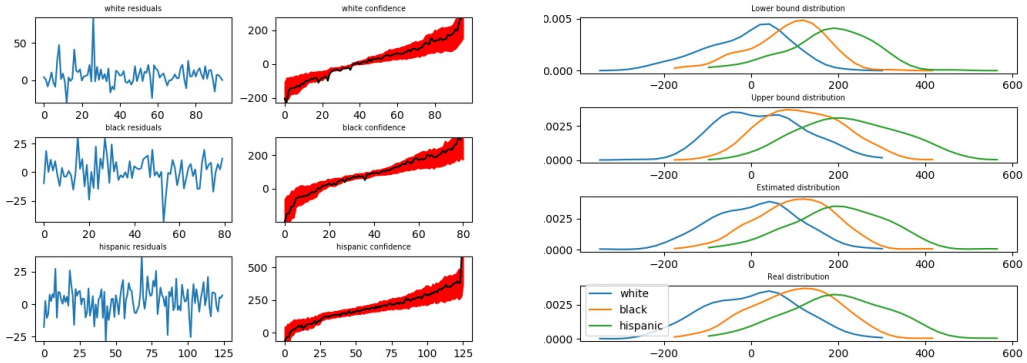
Ethnicity	Precision	Recall	F1-score
White	.95	.72	.82
Black	.15	.49	.23
Hispanic	.38	.80	.52
Mean	.49	.67	.52

Table 5. Experimental result of the disparity estimation for each situation 1 described in 2 using first name and last name as a proxy with increasingly higher noise (Ni) value.

Ni	DD_{True} b-h b-w h-w	DD_{Pred} b-h b-w h-w	DD_{low} b-h b-w h-w	DD_{up} b-h b-w h-w	$MAPE$ b h w	CV b h w	$\bar{CV}$
0.0	97.9 196.7 98.7	91.5 184.7 93.2	53.0 138.8 49.4	130.9 230.6 140.0	0.3 0.1 0.1	0.3 0.5 0.3	.36
0.2	97.9 196.7 98.7	97.5 194.4 96.9	60.6 148.9 51.7	135.6 239.9 144.0	0.8 0.8 0.5	0.3 0.9 0.3	.5
0.4	97.9 196.7 98.7	92.0 200.5 108.5	53.0 153.0 58.4	132.5 247.9 159.8	1.4 1.3 0.7	0.3 0.9 0.6	.6
0.6	97.9 196.7 98.7	73.2 182.0 108.7	43.6 141.2 66.1	104.1 222.7 151.4	1.5 1.3 1.2	0.6 0.5 0.2	.43
0.8	97.9 196.7 98.7	103.4 187.7 84.3	65.0 147.0 46.9	141.9 228.3 128.7	2.3 1.6 1.3	0.5 1.2 0.1	.63

5. **Discussion**

In the first experiment, the results show that the model is able to perform well in conditions where the proxy is highly correlated with the protected attribute and the number



(a) Residuals and uncertainties *(b)* Distributions

Figure 5. Performance of the BT model using the name proxy with $Ni = 0.0$.

of instances for each protected attribute is balanced. However, as the correlation c_t and the proportion r_t gets lower, the performance predictably deteriorates as the uncertainty CV gets higher and the difference between the bounds DD_{low} and DD_{up} gets larger. This indicates that we can use the model's uncertainty to know when the disparate treatment assessment is reliable. This in turn can help decision-makers plan to either seek out more information to correct the model when the uncertainty is high and the bounds are wide, or to improve their business practices when certainty about the disparity is high. Future work will seek to determine the threshold at which the disparity value CV can be trusted to call a treatment discriminatory.

In the second experiment, we test the performance of the model using the individuals' names as a proxy for their ethnicity, a popular and readily-available proxy choice. We show that our method is very reliable when using this proxy, which we can see by looking at $\bar{CV}$ which tend to be low (around 0.5). Moreover, while the performance of our BT degrades when noise is introduced in the proxy, which was expected, this degradation seems to be bounded at high noise levels, which is a very positive result. Further study will explore this phenomenon at greater depth.

6. Conclusion

In this paper, we propose and analyse a new way to evaluate disparate treatment based on protected group status, as well as the uncertainty on this disparity when an imprecise proxy is used instead of the protected attribute. We also propose a new metrics to derive the lower and upper bounds for the disparity. Our result show that our model is able to achieve reliable disparity estimations in situation where the correlation of the proxy with the real protected attribute is low and the proportions of protected classes are imbalanced, and that it is able to correctly model the uncertainty around the disparity estimation in those situations. We also demonstrate that our method can reliably use the first and last name as a proxy for the protected attribute to uncover disparity in treatments.

In addition to the directions for future work mentioned in Section 5, we also plan to focus next on modelling confounding effects. It could be the case that an apparent discrimination is due to an unseen variable that influence both the outcome y and the features x while also depending on the protected attribute t, which make it seem as if y is dependent on both x and t when it really depends on the missing variable. Future work will explore how to take this possibility into account in our system.

Acknowledgment

This research was made possible thanks to the support of *La Capitale Assurances et Services Financiers* and NSERC research grant RDCPJ 537198-18.

References

[1] *Machine Bias ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] A. By, J. Silberg, and J. Manyika. *Notes from the AI frontier: Tackling bias in AI (and in humans)*. Tech. rep. 2019.

[3] R. Warner and R. H. Sloan. "The Proxy Problem: Fairness and Artificial Intelligence". In: *SSRN Electronic Journal* (Sept. 2019). ISSN: 1556-5068. DOI: 10.2139/ssrn.3441888.

[4] *Using publicly available information to proxy for unidentified race and ethnicity | Consumer Financial Protection Bureau*. URL: https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/.

[5] H. R. Varian. "Causal inference in economics and marketing". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27 (July 2016), pp. 7310–7315. ISSN: 10916490. DOI: 10.1073/pnas.1510479113.

[6] F. Eberhardt. *Causation and Intervention*. Tech. rep.

[7] J. Pearl. "Causal Inference in Statistics : An Overview * ". In: *Statistics Surveys* 0.0000 (). ISSN: 1935-7516. DOI: 10.1214/154957804100000000.

[8] S. Verma and J. Rubin. "Fairness Definitions Explained". In: *IEEE/ACM International Workshop on Software Fairness* 18 (2018). DOI: 10.1145/3194770.3194776.

[9] R. Guo, L. U. Cheng, P. R. Hahn, H. Liu, L. Cheng, and J. Li. "A Survey of Learning Causality with Data: Problems and Methods". In: (2020). DOI: 10.1145/3397269.

[10] K. H. Chu, N. J. Brown, A. Pelecanos, and A. F. Brown. "Simpson's paradox: A statistician's case study". In: *EMA - Emergency Medicine Australasia* 30.3 (June 2018), pp. 431–433. ISSN: 17426723. DOI: 10.1111/1742-6723.12943.

[11] F. D. Johansson, U. Shalit, and D. Sontag. *Learning Representations for Counterfactual Inference*. Tech. rep. 2016.

[12] U. Shalit, F. D. Johansson, and D. Sontag. *Estimating individual treatment effect: generalization bounds and algorithms*. Tech. rep.

[13] P. Schwab, L. Linhardt, and W. Karlen. *Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks*. Tech. rep. URL: https://github.com/d909b/perfect_match..

[14] *Optimal Transport and Wasserstein Distance*. Tech. rep.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Tech. rep. 2014, pp. 1929–1958.

[16] Y. Gal and Z. A. Uk. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani*. Tech. rep. 2016. URL: http://yarin.co..

[17] *Uncertainty in Deep Learning (PhD Thesis) | Yarin Gal - Blog | Oxford Machine Learning*. URL: http://www.cs.ox.ac.uk/people/yarin.gal/website/blog_2248.html#demo.

[18] M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. *Erratum: Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities (Health Serv Outcomes Res Method (2009) 9 (69-83) DOI: 10.1007/s10742-009-0047-1)*. Dec. 2009. DOI: 10.1007/s10742-009-0055-1.

[19] G. Sood and S. Laohaprapanon. *Predicting Race and Ethnicity From the Sequence of Characters in a Name *. Tech. rep. 2018. URL: http://github.com/appeler/ethnicolr.Gauravcanbereachedatgsood07@gmail.com.

[20] *Decennial Census Surname Files (2010, 2000)*. URL: https://www.census.gov/data/developers/data-sets/surnames.html.

[21] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. *The cramer distance as a solution to biased wasserstein gradients*. Tech. rep. 2017.

[22] A. Ramdas, N. Garcia, and M. Cuturi. "On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests". In: *Entropy* 19.2 (Sept. 2015). URL: http://arxiv.org/abs/1509.02237.