Using Data Science to Protect Residential Water Quality

R. Nerenberg, M.D. Lemmon, M. Sisk, E. Clements, and Y. Duan

University of Notre Dame

Project Objectives

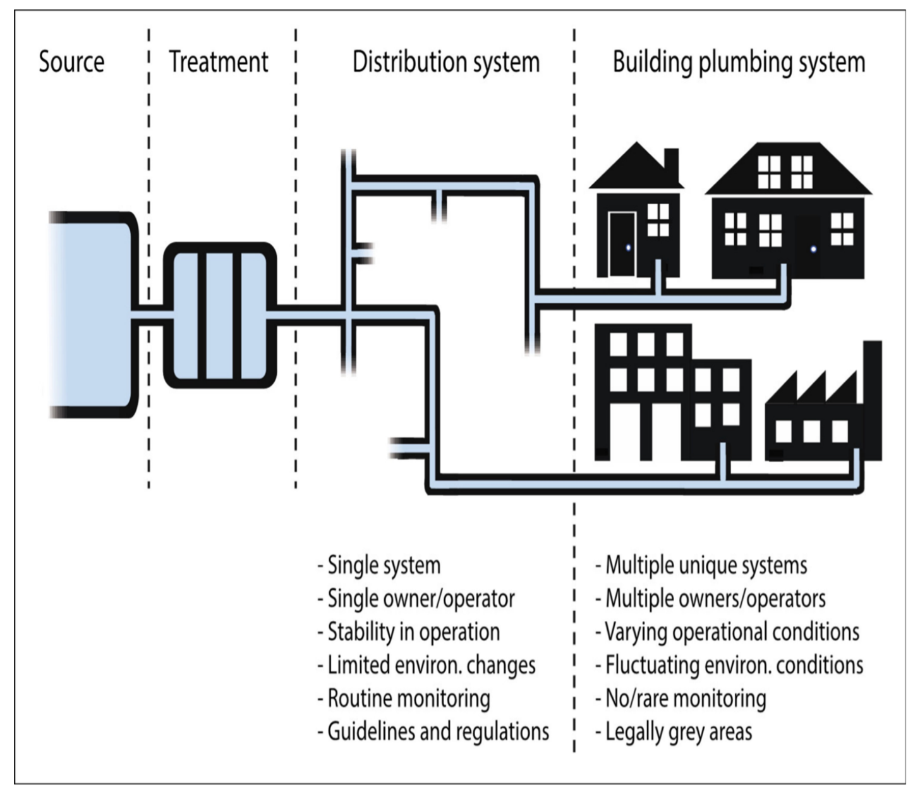


Figure 1. Schematic of water distribution network

Water utilities capture raw water, treat it to EPA standards, and distribute it to users via a piped network. While utilities must comply with EPA standards up to the user's connection, conditions in residential distribution networks can greatly degrade water quality.

Tools are needed to identify homes at risk for water quality problems and to develop community wide strategies for mitigating these problems in a fair and equitable manner.

Federated Learning Approach

Federated Learning takes data from the edge of a distributed cloud system and trains local models that are then sent to a central server to obtain a model for the entire system. In our application, the local models are trained from water quality/usage data streamed from sensors embedded in selected residences of a given community cell (neighborhood). These data streams are used by community centers to train local models that are then uploaded to city hall to assess city-wide residential water quality and to develop mitigation methods addressing water quality issues.

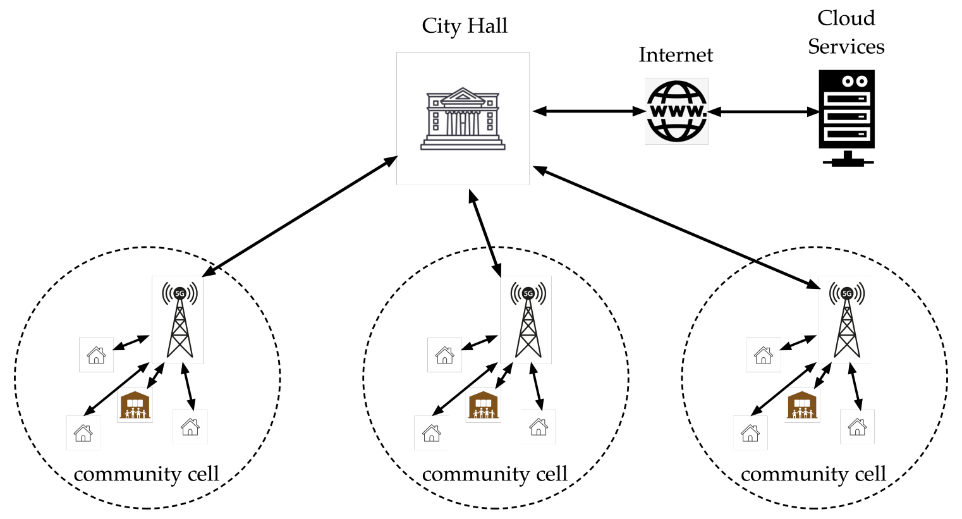


Figure 2. Smart City Federated Learning

Benefits of this Federated Learning Framework include

- preserving the *privacy* of a user's personal water usage,
- reducing the *information rate* that transmitted over the 5G network.

Challenge of this Federated Learning Framework include

- *equitable mitigation policies* across all neighborhoods.

APPROACH: Framework for Fair Federated Learning based on Wassterstein generative adversarial network (WGAN).

WGAN Fair Federated Learning Framework

In the proposed **Fair Federated Learning** framework, each local community center learns the following two models using data which includes individual residence's profile (physical age, occupants) and label (water age).

- A **local classifier** is a dense sequential network that predicts a home's water age based on the residence's profile. The home's water age is used to strategies for reducing water age.
- A residence's profile trains a **local WGAN** that generates "samples" representative of that community's mixture of residences.

Local communities send the local classifiers, η_k, and the local WGAN generator's distribution, P_k, to city hall. The cloud side model, η, is trained using data generated from the mixture of generators, P_k, to minimize the loss function in eq. 1 formed from the empirical mean squared model error and the statistical risk difference

$$\min_{\eta}\left[\frac{1}{N}\sum_{k=1}^{N}(\eta_k(\hat{x})-\eta(\hat{x}))^2+\alpha\,|P(\eta(\hat{x})=1\,|\,\mathrm{SAC})-P(\eta(\hat{x})=1\,|\,\mathrm{SDC})|\right] \qquad (1)$$

where $\{\hat{x}_k\}_{k=1}^{N}$ are samples generated by a mixture of the local WGAN distributions, P_k, SAC are socially *advantage* communities, and SDC are socially *disadvantaged* communities.

WGAN For Local Data Generation

WGAN is used to generate data that has the same distribution as the local community's original dataset. The WGAN consists of a

- **Generator G**: ($R^m \to R^n$) that is a dense sequential network that generating samples that have the same distribution as a prior distribution P_z, and a
- **Discriminator D**: ($R^n \to R$) is a that is dense sequential network that estimates the Wasserstein distance (W distance) between the local real distribution $P_r(x)$ and the generative distribution $P_g(x)$.

The WGAN's objective function is

$$\min_{G}\max_{D}\left[\mathbb{E}_{x\sim P_r}[D(x)]-\mathbb{E}_{z\sim P_z}[D(G(z))]+\lambda\mathop{\mathbb{E}}_{\boldsymbol{x}\sim P_g}\left[(\|\nabla_{\boldsymbol{x}}D(\boldsymbol{x})\|_2-1)^2\right]\right].$$

which may be seen as game between the generator, G, and the discriminator, D.

Preliminary Experiment Results

Our fair Federated Learning framework will be used with the water usage and residence profiles to be generated by this project. As a preliminary test we applied our fair framework to the UCI Adult dataset. The UCI dataset is composed by:

- **14 categorical features**, e.g.:workclass, education, race, sex, age, capital-gain.
- **1 binary label**: income(≤50k, >50k)

We split the dataset to two groups:

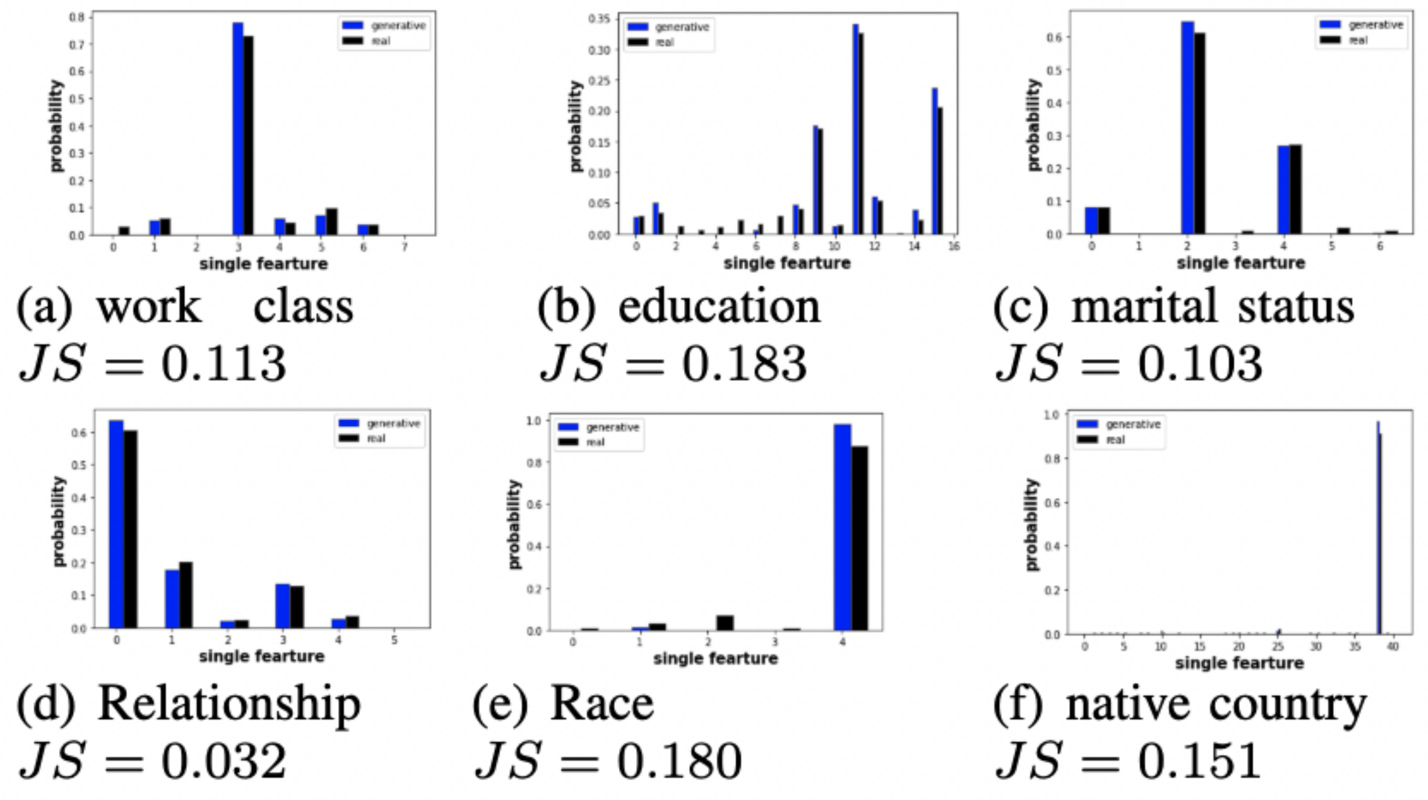
- **Male group**: which represents the socially advantaged community (SAC).
- **Female group**: which represents the socially disadvantaged community (SDC).

We applied the fair federated learning framework to the UCI Adult dataset and obtained the following results.

Local data generation

We trained local WGAN for generating local data. We plot the generative distribution with the real distribution for several features and compare them from two perspectives.

1. **Visual comparison**: similar
2. **JS divergence**: relatively small value.



(a) work class $JS = 0.113$ (b) education $JS = 0.183$ (c) marital status $JS = 0.103$

(d) Relationship $JS = 0.032$ (e) Race $JS = 0.180$ (f) native country $JS = 0.151$

Fair global model training

We trained the global side model and plot the accuracy and risk difference of different α value in equation (1).

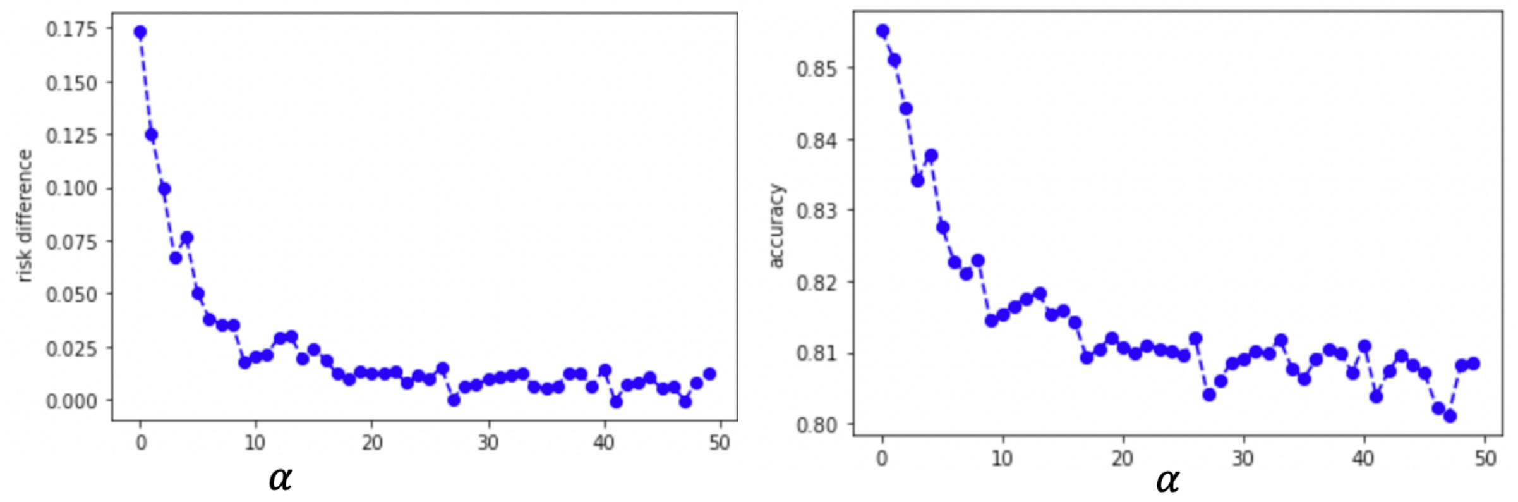


Figure 3. The risk difference and classifier accuracy corresponding to different α

1. When $\alpha = 0$, there is no fairness regularization. The accuracy is 0.85. The risk difference is 0.175, which means the male group has a higher probability to be predicted as having a higher salary. The classifier trained under $\alpha = 0$ is unfair for the female.
2. Classifiers trained under $\alpha \geq 20$ can perform fairly between the male and the female (risk difference is around 0.01), with a 0.04 cost of accuracy (accuracy is 0.81).

Conclusion And Future Work

- We proposed a fair Federated Learning framework and preliminary results with the UCI Adult dataset shows our framework can train models that achieve fairness.
- Future work will use this framework on data generated by simulation models: stochastic water demand generator (SIMDEUM), plumbing model software (EPANET).