Prediction of Pressures as a Result of Air & Liquid Flow

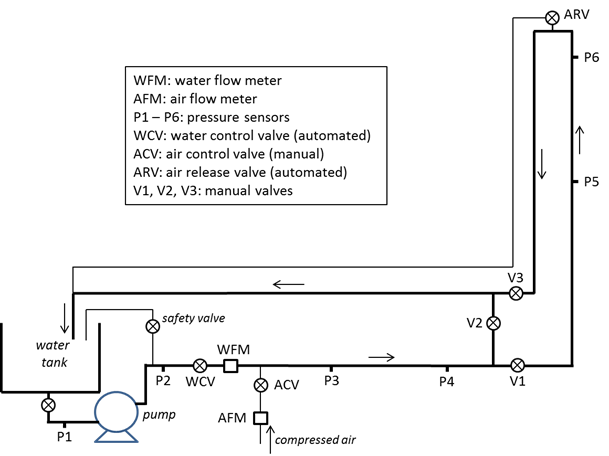
By Group 2 (Lisa Reisenauer, Triton Wolfe, Joshua Randrup, Zhenzi Yu, Yuhan Yang)

## PROJECT BACKGROUND

In a production environment, a significant amount of resources are invested in being able to anticipate phenomena, such as reliability events and process variability, before they occur. Examples would be trying to predict fouling factors and the differential pressures of a system based on process flow rates. In some processes, like the Methacrylic Acid production process patented by Rohm and Haas, where injected air causes two-phase flow to be present, this can be particularly difficult.

## PROJECT OBJECTIVE

A unit operations set-up on Georgia Tech’s campus injects air into a liquid water stream and measures downstream pressures. The schematic diagram is shown as Fig. 1. This project aims to use data from this set-up to create a regression model that can correlate the downstream pressures to the pressures in upstream and the air/water flow rates. Namely, the single target variable is P6 (Fig. 1), features are P1-P5, air/water flow rates.



**Figure 1**. Schematic diagram of two-phase flow

## GENERAL STRATEGY

After data exploratory analysis and cleaning, algorithms that have low complexity will be considered first to build a baseline model. The baseline model can later be optimized by utilizing time-series analysis and modeling techniques. Principle component analysis and feature engineering will also be implemented if necessary, to further improve the model.

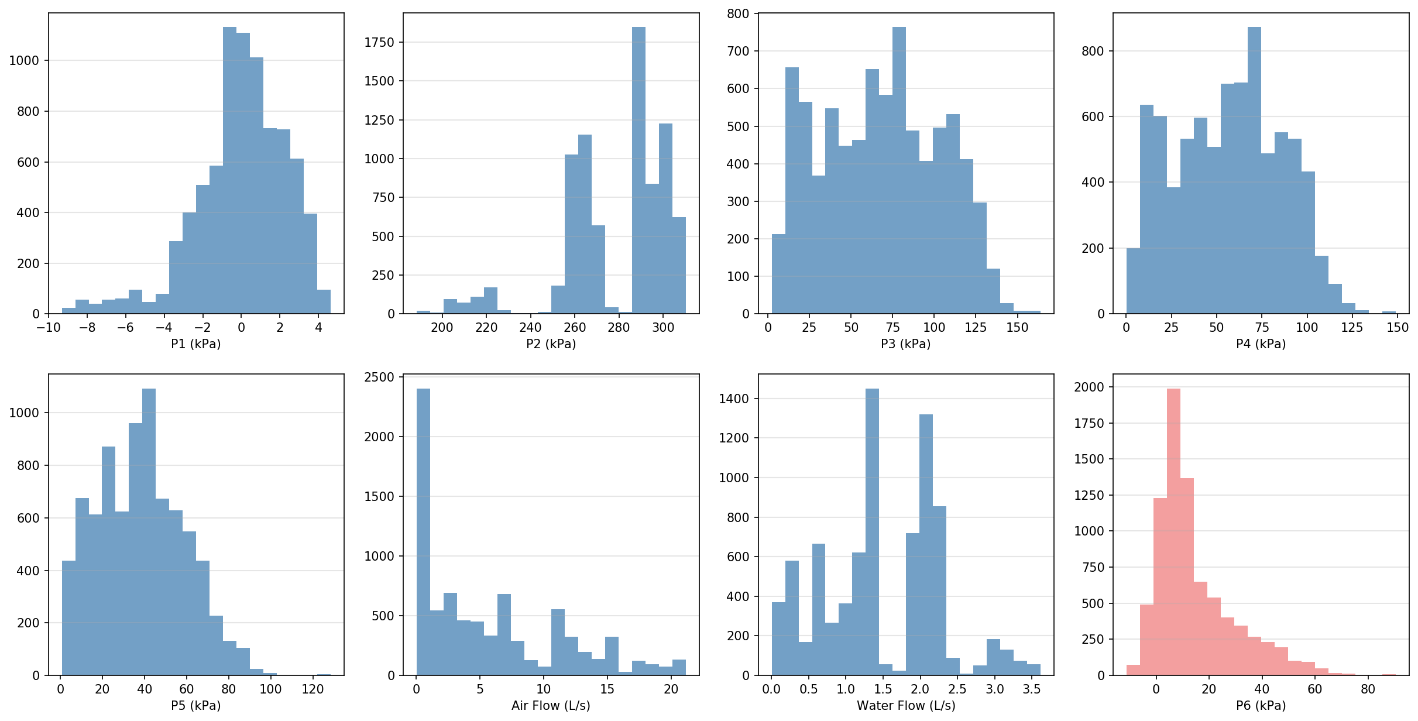
## PROJECT RISKS & CONTINGENCY PLANS

The main risk our project currently faces is potentially insufficient data to reach appreciable generalizability with our model. This would likely arise from lacking variation in certain input conditions such as valve position that would change the amount of fouling present in the system. If such a case did arise and the data was found to have insufficient variations from our desired inputs, we would have to obtain additional data ourselves. Presuming we can schedule a time around those in charge of the two-phase flow mechanism of Georgia Tech’s unit operations lab, obtaining additional data should not be a significant setback.

## DATA EXPLORATORY ANALYSIS

## Feature Distribution

The dataset is provided by the Georgia Tech ChBE Unit Control Lab, which is obtained during the two-phase flow correlation experiment. It consists of 8 parameters: air flow rate, water flow rate and 6 pressures along the pipeline. We will use the pressure-6 as our target variable, while the rest 7 features are the inputs. We have 8057 data points available, which is a good sample for this system, considering the number of features.

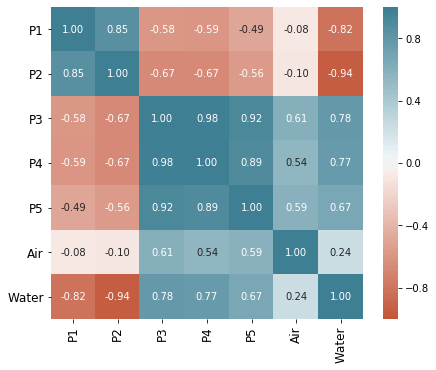


**Figure 2**. Distribution of the variables. Features are colored by bule, target variable colored by coral.

As shown in Fig. 2, after data visualization we find that there are mainly three types of distribution: near normal, random, and damped which indicating a pulsing flow. Moreover, the orders of magnitude of these features vary a lot. Thus we will rescale the features before further processing.

## Correlation Matrix

To further study the relationships between the features, we calculated the correlation matrix of this system. As shown in Fig. 3, P3, P4 and P5 are highly correlated (>0.9) while P2 and water flow rate are anti-correlated (<-0.9). These observations are sensible according to their positions in the pipeline diagram. We could further reduce the feature space’s dimension based on the correlation matrix, however the total number of the features is not large and computational workload is not heavy in this system, we choose to keep all these features to maximum the prediction accuracy.



**Figure 3**. Feature correlation matrix.

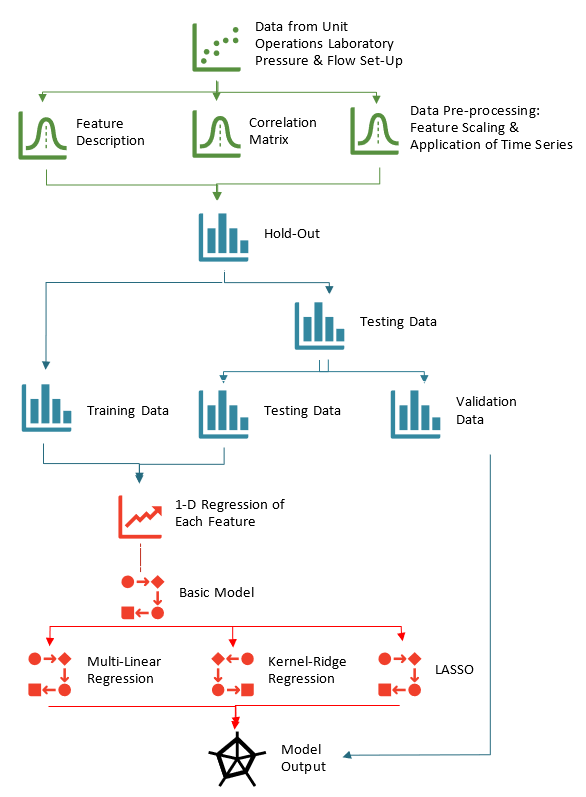
## Single component regression

We also did a 1-D regression to study the correlations between the target variable (P6) and each single feature using a piecewise function and a rbf kernel function (Tab.1). Since the non-parametric models are not good at extrapolation, it is not quite surprising the piecewise linear model gives good R-score on training set, but not for test set. RBF kernel model need further hyperparameter tuning to give better performance. We were expecting high correlation between P6 and P5 since these two sensors are close and in the same pipe section, however the R-score of P5 in both models are not outstanding among other features. This observation agrees with the conclusion in the correlation matrix: P3, P4 and P5 are highly correlated hence have similar performance in the single component regression.

**Table 1.** R-scores of the single component regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Piecewise | | RBF kernel | |
|  | **training** | **test** | **training** | **test** |
| P1 | 0.933 | 0.080 | 0.168 | 0.158 |
| P2 | 0.993 | 0.624 | 0.497 | 0.441 |
| P3 | 0.997 | 0.495 | 0.504 | 0.516 |
| P4 | 0.998 | 0.649 | 0.511 | 0.519 |
| P5 | 0.997 | 0.529 | 0.535 | 0.535 |
| Air | 0.922 | 0.603 | 0.485 | 0.433 |
| Water | 0.832 | 0.516 | 0.395 | 0.364 |

## WORKFLOW

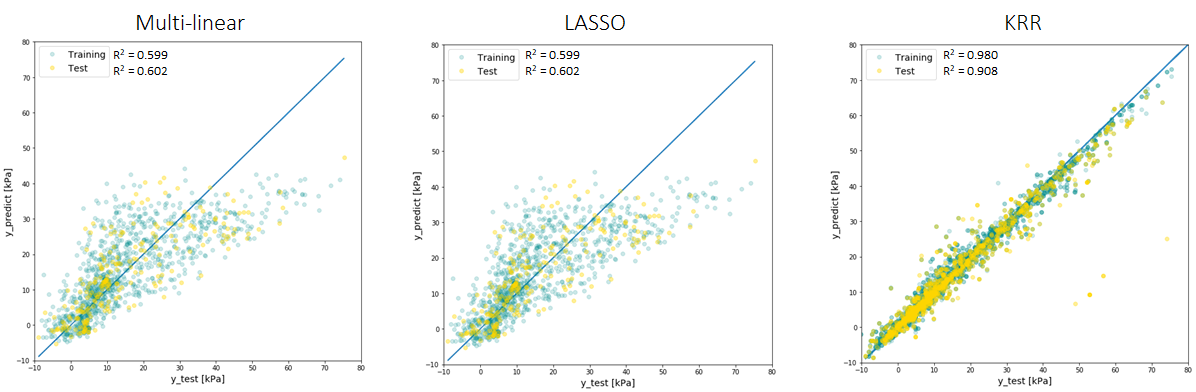


**Figure 4**. Workflow of the project

## BASELINE MODEL

The baseline model pipeline is shown as Figure 4. As a starting point for analyzing our data and determination of likely favorable models, a correlation matrix was created for the various features of our data. Variables P3, P4 and P5 show a relatively linear relationship; the 1-D regression analysis indicates a single feature is not enough to give accurate prediction of our target variable (P6).

Linear, non-linear, non-parametric models and time-series models were tried out at this exploratory stage. Specifically, the models implemented were multi-linear regression, Kernel Ridge Regression, Lasso, and time-series. As shown in Fig. 5, the multi-linear model got comparable r2 score for training set and test set. But the model tends to underestimate the P6 in the high-pressure region. The KRR model was relatively successful, as it was able to predict the test set quite well based on the r2 scoring metric. One downside of the model was that it required significant tuning of the hyperparameters, potentially signifying limited applicability outside the range of training data. The Lasso model showed surprisingly poor performance, given how well the KRR model performed. The r2 score of its prediction was relatively constant around 60%, showing little variation with hyperparameters tended towards zero until producing errors. Given that the data is a time-series dataset, we also calculated autocorrelation using the time-series model and we found that 2 prior points having significant partial autocorrelations, but no long-term seasonal variations were visible. This make senses because we expect the system reaches equilibrium relatively quick. Also we see some scatter data points prior to 5 steps before also have high partial correlation, this probably due to the experiment design flaw. This is just a sample test for the incorporation of time-series theory while more details will be potentially included in the future work. Moreover, as a baseline model the KRR model’s success definitely bears continued attention and optimization as we progress.

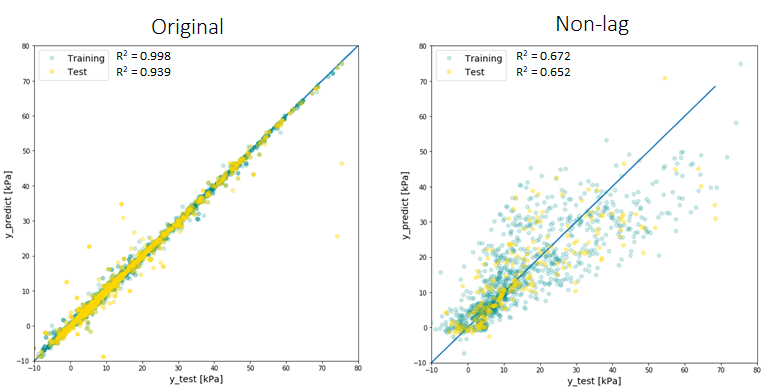


**Figure 5**. Comparison of different models

## MODEL IMPROVEMENT

## Hyperparameter tuning

As a winning model in last section, the KRR model was chosen to be improved by tuning the hyperparameters. Wide ranges of hyperparameters and a secondary GridSearch were employed in the tuning step. The final optimal hyperparameters are: alpha = 0.01 and sigma = 0.35, which increase the R score of the training set to 0.998 and R score of the test set to 0.939.



**Figure 6**. Comparison of original data and non-lag data

## Non-lag data

As can be seen in Fig.6, we implemented a modified dataset named no-lag data. This data is created to solve the problems we observe in our original data, namely 1. there exists lag between pressure change and flow change during the experiment for some data points and 2. There exists lots of replicated data points.

Note that instead of being based on data analysis theory, we identified this issue based on chemical engineering knowledge. The data collected with lag is different than that collected at steady state; however, including replicated data points has no contribution towards a meaningful model. We dealt with this problem by removing any duplicate data points, opposed to factoring in a lag. After doing this, the number of data points decreased to roughly 1500.

From the regression result, we find that the R score gets worse in KRR model. This may be mainly due to the deletion of replicated data points that were being weighted more heavily and consequently fit more accurately. Moving forward, we will try to come up with a better way to deal with the lag data issue.

## BIBLIOGRAPHY

Curtis Ingstad Carlson, J. H., Michael Stanley DeCourcy, H. T., & Jamie Jerrick John Juliette, H.

T. (2007, August 07). United States Patent No. US 7.253,307 B1.