Data Exhibition: Modeling Turbulence

Mihir Dutta, Michael Nicholson, and Judin Thomas

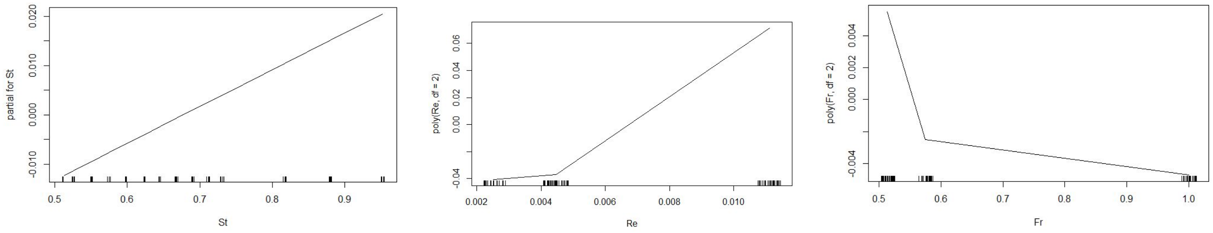
Introduction:

Across virtually all industries, turbulence plays a role in the real-world applications of science and engineering. We feel it on a flight to New York, see it in the distribution of our coffee creamer and marvel at its ability to make a snow globe a bit more magical. However, despite its importance, modeling it at any usable scale has proven largely unattainable. Our objective in this analysis is to provide a machine-learning based solution to this problem. Specifically, we will be attempting to use Reynolds number, Froud number and Strokes number to predict the spatial distribution and clustering of particles. More plainly, we seek to use the quantification of turbulence, gravity, and particle size to predict the clustering of particles in a closed system, for example, the clustering of water droplets in a cloud. While this clustering can be directly simulated, the process requires significant computational power and processing time. Our model will allow much quicker and computationally manageable estimates to be produced.

To achieve this goal, we will explore two model types: trees and general additive models. We believe trees will allow us to capture the strong interactions of the inputs and GAMs will give us a good mix of interpretability and predictive accuracy. We limit our focus to two model types to limit familywise error from fitting too many models. From here, we will use a mix of cross-validation, ANOVA and intuition of the problem at hand to choose our final predictive model. Below, we will discuss our conclusions from each model before ultimately choosing the model we believe is best suited for the problem at hand. In order to simplify the analysis, we use the inputs to predict the first four moments of the distribution of clustering in place of the underlying probability distribution. This significantly reduces the complexity and computation required to fit the model while sacrificing little of the practical use of the output.

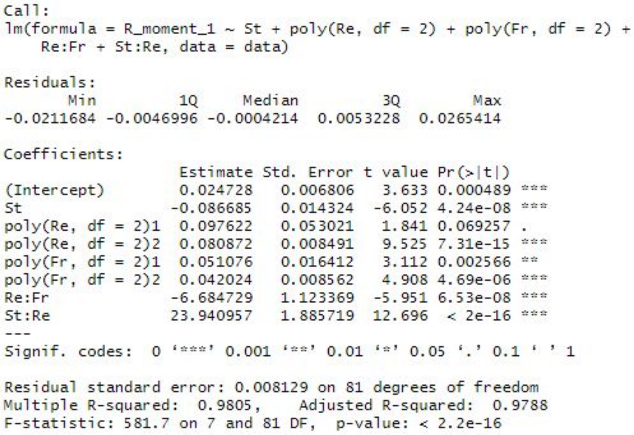
Methodology:

For the GAM model, we began with a simple multiple regression model and added complexity by adding polynomial degrees and interactions in a stepwise fashion, ultimately using ANOVA for model selection. However, before fitting the model we transformed the data to eliminate infinite values and to standardize our inputs. We performed logistic transforms on Fr and St and inverted Re. This resulted in all variables ranging between 1 and 0 with non-zero variation. We elected to invert Re in place of the logistic transform to ensure all the values were not clustered closely around one. We first ran five models of increasing complexity to determine the optimal degree of each input. We did this separately for each of the four moments we wished to predict and used ANOVA analysis for model selection. Once we found the optimal degree, we iteratively added interactions to determine how many to include in the model, again using ANOVA for model selection. We only performed this process twice to avoid overfitting by fitting many models.



GAM plots for St, Re, and Fr for the non-interaction model of the first moment

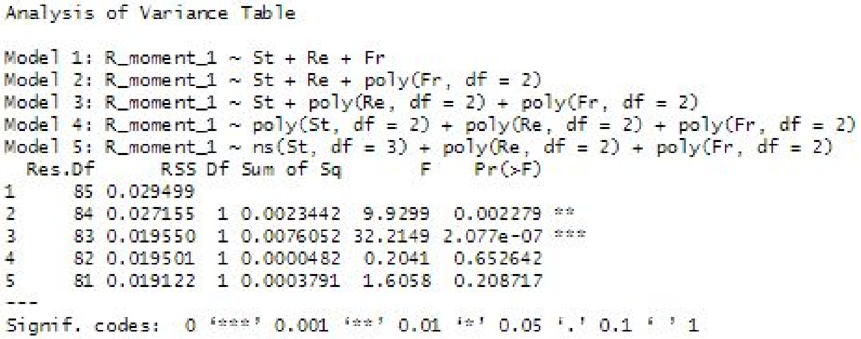
For the first moment, we found the optimal degree to be 1 for St and 2 for Fr and Re. Note, since we transformed the data before fitting, a degree of 1 does not imply a linear relationship. However, interactions within gravity, particle size and turbulence were also certainly a factor in the formation of clusters. The model reflected this fact with the ANOVA showing interactions between Re and Fr as well as St and Re to be significant. For the other moments, we found the optimal degree to be 1 for St and Re and 2 for Fr. For the model with interactions, the second-moment model was significant with interactions between Re and Fr as well as St and Re. For the third and fourth moments, all pairwise interactions were significant at around .1 or lower. Scientifically, the negative beta of Re (which was inverted so that a positive beta implies a negative effect on cluster volume) and Fr (which was logistically transformed) implies that as gravity increases, the negative effect of Re on clustering decreases. For example with rain, this implies that with higher turbulence the droplets are smaller (less clustered) because the turbulence works to break them apart. However, as gravity increases the downward pull partially counteracts this effect and results in larger droplets (more clustered) falling to earth. For St and Re, we find the opposite effect. As the Stokes number increases, the negative effect of turbulence on clustering increases. Scientifically, the larger the particles the greater the effect of turbulence in a system.



```
Call:
lm(formula = R_moment_1 ~ St + poly(Re, df = 2) + poly(Fr, df = 2) +
    Re:Fr + St:Re, data = data)

Residuals:
      Min         1Q     Median         3Q        Max
-0.0211684 -0.0046996 -0.0004214  0.0053228  0.0265414

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.024728   0.006806   3.633 0.000489 ***
St                -0.086685   0.014324  -6.052 4.24e-08 ***
poly(Re, df = 2)1  0.097622   0.053021   1.841 0.069257 .
poly(Re, df = 2)2  0.080872   0.008491   9.525 7.31e-15 ***
poly(Fr, df = 2)1  0.051076   0.016412   3.112 0.002566 **
poly(Fr, df = 2)2  0.042024   0.008562   4.908 4.69e-06 ***
Re:Fr             -6.684729   1.123369  -5.951 6.53e-08 ***
St:Re             23.940957   1.885719  12.696  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008129 on 81 degrees of freedom
Multiple R-squared:  0.9805,    Adjusted R-squared:  0.9788
F-statistic: 581.7 on 7 and 81 DF,  p-value: < 2.2e-16
```

GAM model output for first moment

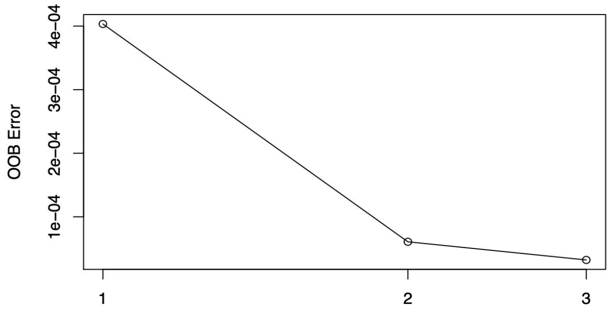
Overall, we believe the GAM framework works well with the data. It estimates intuitive relationships between the three inputs and the first four moments when using the polynomial fits while also allowing us to achieve high predictive accuracy by adding interaction effects. Overfitting, however, is still an issue since around 10 models are fitted and tested for each output. Using the ANOVA framework reduces this potential issue, but fitting multiple complex models still reduces our certainty in the true accuracy in our models, since it increases the odds at least one fit performs well by chance.



```
Analysis of Variance Table

Model 1: R_moment_1 ~ St + Re + Fr
Model 2: R_moment_1 ~ St + Re + poly(Fr, df = 2)
Model 3: R_moment_1 ~ St + poly(Re, df = 2) + poly(Fr, df = 2)
Model 4: R_moment_1 ~ poly(St, df = 2) + poly(Re, df = 2) + poly(Fr, df = 2)
Model 5: R_moment_1 ~ ns(St, df = 3) + poly(Re, df = 2) + poly(Fr, df = 2)
  Res.Df      RSS Df Sum of Sq       F    Pr(>F)
1     85 0.029499
2     84 0.027155  1 0.0023442  9.9299  0.002279 **
3     83 0.019550  1 0.0076052 32.2149 2.077e-07 ***
4     82 0.019501  1 0.0000482  0.2041  0.652642
5     81 0.019122  1 0.0003791  1.6058  0.208717
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Table for first-moment GAM

We also created a bagging model to analyze the importance of the three predictors for calculating the moments. We used the tuneRF function to identify the optimal parameter value which minimizes the out-of-box error for our tree model. We found the optimal value for the “mtry” parameter is three which is equal to the number of predictors so our model is actually a bagging model. Then we created four bagging models to predict the first, second, third, and fourth moments.



tuneRF output for first moment model

In order to determine which variables are the most important, we decided to analyze feature importance within our bagging model. The feature importance analyzes out-of-bag MSE to determine how important each of our predictor variables is in the predictions by the bagging model. What we find is that for predicting the first moment Re is much more important than the rest of the variables. Re also seems to be a very important predictor as it has an IncMSE% of 93.9%. For predicting the second moment it seems that Re and Fr are more important than St and Re and Fr seem to be nearly equally important. For predicting the third moment Re and Fr also seem to be the important prediction while St is not as important. For the fourth moment, it seems that Re and Fr are near-equally important(Re-48.1% and Fr-48.5%) while St is not as important.

This suggests that the Reynolds number is the main predictor for the first moment which suggests turbulence is highly related to the intensity of the turbulence. For the other three moments, the IncMSE% drops to around 40%-50% for the important predictors and Reynolds number and Froude constant seem to be important predictors for the second, third, and fourth moments. This suggests that it is harder to predict the second, third, and fourth moments compared to the first moment. Since the Reynolds number and Froude constant are the important predictors for the second, third, and fourth moments, this suggests that these moments depend on the "turbulence" and "gravity" of the fluid. Scientifically, this result also implies that turbulence alone can be predictive on the mean volume of clusters, but for higher moments like the variance of these clusters, other factors like gravity and particle size become increasingly impactful.

Results:

After fitting our Bagging and GAM models, we split our data into test and training sets to compare test MSE. We set the seed to 1234 in order to ensure the 'random' split was the same for both models. We found our Bagging model had significantly lower MSE than our GAM on the holdout set and thus decided to use this model for our final predictions. We also fitted the Bagging model using both raw moments and the log of the moments as outputs. We found the hold out MSE was lowest for the raw moment model, as seen below. Thus, we used this model for our final predictions. These predictions can be found in a separate CSV file.

Moment	Bagging MSE - raw moments	GAM MSE - raw moments	Bagging MSE - log(moments)
1st	$1.165 * 10^{-5}$	$1.131 * 10^{-4}$	28.124
2nd	3,481	42,211	81,832
3rd	$3.785 * 10^{11}$	$3.233 * 10^{12}$	$5.412* 10^{13}$
4th	$2.534 * 10^{19}$	$2.52 * 10^{20}$	$2.55 * 10^{20}$

We believe the outperformance of the Bagging models potentially stems from three factors: potential overfitting of the GAM, strong interaction effects between the three variables and non-linearity in the relationship between the inputs and particle clustering. Despite the results of ANOVA, we fit many GAM models for each moment. It is possible the numerous and increasingly complex fits led to overfitting in the data. This is certainly problematic for prediction, but in the context of the problem at hand, it is even more worrying. When fitting our model, we have a small subset of the possible values for St, Re, and Fr since each datapoint must be directly simulated. Thus, in order to produce meaningful predictions, our model must be robust to both interpolation and extrapolation. A model that can only take in the three values we are given for Re and Fr, for example, is near meaningless in a real-world setting.

Regarding interactions, it is important to remember what we are modeling. The effect of turbulence, gravity and particle size are not independently related to clustering. Larger or smaller sized particles, for example, will be affected by an increase in turbulence differently. The three variables interact simultaneously in the ultimate clusters that are formed. The formation of waterdrops in clouds is an example of this. The turbulence caused by wind, the particle size of the droplets and the difference in gravity based on altitude are all factors in the ultimate volume of the droplets and whether they fall as rain. This scientific intuition is supported by our results. Within the GAM framework, the additive models performed much poorer than the models including interaction.

For trees, the various splits and beaches allow for a natural form of interaction to be captured. For example, the effect of Fr on the four moments when Re is 398 versus when it is 90 is different. However, while the tree structure of our bagging model works well for predicting the values of Reynolds, Froud and Strokes number we are given, it may not capture these interactions as well for new values of these inputs. When deciding where split in the trees, it does not have information for the range of values for the inputs we do not see. For example, with Re there is a large jump 90 to 224. A tree in our Bagging model may split at values greater than 90 when say 120 may be the better split. That is the model only truly captures interaction for values of the inputs it has seen. In practice the framework likely still holds, but it certainly adds uncertainty to the true accuracy of the model.

Concerning linearity, our GAM modeled performed better with lower degree polynomials on the transformed inputs than higher degrees but was significantly outperformed by the random forest. There is certainly complexity that is missed in the model, but it would seem this flexibility is not best found in higher-order models. Rather it is interactions and other nonlinear relationships that best capture the complexity in the underlying data. Though it is also possible, as mentioned previously, that the Bagging is simply good at mapping the sparse data points in the simulated training set and would underperform when new variables of Reynolds, Froud and Strokes number are used in a real-world application.

Conclusion:

Our goal with this analysis was to find a model to determine the probability distribution of clustered particles in a timely and computationally efficient manner using the intensity of turbulence, the size of the particles and the magnitude of gravity as inputs. We decided to fit a GAM to give interpretable estimates of the relationship between the inputs and the first four moments of the probability distribution and a Bagging model to better capture the interactions of the inputs. After fitting our Bagging and GAM models, we found the Bagging models performed best on the hold set. In addition, we believe the Bagging models better captured the interaction of the three inputs resulting in better predictive accuracy on unseen data points. For this reason, we decided to use our fitted Bagging model for our final predictions.

We believe our Bagging model performs well on the data and feel confident in our predictions on the test set. However, there are still some potential uncertainties in our predictions on real-world data. Since our end goal is to predict clustering in real-world settings, it is important that our model is robust to both interpolation and extrapolation. The need for extrapolation is complicated by the limited number of values we observe for the inputs. Our model works well on the values in the model, but it is impossible to precisely predict its behavior on entirely new data points (especially if new values are imputed for all three inputs). Also, it is important to note that our model works best for the first moment, but predictive accuracy falls for the lower moments. In practice though, the first moment is likely the most important for application making this result less problematic for real-world applications. Our Bagging feature importance analysis combined with our GAMs also give us some insight into the interpretation. From our feature importance analysis and our GAM models, we can conclude that the Reynolds number is the most important predictor for particle clustering, especially for the first moment.

Altogether, we believe the Bagging model captures the complexity of the three inputs well without significant overfitting. It is less interpretable than is ideal for scientific inference but this trade-off results in better predictive accuracy. Ultimately, we decided the loss of interpretability was worth the additional accuracy. Our model in no way ‘solves’ the issue of turbulence, but instead provides a predictive framework which reduces the computational needs of computing its effect on clustering and thus contributes to the field in this way.