Case Study: Turbulence

Introduction

Our goals are as follows: For a new parameter setting of (Re, F r, St), predict its particle cluster volume distribution in terms of its four raw moments. Inference: Investigate and interpret how each parameter affects the probability distribution for particle cluster volumes.

Methodology

Linear Modeling

Our univariate exploratory data analysis of Re, Fr, St, and each moment revealed that the R moments are heavily right skewed, which poses a problem to linear regression. We applied log transformations on each moment to obtain more normally distributed variables. The log-transformed R moments are approximately normal, and it appears that each R moment variable has somewhat of a linear relationship with St.

Accordingly, we fit a basic linear model onto each log-transformed response variable. Linear models were evaluated based on adjusted R^2 values and Pr(>|t-value|) for coefficient estimates. We chose to treat Re and Fr as factors or categorical variables, as Re only takes on the values of 90, 224, and 398; Fr only takes on the values 0.052, 0.3, and infinity. While the adjusted R^2 value for R_moment_1 was very high at 0.9949, subsequent moments exhibited decreasing adjusted R^2 values, with R_moment_4 having an adjusted R^2 value of 0.6518. We explored multicollinearity through VIFs for each model, which were very low. We also explored the addition of interaction terms to the model. The only interaction term which was significant for all R_moments was the interaction between Re and Fr. Therefore, we constructed the linear models with an interaction between Re and Fr:

$$log(R_{moment1}) = -2.273 + 0.25(st) - 3.816(Re_{224}) - 5.988(Re_{398}) - 0.263(Fr_{0.3}) - 0.329(Fr_{\infty}) + 0.221(Re_{224} * Fr_{0.3}) + 0.402(Re_{224} * Fr_{\infty}) + 0.502(Re_{398} * Fr_{\infty})$$

$$log(R_{moment2}) = 5.187 + 0.834(st) - 7.434(Re_{224}) - 11.384(Re_{398}) - 6.416(Fr_{0.3}) - 6.652(Fr_{\infty}) + 4.387(Re_{224} * Fr_{0.3}) + 4.718(Re_{224} * Fr_{\infty}) + 7.076(Re_{398} * Fr_{\infty})$$

$$log(R_{moment3}) = 13.398 + 1.174(st) - 11.164(Re_{224}) - 17.030(Re_{398}) - 12.478(Fr_{0.3}) - 12.772(Fr_{\infty}) + 8.365(Re_{224} * Fr_{0.3}) + 8.772(Re_{224} * Fr_{\infty}) + 13.371(Re_{398} * Fr_{\infty})$$

$$log(R_{moment4}) = 21.695 + 1.469(st) - 14.906(Re_{224}) - 22.715(Re_{398}) - 18.471(Fr_{0.3}) - 18.811(Fr_{\infty}) + 12.276(Re_{224} * Fr_{0.3}) + 12.756(Re_{224} * Fr_{\infty}) + 19.568(Re_{398} * Fr_{\infty})$$

Adding the interaction term between Re and Fr improved the fit of the model according to the adjusted R^2 values, which are much higher for every moment. With this new interaction term included, the adjusted R^2 value for R_moment_1 was slightly higher than before at 0.9966, and increased for moment 2 at 0.8909, moment 3 at 0.8770, and moment 4 0.8809 respectively.

To analyze predictive performance of our models, we split data into training and testing sets to evaluate the predictive ability of the models we explored. The linear models with the interaction term for Re and Fr outperformed any other linear model, producing lower test MSEs for every moment of R. Having this interaction term significantly improved the test MSEs of the linear model.

Splines and GAM

For each of the four moments, we fit a generalized additive model with a smoothing spline in the interest to make the model significantly complex to capture non-linearities in the data. Without interaction terms, the MSE of these models were similar to the test MSE for least squares regression for all moments. We added interaction terms to the model for each pair of predictors to extend the traditional GAM structure which does not allow for interaction effects. Although interactions with non-parametric smooth terms are not fully supported by this model, the gam() function still allows for parametric interactions. Thus, although adding interaction effects makes the model slightly more computationally expensive, it affords additional insight into the relationship that the predictors have with one another and improves prediction accuracy. This model performs significantly better with regard to both adjusted R^2 values and test MSE.

As mentioned previously, a smoothing spline was used to capture the non-linearities between St and the higher order moments; this was not necessary for the first moment as this relationship was approximately linear. This smoothing spline is a thin-plate regression spline— which simplifies to a cubic spline in this one-dimensional case— and has many attractive features. Thin-plate regression splines circumvent issues with knot placement in conventional spline modelling. As the true regression function that describes the relationship between St and the higher order moments seems to be highly complex, using splines allow for automated model specification without a priori knowledge of this relationship.

Alternate model selection and regularization methods

As criteria, we used Adjusted R^2 values, which we try to maximize, and mean squared errors of test sets, which we try to minimize. We applied other linear model selection methods such as principal components regression, partial least squares, but none produced fits or mean squared errors of the test set lower than the linear models represented above using least squares. Given enough evidence for nonlinearity in higher moments' behaviors in our EDA and modeling performance, we also tried a regression tree, random forest, box cox transformations, and polynomial regressions with various degrees of freedom. The model fits and predictive performances of these models were poor, or similar to the linear regression model, in terms of the MSE of the test set.

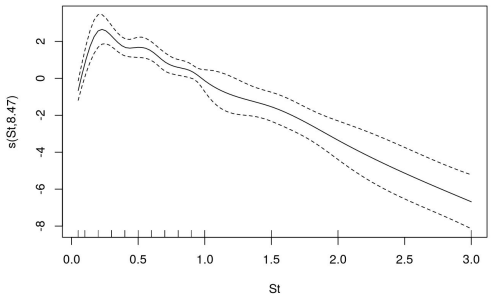
Final Results & Discussion

Based on the test MSE's for least squares and the GAM model with interaction terms, we decided to use the least squares model with interaction term for the first moment, and the GAM models for the higher moments. As per the definition of the different moments, only the first moment, E(X), involves the average value of a linear term, whereas the other moments involve predicted values of higher degrees.

Accordingly, our final model for the first moment is as follows. This yielded a test MSE of 0.00882.
log(E(X)) = -2.2.738 + 0.2443*St - 3.7886I(Re = 224) - 5.9965I(Re = 398) - 2.44I(Fr = 0.3) - .3315I(Fr = Inf) + 0.1542*I(Re = 224)*I(Fr = 0.3) + 0.3826I(Re = 224)*I(Fr = Inf) + 0.5011*I(Re = 398)*I(Fr = Inf)
(Note: no observations in the training set had both Re = 398 and Fr = 0.3, so that interaction term was not calculated by R. Interaction terms with St were not included because none were deemed significant.)

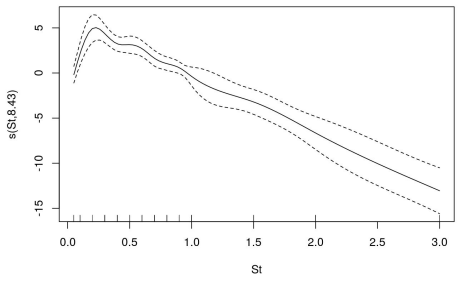
For the second moment, the model giving log(E(X^2)) is the GAM model with a smoothing spline and interaction terms, given below. Plots of the residuals against the fitted values were used to additionally evaluate the goodness-of-fit on the models; there was no flagrant patterning of residuals. On the test data, this model yielded an MSE on the log-scale of 0.861.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5466     0.3752    6.79  1.2e-08 ***
Re224        -7.1159     0.3758  -18.93  < 2e-16 ***
Re398       -11.0853     0.5222  -21.23  < 2e-16 ***
Fr0.3        -6.8740     0.4918  -13.98  < 2e-16 ***
FrInf        -6.1737     0.4269  -14.46  < 2e-16 ***
Re90:St       3.7793     0.3544   10.66  1.6e-14 ***
Re224:St      3.6394     0.3683    9.88  2.2e-13 ***
Re398:St      3.2667     0.3846    8.49  2.7e-11 ***
Fr0.3:St      0.5126     0.3857    1.33     0.19
FrInf:St     -0.0742     0.2530   -0.29     0.77
Re224:Fr0.3   4.0664     0.4728    8.60  1.9e-11 ***
Re398:Fr0.3   0.0000     0.0000      NA       NA
Re224:FrInf   4.1555     0.4611    9.01  4.4e-12 ***
Re398:FrInf   6.7440     0.5552   12.15  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df  F p-value
s(St) 8.47    8.8 17  <2e-16 ***
```



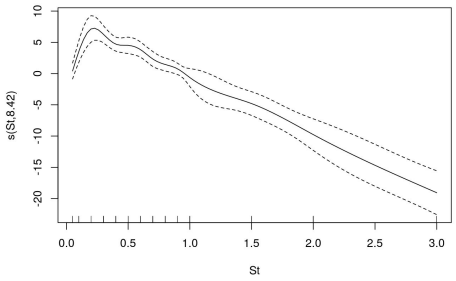
For the third moment, the model giving log(E(X^3)) is the GAM model with smoothing spline and interaction terms, given below. On the test data, this model yielded a MSE on the log-scale of 2.29.

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.005      0.653   12.25  < 2e-16 ***
Re224        -10.601      0.657  -16.12  < 2e-16 ***
Re398        -16.431      0.913  -17.99  < 2e-16 ***
Fr0.3        -13.064      0.860  -15.19  < 2e-16 ***
FrInf        -11.789      0.747  -15.79  < 2e-16 ***
Re90:St        7.160      0.616   11.61  7.3e-16 ***
Re224:St       6.882      0.641   10.73  1.3e-14 ***
Re398:St       6.233      0.671    9.30  1.6e-12 ***
Fr0.3:St       0.686      0.675    1.02     0.31
FrInf:St      -0.300      0.443   -0.68     0.50
Re224:Fr0.3    7.856      0.827    9.50  8.1e-13 ***
Re398:Fr0.3    0.000      0.000      NA       NA
Re224:FrInf    7.873      0.807    9.76  3.3e-13 ***
Re398:FrInf   12.763      0.971   13.14  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df    F p-value
s(St) 8.43   8.79 19.9  <2e-16 ***
```



For the fourth moment, the model giving log(E(X^4)) is the GAM model with smoothing spline and interaction terms, given below. On the test data, this model yielded a MSE on the log-scale of 4.1.

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.697      0.903   15.17  < 2e-16 ***
Re224        -14.150      0.911  -15.54  < 2e-16 ***
Re398        -21.858      1.265  -17.28  < 2e-16 ***
Fr0.3        -19.135      1.191  -16.06  < 2e-16 ***
FrInf        -17.359      1.034  -16.78  < 2e-16 ***
Re90:St       10.339      0.851   12.15  < 2e-16 ***
Re224:St       9.950      0.886   11.23  2.5e-15 ***
Re398:St       9.050      0.927    9.76  3.3e-13 ***
Fr0.3:St       0.802      0.934    0.86     0.39
FrInf:St      -0.520      0.613   -0.85     0.40
Re224:Fr0.3   11.608      1.145   10.13  9.3e-14 ***
Re398:Fr0.3    0.000      0.000      NA       NA
Re224:FrInf   11.563      1.117   10.35  4.5e-14 ***
Re398:FrInf   18.705      1.345   13.90  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df    F p-value
s(St) 8.42   8.78 22.5  <2e-16 ***
```



In general, for all three general additive models, as Re increases, the log expected value of each of the moments decreases. Since the particle clustering value is nonnegative, this indicates a general decrease in magnitude accompanying an increase in Reynolds number. Furthermore, as Fr increases, the log expected value of

each of the moments decreases. For similar reasons, this indicates an expected decrease in magnitude of the expected moment values with an increase in the gravitational acceleration.

Across the three higher moments, the interaction effects between any Re indicator and St was positive; this indicates that while particle characteristic has a general negative correlation with the log moments after about St=0.9 when s(St) dips below the x-axis (as shown in the visuals), this being coupled with a high value of Reynolds number (i.e. high fluid turbulence) mitigates this effect. Furthermore, for all four models, even though a higher Reynold's number or gravitational acceleration alone have a negative effect on all four log expected moments, interaction effects with the two variables indicate that as both take on larger values at the same time, the coefficients take a positive effect, partially opposing the coefficient of the original negative effect. Thus, this may indicate that the proportion of Re and Fr together, with them either being both high or both low, mitigates the negative association on the log moment of a negative Re or a negative Fr on its own.

With the models for moments 2, 3, and 4, we fit a spline model on the St variable. From the plots above, we can clearly see that the effects of St on the respective response variables at each moment are non linear. Using this spline fit on St, we are able to better capture the non-linear effects of St, while also considering the different interaction effects that are possible.

We must take into consideration the effect of St for each moment. In moments 2, 3, and 4, we can see that increases in St for values between 0.0 and 0.3 have an increasing effect on the respective response variable at that moment, any increase in St past ~0.3 starts to have an increasingly larger negative effect on the response. However, we see a different pattern for moment 1 in our least squares regression model as St has a positive effect on the response for all increases in St.

The low test errors and the large adjusted R^2 values for all four of our chosen models indicate high predictive power for our models. For our first model, we also achieve high interpretative value, due to being able to isolate effects from the feature coefficients. One trade-off, however, for larger moments was that in using the smoothing spline GAMs models automatically conducted by R, we lose some interpretability in order to minimize test MSE of our predictions. However, due to the real world applicability of our models and the results we obtain, we decided to prioritize predictability over interpretability for the last three moments, gaining most of our inferences from the first moment.

The uncertainty of the coefficients for the higher order models is demonstrated by the corresponding R output and s(St) plots. For the first moment, all p-values for the included terms were deemed significant, except for the interaction term between Re=224 and Fr=0.3, which had p=0.1. Thus, we can conclude that the relevant features have a significant statistical effect on the particle distribution in terms of the corresponding moment, although the exact coefficients likely range around the point estimate by the outputted standard errors.

Although these are the chosen models at each moment with the lowest MSE, one factor to consider is that our testing dataset only has 18 observations. Because of this lack of testing data, it may be difficult to determine whether variance in test MSEs for different models are significant, as well as whether we would see similar test MSE results for a different combination of 18 test set observations.

Conclusion

All three predictor variables display some significant statistical effect on particle clustering. In general, larger particle characteristics (i.e. size, density) lead to higher first moments. Higher fluid turbulence (Re) and gravitational acceleration lead to lower values of all four moments, but this is partially mitigated if they are both high as evidenced by the coefficients of the interaction effects, rather than one of these two values being individually high. Furthermore, in higher moments we see evidence of non-linearities between the variables.