Case Study

Abdel Shehata

2022-10-26

Introduction and Data

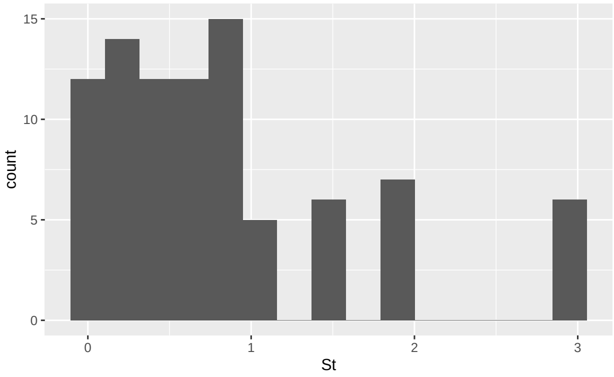
Introduction

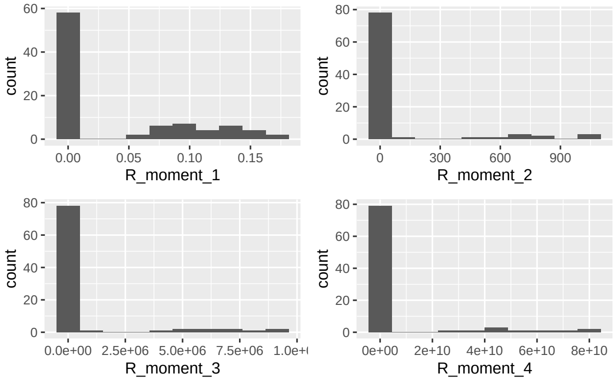
In fluid mechanics flows are either turbulent or laminar. Turbulent flows is characterized by random and chaotic motion, whereas laminar flow is predictable and orderly. Turbulent flow has various applications in air pollution, chemical reactions and heat transfer. In an idealized turbulence the clustering of particles is affected by fluid turbulence (Reynolds number Re), gravitational acceleration (Froude number Fr) and particles' characteristics (Stokes number St). We wish to develop a model that predicts the first four raw moments ($\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathbb{E}[X^3]$, $\mathbb{E}[X^4]$) of a particles cluster volume distribution based off the clustering's Reynolds, Forude, and Stokes number.

Data Introduction

The data which we will use to train our model consists of n = 89 tuples which each represent simulations conducted at a different parameter setting (Re, St, Fr). Each tuple contains the first four moments of the particle cluster volume distribution in addition to the parameter settings.

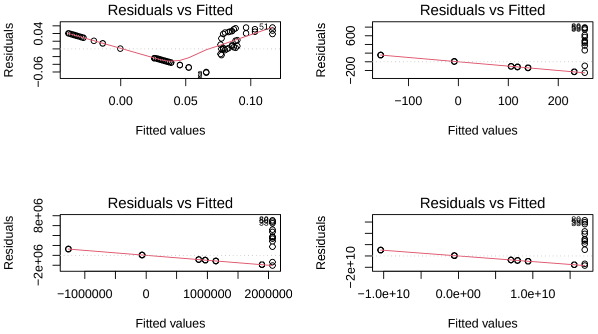
Exploratory Data Analysis

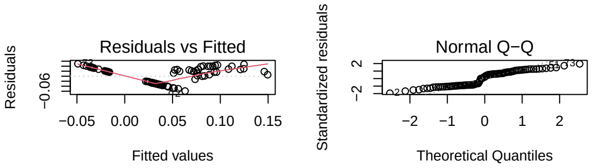


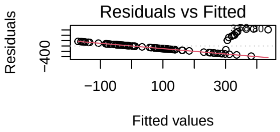


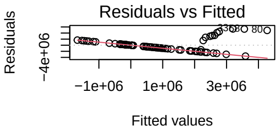
Here are the linear models we obtained for the moments (without interactions):

$$\hat{R}_1 = 0.0102 + 0.01353 * St - 0.0003798 * Re$$
$$\hat{R}_2 = 299.6593 - 0.8473 * Re - 10.2317TFr$$
$$\hat{R}_3 = 2442265 - 6905 * Re - -83580TFr$$
$$\hat{R}_4 = 2.008 * 10^{10} - 5.677 * 10^7 * Re - 6.872 * 10^8TFr$$









$$\hat{R_1} = 9.822 * 10^{-2} + 3.398 * 10^{-2} * St - 2.534 * 10^{-3} * TFr - 3.176 * 10^{-4} * Re$$
$$-1.002 * 10^{-4} * St * Re + 9.098 * 10^{-6} * TFr * Re$$

$$\hat{R_2} = 327.50288 + 46.54518 * St - 36.88062 * TFr - 1.19146 * Re$$
$$-0.11802 * TFr * Re$$

$$\hat{R_3} = 2525699.0 + 556624.1 * St - 252310.1 * TFr - 9805.7 * Re$$
$$-54692.6 * St * TFr + 953.2 * TFr * Re$$
$$\hat{R_4} = 1.528 * 10^{10} + 1.050 * 10^{10} * St - 1.915 * 10^{10} * TFr - 5.598 * 10^{7} * Re$$
$$-5.176 * 10^{8} * St * TFr - 2.662 * 10^{7} * St * Re + 7.7304 * 10^{6} * TFr * Re$$

Model Evalutation

Linear RMSE	No Interactions	Interactions
R__1	0.0349	0.0340
R__2	237.411	222.3977
R__3	1991432	1870992
R__4	16757102688	15946689806

Linear R-Squared	No Interactions	Interactions
R__1	0.6054832	0.6282726
R__2	0.1716913	0.2754936
R__3	0.161867	0.2650636
R__4	0.1539926	0.2466392

Complex Model

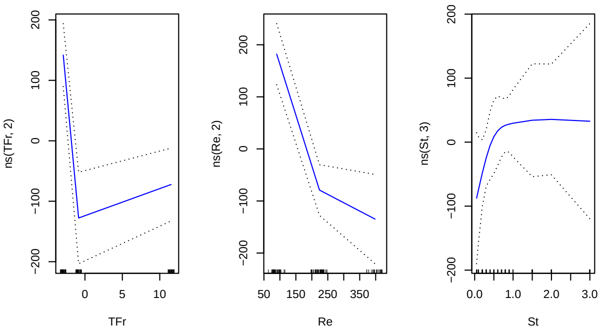
Lets First start start by evaluating a gam model to see if there is a complex relationship between our response variables and predictor variables. For this section of the project, I will be only employing the simpler lm function to keep all the models in the same format.

Gam_Model	Adjusted.R.Squared	RSS
gam1	0.9243735	1.910310e-02
gam2	0.4170109	3.237061e+06
gam3	0.4014772	2.310425e+14
gam4	0.3887645	1.654812e+22

As we can see, a Gam Model performs better with all the moments than linear regression. Nevertheless, as we can see the model only performs adequately with the first moment with an adjusted r square of 0.9244 and an RSS of 0.01536 compared to an average of 0.4 for the other models. This likely is due to the lack of interaction factors in our model, which appear to affect the second, third and fourth moment more than the first. Thus, we might need to add some interaction values to our polynomial model.

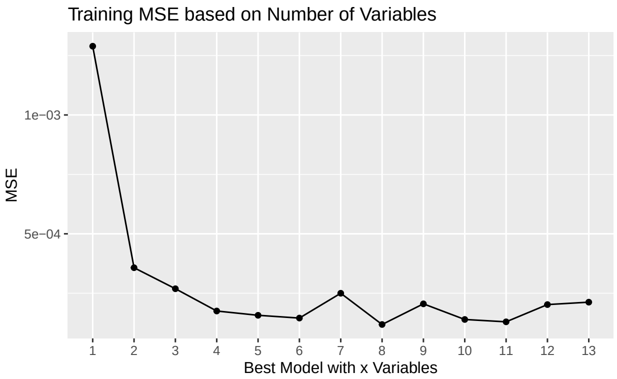
Note: 2 degrees were chosen due to the number of unique values of in our data. Only 3 degrees were chosen for St since it has multiple unique values.

Plots Since the first GAM performed particulary well, it might be worthwhile to explore the relationship between our response varible and predicator varibles using plots.

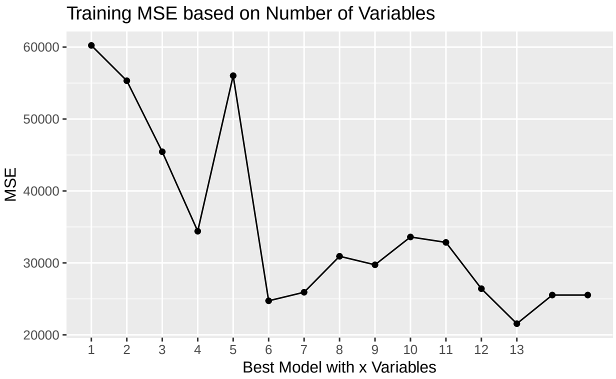


As we can see from the plots, the first moment appears to experience a steep drop for both TFr and Re from the first to the second observation. The decrease becomes much less significant from the second observation to the third observation for both Tfr and Re. The relationship between St and the first moment appear to be roughly linear and increasing.

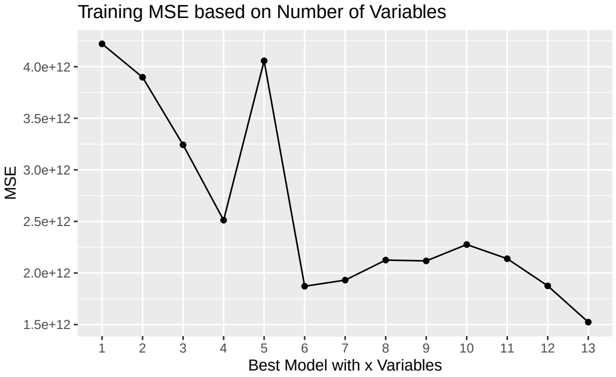
Best Degree Model We can utilize a sequential stepwise selection method with a full model. Our full model would utilize the max number of interactions with a polynomial degree of 2 for Tfr and Re (Max number of degrees based on unique values). We will use one for St since throughout our linear models and GAM models, it appears that the relationship between the St and the moments is approximetly linear.



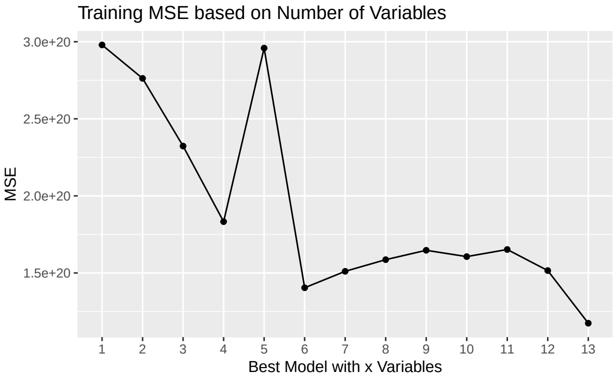
As we can see from the plot, there appears to be in MSE until about the 5 degrees best model. Then the MSE increases again until about the model with 9 variables, where it decreases again. Since, for the first moment, we are focusing more on inference and the error appears to be neigable between the fifth model and later models,we will choose the fifth model as our best model.



As we can see from the plot, the model does best is the one with 12 variables. However, this model is most likely not very understandable since it does include a variety of interactions including the three way interaction. So the model with 9 variables might be better for some inference. Nevertheless, it does appear that the higher the moment the harder it is to predict.



As we can see, the complex model (the almost full model with 12 variables) appears to be doing significantly best. However, since the errors are so big, it might be worth exploring the difference in error between including 12 variables and 8 variables in our final model.



Interestingly, for the fourth moment, the full model appears to do the best. However, the model with 8 variables might of interest to explore since it might be better for inference while not scaificing a ton of accuracy.

Results (Final Models)

```
##
## Call:
## lm(formula = R_moment_1 ~ I(Re^2) + St + Re + TFr + St:Re, data = data_train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033233 -0.007660 -0.000093  0.005800  0.026518
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.074e-01  6.391e-03  32.459  < 2e-16 ***
## I(Re^2)      2.576e-06  1.100e-07  23.408  < 2e-16 ***
## St           3.394e-02  3.304e-03  10.274  < 2e-16 ***
## Re          -1.520e-03  5.563e-05 -27.327  < 2e-16 ***
## TFr         -6.816e-04  1.972e-04  -3.456 0.000868 ***
## St:Re       -1.072e-04  1.426e-05  -7.517 5.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01196 on 83 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9542
## F-statistic: 367.4 on 5 and 83 DF,  p-value: < 2.2e-16
```

As we can see a model with 5 variables is enough to predict the first raw moment accurately. Interestingly enough, similar to the linear models TFr doesn't appear to have a significant relationship between it and the first raw moment.

```
##
## Call:
## lm(formula = R_moment_2 ~ I(TFr^2) * I(Re^2) * I(St^2) + St *
##     Re * TFr, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -312.93  -69.69   10.39  100.46  188.99
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.215e+02  9.997e+01   1.215 0.228140
## I(TFr^2)                    1.298e+01  1.536e+00   8.449 1.83e-12 ***
## I(Re^2)                     1.213e-02  1.987e-03   6.104 4.37e-08 ***
## I(St^2)                    -1.476e+02  3.246e+01  -4.547 2.08e-05 ***
## St                          5.621e+02  1.188e+02   4.732 1.04e-05 ***
## Re                         -4.259e+00  7.968e-01  -5.345 9.63e-07 ***
## TFr                        -1.700e+02  2.304e+01  -7.378 1.93e-10 ***
## I(TFr^2):I(Re^2)           -1.275e-04  2.345e-05  -5.440 6.59e-07 ***
## I(TFr^2):I(St^2)            1.703e+00  4.013e-01   4.245 6.27e-05 ***
## I(Re^2):I(St^2)             1.384e-03  4.026e-04   3.437 0.000967 ***
## St:Re                      -1.841e+00  4.917e-01  -3.744 0.000356 ***
## St:TFr                     -7.799e+01  1.683e+01  -4.633 1.51e-05 ***
## Re:TFr                      5.538e-01  1.128e-01   4.909 5.31e-06 ***
## I(TFr^2):I(Re^2):I(St^2) -1.611e-05  7.195e-06  -2.240 0.028123 *
## St:Re:TFr                   2.434e-01  7.740e-02   3.145 0.002390 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127.7 on 74 degrees of freedom
## Multiple R-squared:  0.8001, Adjusted R-squared:  0.7623
## F-statistic: 21.15 on 14 and 74 DF,  p-value: < 2.2e-16
```

The final model for the second raw moment is a model with 10 variables that has up to two interaction levels.

```
##
## Call:
## lm(formula = R_moment_3 ~ I(TFr^2) * I(Re^2) * I(St^2) + St *
##     Re * TFr, data = data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2655800  -576015   103704   820074  1761835
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.540e+05  8.392e+05   0.779 0.438259
## I(TFr^2)                  1.050e+05  1.290e+04   8.142 6.98e-12 ***
## I(Re^2)                   9.754e+01  1.668e+01   5.848 1.26e-07 ***
## I(St^2)                  -1.312e+06  2.725e+05  -4.816 7.57e-06 ***
## St                        5.100e+06  9.973e+05   5.114 2.39e-06 ***
## Re                       -3.314e+04  6.689e+03  -4.954 4.47e-06 ***
## TFr                      -1.337e+06  1.934e+05  -6.913 1.42e-09 ***
## I(TFr^2):I(Re^2)         -1.030e+00  1.968e-01  -5.235 1.49e-06 ***
## I(TFr^2):I(St^2)          1.526e+04  3.368e+03   4.530 2.22e-05 ***
## I(Re^2):I(St^2)           1.253e+01  3.380e+00   3.707 0.000403 ***
## St:Re                    -1.683e+04  4.128e+03  -4.077 0.000114 ***
## St:TFr                   -7.091e+05  1.413e+05  -5.019 3.48e-06 ***
## Re:TFr                    4.351e+03  9.470e+02   4.594 1.75e-05 ***
## I(TFr^2):I(Re^2):I(St^2) -1.461e-01  6.040e-02  -2.418 0.018051 *
## St:Re:TFr                 2.229e+03  6.498e+02   3.430 0.000991 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1072000 on 74 degrees of freedom
## Multiple R-squared:  0.7973, Adjusted R-squared:  0.759
## F-statistic:  20.8 on 14 and 74 DF,  p-value: < 2.2e-16
```

Our Final model for Moment 3 will be the full model with 13 variables (ST is in the model due to the hierarchy principle) . This model has all interaction variables up to 3, since they appear to be significant (P<0.05)

```
##
## Call:
## lm(formula = R_moment_4 ~ I(TFr^2) * I(Re^2) * I(St^2) + St *
##     Re * TFr, data = data_train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.282e+10 -5.316e+09  4.454e+08  6.286e+09  1.628e+10
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.072e+09  7.046e+09   0.436 0.664166
## I(TFr^2)                  8.555e+08  1.083e+08   7.901 1.99e-11 ***
## I(Re^2)                   7.914e+05  1.400e+05   5.651 2.81e-07 ***
## I(St^2)                  -1.145e+10  2.288e+09  -5.004 3.68e-06 ***
## St                        4.534e+10  8.373e+09   5.415 7.29e-07 ***
## Re                       -2.610e+08  5.616e+07  -4.647 1.43e-05 ***
## TFr                      -1.062e+10  1.624e+09  -6.541 6.97e-09 ***
## I(TFr^2):I(Re^2)         -8.383e+03  1.653e+03  -5.073 2.82e-06 ***
## I(TFr^2):I(St^2)          1.345e+08  2.828e+07   4.757 9.49e-06 ***
## I(Re^2):I(St^2)           1.112e+05  2.838e+04   3.917 0.000198 ***
## St:Re                    -1.507e+08  3.466e+07  -4.347 4.33e-05 ***
## St:TFr                   -6.333e+09  1.186e+09  -5.338 9.91e-07 ***
## Re:TFr                    3.450e+07  7.951e+06   4.340 4.46e-05 ***
## I(TFr^2):I(Re^2):I(St^2) -1.302e+03  5.071e+02  -2.567 0.012273 *  
## St:Re:TFr                 2.002e+07  5.456e+06   3.669 0.000456 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.998e+09 on 74 degrees of freedom
## Multiple R-squared:  0.7963, Adjusted R-squared:  0.7578 
## F-statistic: 20.66 on 14 and 74 DF,  p-value: < 2.2e-16
```

Our Final model for Moment 4 will be the full model with 13 variables (ST is in the model due to the hierarchy principle) . This model has all interaction variables up to 3, since they appear to be significant (P<0.05).

Discussion & Conclusion

References