USING DATA SCIENCE TO PROTECT TAP WATER QUALITY

Final Project Report to the Lucy Family Institute for Data and Society

June 1, 2023

This report summarizes the activities carried out as part of a Lucy Family Institute seed grant to **Rob Nerenberg**, Professor, CEEES, **Mike Lemmon**, Professor, EE, and **Matt Sisk**, Associate Professor of the Practice, Lucy Family Institute. The one-year grant started on June 1, 2022.

The goal of the research was to use data science to address the safety of public water supplies. This is especially important for the most vulnerable urban populations, low-income and older community members. It addresses one of the themes identified at the Lucy Institute's inaugural symposium on October 27, 2021: "Data Science for Tackling Health Inequities & Disparities."

The overarching concern is water stagnation in premise (private) plumbing systems, also referred to as "water age". During this stagnation time, chlorine residuals can dissipate, allowing for microbial growth. Stagnation also can result in higher concentrations of lead and copper. Since water age is linked to both plumbing system configuration, which often is unknown, and water usage, which is highly stochastic and dependent on house occupancy, it is difficult to predict water quality problems. Water age can be reduced by periodic flushing, but excess flushing can lead to water wastage, which is an increasing concern in many part of the US and the world.

In this research project, two studies were included. The goal of Study 1 was to use data science to help identify homes with greater risks of high water age. The goal of Study 2 was to use data science to develop strategies to mitigate risks. Each is briefly described below, followed by a summary of outcomes and future steps. The Study 2 Final Report was published at the project's website: https://www3.nd.edu/~lemmon/projects/Lucy/

STUDY 1 - IDENTIFYING POTENTIAL RISK

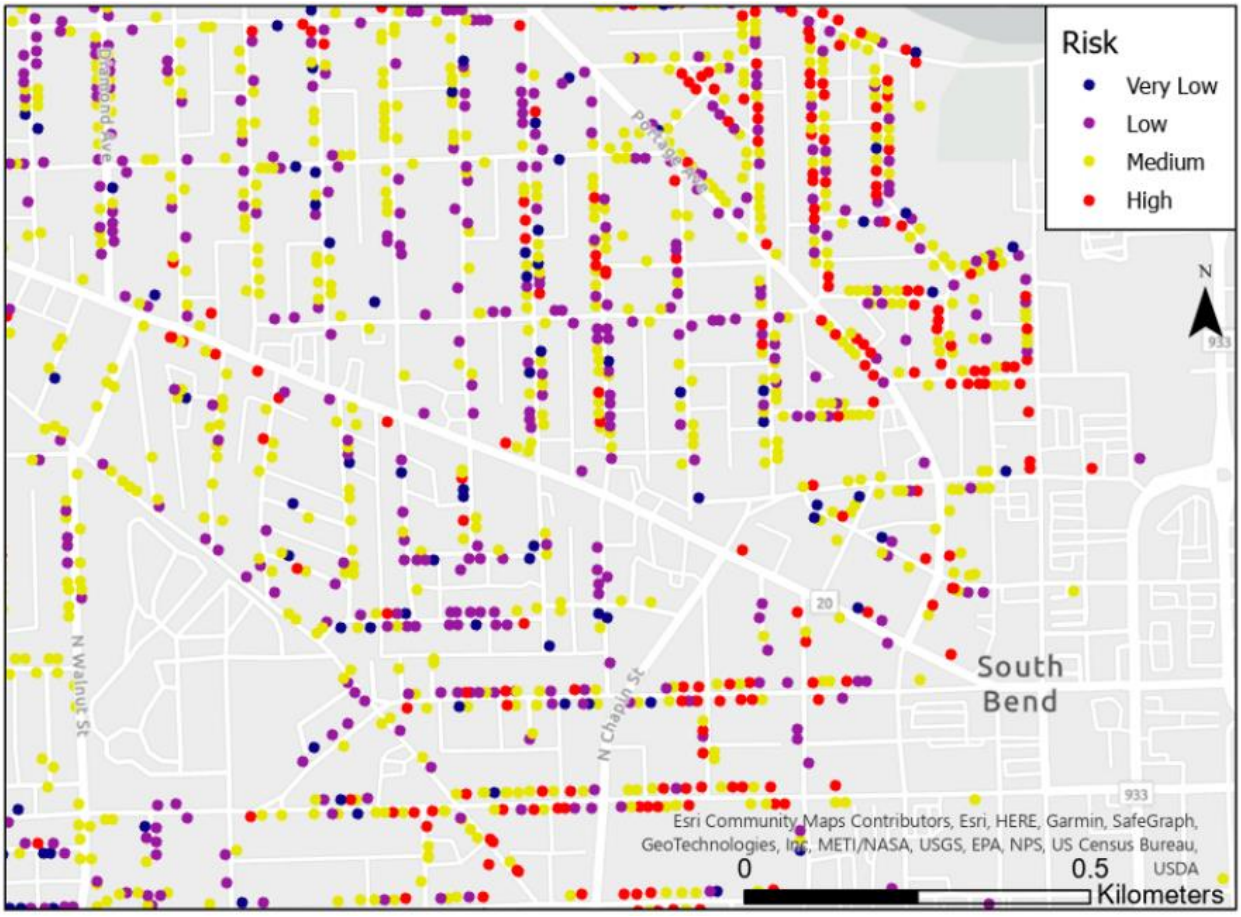
This study used a novel approach to identify homes with potential for high water ages: coupling existing GIS records of home age and size, and water meter records from the utility. Larger, older homes with low water consumption are likely to have high water age and poor water quality. Data from the City of South Bend was used as a case study.

Proposed Deliverable: in order to assess the relative age of water at the individual household level, data on water usage were acquired from the City of South Bend using an existing data-sharing agreement. These data are set up with API access, which gives a monthly updated view of accounts in the utility system, consumption and billing amounts. This database contained over 7 million records of consumption for around 120,000 individual accounts. These accounts were then tied to around 43,000 addresses (because there are frequently old accounts for non-current owners/renters). Once these data were cleaned (removing inactive accounts, sewage-only charges, commercial properties, and cases where the last bill was over a year old), there were around 40,000 accounts left. This matches well with the total number of households in the City of South Bend.

This cleaned dataset allowed us to compare the average use of water for property and compare it to the recent usage. This inferred a higher risk of stagnant water for those properties with a lower current use. Because each account is associated with a service address, these were also tied to property records

indicating the size of the structure, its construction age and years since last renovation, all characteristics that can be used to model the risk of stagnant water in the household.

Outcome: as a proof of concept, the following map shows risk at the household level. Risk is defined as larger households (> 2000 ft^2) with lower than average water use and built before 1979. Moving forward, this definition of risk can be modified based on the results of other studies into the factors leading to stagnant household water.



STUDY 2 - MITIGATING RISK

Periodic flushing of residential plumbing can improve water quality and mitigate health risks to residents. This can be done manually, or with automated purging devices. Optimal flushing schedules, however, are site specific, depending on plumbing configuration and water usage patterns. Providing a single flushing schedule for all homes would be inefficient, probably leading to excessive flushing in some homes and insufficient in others. Water wastage due to excessive flushing would be a concern in areas with water scarcity.

We proposed using federated learning concept to determine optimal flushing schedules based on the type of neighborhood. In particular, this study proposed a plan that first trained models predicting water age for a statistical sampling of community residences and then uses federated learning to identify aggregated models from which neighborhood-specific purging schedules can be obtained. By using federated learning, we can statistically sample a small number of households while developing

models that are tailored to each neighborhood. The proposed work would first develop the proposed federated learning framework, then use simulation studies to assess its ability to fairly reduce residential water age in a manner that was statistically agnostic to neighborhood sensitive attributes such as racial or socio-economic makeup.

Task 1 – Training Local Models for Residential Water Age:

Proposed Deliverable: a neural network model predicting water age of a residential building given its water distribution network and pattern of water usage. That model would be used to create the training dataset used in Task 2.

Outcomes: The Lemmon group randomly generated 10,000 residential water distribution networks. This data was to be used with hydraulic simulation models to generate water age targets for the training dataset. The Nerenberg group developed a hydraulic and water quality model, using chlorine concentration as a water quality indicator. Unfortunately, the Nerenberg group was not able to scale up from a hydraulic model for a single home to simulating the 10,000 homes from the Lemmon group in time for this study. Lacking this simulation data, the Lemmon group modified this task to work with an existing data set (UCI Adult Dataset: https://archive.ics.uci.edu/ml/datasets/Adult) for individual income prediction. The income level was taken as a surrogate for water age and treated race as the sensitive attribute. The resulting income/race dataset was randomly partitioned into 7 distinct neighborhood datasets, each with differing racial makeup and income profiles. Those datasets were used to train neural network models that used the dataset's categorical inputs to predict race and income level.

Task 2 – Federated Learning of Neighborhood Residential Water Age:

Proposed Deliverable: A generative adversarial network (GAN) that generates "fake" data from neighborhood households in a way that preserves the privacy of individual users. Those GANS would form the local models of the federated learning framework used in task 3.

Outcomes: This task was modified to make use of the income/race dataset created in task 1. Because task 1's dataset had categorical inputs, this task used a Wasserstein GAN (WGAN). This task demonstrated that the WGAN successfully learned the income and racial distributions of each neighborhood with a small JS divergence and so could be used as a local model in a federated learning framework.

Task 3 - Simulation Study of Water Purging Schedules:

Proposed Deliverable: a neighborhood water purging schedule that is statistically fair.

Outcomes: This task was modified to make use of the income/race WGANs generated in task 2. The original task was to determine water purging schedules, but because of changes in task 1's dataset, we modified the problem to fairly allocate improvement grants to economically disadvantaged neighborhoods. ``Economically disadvantaged'' meant that the resident had an income below the poverty line of $20,000/year. ``Fair'' meant that the decisions also minimized risk difference with respect to racial statistical parity. ``Grant allocation'' served as a surrogate for water purging schedules. This task demonstrated that the proposed federated learning framework could indeed enhance the statistical parity of grant decisions with regard to neighborhood racial makeup. Details of this task's

results will be found in the study's final report published at the project's website: https://www3.nd.edu/~lemmon/projects/Lucy/

PRODUCTS

1. R. Nerenberg, M. Sisk, M.D. Lemmon, E. Clements, Y. Duan, ``Using Data Science to Protect Residential Water Quality'', Lightning Talk and Poster, Lucy Family Institute for Data and Society Annual Fall Symposium, October 11, 2022.
2. R. Nerenberg, M. Sisk, M.D. Lemmon, E. Clements, Y. Duan, ``Using Data Science to Protect Residential Water Quality'' Project Update presented to Katie Liu (Lucy Family Institute), March 30, 2023
3. Yuying Duan and M.D. Lemmon, ``Fair Federated Learning for Deciding Neighborhood Improvement Grants'', Study 2 Final Report for Lucy Family Institute project – Using Data Science to Protect Residential Water Quality, May 15, 2023.

FUTURE PLANNED DELIVERABLES:

1) M.D. Lemmon and L. Montestruque, "SCC-IRG Track 2 – Project Overview – Transfer Learning for Fair and Scalable Flood Management in Smart Connected Cities", White Paper for NSF-SCC Project, University of Notre Dame and HydroDigital LLC, May 1, 2023.

   OVERVIEW: The U.S. National Climate Assessment predicts the intensity and frequency of extreme precipitation events will increase by 15% over the midwestern United States. Flood damage costs in the US were nearly $17 billion/year between 2010 and 2018 and this flooding will certainly worsen under future climate scenarios. There is, therefore, a critical need for tools that midwestern cities can use to mitigate the flooding risks created by these extreme storms. This project's objective is to develop those tools for smart cities that use Information, Communication and Control (ICC) technologies to meet its citizens' needs.

   This project uses an embedded sensor network called CSOnet to meet its objectives. CSOnet embeds Internet-connected sensors and actuators into a city's underground wastewater (sewer) network to reduce combined sewer overflow (CSO) events during rainstorms. Successful commercialization of CSOnet technology has led to CSOnet deployments in a number of U.S. cities. This project will work with these CSOnet-enabled cities to train deep learning models that predict the likelihood of surface and basement flooding in each city neighborhood. These models would then be used for the real-time control and long-term management of urban flooding.

   This project addresses technological and social science challenges that face smart cities using deep learning to enhance their resilience to extreme climate. The main technological challenges involve training models that can be used to control wastewater flow and whose knowledge can be easily transferred between cities of all sizes. That challenge will be overcome using transfer learning in which a pre-trained flood model is retuned using real-time data from the city's CSOnet system. The main social science challenge involves training models that can recommend which neighborhoods should receive wastewater infrastructure (I/F) improvements, and to do so in a manner that is perceived as fair with respect to a sensitive attribute such as neighborhood racial or socio-economic makeup.

That challenge will be overcome using models that employ a fair learning representation layer to statistically decorrelate neural network decisions from a chosen sensitive neighborhood attribute.

2) Emily Clements and Robert Nerenberg, “Assessing the Impact of Stochastic Demands and Water Purging Devices on Chlorine Residuals in Premise Plumbing Systems.” Manuscript in final stages of preparation, May 31, 2023

3) Matt Sisk, Emily Clements, and Robert Nerenberg, “Coupling GIS records of home characteristics with water meter records to identify homes with potential water quality risks.” Proposed manuscript.