Statistical analyses of a closed-circuit reverse osmosis system (E2.9)

Energy and resource recovery

Re-Inventing the Nation's Urban Water Infrastructure (ReNUWIt)

Jacob Naranjo[1], Kate Newhart[1], Tzahi Cath[1]
[1]Colorado School of Mines

Background

- In partnership with Re-Inventing the Nation's Urban Water Infrastructure (ReNUWIt), Colorado School of Mines (CSM) has a closed-circuit desalination (CCD) system that continuously operates an reverse osmosis (RO) system in CCD mode for the effluent of the sequencing batch membrane bioreactor (SB-MBR) testbed located at Mines Park on campus. CCD is a novel configuration and operation of RO systems and can achieve high water recovery at lower energy demand and lower membrane fouling. Unlike traditional desalination systems that require more membranes to achieve greater water recovery, CCD is only limited by time running the cycles. This approach achieves greater more water recovery in less time.

- In CCD system, a variety of sensors are used to monitor the process. These sensors measure values such as:
  - volume
  - temperature
  - conductivity
  - pressure
  - flow rate
  - salt concentration
  - flux
  - power

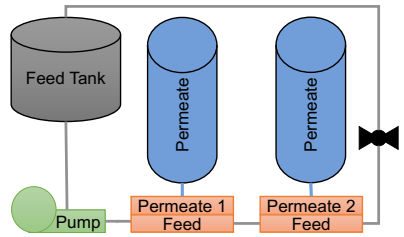

Figure 1 . Flow schematic of the CCD bench-scale system

- To keep the system operating smoothly good process control is needed. Monitoring the system with various sensors is needed to maintain an efficient system.

- In order to better understand the relationships between variables in membrane systems, different data analysis methods were compared in order to predict variables that are otherwise too difficult to detect in real time.

Approach

Three methods were used to analyze the process data produced by a CCD system: pair plots with correlations, multiple linear regression lines, and principal component analysis (PCA).

- Pair plots with linear correlations

The purpose of pair plots is visualize relationships between two process variables. To quantify relationships, linear correlation functions were used. These values represent how closely the data is related. The closer the value is to one the more likely there is a relationship between them and vice versa.

- Multiple linear regression model

Multiple linear regression is a powerful data analysis that accounts for the relationships among multiple process variables to predict an output variable. The linear regression model built in this work is: Volume ~ Cond + Flow + Salinity. This equation relates the conductivity, the flow, and the total dissolved solids (i.e., salt concentration) to predict volume.

- Principal component analysis (PCA)

PCA identifies groups of process variables that account for some fraction of variation in a dataset. These "principal components" can model data in fewer dimensions and can still represent majority of the data.

Results

Table 1: Strength of linear correlations between process variables

	Vol	Cond	Flow	Salinity
Vol	1.000	0.003	0.104	0.289
Cond	0.003	1.000	0.030	0.954
Flow	0.104	0.030	1.000	0.061
Salinity	0.289	0.954	0.061	1.000

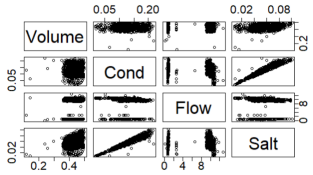

Figure 2: Process variables plotted pairwise to compare trends

Pair Plots with Correlations

In Figure 2, looking diagonally (from left to right, up to down) there is the name of the process variable. Looking to the left or right from the data set that data set's values will be the y-axis. Looking above or below the data set that data set's values will be the x-axis. If the data is randomly scattered across the range of values, there is a low correlation. Whereas, a data that looks in a straight line there is a higher correlation. A linear regression model can be made of each variable set and the strength of the correlation qualified. In Table 1, a 0 indicates no linear correlation whereas a 1 is a perfect linear correlation.

Multiple Linear Regression Lines

The use of multi linear regression lines provides more statistical information than visual. The first output shows a brief information on the data with the min, first quartile, median, third quartile, and Max.-Next the statistics show how effective the data is at estimating the predicted value. This is shown mostly in the Pvalue, the higher the value the better the association. Since we can see some of the Pvalues are very small they do not have much association with predicting the output value. Lastly, we see the R-squared value which, like the correlation values, the closer to one the better the correlation. For this specific model we see a R-squared value of 0.9047 which shows a reasonably good correlation, this can be because all the other values are relative to the amount of water that is in our system.

Table 2: Summary of multiple linear regression model

Min	1Q	Median	3Q	Max
-0.16	-0.0053	-0.00003	0.0051	0.098
	Est.	Std. Err	Tvalue	Pr(>|t|)
Intercept	4.3e-1	1.9e-3	232.5	<2e-16
Cond	-2.9	3.7e-2	-80.8	<2e-16
Flow	2.1e-5	1.3e-4	0.17	0.88
Salt	6.9	8.2e-2	84.7	<2e-16
Residual std err: 0.0124 on 764 dof				
Multiple R^2: 0.9047, Adjusted R^2: 0.9043				
F-stats: 2418 on 3 and 764 dof, p-val<2.2e-16				

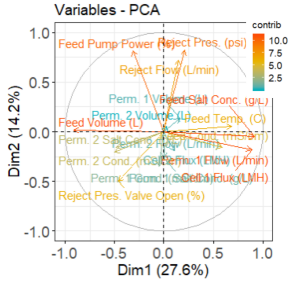

Figure 3: PCA Plot

Principal Component Analysis (PCA)

The first step of this analysis is breaking the wide range of data into principal components. The principal components are organized by their various and each one represents a portion of the data. The first two principal components, PC1 and PC2, have 27.69% and 14.19% of the data, respectively. Although this does not represent the entirety of the data it does represent a majority with 41.88%. The plot shows an arrow (vector) with a dataset associated with it. What we want to see which variables are the strongest along each component, this will suggest a relationship between the variables. For example, looking in the positive direction of the y-axis (PC2) we see the two longest vectors are Feed Pump Power and Reject Pressure. Logically it makes sense that the higher the pump power that there will be a higher reject pressure and vice versa. We can additionally make other relationships based on this plot and use logic or other data analysis to confirm it.

Conclusions

Data analysis can be a very tedious task but, in the long run it is incredibly useful and necessary for understanding data. Especially in a system like the CCD RO pilot where there is a wide range of data it is very important to be able to organize it in a way that can be viewed and understood. In using these three methods, we can view correlations between different data sets. This can help us better understand how the overall system and how various parts of the system can affect others. With each of method there were pros and cons to using them, as seen in Table 3.

Table 3: Review of linear data analysis methods used for the CCD system

	Pair Plots with Correlations	Multiple Linear Regression	Principal Component Analysis
Pros	- Simple to view and use. - Quantifies correlation	- Can view multiple data sets.	- View wide range of data values. - Easy visuals.
Cons	- Only can only compare two data sets with one another. - Only shows linear correlations	- Only used for linear correlations. - Difficult to set up effectively.	- Only suggests correlations, need other methods to quantify them.

Next Steps

- In further development of this research, these data analysis methods can be used on the current system and can be used to make decisions on the systems overall status or "health."

- Ideally this data analysis will compare all data collected by all sensors throughout the various locations within the system, but as discussed before the analysis becomes increasingly more difficult and time consuming the more values added. However, the results will tell us more about the entire system the more data set we can use in our analysis.

- Once we can make relationships between the sensors we can try and use this knowledge to make predictions on future reading for the system. This will ideally let the system run smoothly with less failures.

- Since the system on campus is ever changing due to different methods and technologies the data analysis ran should be as well. This means the analysis should be able to take on more and different data values easily

- There are an abundance of data analysis methods that are available to use and some may be more beneficial and provide more insight into our system. Being able to use more analysis methods will only strengthen the operation of this system.

Acknowledgements

RRS program, ReNUWIt, industry-advisory board member, Aqua-Aerobics, Inc., and the Colorado School of Mines for supporting this research.

Research Scholar Contact Information Jacob Naranjo | jacobnaranjo@mines.edu

Supported by the National Science Foundation under EEC-1028968