Developing Data Cleaning and Familiarization Tools for StreamPULSE Users

Undergraduates:
Jin Cho, Yuval Medina, Vivek Sahukar

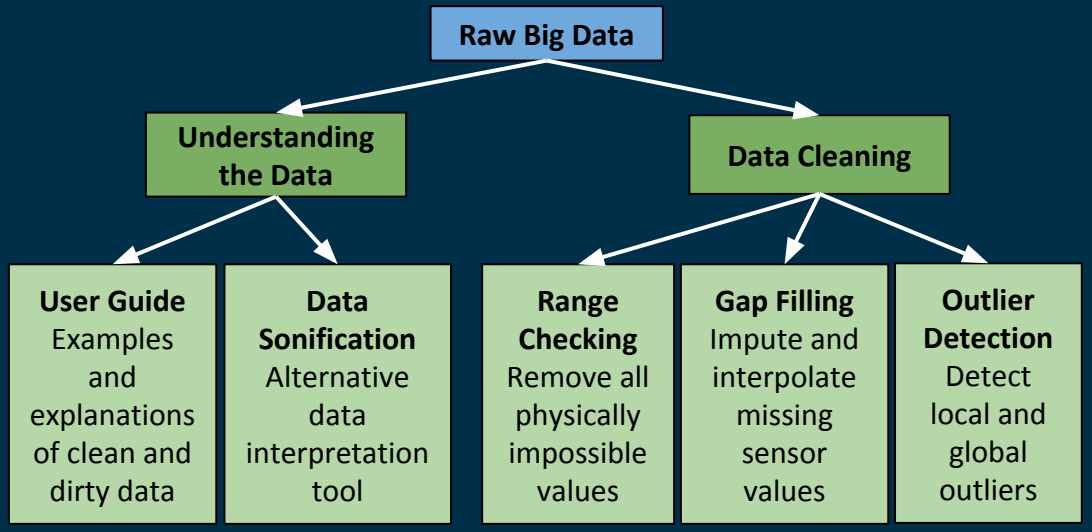
Project Managers:
Alice Carter, Mike Vlah

Faculty Leads:
Dr. Emily Bernhardt, Dr. Jim Heffernan

OVERVIEW

At StreamPULSE, where stream metabolism is modeled using oxygen, using a clean and complete dissolved oxygen curve is essential for accurate models. However, raw sensor data is often incomplete or dirty and cannot be used to accurately model metabolism. Additionally, the interpretation and cleaning of large data sets worldwide can be very difficult and creates the risk of misuse and misinterpretation.

We created tools to help users understand their data and metabolism modeling, and to facilitate data cleaning to prepare for modeling. The tools developed for detecting outliers still have to be synthesized to optimize the accuracy of detection. Finally all of these tools will be incorporated onto a user-friendly pipeline.

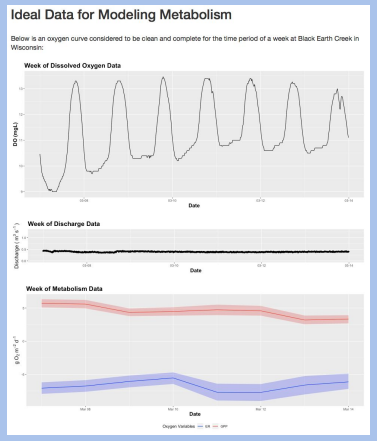


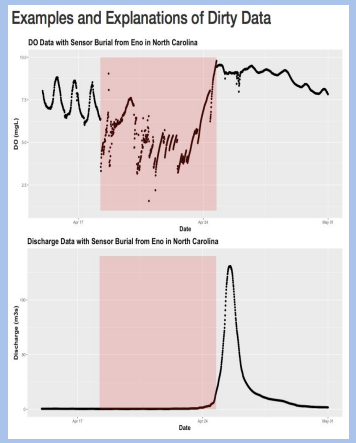
Data Familiarization

Stream Metabolism User Guide

Not all users of *StreamPULSE* are familiar with the complexities of dealing with raw stream sensor data and modeling stream metabolism.

We created a ***guide*** to cover the significance and complexities of modeling metabolism and obtaining clean data to create those models. With examples and explanations, the guide shows users what **dirty data** they may encounter and the **clean data** they should look for. This will be accessible on the *StreamPULSE* portal where users go to upload and clean their data.



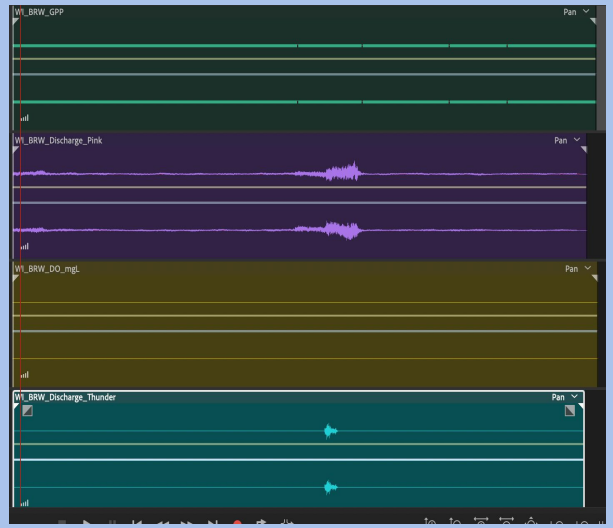


Data Sonification

The data analysis river biologists and technicians have to do is often time-intensive and repetitive. We created ***data sonification*** tool that uses audio and music to facilitate a more intuitive, more enjoyable, and less time-consuming way of perceptualizing data.

Being able to *hear* what river systems sound like also has great potential in expanding the **outreach** of the StreamPulse project.

The extremely malleable sound pipeline, created using SuperCollider and Python, is made to be continuously changed based on the input of river biologists and the *StreamPULSE* community. Intuition, after all, is highly objective.

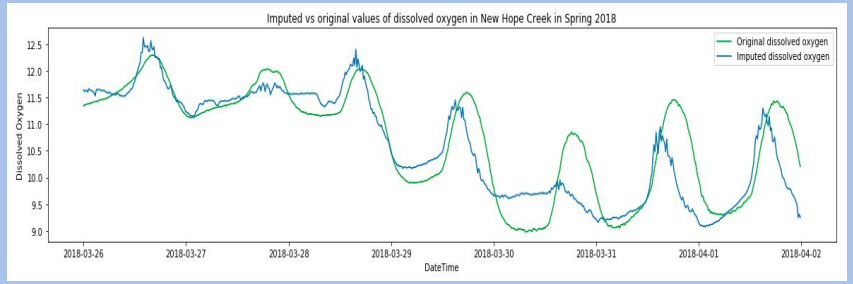


Data Cleaning

Gap Filling:

Simple Imputation Methods

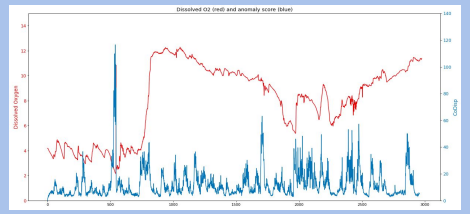
We used variables highly correlated with dissolved oxygen to fill missing values in the raw dissolved oxygen data. Our simple imputation methods proved more accurate than recurrent neural networks for gap filling, even with periodic trends in the time series.



Outlier Detection:

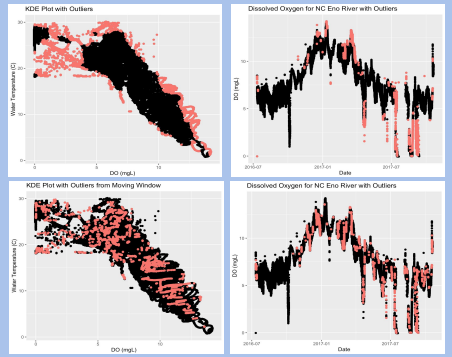
Robust Random Cut Forest (RRCF)

Each data point (in red) was assigned an anomaly score (in blue) with significantly higher scores indicative of outliers. We used this method to detect both local and global outliers with a threshold determining which scores to flag as outliers.



Kernel Density Estimation (KDE)

KDE plots were developed between DO and water temperature to flag outliers identified as points lying outside the 95% contour line. This was applied to both the whole data set (top row) and a moving window of 2,300 points (bottom row) and plotted along the DO time series data.



Moving Standard Deviation Window

We flagged points lying outside a standard deviation range of 3.1 within a sliding window of 900 points as outliers.