# A Simulation Study

## A.1 Simulation Setup

We simulate the causal relationships of the fictitious credit application example depicted in Figure 1, using the R package `simcausal` [57]. In the FiND world, we eliminate the PA's effect by setting amount $X_A$, debt $X_D$, and the target $Y$ of $A = a$ to their corresponding values among the $A = a'$ distributions $X_A^*$, $X_D^*$ and $Y^*$:

**Real world:**

$$\begin{aligned}
A &\sim \mathrm{B}(\pi_A)\\
X_C &\sim \mathrm{Ga}(\alpha_C, \beta_C)\\
X_A|X_C,A &\sim \mathrm{Ga}(\alpha_A(X_C,A), \beta_A(X_C,A))\\
X_D|X_C,A &\sim \mathrm{B}(\pi_D(X_C,A))\\
Y|X_A,X_D,X_C,A &\sim \mathrm{B}(\pi_Y(X_A,X_D,X_C,A))
\end{aligned}$$

**FiND world:**

$$\begin{aligned}
A &\sim \mathrm{B}(\pi_A)\\
X_C &\sim \mathrm{Ga}(\alpha_C, \beta_C)\\
X_A^*|X_C &\sim \mathrm{Ga}(\alpha_{Am}(X_C,a'), \beta_A(X_C,a'))\\
X_D^*|X_C &\sim \mathrm{B}(\pi_D(X_C,a'))\\
Y^*|X_A^*,X_D^*,X_C &\sim \mathrm{B}(\pi_Y(X_A^*,X_D^*,X_C,a'))
\end{aligned}$$

In both worlds, the PA $A$ is generated by a Bernoulli distribution with success probability $\pi_A = 0.5$, while the confounder is Gamma distributed with $\alpha_C = 3.26$ and $\beta_C = 10.91$. For $\alpha_A$ and $\beta_A$, we take linear combinations of the features in combination with a log link, and for $\pi_D$ and $\pi_Y$ a logit link. We simulate datasets of size $N = 10,000$ for each world, where we use the same seed for both worlds to assure comparability and perform 25 iterations.

## A.2 Approximating the FiND world

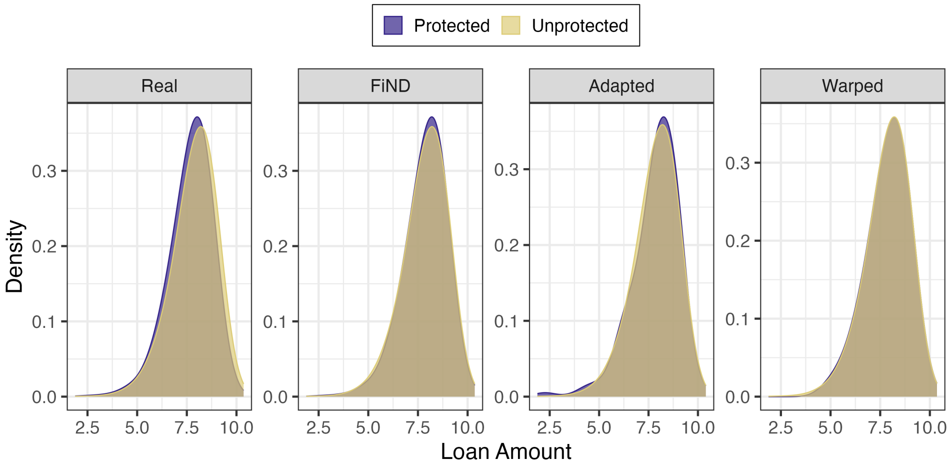

Fig. 5. Distribution of $X_A$ in simulated real, FiND, adapted and warped world per protected $a$ and unprotected group $a'$

Table 2. Distribution of $Y$ and $X_D$ in simulated real, FiND, adapted and warped world per protected $a$ and unprotected group $a'$

| | **Real** | **FiND** | **Adapted** | **Warped** |
|---|---|---|---|---|
| $P(Y=1)$ | $a=0.755$<br>$a'=0.569$ | $a=0.569$<br>$a'=0.569$ | $a=0.585$<br>$a'=0.569$ | $a=0.569$<br>$a'=0.569$ |
| $P(X_D=1)$ | $a=0.732$<br>$a'=0.927$ | $a=0.928$<br>$a'=0.927$ | $a=0.922$<br>$a'=0.927$ | $a=0.927$<br>$a'=0.927$ |

Table 3. Fairness and performance metrics in real, FiND, adapted, and warped world for the simulation study in Section 4.2 alongside their 95% confidence intervals. All predictors are trained and evaluated using data from the same world, e.g., trained and evaluated on real world, trained and evaluated on FiND world, etc.

| World | Fairness | | | | Performance |
|---|---|---|---|---|---|
| | DP | FPR | FNR | PPV | AUC |
| Real | $0.825_{[0.792,0.862]}$ | $0.782_{[0.750,0.836]}$ | $0.954_{[0.926,0.983]}$ | $0.986_{[0.955,0.998]}$ | $0.887_{[0.895,0.899]}$ |
| FiND | $0.987_{[0.964,0.999]}$ | $0.989_{[0.963,1.000]}$ | $0.991_{[0.976,0.999]}$ | $0.988_{[0.957,0.999]}$ | $0.897_{[0.895,0.899]}$ |
| Adapted | $0.982_{[0.959,0.997]}$ | $0.972_{[0.942,0.995]}$ | $0.984_{[0.954,0.997]}$ | $0.975_{[0.952,0.999]}$ | $0.886_{[0.883,0.889]}$ |
| Warped | $0.982_{[0.959,0.996]}$ | $0.964_{[0.938,0.992]}$ | $0.974_{[0.943,0.998]}$ | $0.971_{[0.943,0.996]}$ | $0.893_{[0.888,0.899]}$ |

# B HMDA Experiment

## B.1 Data Setup

We encode and filter the 2022 Home Mortgage Disclosure Act (HMDA) data of the state Wisconsin in the following way[9]:

- $Y$: Binary target indicating loan approved (1) or not approved (0). The original variable "action taken" has eight categories and encodes the status of the loan.
- $A$: Binary PA race with levels $a$ Black applicant or $a'$ non-Hispanic White applicant.
- $X_A$: Numerical variable of the amount of the covered loan, log-transformed.
- $X_P$: Binary variable indicating the purpose of the loan, (1) home purchase or not (0). The original variable has four categories.
- $X_D$: The debt to income ratio, with binary category (1) high ratio or not (0).
- $X_C$: The joint confounders age and gender. Binary age indicates (1) age above 62 or not (0). Binary gender indicates (1) female or not (0). (Note that gender is assumed to be binary purely for simplicity reasons and does not reflect the authors' personal view.)

## B.2 Approximating the FiND world

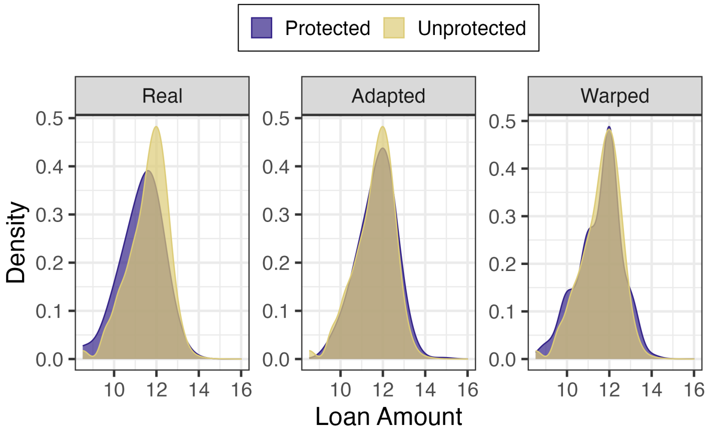

Fig. 6. Distribution of $X_A$ on real-world, adapted and warped HMDA data per protected $a$ and unprotected group $a'$

[9] A detailed description of all variables is provided here: https://ffiec.cfpb.gov/documentation/publications/loan-level-datasets/lar-data-fields

Table 4. Distribution of $Y$, $X_P$ and $X_D$ on real-world, adapted and warped HMDA data per protected $a$ and unprotected group $a'$

| | **Real** | **Adapted** | **Warped** |
|---|---|---|---|
| $P(Y=1)$ | $a=0.486$<br>$a'=0.679$ | $a=0.677$<br>$a'=0.679$ | $a=0.679$<br>$a'=0.679$ |
| $P(X_P=1)$ | $a=0.347$<br>$a'=0.390$ | $a=0.398$<br>$a'=0.390$ | $a=0.384$<br>$a'=0.390$ |
| $P(X_D=1)$ | $a=0.298$<br>$a'=0.371$ | $a=0.335$<br>$a'=0.371$ | $a=0.371$<br>$a'=0.371$ |