Detecting Lead Contamination in Tap Water, a Data-Driven Approach

Armin Sabouri, Tuna Sogut, Renato Mancuso

***Abstract*—Leaching of lead and other contaminants into tap water poses a significant health risk and is expensive to accurately and efficiently detect in real time. We have utilized commercially available low-cost water metric sensors to represent water samples in an n-dimensional space and leveraged classification models to detect lead contamination in tap water. Experimental results suggest that our novel methodology is an effective and cost efficient way to detect lead contamination. Furthermore, our data driven approach shows promise of being applied to other heavy metal contaminants and organic pollutants while still using the same hardware.**

I. Introduction

Drinking water contamination remains a widespread issue in both the developed and developing world. According to a recently published report [1], nearly 77 million people in the U.S. — about a quarter of the total population — in 2015 alone were served by water delivery systems reported in violation of the safety thresholds established by the Safe Drinking Water Act. Common contaminants include lead, copper, nitrates, radionuclides, arsenic, and pathogenic bacteria. Exposure to these contaminants has been linked to cancer, to chronic diseases, as well as to miscarriages and birth defects. According to a UNICEF/World Health Organization (WHO) report [2], about 2.1 billion people globally lack access to safely managed drinking water. The problem of poor water quality is hard to address because: (1) water quality testing requires expensive equipment only available in specialized facilities; and (2) testing at the source is often not sufficient because contaminants and pollutants frequently infiltrate last-mile distribution channels, meaning water that is measured to be safe at distribution centers is prone to becoming unsafe by the time that water reaches the tap. Our goal is to develop a new technology to detect harmful pollutants and contaminants in drinking water. Three primary design goals for our technology are (1) low production and maintenance costs; (2) re-usability of the solution for repeated measurements over time; and (3) ability to produce water quality assessments in real-time. Because no affordable and reusable sensors exist to detect a number of common contaminants (e.g. lead, copper, nitrates), we use a data-driven approach instead. In a nutshell, our approach consists in using an array of inexpensive, commercially available sensors to collect multiple metrics (e.g. pH, conductivity, ...). The variation of such metrics is correlated in different magnitudes to the presence of common pollutants, meaning pollutants can have their own signature, which we can detect by evaluating such metrics. Thus, we create in-lab contaminated water samples with known characteristics and a specific lead concentration to build a training data-set. We are then able to classify new water samples that have unknown characteristics by leveraging established data analysis and classification techniques. As a first stage, we specifically focus on detection of lead contamination in tap water. The overarching challenges of detecting lead contamination is noted in figure 1. The contamination surface of water supply systems is large as there are many different sources of potential contamination. Lead can leach into a water supply at the source, during distribution, or in the last-mile delivery to residences. The size of the contamination surface introduces great complexity, therefore, it is far more efficient to monitor and check for contamination at the point of delivery: residences. Contamination in residences can provide conclusive information on if the water is contaminated and to what degree. For human health, and especially infants, it is critical that the detection must be in the magnitude of tens of ppb concentration. The exposure surface is an even larger domain in which a given individual's exposure is dependent on which water sources they drink from and how often, and how many additional sources of lead they are exposed to. There is also the added complexity that lead exposure is far more detrimental and significant for infants than for adults. Following a similar logic as before, the complexity of the exposure surface means that efficient and conclusive testing can only be done at the very end of this sequence: blood lead levels. With that being said, our research focuses exclusively on the detecting lead contamination in residential tap water.

II. Related Works

The requirements for our device stem from the challenges that lead contamination detection presents. Lead poisoning occurs from a buildup of lead in the body, meaning that an individual has to consistently consume lead in order to get lead poisoning. Therefore, an effective lead sensor needs to operate for long periods of time without human intervention to accurately determine if a given water source is consistently lead contaminated. Lead contamination is mainly a concern in lower income communities due to how widespread lead plumbing is in older homes. Therefore, to feasibly combat lead contamination on a large scale, an effective lead sensor must also be inexpensive, costing on the order of tens of dollars as opposed to hundreds. Last but not least, an effective lead sensor must be precise enough to detect lead presence at 15ppb and have a low false positive rate as there could be many other harmless metals in drinking water, such as iron. Currently, there exists no method or device for detecting lead contamination that satisfies all these criteria.

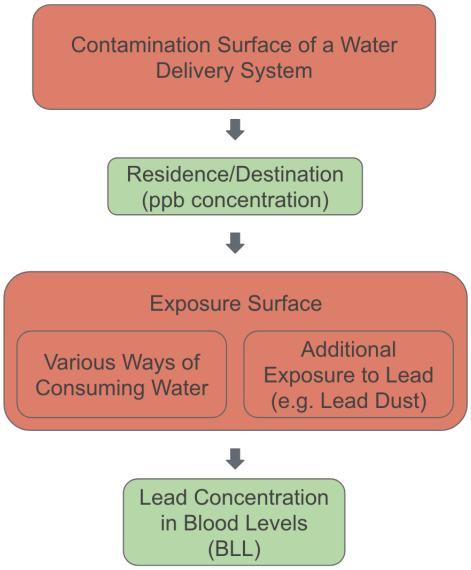

Fig. 1. Green boxes represent points of convergence where efficient and conclusive detection is feasible. Red areas represent contamination/exposure surfaces too large to make conclusive and efficient testing feasible.

A. lab testing

The currently established method to test for lead contamination in tap water is to collect and send samples to a laboratory with specialized instruments. While this method is precise, it is also costly. In order to monitor the quality of water over a long-term period lab testing requires routine collection and processing of samples, which means that the cost of testing continues to build up over time. This makes laboratory testing inaccessible to lower-income populations, who are most at risk for lead contamination. The following works propose different solutions to address the same problem we aim to solve.

B. Lead Detection via Platinum Electrodes

Researchers from the University of Michigan, Ann Arbor, have fabricated a platinum-based sensor that detects lead when lead oxide present in water is deposited in the gaps between the sensor's electrodes [4]. In terms of the requirements for an effective lead sensor, the platinum-based sensor is reusable and low cost. Furthermore, it is very precise as the two-electrode sensor with 5 m gaps can detect lead ions at 15 ppb with no false responses in 3 days. The only aspect that needs improving is the time component. Because lead detection occurs when the gap between the electrodes is bridged by the lead oxide being deposited, it can take several days for the detection to occur.

C. Lead Detection via Differential Pulse Voltammetry

An alternative is to use an electrochemical sensor with under-potential deposited bismuth sub-layer on a thin gold film to cause a reaction with the lead present in the water and use differential pulse voltammetry (DPV) to gauge the concentration of lead[5]. One issue is that their device is a single use sensor that goes through an irreversible reaction that renders the electrode incapable of reacting with lead. Thus this sensor is single-use, and while the cost of each sensor is $2, the aggregated cost over a long period of time builds just like labratory testing. However, the central issue is their precision. The researchers claim their detectable concentration of lead ions to be between 8×10^{-7} M and 5×10^{-4} M. The low end of their range is too high to be considered effective for residential use, and as seen in figure 2, their low end corresponds to 265 ppb, which is over 17 times EPA's action threshold for lead. While it is still important to be able to detect lead in the 100s ppb range, failing to detect lower concentrations of lead can still pose significant health risks for children.

$$8 \times 10^{-7} M = \frac{0.0000008\ mol}{liter}$$

$$\text{Lead Nitrate } (Pb(NO_3)_2) \text{ molar mass} = \frac{331200mg}{mol}$$

$$\frac{331200mg}{mol} \times \frac{0.0000008mol}{liter} = \frac{.26496mg}{liter} = \frac{.265mg}{liter}$$

$$\frac{.265mg}{liter} = \frac{.265mg}{1{,}000{,}000mg} = .265\ ppm = 265\ ppb$$

Fig. 2. Steps of calculating ppb from mol concentration

Molar mass of lead nitrate multiplied by how many moles solute exist in 1 liter of water equals how many milligrams of lead nitrate exist per liter of water. Because 1 liter of water equates to 1 million milligrams of water, this result can be restated as .265 ppm, which equals 265 ppb

TABLE I
SENSOR REQUIREMENTS

Detection Methodology	Reusable	Real-Time	Cost	Precision
AquaWatch	Yes	Yes	Low	TODO
UMich Lead Sensor	Yes	No	Low	High
Lab	No	No	Highest	High
Bismuth-DPV Sensor	No	Yes	Medium	Medium

III. MOTIVATION

Lead is a highly toxic contaminant that can cause irreversible neurological, hematological, gastrointestinal, cardiovascular, and renal damage, especially for infants and pregnant women. According to a WHO report, lead exposure accounts for roughly 0.6% of the global burden of disease [3], with the highest burden in developing regions. Because lead exposure can cause irreversible damage, its effects can echo throughout victims lives by impacting their earning ability, health, and quality of life. Given the extensive health risks caused by lead poisoning and the lack of effective methods to identify lead exposure, our device will have an immediate and tangible impact on preventing lead-related health complications. In

this project, we target the development of a real-time water quality measuring device able to detect lead contamination in residential tap water at concentrations where lead first starts posing health risks – i.e. at 15 ppb (official EPA limit) and below.

IV. Overview/Approach

The AquaWatch device uses multiple sensors each of which monitors a unique property of the water rather, as opposed to using one expensive lead sensor. By utilizing a simple and cost-efficient micro-controller that collects sensor readings, a user can sample their water and receive results in real-time. This system is preferable to the alternative, sending samples to a third-party and waiting days or weeks for the results. A data-driven approach ensures that even after the sensors have been used past their lifetime ($>$ year), the data collected across all devices is still useful for determining the general quality of a water sample and the presence of lead.

A. General Water Quality

Our team first considered AquaWatch as a cost-efficient, real-time and fully commercialized product which would inform its users of the quality of their tap water at the destination. Every time the device is used 30-50 samples are taken from each sensor and are averaged and saved under a users portfolio. The web application is then capable of informing the user whether their water quality meets the FDA's standard of water quality across various metrics. A sample of water is considered unsafe for consumption if any of the sensors obtain readings outside of the established healthy ranges.

In the future iterations of our web service, we plan to combine our user's samples, their geographic location, and a well maintained city plumbing API to build a model that can provide non-users an estimation of water quality of their neighborhood or home using what we know from their neighbors.

B. Lead Detection

Our method for building reliable models capable of detecting low levels of lead in tap water involves collecting a statistically significant number of samples while being able to leech a detectable amount of lead in our sample water over a considerable amount of time in a laboratory setting. For our experimentation, we kept a reservoir of water and added a piece of flat pipe metal sheet to emulate actual piping. To extract lead ions from our sheet our team decided to use vinegar as an acidic agent, adjusting the acidity of the water to 4 pH. Over time we survey the water and create a R^d space that is correlated with a concentration of lead contamination. Lead contamination is defined by how long a sample of lead has been left in the water.

For the rest of this reading consider a data-point to be a point comprised of all d sensor readings, a sample to be N data points, a sample of water to be M cups of market pantry distilled water, a sample of flat metal sheet to be L cm^2, a sample amount of vinegar to be V, the frequency in which we sample to be t, and total sample time T.

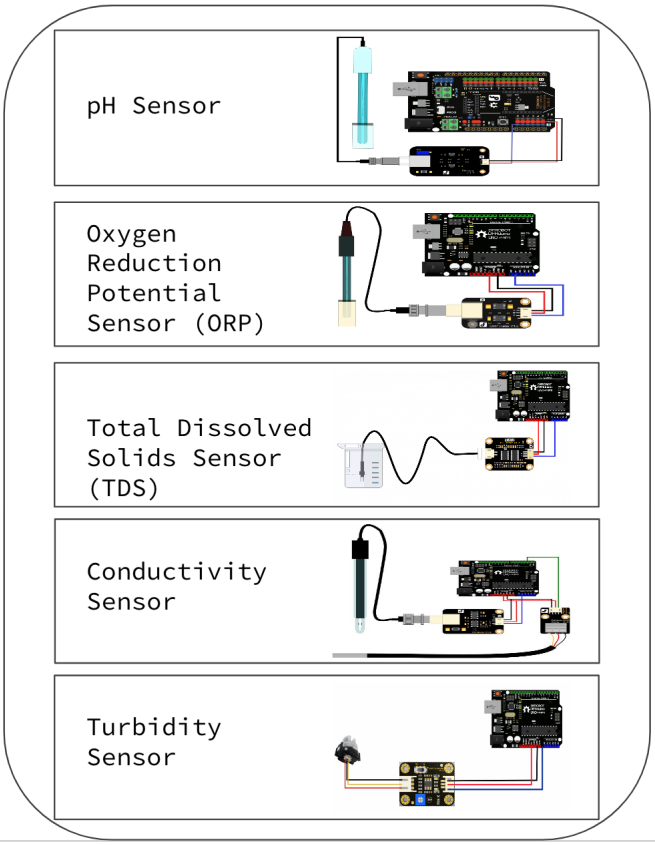

Fig. 3. Schematic of each of the 5 sensors

V. Implementation Details

A. Micro Controller

The AquaWatch device is comprised of 5 sensors that each collect one of the following metrics of water quality: pH, conductivity, turbidity, oxygen reduction potential (ORP) and total dissolved solids (TDS). Each sensor is connected to an Arduino Mega with a Gravity IO expansion shield attached. Because the sensors are part of the Gravity family of sensors manufactured by DF Robot, interfacing them with the Arduino Mega is seamless. The sensors are fed 5V straight from the Arduino, and sensor readings are done in a sequential order.

An important early design decision our team had to make was whether we wanted AquaWatch to interface via a PC and serial port connection or a network device that can work in conjunction with the IO Expansion shield. Because we decided to prioritize collecting large samples from the device at once in a timely and reliable manner, our team chose to perform IO operations from the serial port; however, for a widely distributed commercial product our team built out network capabilities so AquaWatch can make outgoing network requests to a web application.

B. Data Collection

Our data collection method can be described in two parts. First, the AquaWatch is burdened with the task of collecting and formatting a sample such that on the receiving side of the serial port, the package can be read in a common

format. For our use case, we decided to format our data in a JSONable format which essentially mimics a matrix. $\underset{5\times N}{\text{Sample}} = \begin{bmatrix} PH & ... \\ Cond. & ... \\ ORP & ... \\ TDS & ... \\ Turb & ... \end{bmatrix}$. The second portion is to reliably maintain a repository of all samples taken and to include certain meta-information from each sample, including a label indicating if the sample is contaminated, a time stamp, the d dimensional center point of the cluster, and the d dimensional standard deviation of the cluster. The repository described is also a JSONable file. For a particular experiment with adjusted variables, we would make an update a specific JSON file which we refer to as the model.

The largest constraint within our data collection process is the capabilities of our sensors. The pH, conductivity, and ORP need to be immersed in a 1413μS/cm solution for at least two minutes for every two hours of sampling. In our first attempts to use AquaWatch as a research-oriented device, someone needed to be present to move the sensors when needed and to ensure that non-malformed data gets propagated to the model.

For our second attempt, we programmed a mechanical arm made up of five servos and an Arduino Uno to automate all future sampling. Currently, the arm routinely moves the sensors from the lead-contaminated test medium to a reservoir of 1413μS/cm solution in which the ORP, conductivity, and pH sensors are rinsed. To ensure the arm's movement is in sync with the sensors, we keep track of the arm's state. At present, the arm has two states. [1] The arm is in the test medium, which is when the sensors need to begin sampling. [2] The arm is outside the test medium, meaning the sensors must not sample. As a secondary check to help us keep track of the arm's state, we created a simple setup where wires connected to a GPIO and a power pin make contact only when the arm is in the first state. This mechanical check on the arm's state provides a secondary feedback. This redundancy is critical because if samples were accidentally gathered while the sensors were not immersed in the test medium, those readings could potentially skew our data. The listening program that synchronizes the arm's state with the sensors is noted under Algorithm 1.

Algorithm 1 Listening Program

```
procedure LISTEN
    while 1 == 1:
    GPIOStatus = readFromArm()
    if Int(GPIOStatus) == 1
        do:
            currentSample = readFromAquaWatch()
        while isMalformed(currentSample)
        addToModel(currentSample)
    thread.sleep(t)
```

Lastly, there is a web interface that reads and displays all the content that has been recorded as well as displaying useful graphs and visuals. The web interface also displays log files necessary to identify any problems the two devices are having. Connecting to http://192.168.1.20:5000 should show a live feed of our sampling if the an experiment is in progress.

C. Adding to the model

Each sample is considered a d dimensional cluster of N points. Within each cluster, we pre-compute certain useful properties. First, using the least squares approach to determine the center-point of a cluster.

$$\frac{\delta x_\alpha}{\delta} \sum_{i=0}^{N} (x_\alpha - x_{\alpha_i}) = 0$$

Where α is a specific dimension/feature of a cluster, for example, TDS or pH. Second, we take note of the standard deviation of a cluster or in other words how sparse the cluster is:

$$\hat{S} = \sqrt{\frac{(\sum y_n - m)^2}{N-1}}$$

Where

$$m = \frac{\sum_d^{i=0} X_i}{d}$$

and where

x = center of a dimension within a cluster

Third, we compute the point within a cluster that is the closest to all other points in $O(n^2)$ Euclidean distance measurements. Lastly, it is important to note whether this cluster is associated with a contaminated experimentation set or whether we are collecting normal tap water to train our models against.

For every experiment set, we separate two models. One where all values within a data point are normalized according to the min-max method: $z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$. Another data-set includes the original values from the sensors. Both models are saved as separate files. For the rest of this reading we will refer to one data-set as normalized and the other as absolute.

D. Querying against model

When we query a data point against a data-set, we are checking to see if the query lies close enough to a lead-contaminated cluster. If so, we can infer that the point is from a potentially contaminated source. A data point is defined by $\begin{bmatrix} PH \\ Cond. \\ ORP \\ TDS \\ Turb \end{bmatrix}$

Our querying method can be explained in two steps. First, in $O(n)$ Euclidean distance measurements we identify the cluster which the data point is closest to using the clusters center points. Second, we temporarily add the point to the cluster. At this point, we re-evaluate the standard deviation. If the difference of standard deviation is $> \sigma$ then we consider the point to be sufficiently close to be considered part of the cluster. The logic can be further explained as

Algorithm 2 Querying points

```
procedure FINDCLOSESTCLUSTER( dataPoint )
    distances = []
    for cluster in data-set:
        distances += euclideanDistance(datapoint,
        cluster["centerPoint"])
    distances = sort(distances, "acendingOrder");
    return distances[0]
procedure ISINCLUSTER( dataPoint, cluster )
    prevStd = cluster.standardDeviation
    tempCluster = cluster + dataPoint

    return(prevStd - tempCluster.standardDeviation > σ)
```

VI. Experiments and Results

In our experiments, our device was comprised of five analog sensors: pH, conductivity, TDS, ORP, and turbidity ($d = 5$). The turbidity sensor measures scattered light at 90 degrees from the incident light beam as NTUs. The TDS sensor measures ppm concentration of all dissolved solids. The sample size used is fifty data-points ($N = 50$). For every 102 cm^2 (L = 102 cm^2) we use three cups of distilled water (M=1) and add 2.5 ML of distilled white vinegar (V = $2.5ML$) to make the sample acidic. This acid solution is required for the lead leeching to occur. Lastly, we sample every 15 minutes (t=15min) for two hours.

For our training data-set, we use an absolute data-set containing four uncontaminated samples and eight contaminated samples. All models, excluding our clustering method, use 70% of the data-set to train the model. Our testing data-set includes the training set and six more uncontaminated samples.

A. Linear Regression

Our linear regression model aims to predict the coefficients of the following equation:
$y = pH \cdot x_0 + Cond \cdot x_1 + TDS \cdot x_2 + ORP \cdot x_3 + Turb \cdot x_4$
Where y is a boolean indicator of the presence of lead contamination. Using modules from Sci-kit-learn, our linear regression model performed with a 94% accuracy on the training data and 0% accuracy on the testing data-set, meaning our model has been completely over trained and biased towards the training data-set.

B. Clustering Method

When setting $\sigma = 0.7$ and using the querying method mentioned in the section above, our accuracy came out to 83.33%. When testing against only tap water, our accuracy came out to 100%. A possible explanation for this outcome is the distance factor of uncontaminated distilled water and contaminated distilled water after the two-hour mark (T). Center point for a uncontaminated distilled cluster:

$$\begin{bmatrix} 6.84\ (\text{pH}) \\ 1.55(Cond.) \\ 179.5(ORP) \\ 179.98(TDS) \\ 232.72(Turb) \end{bmatrix}$$

Center point for a un-contaminated distilled cluster:

$$\begin{bmatrix} 7.41\ (\text{pH}) \\ 1.53(Cond.) \\ 334.12(ORP) \\ 167.12(TDS) \\ 216.18(Turb) \end{bmatrix}$$

Looking at these two points, we can observe that the real differentiating factor is Oxygen Reduction Potential (ORP). Looking at the graphs in the results section we notice that our lead/vinegar contamination clearly affects the water, however, a sample time of two hours is simply too short for making clear distinctions between distilled uncontaminated and contaminated water. A possible solution is to ignore the cluster which is labeled differently and have a distance $< \mu$ in our preprocessing stage and to increase T.

C. Logistic Regression

Using modules from Sci-kit learn, our logistic model performed with an 88% accuracy on the training data and 91.6% accuracy on the testing data-set.

D. Refining our models

In the next cycle of training and testing, we plan to not only train our models with the absolute values provided from the sensors but the normalized ones as well. We essentially create a space R^{10} where each cluster would correlate to a matrix:

$$\underset{10x50}{\text{Cluster}} = \begin{bmatrix} PH & ... \\ Cond. & ... \\ ORP & ... \\ TDS & ... \\ Turb & ... \\ PH\_norm & ... \\ Cond.\_norm & ... \\ ORP\_norm & ... \\ TDS\_norm & ... \\ Turb\_norm & ... \end{bmatrix}.$$

The motivation for containing both absolute and normalized entries in R^{10} space is to allow for broad ranges of values that may vary within our features. Normalization would then ensure that all features equally contribute when computing classifying variables such as distance.

E. Using Time as a Feature

In our future work, we plan to use "time after lead has been added" as a feature and label. This would provide additional clarity on the concentration of lead included in a sample. However, this provides an additional complexity. If we were to train our models with a known time feature, query points would also be expected to have time as a known feature. A potential solution is to use another learning technique to

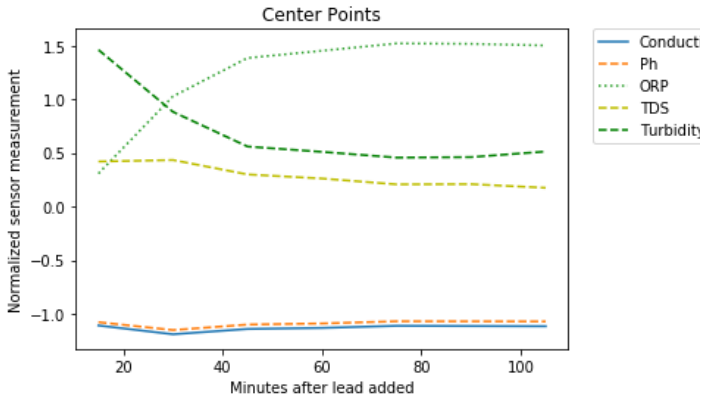

Fig. 4. Data-set has been normalized using min max method

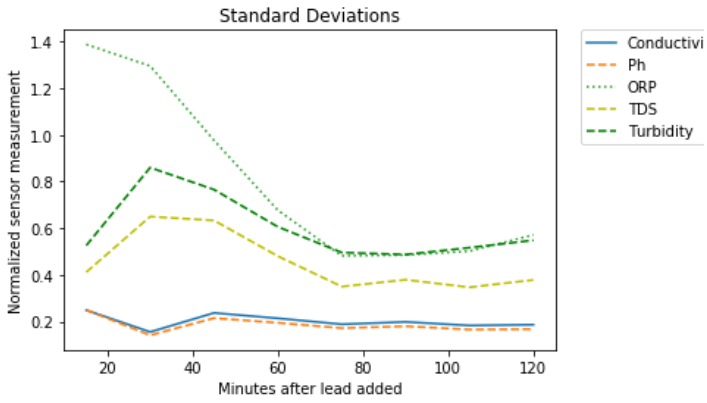

Fig. 5. Data-set has been normalized using min max method

decipher a users time feature and then query against a specific model.

F. Testing

Testing was conducted in two ways. [1] We prepared multiple test mediums by adding vinegar to filtered water and sampled how a piece of pure lead leached into it over a period of two hours. The purpose of this testing was to model how lead leaches into tap water by imitating the actual conditions of lead plumbing. However, because this active lead leaching increases the lead concentration over time, we have no way of knowing what the actual lead concentration is at any given point, only what that concentration looks like in our five-dimensional space. This is where our second method of testing comes into play. By sampling various concentrations of lead nitrate solutions, we obtain data on how specific lead concentrations - 10, 20, 30, 300, 3000ppb - appear in our 5-dimensional space. Finally, by cross-referencing the data points from our initial lead leaching tests with the data points from our known concentration lead nitrate tests, we are able to derive a range of concentration for the lead leaching at given points in time.

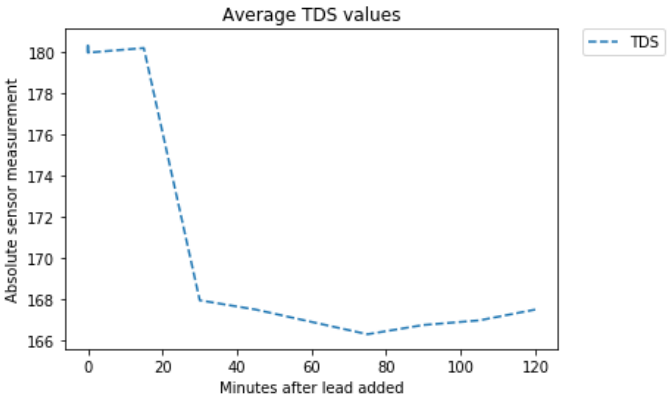

Fig. 6.

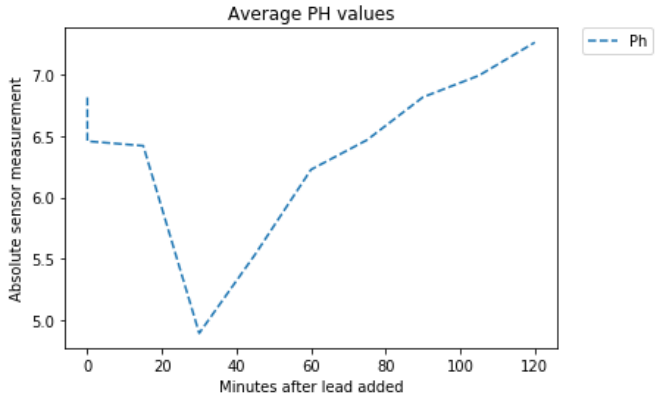

Fig. 7.

VII. Future Works

A. Distributable Product

There are several more requirements our device would need to satisfy before it can become a product. For starters, we must fabricate a custom mold to keep the water sensitive portions of the device away from water while allowing the sensors to be immersed in a liquid.

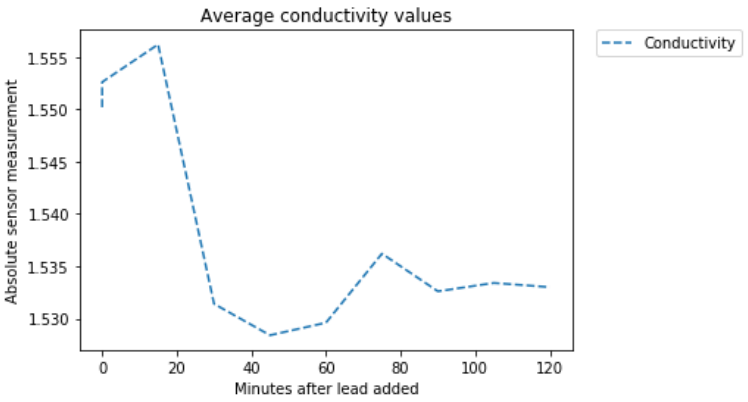

Fig. 8.

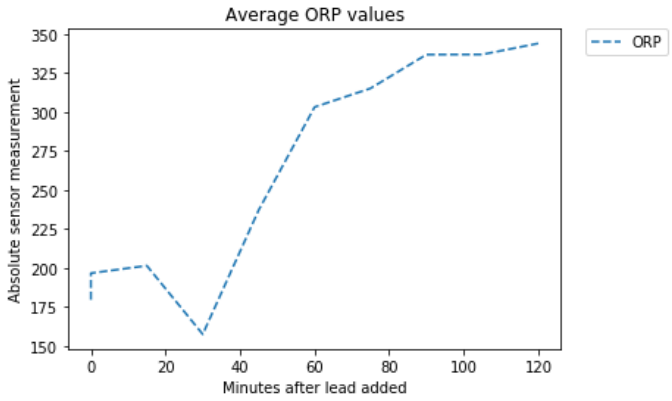

Fig. 9.

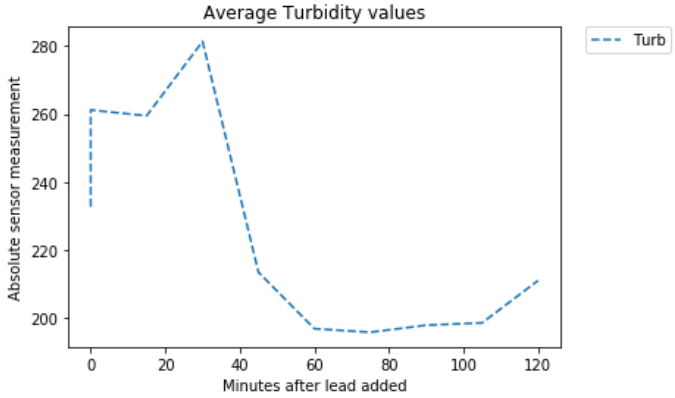

Fig. 10.

B. Providing ppb estimation

Towards the end of this iteration of research, we worked with the BU chemistry department to create several samples of lead nitrate with known concentrations. These concentrations were 10,20,30,300,3000 ppb. We plan to maintain a model for each concentration and use our existing data-sets to map the "time after lead added" feature to the actual concentration of lead. Specifically, we would have two center points – one labeled and one where time after lead added is the only indicator of the concentration of lead. The same querying method described in the previous section would be used to correlate our time feature to a labeled concentration of lead. Essentially, after adding the labeled center to the closest cluster, does the standard deviation increase by δ.

C. Applying methodology to other contaminants

A major benefit of our device and methodology is that it can be generalized to various other contaminants. Ultimately, our device determines if the tested water is safe for consumption by checking for a contaminant specific signature in an n-dimensional space where each dimension is a metric pertaining to the test waters characteristics.

While we have focused our efforts on lead contamination detection in this paper, we believe there is a significant need for a low-cost, durable, and easy to operate water quality sensor for detecting common pollutants and contaminants. Upgrading the United States water infrastructure to eliminate the need to worry about contaminants and pollutants is unfeasible. The American Water Works Association estimates that the cost to simply maintain our water infrastructure will be $1 trillion over the next 25 years, let alone make upgrades. Hence, there is a need for a destination-level water testing method that is cost-efficient enough to be accessible by populations with lower economic status most at risk of being affected by water contamination and pollutants. The core benefit of AquaWatch is that it is a generalized device capable of being trained and used to detect a variety for contaminants and upgraded with contaminant specific sensors if necessary.

VIII. Conclusion

The AquaWatch device coupled with our data-driven methodology can help detect lead contamination in tap water and has the potential to be trained on other harmful organic pollutants and heavy metal contaminants. Because the device is decentralized, durable, economical, and reusable, it has applications in preventing water-related crises and outbreaks in both the developing world and the developed one. While the developed world has safe and drinkable water, for the most part, toxic lead exposure from lead plumbing is still a concern in the US. However, even though our sensor is reusable, further research and development are required to test and guarantee the durability of our sensors being immersed in water for an extended period of time. Furthermore, a procedure and mechanism must be developed for the sensors to automatically be rinsed by the 1413μS conductivity solution. We believe our data-driven methodology can catalyze additional research into detecting contaminants and pollutants by utilizing commercially available sensors and using them to model water samples in an n-dimensional space and analyzing contaminant specific signatures to detect the levels of contamination pollutants.

References

[1] [1] Kristi Pullen Fedinick, Mae Wu, MPhil, Mekela Panditharatne, Erik D. Olson, Threats On Tap: Widespread Violations Highlight Need For Investment In Water Infrastructure And Protections, NRDC Report, May 2017.

[2] [2] UNICEF/WHO, Progress on Drinking Water, Sanitation and Hygiene, Report, 2017

[3] [3] Merit Kennedy, Lead-Laced Water In Flint: A Step-By-Step Look At The Makings Of A Crisis, NPR, April 20, 2016.

[4] [4]Lin, Wen-Chi, et al. A Drinking Water Sensor for Lead and Other Heavy Metals. Analytical Chemistry, vol. 89, no. 17, 2017, pp. 87488756., doi:10.1021/acs.analchem.7b00843.

[5] [5]Yifan Dai, et al. A Simple, Cost-Effective Sensor for Detecting Lead Ions in Water Using Under-Potential Deposited Bismuth Sub-Layer with Differential Pulse Voltammetry (DPV). Sensors, vol. 17, no. 5, 2017, p. 950., doi:10.3390/s17050950.

[6] American Water Works Association, "Buried No Longer: Confronting America's Water Infrastructure Challenge", [FILL IN]