Case Study Report

06 November, 2022

Introduction

The objective of this case study is to develop a predictive model to predict the distributions of particle clusters in turbulence from three predictors: fluid turbulence characterized by Reynold's number Re, gravitational acceleration Fr, and particle characteristics (size or density which is quantified by Stokes number St). We also want to understand how does each of the three parameters affect the distribution of particle cluster volumns.

Developing an understanding of turbulence is important because the effects of turbulence are present in a wide variety of problems. For example, the distribution of ash in the atmosphere is controlled by atmospheric turbulence, which has many implications for the environment and aviation. Turbulence also controls the population dynamics of plankton, which play an important role in the carbon cycle. On a more cosmological and atmospheric level, turbulence plays a central part in the dispersion of ash in the atmosphere which has many implications for the environment and aviation as well as the thermodynamics of clouds, radioactive properties, and the rate at which droplets grow to form rain.

We'll build a supervised machine learning model to give prediction on the complex physical phenomenon, and interpret the variables' relationship. We first conducted several necessary transformation on the variables. After trying several methods, we decided to use a linear model with interaction terms, and we performed 5-folds cross validation on the linear model to assure that our model has good predicting ability.

The model we have decided to use to predict turbulence is a linear regression with all three variables and interaction terms, shown below:

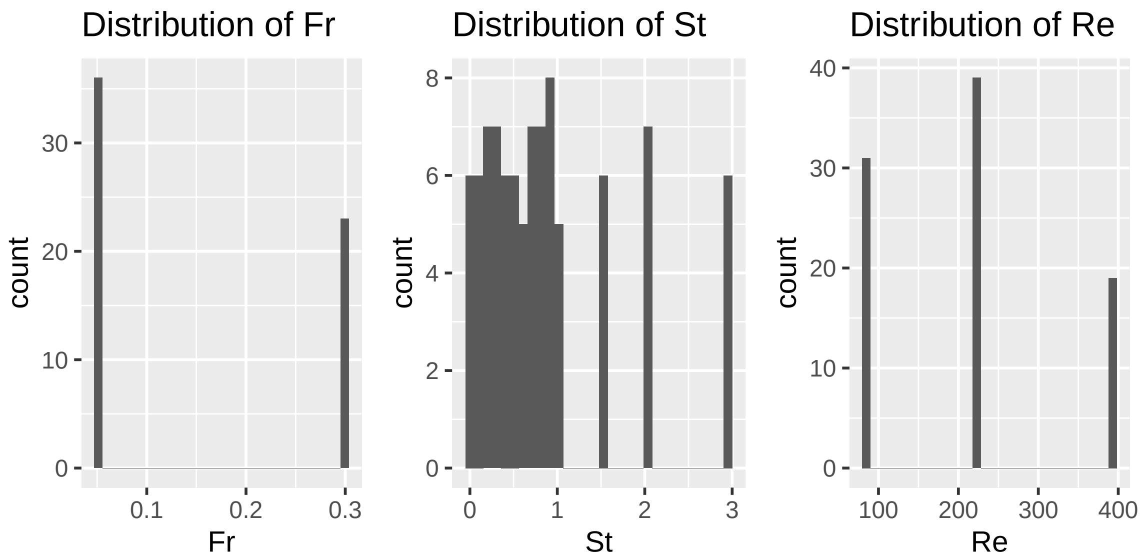
$$St + Re_{category} + Fr_{transformed} + Fr_{transformed} \times Re_{category} + St \times Re_{category}$$

where $Re_{category}$ is a categorical variable that classifies the training Re into three categories ($Re = 90$: Low, $Re = 224$: Medium, $Re = 398$: High) and $Fr_{transformed}$ is the inverse logit of Fr to handle infinity values present in the data. The model includes interactions between gravitational acceleration and fluid turbulence, and between fluid turbulence and particle characteristics.

Methodology

EDA

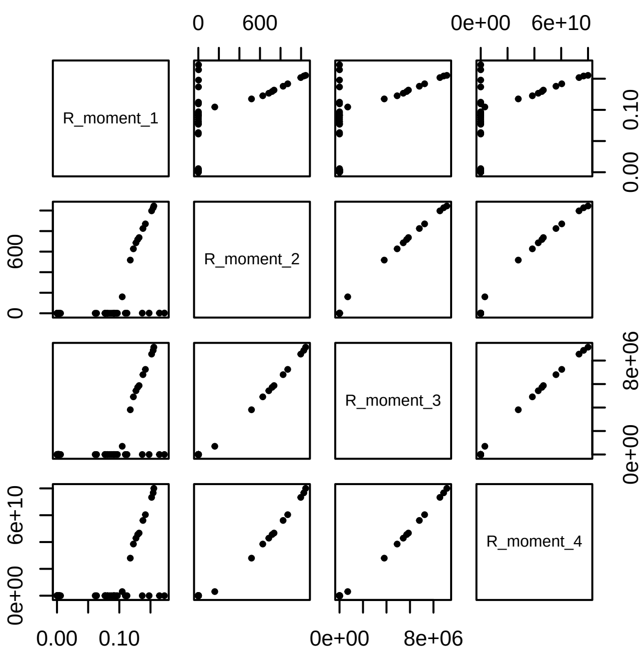
After loading the data, we performed exploratory data analysis on all three predictors and four moments.



From observing the data set and the histogram plots, we can see that there are some data transformation needed. First, Fr has Inf value, which can't be quantified, so Fr only has two values in the histogram. We have two options: 1.performing logit transformation on Fr to transform Inf to 1; 2.transforming Fr to categorical variables.

We can do this because the physicist only need to predict Fr on these three levels, so we don't need to consider extrapolation. Second, Re only has three levels. We also decided to convert Re to categorical variables, because of the same reasons as Fr.

We also found that the moments are highly linearly correlated:



It is worth-noticing that since we are trying to understand the probability distribution of the particles in the flows, we may want to look into the central moments instead of the raw moments here, because central moments give us a more meaningful interpretation of the probability distribution. However, since the 1st central moment is always 0, we need to predict 1st raw moment and other three central moments separately. Since `C_moment_2`, `C_moment_3`, and `C_moment_4` are highly linearly correlated, we decided to fit a model on `C_moment_2`, which will give us the relationship of the predictor variables on the other moments due to the high linear correlation between the moments.

Linear Model (Inference)

We started by building a preliminary linear model:

$2nd\ \widehat{Central}\ Moment = \hat{St} + \widehat{Re_{category}} + \widehat{Fr_{category}} + \epsilon$

We find that this basic model has very low Rsquared, so we decided to increase the model complexity by adding interaction terms.

Since the 1st central moment is always 0, we need to predict 1st raw moment directly. We fitted models on the same model to predict second, third and fourth moments due to the collinearity as explained above:

Here we perform a 5-fold cross validation on the model. We chose 5 folds over 10 because the limited data available.

Model	RMSE	Rsquared	MAE
R_M1	1.000000e-02	0.978	5.000000e-03
C_M2	1.110450e+02	0.882	4.657600e+01
C_M3	9.534512e+05	0.820	3.988096e+05
C_M4	7.653045e+09	0.808	3.356711e+09

Ridge Model

We also explored shrinkage methods such as ridge regression. We chose ridge because we don't need to perform However, since we know that the three predictors are all active so that we do not need predictor selection, so we attempted to fit a ridge regression model.

RidgeModel	RMSE
C_M2	1.972020e+02
C_M3	1.693927e+06
C_M4	1.420521e+10

We can see that the linear model performs better than the ridge model over all, so we 'll just use the linear model for prediction. Moreover, it is simpler to interpret the inference result with the linear model. So we can also use the linear model for interpreting the relationship between variables.

Log-Transformed Model (Final Model)

We decided to log transform the second, third and fourth moment response variables in order to restrict the predictions of values to positive only, since the second, third and fourth moments cannnot be negative

Then we fitted linear regression model on the original first moment response variable, and linear regression model on the log-transformed response variables for second, third and fourth moments.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.000	0.008	0.026	0.979
St	0.000	0.003	0.023	0.982
Re_categoryLow	0.122	0.010	12.272	0.000
Re_categoryMedium	0.001	0.009	0.159	0.874
Fr_transformed	0.000	0.009	0.013	0.989
Re_categoryLow:Fr_transformed	-0.052	0.012	-4.386	0.000
Re_categoryMedium:Fr_transformed	0.001	0.011	0.049	0.961

	Estimate	Std. Error	t value	Pr(>|t|)
St:Re_categoryLow	0.028	0.003	7.995	0.000
St:Re_categoryMedium	0.001	0.004	0.231	0.818

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.423	0.480	-11.301	0.000
St	0.295	0.387	0.763	0.448
Re_categoryLow	3.824	0.662	5.775	0.000
Re_categoryMedium	1.130	0.650	1.738	0.086
Fr_categoryLow	-0.229	0.580	-0.395	0.694
Fr_categoryMedium	-0.067	0.496	-0.135	0.893
Re_categoryLow:Fr_categoryLow	6.886	0.794	8.676	0.000
Re_categoryMedium:Fr_categoryLow	2.175	0.755	2.880	0.005
Re_categoryLow:Fr_categoryMedium	0.269	0.753	0.357	0.722
St:Re_categoryLow	0.684	0.465	1.469	0.146
St:Re_categoryMedium	0.651	0.472	1.377	0.172

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.474	0.778	-3.179	0.002
St	0.315	0.627	0.502	0.617
Re_categoryLow	2.874	1.074	2.677	0.009
Re_categoryMedium	0.573	1.054	0.543	0.589
Fr_categoryLow	-0.289	0.941	-0.307	0.759
Fr_categoryMedium	-0.077	0.804	-0.096	0.924
Re_categoryLow:Fr_categoryLow	13.060	1.287	10.146	0.000
Re_categoryMedium:Fr_categoryLow	4.303	1.225	3.514	0.001
Re_categoryLow:Fr_categoryMedium	0.307	1.222	0.251	0.802
St:Re_categoryLow	1.116	0.755	1.478	0.143
St:Re_categoryMedium	1.002	0.766	1.308	0.195

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.475	1.042	0.456	0.650
St	0.335	0.840	0.399	0.691
Re_categoryLow	2.102	1.437	1.463	0.148
Re_categoryMedium	0.096	1.411	0.068	0.946
Fr_categoryLow	-0.349	1.260	-0.277	0.782
Fr_categoryMedium	-0.097	1.076	-0.090	0.929
Re_categoryLow:Fr_categoryLow	19.153	1.723	11.117	0.000
Re_categoryMedium:Fr_categoryLow	6.421	1.639	3.917	0.000
Re_categoryLow:Fr_categoryMedium	0.348	1.635	0.213	0.832
St:Re_categoryLow	1.488	1.010	1.472	0.145
St:Re_categoryMedium	1.304	1.025	1.272	0.207

The final equations are:

$\hat{R\_moment\_1} = 0.122 \times Re\_categoryLow + 0.001 \times Re\_categoryMedium - 0.052 \times Re\_categoryLow * Fr\_transformed + 0.001 \times Re\_categoryMedium * Fr\_transformed + 0.028 \times St * Re\_categoryLow +$

$0.001 \times St * Re\_categoryMedium$

$\hat{C\_moment\_2} = \exp(-5.423 + 0.295 \times St + 3.824 \times Re\_categoryLow + 1.130 \times Re\_categoryMedium - 0.229 \times Fr\_categoryLow - 0.067 \times Fr\_categoryMedium + 6.886 \times Re\_categoryLow * Fr\_categoryLow + 2.175 \times Re\_categoryMedium * Fr\_categoryLow + 0.269 \times Re\_categoryLow * Fr\_categoryMedium + 0.684 \times St * Re\_categoryLow + 0.651 \times St * Re\_categoryMedium)$

$\hat{C\_moment\_3} = \exp(-5.423 + 0.295 \times St + 3.824 \times Re\_categoryLow + 1.130 \times Re\_categoryMedium - 0.229 \times Fr\_categoryLow - 0.067 \times Fr\_categoryMedium + 6.886 \times Re\_categoryLow * Fr\_categoryLow + 2.175 \times Re\_categoryMedium * Fr\_categoryLow + 0.269 \times Re\_categoryLow * Fr\_categoryMedium + 0.684 \times St * Re\_categoryLow + 0.651 \times St * Re\_categoryMedium)$

$\hat{C\_moment\_4} = \exp(-5.423 + 0.295 \times St + 3.824 \times Re\_categoryLow + 1.130 \times Re\_categoryMedium - 0.229 \times Fr\_categoryLow - 0.067 \times Fr\_categoryMedium + 6.886 \times Re\_categoryLow * Fr\_categoryLow + 2.175 \times Re\_categoryMedium * Fr\_categoryLow + 0.269 \times Re\_categoryLow * Fr\_categoryMedium + 0.684 \times St * Re\_categoryLow + 0.651 \times St * Re\_categoryMedium)$

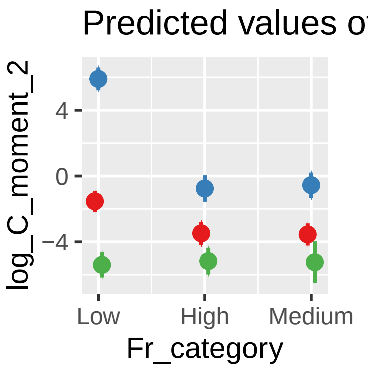
Results

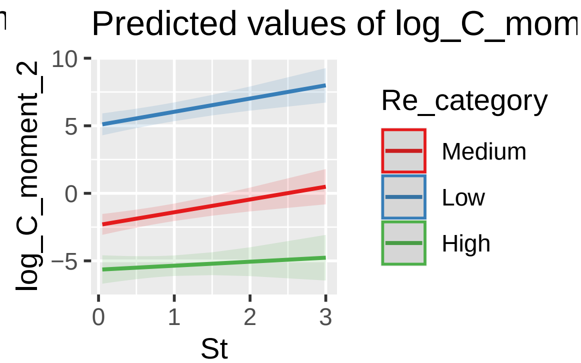
We made predictions on the hold-out set in data-test.csv, and generated a csv file containing the predictions for the first, second, third and fourth moments.

St	Re	Fr	Predicted_R_M1	Predicted_R_M2	Predicted_R_M3	Predicted_R_M4
0.05	398	0.052	1.00027	1.00410	1.075550e+00	2.431870e+00
0.20	398	0.052	1.00028	1.00428	1.079160e+00	2.504640e+00
0.70	398	0.052	1.00031	1.00494	1.092480e+00	2.775010e+00
1.00	398	0.052	1.00033	1.00538	1.101520e+00	2.960040e+00
0.10	398	Inf	1.00033	1.00521	1.101550e+00	3.039210e+00
0.60	398	Inf	1.00037	1.00600	1.118630e+00	3.405890e+00
1.00	398	Inf	1.00039	1.00671	1.134340e+00	3.746200e+00
1.50	398	Inf	1.00043	1.00772	1.156930e+00	4.240100e+00
3.00	398	Inf	1.00053	1.01175	1.250160e+00	6.322170e+00
3.00	224	0.300	1.00477	1.22761	8.848600e+00	2.505945e+02
0.10	224	Inf	1.00247	1.01996	1.222890e+00	3.869470e+00
0.50	224	Inf	1.00283	1.02758	1.362800e+00	6.318300e+00
0.40	90	0.052	1.11229	233.65857	9.305267e+05	4.005530e+09
1.00	90	0.052	1.13117	419.49112	2.194687e+06	1.195146e+10
0.05	90	0.300	1.09789	1.46518	4.196160e+00	3.073924e+01
0.30	90	0.300	1.10562	1.55425	5.336400e+00	4.593703e+01
0.60	90	0.300	1.11496	1.68830	7.304840e+00	7.515832e+01
0.80	90	0.300	1.12123	1.79861	9.127740e+00	1.048709e+02
0.40	90	Inf	1.08429	1.47459	4.890820e+00	4.224337e+01
0.50	90	Inf	1.08733	1.51195	5.411150e+00	4.974356e+01
0.60	90	Inf	1.09039	1.55251	6.005000e+00	5.864196e+01
1.50	90	Inf	1.11826	2.12802	1.709332e+01	2.677197e+02
2.00	90	Inf	1.13405	2.71778	3.239994e+01	6.347067e+02

Since the 1st central moment is always 0, we did not build a model for it. Instead, we directly predicts the 1st raw moment. There is a distinction between the three parameters' effects on mean and other three moments. When predicting the 1st raw moment, we did not transform Fr into categorical variables, and the interaction between Re and Fr has a significant negative, though weak, effects on the value of the mean.

The effects of three parameters are similar over other three central moments. So we'll take the 2nd central moment (variance) as a representative example. Some major observations from the results: First, Re is expected to have a negative relationship with the variance. The lower the Re, the larger the 2nd central moment. Second, St is expected to have a positive relationship with the 2nd central moment. Third, while Fr has a negative relationship with the variance, such negative effect is small when Fr number is high, and lower Fr number has stronger negative effects on the variance. It is worth-noticing that Fr does not have a significant main effect on the variance, given its high p-value. However, Fr's effects become significant in the interaction terms. 4.The interaction terms between Re and Fr has very strong positive effects on the 2nd central moment.





Specifically, based on our modeling results, the two most significant terms are Re and the interaction between Re and Fr. Turbulence with Low Re is expected to have 3.82 unit higher variance than turbulence with High Re on average holding all else constant. The interaction between Low Re and Low Fr has strong positive effects on the variance. If the turbulence has low re and low fr, it is expected to have 13.06 unit higher variance than turbulence with Low Re and High Fr, holding all else constant. This result aligns with our prediction outcome–with 90 Re and 0.052 Fr, the distribution of particle cluster has incredibly high (419.49) variance.

We can now interpret the three parameters' effect in the physical context. Since Re (the Reynolds number) quantifies fluid turbulence, we can induce that the particle cluster volume distribution in turbulence has low uncertainty when Re is low. We can conclude that Laminar flows have low Re number, because the particle distribution is more orderly, regular, predictable. On the other hand, Turbulent flows have high Re number, because high Re is associated with high variance, thus the flows are more random and irregular.

St (the Stokes number) is the ratio of the particle's momentum response time to the flow-field time scale. By definition, a larger Stokes number represents a larger or heavier particle. Our results demonstrate that particles with high St have greater impact on the turbulence. For small St, the particles will mostly follow the fluid motion, thus more predictable; for high St, the carrier fluid will have very limited influence on the particle motion, thus more unpredictable. We can conclude that Turbulent flows have high St, and Laminar flows have low St.

Fr (the Froud number) is the ratio of average flow velocity to the wave velocity in shallow water. So high Fr means fast rapid flow, and low Fr means slow tranquil flow. In our result, Fr in general has a negative effects on the variance, but such negative effects decreases while Fr increases. In other words, flows with high Fr is more unpredictable, and flows with low Fr is more orderly. Therefore, we can conclude that Turbulent flow has high velocity, thus high Fr; Laminar flow has low velocity, thus low Fr.

The interaction between Re and Fr is significant in our results, so we can conclude that Re and Fr combining in very important in affecting flow's motion, while the St is less significant.

Conclusion

Since our response variables, where highly linearly correlated, we came to the conclusion that the model which best fit the first moment would also be a good fit for the other models. We saw that after transforming Fr and Re into categorical variables, we greatly increased the Test MSE of our models. Taking the second central moment as an example, we saw some statistically significant interaction terms between Re and Fr, which, because of the hierarchy principle, means we included the order 1 terms Re_category and Fr_categrory as well. We predicted that the variance of the particle cluster would increase significantly if the gravitational acceleration and Reynold's number jointly decreased. When fitting a model only with the linear terms St, Re_category, and Fr_category, we saw that there was a statistically significant increase in the variance when the Reynold's number and gravitational acceleration decreased independently, but to a much lesser extent than with the interaction term.

In this study, we analyzed the effect of Re, Fr, and St to the distribution of particle cluster volumns. It is important to note some limitation of our results: First, we have a relatively small training dataset (<100). This might result in issues relating to generalization and data imbalance. Second, we only have three levels for Re and Fr. There will be a serious issue of bad extrapolation if we use this model to predict tuples with Re and Fr outside of these three levels. Therefore, this model is only reliable with these three levels of Re and Fr. Last but not least, although our models has high Rsquared for all 4 moments, our model is a linear model with couple interaction terms, but based on our EDA, the relationship between Fr/Re and the moments is not linear. So we think there should be more complex models that can better describe the relationship between variables, thus produce a more accurate prediction result.

We believe that our model provides good prediction and on the probability distribution of particle cluster volumns, and interpretation on how Re, Fr, St affect the flows. We conclude that Turbulent flows have high St, high Re, high Fr; Laminar flows have low St, low Re, low Fr. Our results also reveal that the interaction between Re and Fr has significant effects on flows' motion.

Moreover, we think that there might be other omitted variables that affect the flow motion. We are curious about the threshold where flows transitioning from laminar to turbulent. In future study, we hope to explore other potential predictor variables and the critical point that decides flow's type.

Citations

https://www.sciencedirect.com/topics/engineering/stokes-number

https://www.sciencedirect.com/topics/engineering/froude-number

https://www.sciencedirect.com/topics/engineering/reynolds-number