Case Study: Turbulence

Julia Rosner, Emily Mittleman, Yuzhe Gu, Ashley Chen

2022-11-3

Introduction

Understanding and predicting turbulence in fluid motion is incredibly important to a vast range of problems, such as air pollution, population dynamics, and weather. However, turbulence looks random, irregular, and unpredictable, making it difficult to understand. In order to achieve better insights into understanding and predicting particle clusters in turbulence, we will investigate the effects of three parameters representing properties of turbulent flow and particles on the first four raw moments of the probability distribution for particle cluster volumes. We will be exploring how fluid turbulence (quantified by Reynolds number or Re), gravitational acceleration (quantified by Froude's number or Fr) and particles' characteristics (for example size and density which is quantified by Stokes number or St) affect the spatial distribution and clustering of particles in an idealized turbulence.

Thus, our goals are as follows: For a new parameter setting of (Re, Fr, 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.

We created two models; one is a more simple model whose key purpose is for inference to interpret the effect of Re, Fr, and St on particle cluster volume distribution. The other model is more complex, which allows for greater predictive performance in predicting particle cluster volume distribution for a new parameter setting of (Re, Fr, St).

Methodology

Inference 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. We did not change the first raw moment to the central moment, because the mean of the distribution conveys more meaningful information than the central moment, which is always zero. However, we converted the second, third, and fourth raw moments to central moments for inference so that our results are more interpretable.

In addition, despite their numerical natures, Fr and Re only contain three levels ($Fr \in \{0.052, 0.3, \infty\}$ and $Re \in \{90, 224, 398\}$). We decided to make them categorical variables for our inference model for 2 main reasons. Firstly, treating them as numeric variables puts our model at risk for extrapolation due to lack of data at many levels of Re and Fr, making our model unable to learn the trends around such regions; and secondly, we believe that these categories could carry real life significance. For instance, Fr = 0.3 is representative of cumulonimbus clouds and 0.052 is representative of cumulus clouds. Focusing on observations collected at such specific levels may lend unique insights into practical problems.

A closer examination of the data revealed interesting interactive patterns among the independent and dependent variables. Specifically, St appears to assume a strong, non-linear relationship with each of the moments, which appeared logarithmic, so we log-transformed St. Also, while treating Fr variable as categorical, we observed that there is a linear and decreasing relationship between gravitational acceleration and R moments. Lastly, there is a linear, decreasing relationship between Reynolds number and R moments. We also explored multicollinearity through VIFs for each model, which were very low.

Our goal for inference was to build a model that was both accurate and interpretable. In testing different models, we found that the linear model, while easy to interpret, performed quite poorly. On the other hand, the spline performed quite well, and ever better than the cubic polynomial, but was difficult to interpret. We compared different degrees of polynomial regression models using the validation set approach, and found that for all four moments, the test MSE began to flatten at the third degree. While higher degree polynomial models performed slightly better, we choose degree three avoid the issue of overfitting. Plots of our model predictions on the test set show that the predicted curves fit well to the testing points. We also performed 5-fold cross validation, which showed fairly low RMSE values and high R^2 values, further increasing confidence in the model's performance. Our final models for each of the moments for inference is a cubic polynomical with interaction terms between Re and Fr:

$$\begin{aligned} log(R_{moment1}) = &-2.048 + 2.002(log(St)) + 0.161(log(St))^2 + 0.446(log(St))^3 - 3.824(Re_{224}) - 6.015(Re_{398}) \\ &-0.286(Fr_{0.3}) - 0.317(Fr_{\infty}) + 0.238(Re_{224} * Fr_{0.3}) + 0.389(Re_{224} * Fr_{\infty}) + 0.504(Re_{398} * Fr_{\infty}) \end{aligned}$$

$$\begin{aligned} log(C_{moment2}) = &6.005 + 9.517(log(St)) - 4.885(log(St))^2 + 3.907(log(St))^3 - 7.520(Re_{224}) - 11.673(Re_{398}) \\ &-6.679(Fr_{0.3}) - 6.562(Fr_{\infty}) + 4.612(Re_{224} * Fr_{0.3}) + 4.629(Re_{224} * Fr_{\infty}) + 7.157(Re_{398} * Fr_{\infty}) \end{aligned}$$

$$\begin{aligned} log(C_{moment3}) = &14.567 + 14.025(log(St)) - 8.234(log(St))^2 + 6.315(log(St))^3 - 11.292(Re_{224}) - 17.490(Re_{398}) \\ &-12.888(Fr_{0.3}) - 12.625(Fr_{\infty}) + 8.726(Re_{224} * Fr_{0.3}) + 8.632(Re_{224} * Fr_{\infty}) + 13.498(Re_{398} * Fr_{\infty}) \end{aligned}$$

$$\begin{aligned} log(C_{moment4}) = &23.168 + 18.0035(log(St)) - 22.202(log(St))^2 + 8.469(log(St))^3 - 15.075(Re_{224}) - 23.325(Re_{398}) \\ &-19.018(Fr_{0.3}) - 18.619(Fr_{\infty}) + 12.755(Re_{224} * Fr_{0.3}) + 12.576(Re_{224} * Fr_{\infty}) + 19.746(Re_{398} * Fr_{\infty}) \end{aligned}$$

Prediction One of our goals for this case study was to make a model that is specialized in making accurate predictions, meaning that we are not concerned with inference for this particular model. Because we are not concerned with interpreting this model, we explored more complex models because they tend to have better predictive performance than simple models.

Despite certain advantages in modeling Re and Fr as categorical variables, we recognize the need to extrapolate beyond the three levels of Re and Fr to model real life circumstances and enable a wider range of prediction. In order to predict values for new observations between and/or outside the ranges of values we have seen, we had to make a model that could both interpolate and extrapolate. Instead of using categorical values for our predictor variables Fr and Re, we used numerical values so that our model is more generalizable to accept numerical values between or outside the ranges of Fr and Re that were given in the training data. In order to make Fr a numerical variable, we needed to perform a transformation on it to make Fr take on finite values for modeling. To do this, we transformed Fr so that the new value is equal to $1 - 2^{-Fr}$, so that the new Fr takes on values between 0 and 1.

After exploring linear and polynomial models both with and without interaction terms, we were still not satisfied with the accuracy of these models. Since the performance of our simpler models still showed much room for improvement, we further explored a Generalized Additive Model (GAM) using cubic splines with an interaction between Fr and Re. We expected the splines to perform better compared to polynomial regression because instead of fitting a single polynomial function to the entire dataset, spline interpolation fits a piecewise continuous function composed of many polynomials to model the data set. Splines give a more flexible model when the true regression function changes rapidly in certain regions but not others, which is why we expected a GAM using splines to perform better than polynomial models.

We did analysis on GAM models using cubic splines with and without interaction terms, and as we expected, the model including the interaction between Fr and Re outperformed the model with no interaction. We did an 80/20 training-test data split in which we trained the model on 80% of our data, and then validated it on the remaining 20% of unseen data to evaluate its predictive performance. In the graphs of the model's predictions, we saw that the trained model was able to approximate the new data extremely well, even when the three predictors were far away from any of the observations the model was trained on. This gave us confidence in our model that it is able to predict new datapoints very well without being overfit, and so we selected this GAM model to continue with.

To analyze the predictive performance of our model, we performed 5-fold cross validation on our GAM model using cubic splines with an interaction between Fr and Re. After confirming using 5-fold cross-validation that our model outperformed any other model we created thus far, as it had extremely high R^2 values and it produced lower test MSEs for all four R moments as compared to any other models, we trained our final model. We obtained the final model by training the GAM on all of our training data found in data-train.csv. We conclude our search for and training of our predictive model, and move on to discuss the results we found from our final predictive GAM model.

Results

Inference Results From our inference models, we can gain insights into how Re, Fr, and St affect the probability distributions for particle cluster volumes. We notice that a larger Reynolds number leads to lower predictions for all four moments. As the ratio of fluid momentum force to viscous shear force increases, the mean of the distribution would decrease, the variance would decrease, the distribution will skew more to the right, and the tail will be lighter. This is in line with our knowledge that higher Reynolds numbers correspond to greater turbulent behavior. Likewise, an increase in gravitational acceleration decreases all four moments.

Fr and Re alone both have a negative effect on all four log expected moments, but, the interaction effect of the two variables indicates that as they take on larger values at the same time, their interaction coefficient causes them to take a positive effect instead, partially opposing the coefficient of the original negative effect. While interaction effects are usually less interpretable, in this case it makes sense to keep them in the final model. This is because high speeds cause turbulent flow, so the effect of turbulence (Re) on the distribution is different depending on the gravitational acceleration.For raw moment 1, this suggests that smoother flow at any level of gravitational acceleration may allow larger clusters to form, as they are not broken up by chaotic turbulence. For central moments 2, 3, and 4, this suggests that smoother flow with at any gravitational acceleration seemed to have a positive association with variance, skew, and kurtosis. Finally, the effects of particle size on each moment seems to significantly increase as our moment number increases.

Sources of uncertainty include that Fr and Re were used as categorical variables in our models despite that they take on numeric values in practice. This approach limits the generlizability of our model. Lastly, no observations in the training set had both Re = 398 and Fr = 0.3, so that interaction term was not calculated by R. This could limit the application of our models at observations with those values.

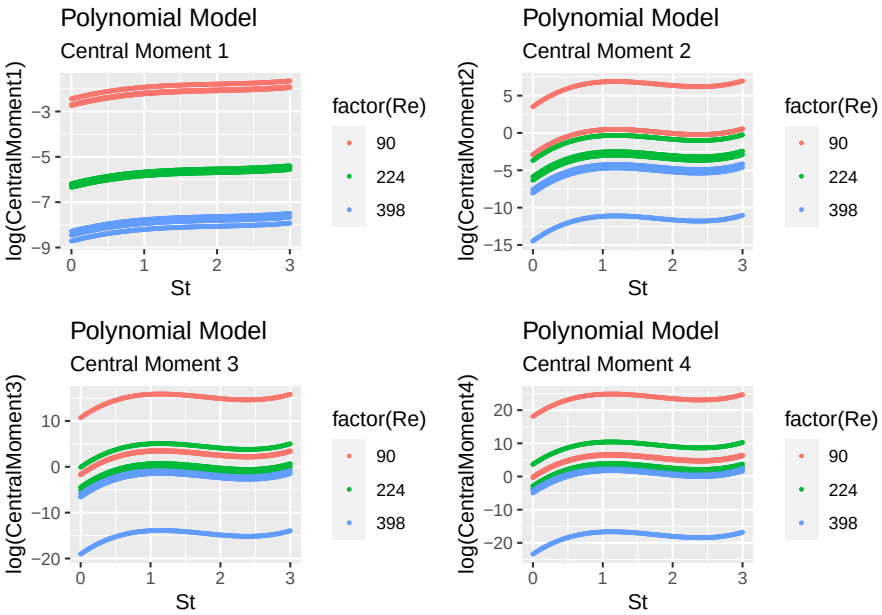


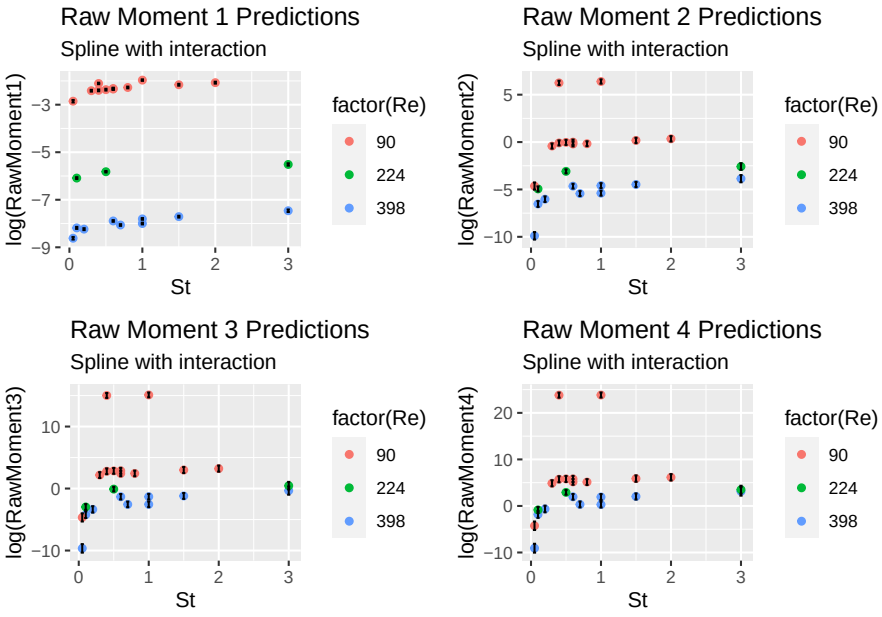
Table 1: Cubic polynomial 5-fold CV

	Moment 1	Moment 2	Moment 3	Moment 4
RMSE	0.098	0.742	1.154	1.827
Rsquared	0.998	0.968	0.951	0.940
MAE	0.069	0.451	0.699	1.026

Prediction Results We used our final predictive GAM model with cubic splines and an interaction between Fr and Re to make predictions for the hold-out testing dataset. When making predictions for the four raw moments based on the three predictors in the testing dataset, we also calculated the standard error of each prediction to find out how uncertain the model is about each of these predictions. The table below shows the maximum standard error for each moment, which are all less than 1. This means that even when the model is the most uncertain about the prediction it makes for a raw moment, it is still relatively confident in its prediction since the standard errors are so low across all predictions for all moments. The graphs below show the standard error bars on top of the points that the model predicted for the test dataset; here we see the same as what our table showed, where the black standard error bars are so small that they don't even extend past the bounds of the point it estimated. Our model has very, very low uncertainty about the points it predicts in the test dataset, which further leads us to believe that our model has very strong predictive abilities. Even though the standard errors are very small, between the moments themselves, the errors are slightly larger for the third and fourth moments compared to the first and second. This is expected because higher moments measure more innate features of the volume distribution and are harder to predict, so it would make sense that the models for the third and fourth moments have slightly more uncertainty.

Table 2: Maximum standard error for validation set predictions

Moment 1	Moment 2	Moment 3	Moment 4
0.054	0.392	0.685	0.948



To evaluate our model's predictive performance, we did an 80/20 training-test data split in which we trained the model on 80% of our data, and then validated it on the remaining 20% of unseen data to evaluate its predictive performance. In the graphs below of the model's predictions, we saw that the trained model was able to approximate the new data extremely well, even when the three predictors were far away from any of the observations the model was trained on. To further assess predictive performance of our model, we performed 5-fold cross validation to find RMSE, R squared value, and MAE across all four raw moments

which is shown in the table below. In an ideal model, RMSE divided by MAE equals 0, and R^2 equals 1, and it can be seen in the table below that our model achieved values very close to these ideals, which speaks to the accuracy of our predictions.

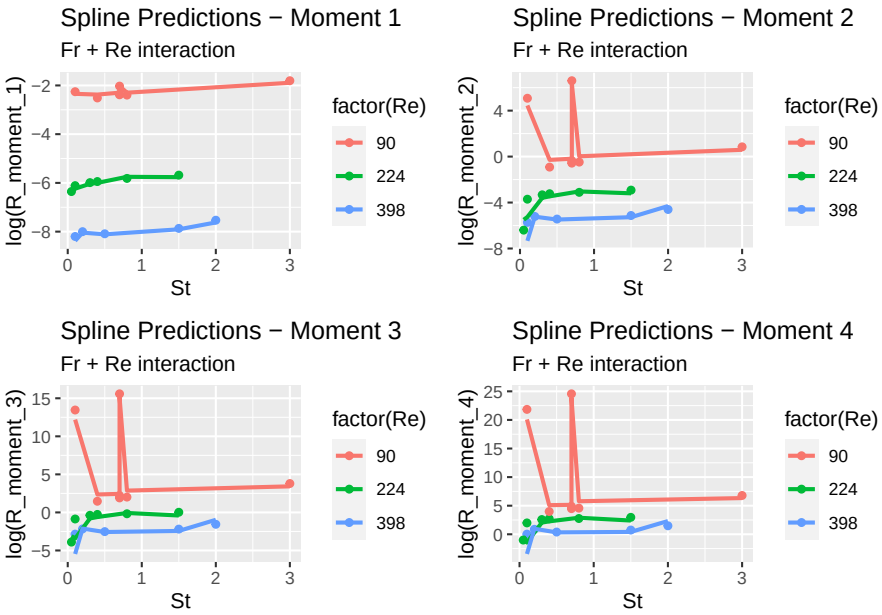


Table 3: Cubic polynomial 5-fold CV

	Moment 1	Moment 2	Moment 3	Moment 4
RMSE	0.098	0.740	1.827	0.098
R^2	0.998	0.968	0.940	0.998
MAE	0.069	0.450	1.026	0.069

Conclusion

In this study, we constructed four models for inference and four models for prediction for the first raw moment and the second, third and fourth central moments (related to mean, variance, skew, and kurtosis) of particle cluster distribution. Our four inference models has polynomial terms and interaction terms between Reynolds' number, Froude's number, and Stokes' number. For our prediction models we used splines constructed Our models were able to explain a large percentage of variation in each of the four moments with reasonable mean squared prediction error for each model. In general, 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. Lastly, St is associated with higher values for all four moments.