Building Tools for Automatically Processing Lab Data

Rania Hakimi, REU Participant, Bunker Hill Community College
Wenjin Zhang, PhD Student, Northeastern University
Amy Mueller, Assistant Professor, Northeastern University

Northeastern University
College *of* Engineering

Abstract

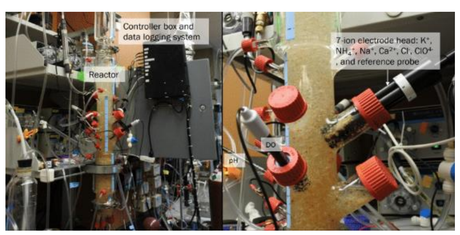
Phosphorus (P) is a major nutrient that wastewater treatment plants are making all efforts to eliminate. The proposed cost-effective way to indirectly quantify P in real-time is through commercially available sensor array. This project aims to building algorithms to automatically extract useful information from available sensors data sheets, thus help build machine learning algorithms for P prediction.

Motivation

- Online sensors for P do not exist. The Mueller lab is bridging this gap by studying chemistry of a P removal reactor.
- This REU project creates tools that automates data extraction to support development of an online P sensor.

Goals

- Creation of a data processing/visualizing tool for Atomic Absorption Spectroscopy (AAS) log files.
- Creation of a tool to extract useful information from logged sensor data.
- Creation of a linear model for P prediction to assess data complexity.



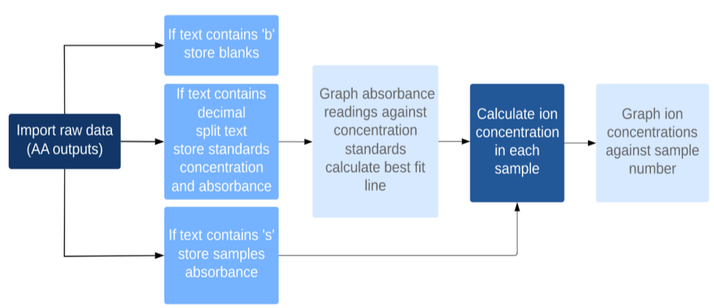
- Sensors include: electrodes for pH, DO, NH_4^+, Na^+, K^+, Ca^{2+}, ClO_4^-, Cl^-, Mg^{2+}.

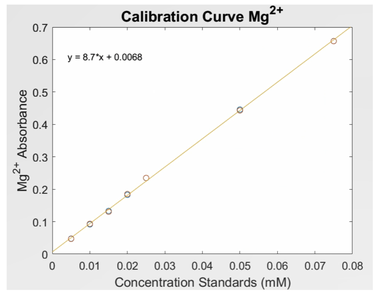
Methodology and Results

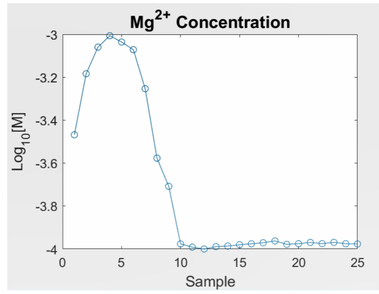
1. Data Processing/Visualizing Tool (AA Data)

Inputs → Algorithm → Outputs

Nab	UNCALp	p	mg/L	0	0.0022	0.0022
W4502:	Sample	uncalibrated				
Nas1	UNCAL	L	mg/L	0	0.2567	0.2567
W4502:	Sample	uncalibrated				
Nab	UNCAL	L	mg/L	7	0.0021	0.0021
W4502:	Sample	uncalibrated				
Nas1	UNCAL	L	mg/L	0	0.2588	0.2588
W4502:	Sample	uncalibrated				
Nab	UNCAL	L	mg/L	3	0.002	0.002
W4502:	Sample	uncalibrated				
Nas2	UNCAL	L	mg/L	0	0.2075	0.2075
W4502:	Sample	uncalibrated				
Nab	UNCAL	L	mg/L	1	0.0025	0.0025
W4502:	Sample	uncalibrated				
Nas2	UNCAL	L	mg/L	0	0.2065	0.2065
W4502:	Sample	uncalibrated				
Nab	UNCAL	L	mg/L	1	0.0025	0.0025
W4502:	Sample	uncalibrated				
Nas2	UNCAL	L	mg/L	0	0.2043	0.2043
W4502:	Sample	uncalibrated				
Nab	UNCAL	L	mg/L	6	0.0022	0.0022

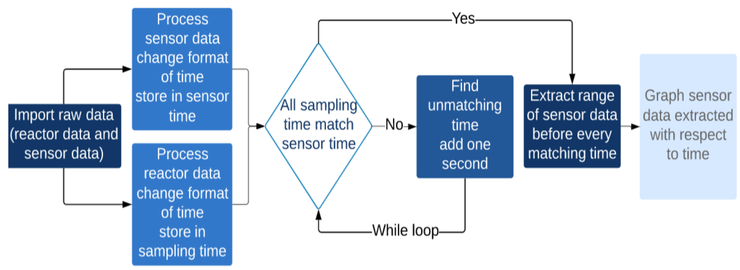


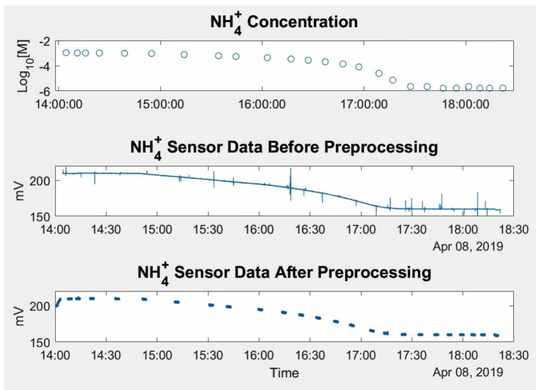




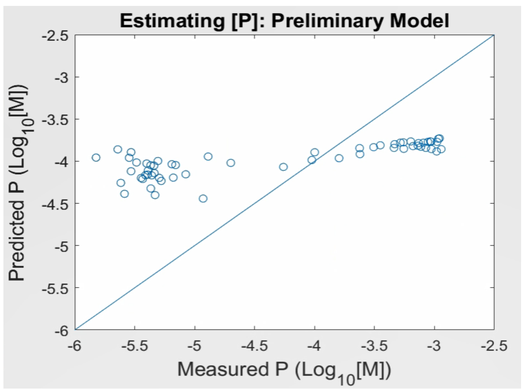
2. Data Extraction Tool (Sensor Data)

20190408	1.555E+09	0	0.19863	0.34698	0.42906	0.34103
20190408	1.555E+09	1	0.19912	0.34714	0.42987	0.34239
20190408	1.555E+09	2	0.199	0.34705	0.42958	0.34201
20190408	1.555E+09	3	0.19887	0.34706	0.42986	0.3418
20190408	1.555E+09	4	0.19787	0.34719	0.42904	0.34122
20190408	1.555E+09	5	0.19791	0.34611	0.4295	0.34169
20190408	1.555E+09	6	0.19838	0.34601	0.43004	0.34128
20190408	1.555E+09	7	0.19838	0.34596	0.42874	0.34096
20190408	1.555E+09	8	0.1992	0.34649	0.42985	0.3414
20190408	1.555E+09	9	0.198	0.34711	0.42877	0.341
20190408	1.555E+09	10	0.19871	0.34591	0.42937	0.34106
20190408	1.555E+09	11	0.19895	0.34648	0.43031	0.3423
20190408	1.555E+09	12	0.19858	0.3472	0.42927	0.3411
20190408	1.555E+09	13	0.19899	0.34623	0.42995	0.34237
20190408	1.555E+09	14	0.1989	0.34697	0.42945	0.3419
20190408	1.555E+09	15	0.19772	0.34674	0.42972	0.34173
20190408	1.555E+09	16	0.19891	0.34617	0.42899	0.34147
20190408	1.555E+09	17	0.19856	0.34704	0.42907	0.34205
20190408	1.555E+09	18	0.19837	0.34699	0.43021	0.34194
20190408	1.555E+09	19	0.19816	0.34687	0.43016	0.34213
20190408	1.555E+09	20	0.19879	0.34617	0.43004	0.34099





3. Preliminary first-order analysis using MLR



- MLR is only valid when the input variables are uncorrelated – sensor data may be correlated.
- Relationship of P to other ions could be non-linear.
- Relationship of P to other ions may be a function of time in the reactor cycle.

$$[P] = B0 + B1 \times ISE_K + B2 \times ISE_{Mg} + E$$

Discussion and Conclusions

- AA sheets have different number of entries. Text processor has been built to extract useful AA data automatically.
- In order to match timestamps in sensor data to sampling data from 10 cycles, loops have been used to deliver this task.
- MLR is a baseline to study the relationship between P concentration and sensor data but it is not robust to capture the non-linear dynamic, in the future advanced models need to be used.

Acknowledgements: Department of Civil and Environmental Engineering at Northeastern University, Boston, MA
Department of Civil and Environmental Engineering at University of Washington, Seattle WA
This work was supported by the National Science Foundation Grant #1559894: NSF REU Data Driven Discovery (D3).