Case Study

Abdel Shehata, Jeremiah Hodges, Francis de Beixedon

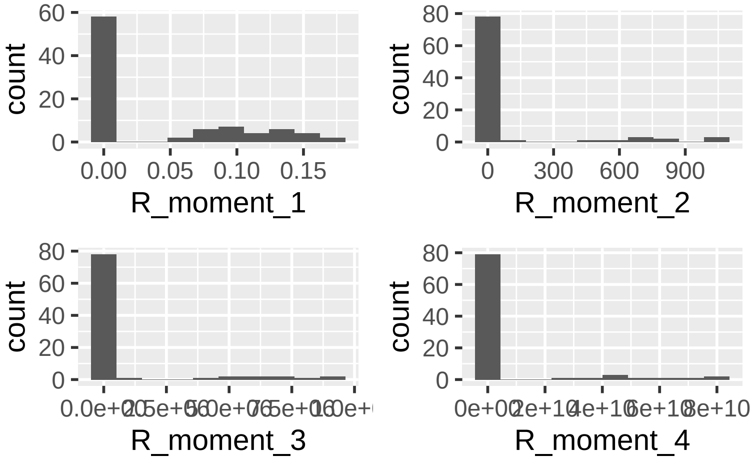
Introduction

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



Upon viewing the data, we encountered several challenges. First the Re and Fr variables only have three unique observed values. For Re we see low (90), medium (224), and high (398) and for Fr we see low (.052), medium (.3), and high (infinite). We decided against treating these variables as categorical because we want our models to be used for extrapolation to minimize the need for expensive mathematical modeling in the future for new values of these predictors. To support this goal, Fr needed to be transformed so it can be machine-readable. We decided to do a logistic transformation on Fr (creating a new variable called TFr) in order to approximate the effects of infinity. Values approaching -3 are roughly zero and values approaching 11.5 are roughly infinite. Additionally, the moment data observations are all raw, so we decided to centralize the second through fourth moments. The first raw moment is useful for interpretation because it tells us about the average amount of turbulence we can expect. However, when it comes to the shape of the probabilistic distribution of turbulence (variance, skewness, and kurtosis) we need to centralize the moments in order to interpret them because the raw moments include locational information (e.g. the first moment). Lastly, we observed that the large majority of the observations of all moments are essentially zero. This means that in most simulations, little clustering was observed. This influenced many models by making coefficients on certain predictors negative. Also, because of the lack of dispersion in response variables, there is probably nonlinearities in the relationships between the three predictors and the responses, and there may be interactions between them where are specific thresholds of values, clustering increases dramatically.

Methodology

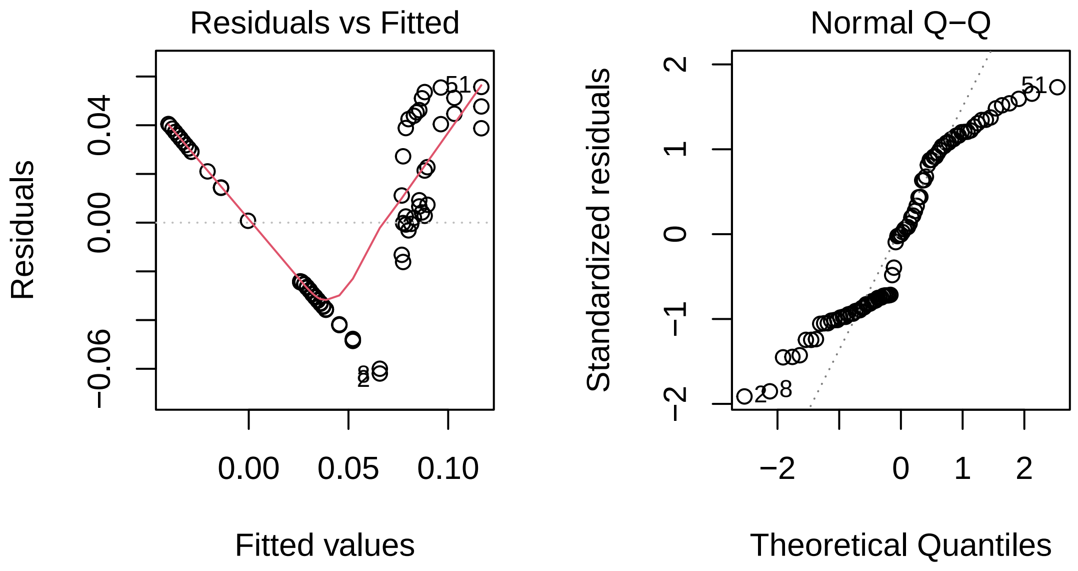
We began by fitting linear models to help us form intuitions to guide more complex and accurate models that will be more useful for prediction and more detailed interpretation. We fit one set of linear models based off of all the given predictors and another set that included all possible two-way interactions. We used bidirectional selection to narrow down variables to only include the most effective ones. Then, we compared the plain linear models to the ones with interactions. | 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 | | It seems that the interactions are increasingly helpful when explaining variation (by the Rˆ2 value). Adding interactions to the first moment model does not help improve fit much. However, for the third through fourth moments, there is a pretty significant increase in Rˆ2 when comparing the strictly linear models versus the ones with interactions. Thus, to improve model fit we should pay attention to interactions and nonlinear relationships between the predictors and response variables.

Result for Linear Model

$$\hat{R_1} = 0.0102 + 0.01353 * St - 0.0003798 * Re$$

This very simple linear model with only two out of the three predictors explains about 62% of the variation of the first moment. The Reynolds number coefficient is small and negative, which contradicts physics theory (Britannica). I believe this is due to the fact that the overwhelming majority of observations had small mean turbulence, so the regression fit a line with negative slope. On average, we just do not often observe turbulence no matter what predictors are used. However, the coefficient on St is slightly larger and positive. I believe this shows that perhaps the most important contributor to increases in the first moment is the size of the particles. In fact, adding interactions or the Fr predictor did not change the Rˆ2 very much, so I believe that St is very important for increasing average turbulence and does so in a linear way. Intuitively, this appears to make sense. Larger particles have a greater chance of bumping into each other and clustering. This regression tells us that the increase in size perhaps increases the chance of bumping and clustering with a constant, linear effect.

Model Evaluation (Linear)



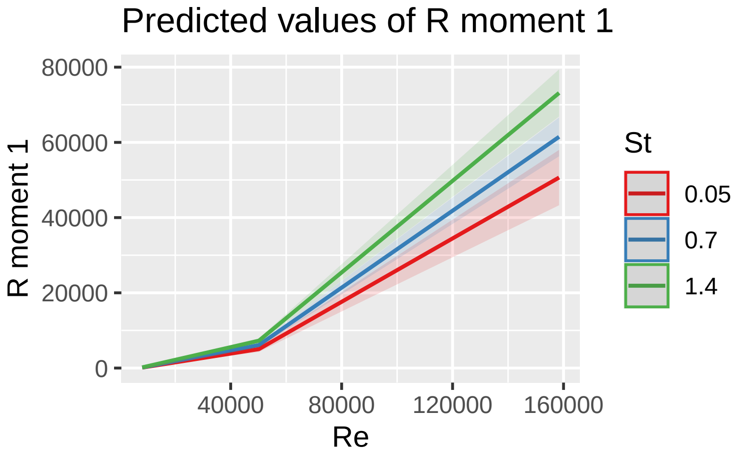
Nonetheless, there is a clear pattern to the residuals plot. First we underestimate, then overestimate, then underestimate again. This is evidence of a potential nonlinear relationship between the variables and the predictors which we will explore next.

Methodology for Complex Models

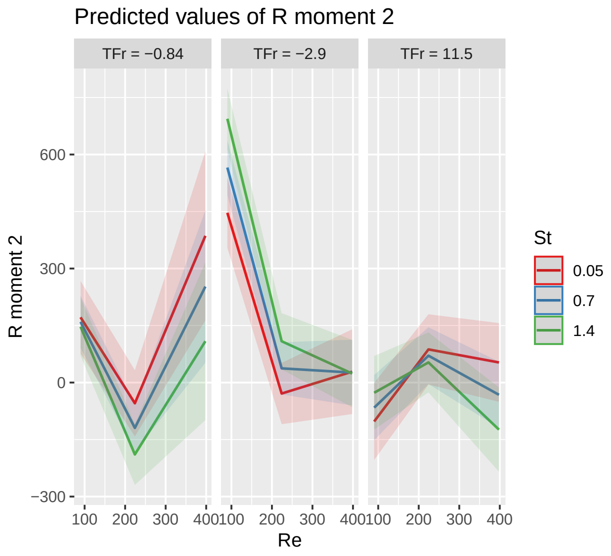
We started with a "full model" that includes the maximum number of two- and three-way interactions between TFr, Re, and St (including up to degree 2), fitted using least squares regression. Because two of the predictors have few unique observations and St already seems to have a linear relationship to the moments (as discovered through our linear regressions), we used only a 2nd degree polynomial. 2nd degree polynomials only need a minimum of 3 unique values to solve. Because the polynomial is relatively low degree, we are not worried about erratic behavior of the model and did not use a spline or smoothing method (when we did attempt to use these, they did not offer much better performance in terms of Rˆ2 and error reduction). To increase our flexibility and capture nonlinear behavior between interactions, we considered many interactions, up to degree two. Nonetheless, from the "full model", we utilized a sequential forward stepwise selection method to select only the variables that decrease mean squared error (estimated by LOOCV) until we believe any added variables overfit the model.

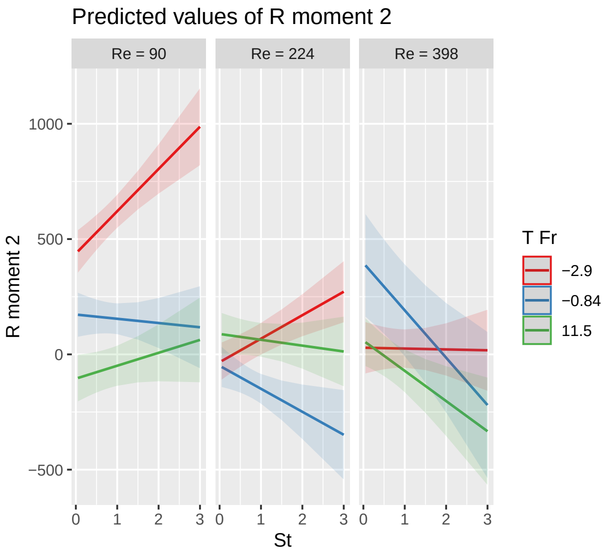
Model Evaluation

$$\hat{R_1} = 0.01882 - 0.0001250 * Re + 0.0000001993 * Re^2 + 0.006143 * St - 0.00004195 * (St * Re) + 0.00000006666 Re^2 * St$$



This plot of the effect of Re on the first moment using the nonlinear model with interactions sheds light on the relationship between Re and the first moment. First, as we determined with the simple linear model, St has a positive relationship with the first moment. The impact of Re on the first moment is higher at each level of Re if St increases. However, this plot shows that as Re grows beyond roughly 50000, its marginal influence on the first moment increases. Thus, at very high values R seems to have a nonlinear relationship to the average expected amount of clustering.



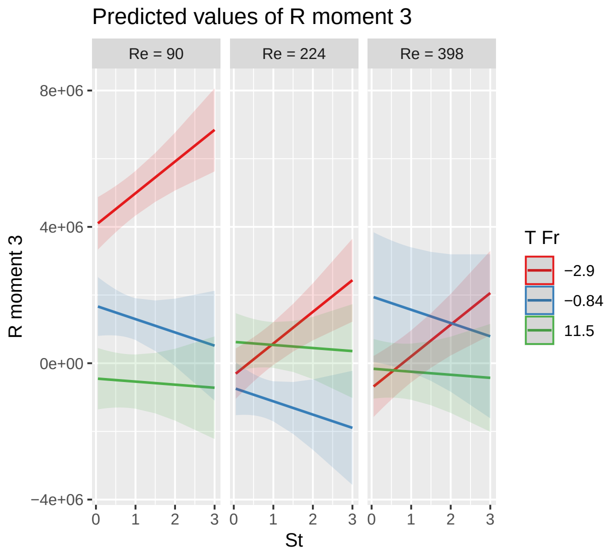


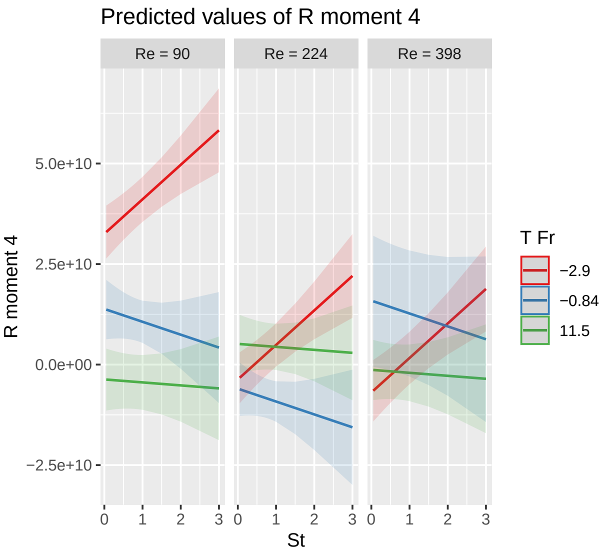
$$\hat{R_2} = 451.9 - 5.346Re - 167 * TFr - 28.38St + 0.00381Re^2 + 8.527TFr^2 + 0.7336(Re * TFr) - 0.6071(Re * St) - 70.9(TFr * St) + 7.216(St * TFr^2) - 0.0000144(TFr^2 * Re^2)$$

First plots: These plots show us a relationship that later holds true in the second and third moments as well. First, we can see by the blue and green lines that regardless of the level of Re and St, higher levels of Fr decrease the variance of clustering. We saw this relationship in the previous set of plots. The most interesting line to observe is the red one that shows what happens to variance as St changes at near-zero Fr. As Re increases, the effect of St converges to essentially nothing. Combined with what we learn about the first moment (that at extremely high levels of Re, average clustering increases), perhaps this means that Re becomes the most important determinant of clustering as it reaches high levels. At very high Re, we may see consistent clustering regardless of the other variables. Nonetheless, if Re is low (90 or 224), St contributes to increases in variance. Thus, at low Re and Fr, not only does St increase average clustering (as we observed in the linear models), it does so with a great deal of variance. This probably reflects the fact that particles colliding does increase in frequency as they grow in size, but collisions are very unpredictable and random events; sometimes they lead to clustering and other times they do not. Second plots: From these plots of the second model, we can first notice a significant amount of overlap of error bounds in all three plots regardless of the level of St. Thus, when it comes to variance, the influence of the size of the particles is not particularly important. We can also see that the convergence of the fits tightens at TFr = 11.5 (Fr = inf). Because Fr = u/sqrt(gL), it makes sense that u (flow velocity) must be relatively large for Fr to be infinite (find any source). At high rates of flow may consistently break up the clusters and thus preventing the occasional examples of high levels of clustering. We also see that high Fr may limit the influence of Re. At near-zero Fr (TFr = -.84), high Re has a strong positive relationship with variance. Additionally, Re seems to exhibit some kind of thresholding behavior past Re = 224 where its relationship with variance reverses past this point. Re = 224 appears to be a special value worth studying more in the future.

$$\hat{R_3} = 4.236*10^6 - (1.429*10^6)TFr - (8.149*10^5)St - (4.711*10^4)Re + (7.72*10^4)TFr^2 + (1.093*10^2)Re^2 - (4.678*10^5)(St*TFr) + (5.905*10^3)(Re*TFr) + (4.616*10^4)(St*TFr^2) - 1.134(TFr^2*Re^2)$$

$$\hat{R_4} = 3.504*10^10 - (1.144*10^10)TFr - (7.011*10^9)St - (3.864*10^8)Re + (6.062*10^8)TFr^2 + (8.955*10^6)Re^2 - (4.185*10^9)(St*TFr) + (4.847*10^7)(Re*TFr) + (4.133*10^8)(St*TFr^2) - (9.301*10^3)(TFr^2*Re^2)$$





Once again, we see essentially the same relationship between St and the third moment as we do between St and the second moment. Thus, higher Fr and higher Re decreases skewness. However, the slope of the red lines is always positive. Thus, in this case, St appears to lengthen the right handed tail even as variance decreases due to Fr and Re. Thus, we can see that in a limited way, particle size increases the chances of rare high-clustering events within certain fixed conditions (where variance is limited overall by Fr and Re). The relationship between the predictors and kurtosis essentially seems identical to their relationship with skewness. Nonetheless, the Re = 398 plot is interesting because in this same plot for the second moment, we saw the effects of all variables converge to zero. However, St always has a positive effect on kurtosis as long as Fr is low. Thus, the size of the particles still makes the tails of the probability distribution of clustering heavier. Similar to our analysis of St's relationship to the third moment, it seems to increase the spread of the distribution within bounds set by the other parameters. Thus, as the other parameters make tail weight smaller overall, higher St makes them as small as possible within those bounds. I believe that this makes sense because Fr and Re seem to be more related to the environment than St, which has to do with the particles themselves. Given certain environmental parameters that limit the amount of variance overall, larger particle size will always increase the chance for rare events of high levels of clustering within the bounds of possibility.

Conclusion:

In conclusion, we have learned that St and the first moment have a mostly positive and linear relationship, although predictive accuracy can increase if we use a more complex model that includes interactions. Through the model with interactions, we saw that extremely high values of Re will increase average clustering. The other moments are best fitted with models that involve complex interactions and exponential terms due to their inherent nonlinearity. In general, high levels of Fr and Re decrease variance, skewness, and kurtosis. These parameters seem to make the results of the simulation more regular and consistent. On the other hand, St pushes for more positive skewness and kurtosis. I believe that this probably has to do with the inherently chaotic and unpredictable nature of collisions between particles. Lastly, with regards to variance, we saw Re = 224 exhibits some kind of thresholding behavior because its effect on variance switches at that point. In addition to this mystery, the specific relationship between Re and Fr is also worth further consideration. Intuitively, as Fr increases I would expect clustering behavior to become more regular because high flow probably cuts short any particularly unusual behavior of particles. However, Re when combined with Fr also seems to concentrate the probability of clustering around the mean. Theoretically, I would expect Re to both increase mean clustering and lead to less predictable behavior. Perhaps our results mean that high levels of Re in fact leads to more clustering, albeit in very consistent ways.