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 Stoke's number St).

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 planktons, 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, radiative properties, and the rate at which droplets grow to form rain. 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.

We first noted that the predictor variables `Re` is clustered at fixed values, with `Re` clustering at 90, 224 and 398 (3 levels).

Table 1: St summary statistics

min	Q1	median	mean	Q3	max
0.05	0.3	0.7	0.86	1	3

Table 2: Re summary statistics

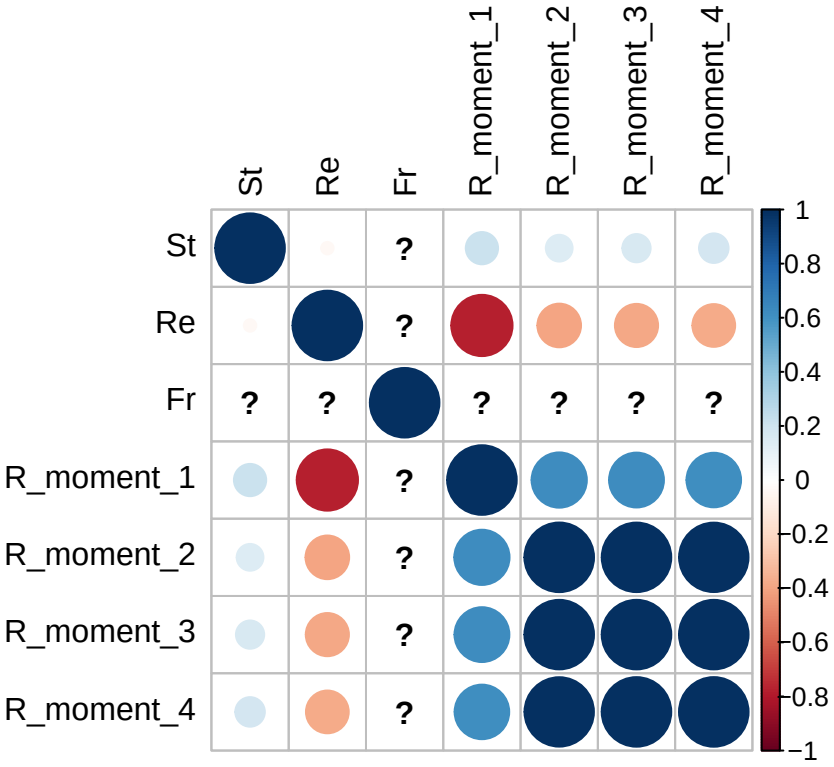
min	Q1	median	mean	Q3	max
90	90	224	214.472	224	398

Table 3: Fr summary statistics

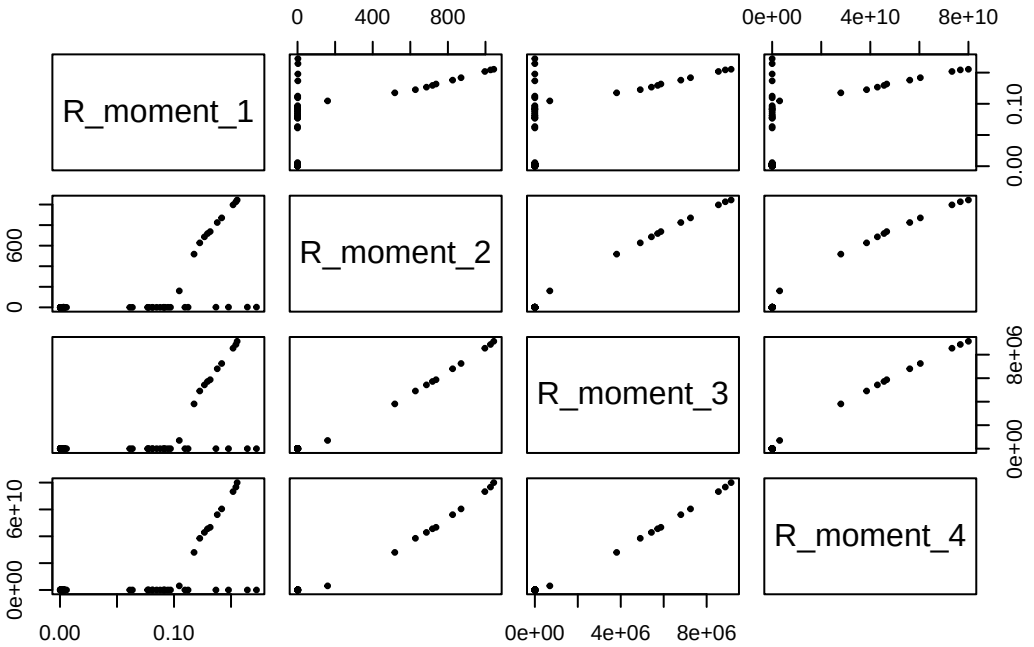
min	Q1	median	mean	Q3	max
0.052	0.052	0.3	Inf	Inf	Inf

We found that the moments are not only highly correlated,

	St	Re	Fr	R_moment_1	R_moment_2	R_moment_3	R_moment_4
St	1.0000000	-0.0316987	NaN	0.2147681	0.1479257	0.1647465	0.1800454
Re	-0.0316987	1.0000000	NaN	-0.7747206	-0.3932344	-0.3844289	-0.3774177
Fr	NaN	NaN	1	NaN	NaN	NaN	NaN
R_moment_1	0.2147681	-0.7747206	NaN	1.0000000	0.6298829	0.6217326	0.6150484
R_moment_2	0.1479257	-0.3932344	NaN	0.6298829	1.0000000	0.9984335	0.9946671
R_moment_3	0.1647465	-0.3844289	NaN	0.6217326	0.9984335	1.0000000	0.9988414
R_moment_4	0.1800454	-0.3774177	NaN	0.6150484	0.9946671	0.9988414	1.0000000



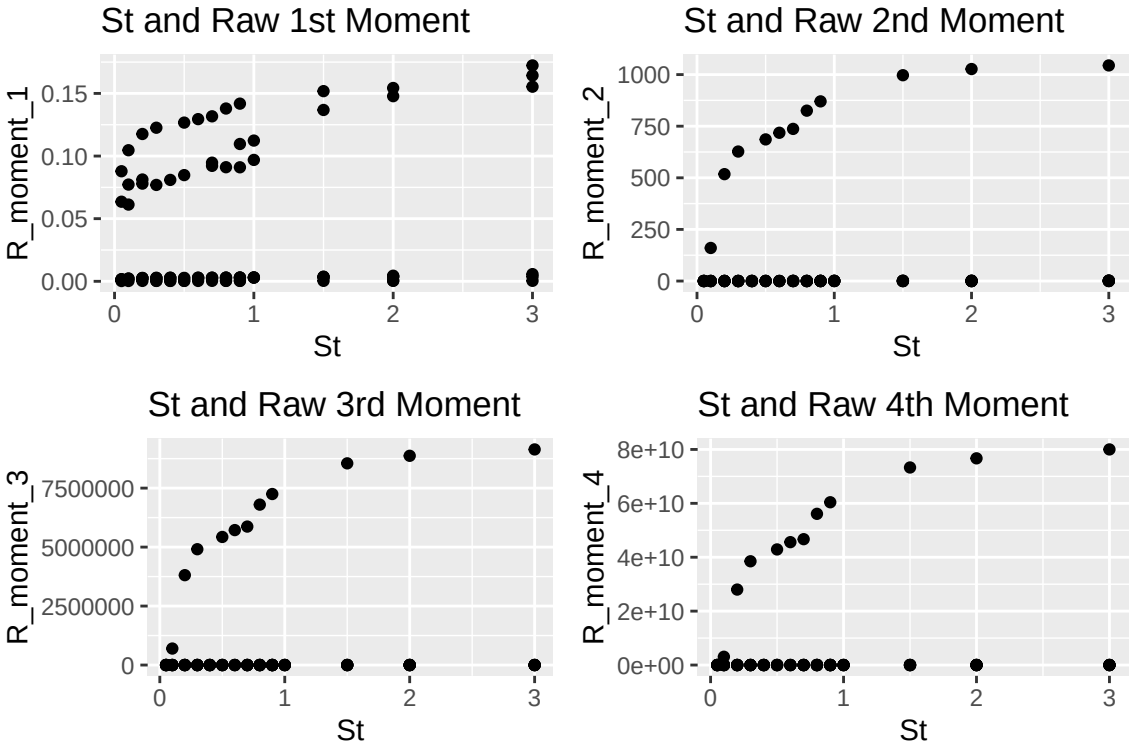
but that they are linearly correlated:

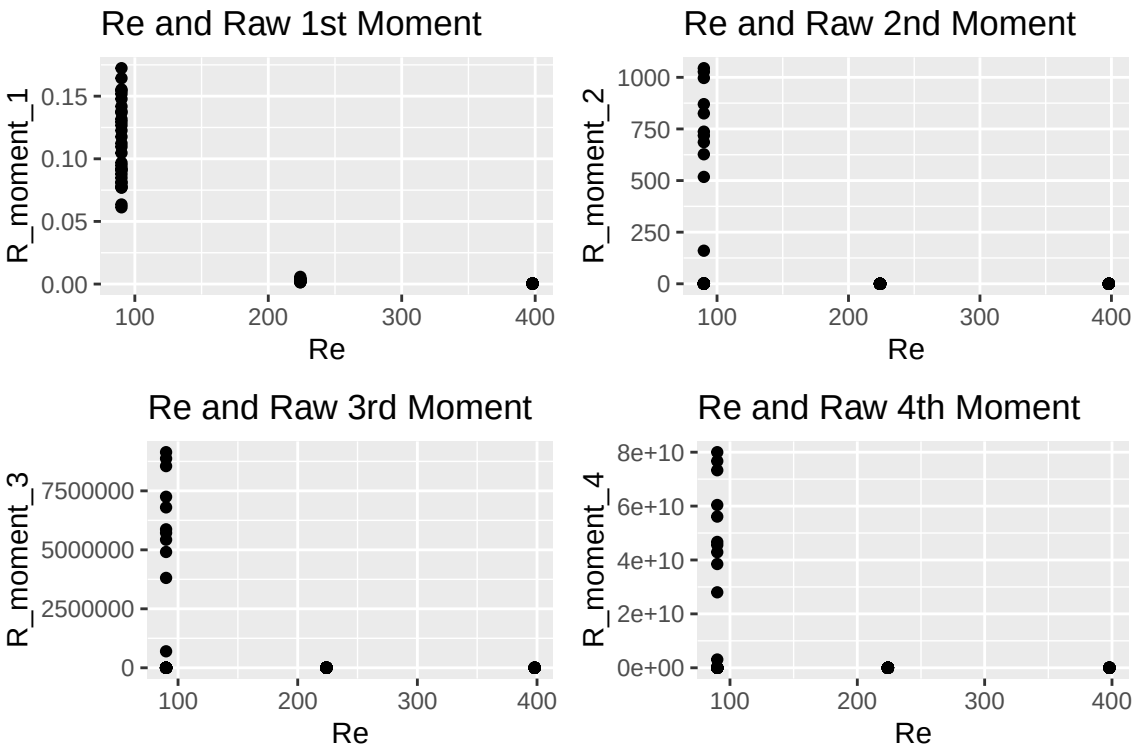


Therefore, we decided to fit a model on `R_moment_1`, which will give us the relationship of the predictor variables on the other moments due to the high linear correlation between the moments.

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

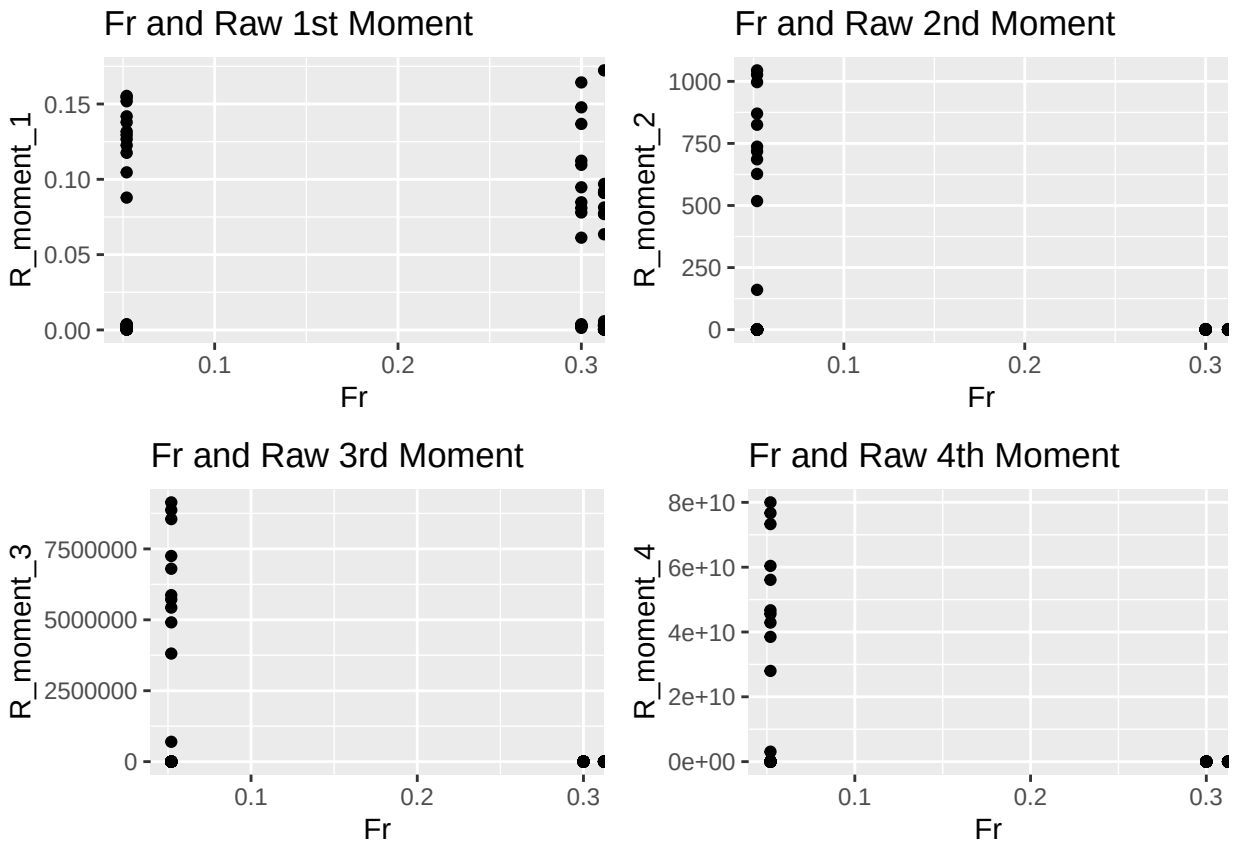




Moreover, we noticed that the gravitational acceleration has infinite values, which is problematic. Therefore, we used inverse logit transform on `Fr` to transform the infinity value into a finite value (Inf transformed to 1):

```
data_train <- data_train %>%
  mutate(Re_category = case_when(Re == 90 ~ "Low", Re==224 ~ "Medium", Re == 398 ~ "High"))%>%
  mutate(Fr_transformed = invlogit(Fr))
```

```
p9 <- ggplot(data = data_train, mapping = aes(x = Fr, y = R_moment_1)) +
     geom_point() + labs(title = "Fr and Raw 1st Moment")
p10 <- ggplot(data = data_train, mapping = aes(x = Fr, y = R_moment_2)) +
     geom_point() + labs(title = "Fr and Raw 2nd Moment")
p11 <- ggplot(data = data_train, mapping = aes(x = Fr, y = R_moment_3)) +
     geom_point() + labs(title = "Fr and Raw 3rd Moment")
p12 <- ggplot(data = data_train, mapping = aes(x = Fr, y = R_moment_4)) +
     geom_point() + labs(title = "Fr and Raw 4th Moment")
grid.arrange(p9, p10, p11, p12, nrow = 2)
```



In order to have better intepretation from the perspective of statistical inference, we convert raw moments to central moments. The first central moment is always 0, so we don't need to convert it.

We then explored shrinkage methods such as ridge regression and lasso. 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.

Ridge Model

We first build ridge models for the 2nd, 3rd, and 4th central moments. We mutate each variables to be binaries.

```
data_train <- data_train %>%
  mutate(Fr_category = case_when(Fr == 0.052 ~ "Low", Fr == 0.3 ~ "Medium", Fr == Inf ~ "High")) %>% mu
```

```
x <- data.matrix(data_train[, c('Re_low', "Re_medium", 'Re_high', 'St', 'Fr_low', "Fr_medium", "Fr_high
```

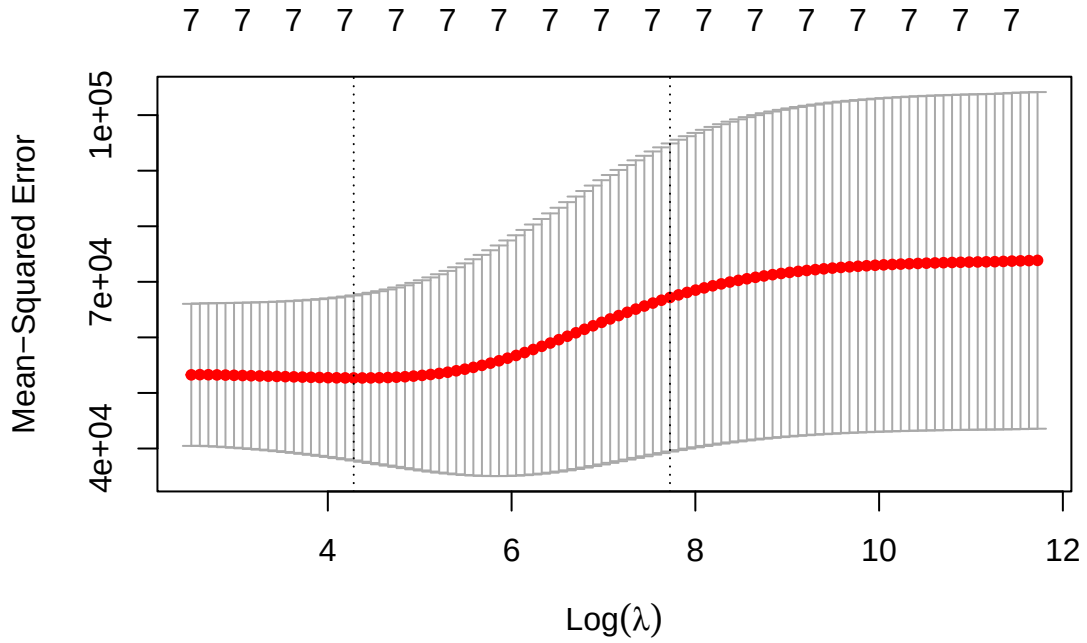
```
set.seed(123456)

train <- sample(1:nrow(x), nrow(x)/2)
test <- (-train)
lambda_seq = 10^seq(10, -2, length = 100)
```

```
y2 <- data_train$C_moment_2
ridge2 <- glmnet(x[train, ], y2[train], alpha = 0, lambda = lambda_seq)
y2.test = y2[test]
```

```
set.seed(123456)
#perform k-fold cross-validation to find optimal lambda value
```

```
cv_ridge2 <- cv.glmnet(x[train,], y2[train], alpha = 0, folds = 5)
best_lambda2 <- cv_ridge2$lambda.min
plot(cv_ridge2)
```




```
ridge.pred2 <- predict(ridge2, s = best_lambda2, newx = x[test,])
rmse_c2 <- sqrt(mean((ridge.pred2 - y2.test)^2))
```

Ridge model for 3rd central moment:

```
y3 <- data_train$C_moment_3
```

```
ridge3 <- glmnet(x[train,], y3[train], alpha = 0, lambda = lambda_seq)
```

```
set.seed(123)
#perform k-fold cross-validation to find optimal lambda value
cv_ridge3 <- cv.glmnet(x[train,], y3[train], alpha = 0, folds = 5)
#find optimal lambda value that minimizes test MSE
best_lambda3 <- cv_ridge3$lambda.min
```

```
y3.test <- y3[test]
ridge.pred3 <- predict(ridge3, s = best_lambda3, newx = x[test,])
rmse_c3 <- sqrt(mean((ridge.pred3 - y3.test)^2))
```

Ridge model on 4th central moment:

```
y4 <- data_train$C_moment_4
```

```
ridge4 <- glmnet(x[train,], y4[train], alpha = 0, lambda = lambda_seq)
```

```
set.seed(123456)
#perform k-fold cross-validation to find optimal lambda value
cv_ridge4 <- cv.glmnet(x[train,], y4[train], alpha = 0, folds = 5)
#find optimal lambda value that minimizes test MSE
best_lambda4 <- cv_ridge4$lambda.min
```

```
y4.test <- y4[test]
ridge.pred4 <- predict(ridge4, s = best_lambda4, newx = x[test,])
rmse_c4 <- sqrt(mean((ridge.pred4 - y4.test)^2))
```

```
RidgeModel <- c("C_M2", "C_M3", "C_M4")
RMSE <- c(rmse_c2, rmse_c3, rmse_c4)

ridgedf <- data.frame(RidgeModel, RMSE) %>% kable()

print(ridgedf)
```

```
##
##
## |RidgeModel |         RMSE|
## |:----------|------------:|
## |C_M2       | 1.972020e+02|
## |C_M3       | 1.693927e+06|
## |C_M4       | 1.420521e+10|
```

The coefficients predicted by ridge:

```
ridge.test2 <- glmnet(x, y2, alpha = 0)
predict(ridge.test2, type = 'coefficients', s = best_lambda2)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)   59.22542
## Re_low       150.04713
## Re_medium    -69.61215
## Re_high     -100.38633
## St            28.96213
## Fr_low       126.57885
## Fr_medium    -90.38756
## Fr_high      -58.75424
```

```
ridge.test3 <- glmnet(x, y3, alpha = 0)
predict(ridge.test3, type = 'coefficients', s = best_lambda3)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept)  514066.7
## Re_low       981428.1
## Re_medium   -490379.9
## Re_high     -607171.3
## St           221532.0
## Fr_low       832500.7
## Fr_medium   -559916.1
## Fr_high     -416927.1
```

```
ridge.test4 <- glmnet(x, y4, alpha = 0)
predict(ridge.test4, type = 'coefficients', s = best_lambda4)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  3525874657
## Re_low       9568215587
## Re_medium   -4522459841
```

```
## Re_high     -6289329957
## St           2500738608
## Fr_low       8087723796
## Fr_medium   -5706626596
## Fr_high     -3817463027
```

Linear Model (Inference)

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:

```
final_model_M1 <- lm(R_moment_1 ~ St+ Re_category+Fr_transformed+Fr_transformed*Re_category+St*Re_categ
final_model_M2 <- lm(C_moment_2 ~ St+ Re_category+Fr_category+Fr_category*Re_category+St*Re_category, da
final_model_M3 <- lm(C_moment_3 ~ St+ Re_category+Fr_category+Fr_category*Re_category+St*Re_category, da
final_model_M4 <- lm(C_moment_4 ~ St+ Re_category+Fr_category+Fr_category*Re_category+St*Re_category, da
```

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

We can see that the linear model performs better than the ridge model over all, so we 'll 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

Observations	89
Dependent variable	R_moment_1
Type	OLS linear regression

F(8,80)	392.74
R^2	0.98
Adj. R^2	0.97

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.

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.

	Est.	S.E.	t val.	p
(Intercept)	0.00	0.01	0.03	0.98
St	0.00	0.00	0.02	0.98
Re_categoryLow	0.12	0.01	12.27	0.00
Re_categoryMedium	0.00	0.01	0.16	0.87
Fr_transformed	0.00	0.01	0.01	0.99
Re_categoryLow:Fr_transformed	-0.05	0.01	-4.39	0.00
Re_categoryMedium:Fr_transformed	0.00	0.01	0.05	0.96
St:Re_categoryLow	0.03	0.00	7.99	0.00
St:Re_categoryMedium	0.00	0.00	0.23	0.82

Standard errors: OLS

Observations	89
Dependent variable	log_C_moment_2
Type	OLS linear regression

F(10,78)	72.83
R^2	0.90
Adj. R^2	0.89

	Est.	S.E.	t val.	p
(Intercept)	-5.42	0.48	-11.30	0.00
St	0.30	0.39	0.76	0.45
Re_categoryLow	3.82	0.66	5.78	0.00
Re_categoryMedium	1.13	0.65	1.74	0.09
Fr_categoryLow	-0.23	0.58	-0.40	0.69
Fr_categoryMedium	-0.07	0.50	-0.13	0.89
Re_categoryLow:Fr_categoryLow	6.89	0.79	8.68	0.00
Re_categoryMedium:Fr_categoryLow	2.17	0.76	2.88	0.01
Re_categoryLow:Fr_categoryMedium	0.27	0.75	0.36	0.72
Re_categoryMedium:Fr_categoryMedium	NA	NA	NA	NA
St:Re_categoryLow	0.68	0.47	1.47	0.15
St:Re_categoryMedium	0.65	0.47	1.38	0.17

Standard errors: OLS

Observations	89
Dependent variable	log_C_moment_3
Type	OLS linear regression

F(10,78)	63.97
R^2	0.89
Adj. R^2	0.88

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

St	Re	Fr	Predicted_R_M1	Predicted_R_M2	Predicted_R_M3	Predicted_R_M4
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

	Est.	S.E.	t val.	p
(Intercept)	-2.47	0.78	-3.18	0.00
St	0.31	0.63	0.50	0.62
Re_categoryLow	2.87	1.07	2.68	0.01
Re_categoryMedium	0.57	1.05	0.54	0.59
Fr_categoryLow	-0.29	0.94	-0.31	0.76
Fr_categoryMedium	-0.08	0.80	-0.10	0.92
Re_categoryLow:Fr_categoryLow	13.06	1.29	10.15	0.00
Re_categoryMedium:Fr_categoryLow	4.30	1.22	3.51	0.00
Re_categoryLow:Fr_categoryMedium	0.31	1.22	0.25	0.80
Re_categoryMedium:Fr_categoryMedium	NA	NA	NA	NA
St:Re_categoryLow	1.12	0.75	1.48	0.14
St:Re_categoryMedium	1.00	0.77	1.31	0.19

Standard errors: OLS

Observations	89
Dependent variable	log_C_moment_4
Type	OLS linear regression

F(10,78)	66.31
R^2	0.89
Adj. R^2	0.88

	Est.	S.E.	t val.	p
(Intercept)	0.48	1.04	0.46	0.65
St	0.33	0.84	0.40	0.69
Re_categoryLow	2.10	1.44	1.46	0.15
Re_categoryMedium	0.10	1.41	0.07	0.95
Fr_categoryLow	-0.35	1.26	-0.28	0.78
Fr_categoryMedium	-0.10	1.08	-0.09	0.93
Re_categoryLow:Fr_categoryLow	19.15	1.72	11.12	0.00
Re_categoryMedium:Fr_categoryLow	6.42	1.64	3.92	0.00
Re_categoryLow:Fr_categoryMedium	0.35	1.64	0.21	0.83
Re_categoryMedium:Fr_categoryMedium	NA	NA	NA	NA
St:Re_categoryLow	1.49	1.01	1.47	0.14
St:Re_categoryMedium	1.30	1.03	1.27	0.21

Standard errors: OLS

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 induce 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 contribute to the dominant effects over the flow's motion, while the St is less significant.

Conclusion

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