Analyzing and predicting particle clustering in turbulence

Yihan Shi, Edna Zhang, Medy Mu

2022-11-07

Introduction

"We are experiencing some turbulence, please fasten your seat belt." Many of us might have heard this radio on the plane and felt bumpy. When we mix paint in water, we can also observe turbulence as the color dissipates. Turbulence is so common and easily observed in daily life, yet its causes and effects are hard to predict. In fluid dynamics, turbulence is "characterized by chaotic changes in pressure and flow velocity". With some knowledge and observation in parameters such as fluid density, flow speed, and the property of particles that cluster inside turbulent flows, we can gain insights into the distribution and clustering of particles in idealized turbulence. In this case study, we will investigate 3 observed features that might contribute to particle distribution in turbulence: Reynolds number (Re), which takes flow speed, viscosity, and density into account; Gravitational acceleration (Fr); Stokes number (St) that quantifies particle characteristics like size, relaxation time, and particle diameter. We use these 3 features to build machine learning models. From the results of multiple polynomial regression, we are able to explain changes in particle distribution; using more complex models such as natural spline and generative additive models, we can extrapolate new data beyond the scope of the known observations.

Methodology

Data Preprocessing

We found that all the variables are numeric. St has a slightly right-skewed distribution, therefore we applied log transformation in a few models to make the distribution less skewed. Re only has 3 possible values in data-train.csv, however, in order to make the model extrapolate better, we decided to keep Re as a numeric rather than a categorical variable. To deal with Inf values in Fr, we applied logit transformation with the below formula.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$$f(x) = \text{output of the function}$$

$$L = \text{the curve's maximum value}$$

$$k = \text{logistic growth rate of steepness of the curve}$$

$$x_0 = \text{the x value of the sigmoid midpoint}$$

$$x = \text{real number}$$

We used 25 as L, because the maximum value of gravitational acceleration in practice is 25 according to research (Ponce & Diaz, 2016). We used 0.3 as midpoint x0, since that's the median value of Fr in our training and testing sets. Then we picked 2 as k, it's a number we picked that produced reasonable transformed Fr values.

We calculated central moments with raw moments in data-train.csv. Then we used both the 4 raw moments and 3 central moments (excluding the 1st central moment as it's 0) as response variables for the models. Through plotting histograms, we found that all the raw moments and central moments have skewed distributions, therefore we applied log transformations to all of the response variables.

Model evaluation

We used 2 metrics to evaluate our models. First, we calculate R-Squared values, which represents the proportion of the variance for a dependent variable that's explained by independent variables in regression models. It's a good measure of how well our model is fitting the training data. Second, we use 5-fold cross validation. It randomly divides the data into 5 equal-sized parts, trains the model on 4 of them and evaluates the model on the left out part. This is done on each part and we take the average of the mean squared errors. This is a good way to estimate the test error, and how well the model can generalize to unseen data.

1) Linear regression models: The linear models didn't perform very well except the linear model for raw moment 1. Most adjusted R-Squared values are around 0.5. All the coefficients are statistically significant, showing that we should include all 3 independent variables. However, the Residual v.s. Fitted and Scale-Location plots clearly demonstrate nonlinearities, thus we need to increase model complexity to address this.

formula	CVmse	adj.R2
log(R_moment_1) ~ logit(Fr) + Re + St	1.426200e-03	0.9282684
log(R_moment_2) ~ logit(Fr) + Re + St	7.460627e+04	0.6270661
log(R_moment_3) ~ logit(Fr) + Re + log(St)	5.300000e+12	0.5201657
log(R_moment_4) ~ logit(Fr) + Re + St	3.710000e+20	0.4347858
log(C_moment_2) ~ logit(Fr) + Re + St	7.379673e+04	0.6279030
log(C_moment_3) ~ logit(Fr) + Re + log(St)	1.563466e+04	0.5115609
log(C_moment_4) ~ logit(Fr) + Re + St	5.226819e+09	0.5247228

2) Linear regression models with interactions: We've added all the possible combinations of interaction effects and found the interaction between Re and logit(Fr) to be the most statistically significant. Therefore we added this interaction term to our original linear regression models. We found that the adjusted R-Squared values all increased, and the cross validation MSE all decreased compared to linear regression models. The red lines in Residual v.s. Fitted and Scale-Location plots got closer to straight lines.

formula	CVmse	adj.R2
log(R_moment_1) ~ logit(Fr) + Re + St + logit(Fr) * Re	1.519200e-03	0.9312442
log(R_moment_2) ~ logit(Fr) + Re + St + logit(Fr) * Re	7.083789e+04	0.7131291
log(R_moment_3) ~ logit(Fr) + Re + log(St) + logit(Fr) * Re	5.270000e+12	0.6251003
log(R_moment_4) ~ logit(Fr) + Re + St + logit(Fr) * Re	3.710000e+20	0.5621174
log(C_moment_2) ~ logit(Fr) + Re + St + logit(Fr) * Re	7.042079e+04	0.7079129
log(C_moment_3) ~ logit(Fr) + Re + log(St) + logit(Fr) * Re	7.639658e+03	0.5790335
log(C_moment_4) ~ logit(Fr) + Re + St + logit(Fr) * Re	1.268310e+09	0.5916249

3) Polynomial models: We attempted another way of adding model complexity by using polynomial models. We selected the optimal polynomial degree in each model through cross validation. Compared to the linear regression models with interactions, adjusted R-Squared values further increased, and cross validation MSE values further decreased. All the models achieved an adjusted R2 of over 0.84, showing that our models are fitting the train data well.

formula	CVmse	adj.R2
log(R_moment_1) ~ logit(Fr) + Re + poly(St, 2) + logit(Fr) * Re	8.737000e-04	0.9347311
log(R_moment_2) ~ poly(logit(Fr), 2) + Re + poly(St, 5) + logit(Fr) * Re	4.606062e+04	0.8986440
log(R_moment_3) ~ poly(logit(Fr), 2) + Re + poly(log(St), 3) + logit(Fr) * Re	4.380000e+12	0.8906483
log(R_moment_4) ~ poly(logit(Fr),2) + Re + poly(St, 2) + logit(Fr) * Re	3.400000e+20	0.8529320
log(C_moment_2) ~ poly(logit(Fr), 2) + Re + poly(St, 2) + logit(Fr) * Re	4.549100e+04	0.8733799
log(C_moment_3) ~ poly(logit(Fr), 2) + Re + poly(log(St), 3) + logit(Fr) * Re	4.014669e+03	0.8439009
log(C_moment_4) ~ poly(logit(Fr),2) + Re + poly(St, 2) + logit(Fr) * Re	8.920932e+08	0.8544865

4) Natural spline models: According to Runge's phenomenon, for certain non-polynomial regression functions, higher-order polynomials (3rd degree and 5th degree in our polynomial regression models) can give poor (highly oscillatory predictions) near domain boundaries. To prevent this danger of polynomial modeling, we used a natural spline on St and selected the optimal degrees of freedom based on cross validation. However, almost all the models (except the model for raw moment 1) had lower values of adjusted R2 and higher values of cross validation MSE. We reasoned that this is because the nonlinearity of logit(Fr) is not addressed in these models.

formula	CVmse	adj.R2
log(R_moment_1) ~ ns(St, df = 7) + logit(Fr) + Re + logit(Fr) * Re	7.587000e-04	0.9429966
log(R_moment_2) ~ ns(St, df = 6) + logit(Fr) + Re + logit(Fr) * Re	7.116176e+04	0.7723868
log(R_moment_3) ~ ns(log(St), df = 3) + logit(Fr) + Re + logit(Fr) * Re	5.280000e+12	0.6592444
log(R_moment_4) ~ ns(St, df = 6) + logit(Fr) + Re + logit(Fr) * Re	3.680000e+20	0.6348304
log(C_moment_2) ~ ns(St, df = 6) + logit(Fr) + Re + logit(Fr) * Re	7.119605e+04	0.7688989
log(C_moment_3) ~ ns(log(St), df = 3) + logit(Fr) + Re + logit(Fr) * Re	8.088620e+03	0.6022442
log(C_moment_4) ~ ns(St, df = 6) + logit(Fr) + Re + logit(Fr) * Re	1.378697e+09	0.6381364

5) Generalized additive models: Therefore we used generalized additive models, as they can fit separate nonlinear functions for each predictor and allow us to investigate and interpret the nonlinear effects in each variable. Notice here that we've included interaction terms in our generalized additive models, which help us produce better prediction results despite being hard to interpret. We found that the generalized additive models produced very promising results, with all the adjusted R2 values larger than 0.88, and lower cross validation MSE compared to all previous models. Therefore, we used generalized additive models to produce predictions of data-test.csv.

formula	CVmse	adj.R2
log(R_moment_1) ~ s(St, 13) + Re + logit(Fr) + logit(Fr) * Re	7.587000e-04	0.9429966
log(R_moment_2) ~ ns(St, 4) + Re + ns(logit(Fr),2) + Re * ns(logit(Fr),2)	3.848408e+04	0.9190544
log(R_moment_3) ~ ns(log(St), 3) + Re + ns(logit(Fr),2) + Re * ns(logit(Fr),2)	3.790000e+12	0.9188423
log(R_moment_4) ~ St + Re + ns(logit(Fr),2) + Re:ns(logit(Fr),2)	2.890000e+20	0.8850566
log(C_moment_2) ~ ns(St, 4) + Re + ns(logit(Fr),2) + Re * ns(logit(Fr),2)	4.449682e+04	0.9156065
log(C_moment_3) ~ ns(log(St), 3) + Re + ns(logit(Fr),2) + Re * ns(logit(Fr),2)	1.511269e+03	0.9626722
log(C_moment_4) ~ St + Re + ns(logit(Fr),2) + Re:ns(logit(Fr),2)	3.220336e+08	0.9800513

Results

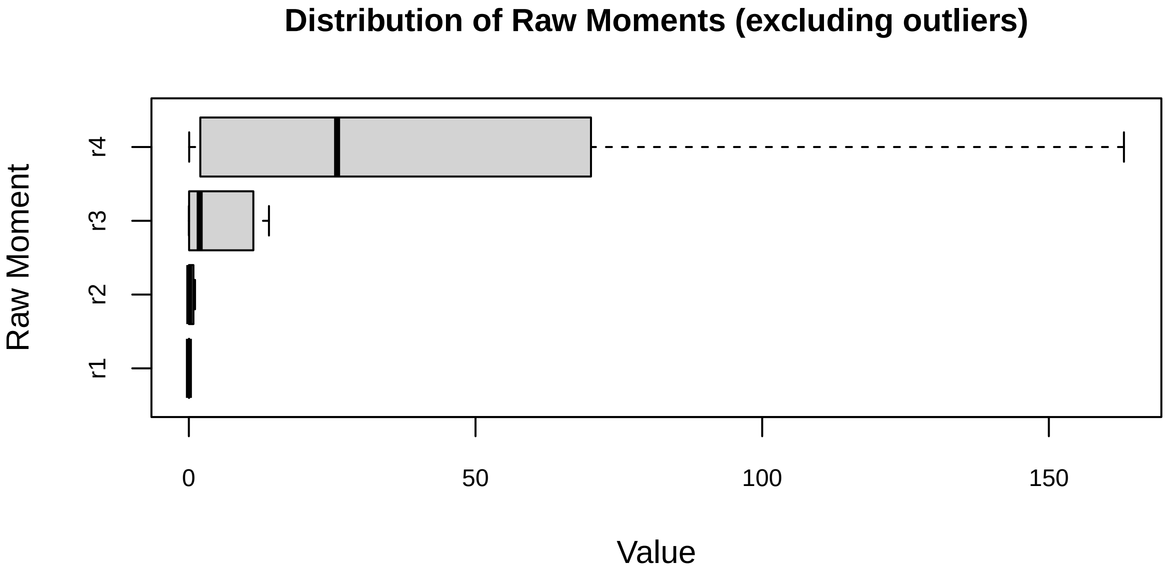
In total, seven models were fitted: 4 models on the first to the fourth raw moments and 3 models on the second to the fourth central moments. The first central moment was not fitted since it is always equal to 0 according to its formulation. Among all seven models, the generalized additive model (GAM) yields the lowest MSE and highest adjusted Rˆ2, suggesting that it outperforms all other models fitted and thus has a high predictive power for both raw moments and central moments due to its high flexibility. However, GAMs lack interpretability compared to simpler models, such as polynomial regression and least square regression. Since polynomial regression produces a better MSE and adjusted Rˆ2 than linear regression, polynomial models were chosen for statistical inference. Moreover, higher-order central moments have more interpretive meaning physically than raw moments. Indeed, while the first raw moment represents the average size of clusters, the second central moment represents the variance of clusters, and the third and fourth central moments are used to define skewness and kurtosis of clusters, respectively after standardization.

According to the final polynomial models, the intensity of turbulent flow, represented by Re, has a negative effect on the mean, variance, and kurtosis of the distribution of particle cluster volumes but a positive effect on the skewness of the distribution. In other words, as the intensity of turbulent flow increases, not only the average size and variance of clusters are expected to decrease, but the extremity of outliers are also reduced, holding all other variables constant. Moreover, the distribution of particle cluster volumes is expected to be more right skewed as the intensity of turbulent flow grows, indicating that most particle clusters have small volumes but with a long right tail of clusters that grow very large. The stronger turbulent flow, the smaller but more concentrated particle clusters with a flatter but long right tail of large particle clusters. Moreover, the effects of turbulent flow are assumed to be approximately linear in this study since adding polynomials to Re did not improve the model performance. The gravitational acceleration, quantified by Fr, and the particle characteristics, such as size and density quantified by St, however, appear to have non-linear effects on higher-order central moments. In particular, St also shows a nonlinear relationship with the first raw moment. The effects of gravity and particle inertia, therefore, have different effects on the response variable of interests compared to the effect of turbulent flow.

Indeed, as the gravitational acceleration increases, the average size of clusters is expected to shrink. The effects of gravitational acceleration on higher-order central moments, however, vary based on the value of Fr due to the addition of the quadratic term. For subcritical or slow fluid with Fr much less than 1, the variance, skewness, and kurtosis of the distribution of particle cluster volume is expected to decrease as the gravity increases. This suggests that as the gravity becomes stronger, the distribution of particle cluster volumes is expected to be less spread out with a flatter long left tail of particle clusters with very small volumes. For fast and supercritical flow with very large Fr, the effect is reversed such that as the gravity increases, the distribution of particle cluster volume becomes more spread out with a fatter long right tail of clusters that are large in size.

Similar to the effects of gravitational acceleration, the effects of particle inertia on the first raw moment and higher-order central moment differ based on the value of St. If the Stokes number is small, that is much less than 1, it means that the particle motion is tightly coupled to the fluid motion. Under this condition, as the particle inertia increases, the average size and variance of clusters are expected to increase accordingly, while the skewness and kurtosis of the distribution of particle cluster volume is expected to decrease, indicating a distribution with a flatter long left tail. If the Stokes number is large, meaning that the particles are not influenced by the fluid, as the particle inertia increases, the clusters are expected to shrink in size but with less variability across clusters. The distribution of particle cluster volumes is also expected to be right skewed with some clusters growing very large as St increases. However, the distribution still has flatter tails as St increases since the effect of the quadratic term on the fourth central moment is very small. In addition to the individual effects of Re, Fr, and St, there are statistically significant interaction effects between Re and Fr in all polynomial models. Specifically, the coefficients associated with these interaction effects are positive in all models, suggesting that the stronger the gravity is, the stronger the effects (more positive) of the intensity of turbulent flow on the response variable of interest.

For predictions, as mentioned above, GAMs were chosen to predict the first to the fourth raw moments in the test set. According to the summary statistics and boxplots, there are two outliers with very large predicted values in the second, third, and fourth raw moments. After a closer examination, the combination of the predictors of interest for these two outliers did not match any observations in the training set. Our GAM models thus might fail to accurately predict those unseen data, pointing to the potential problem with overfitting. However, there are also outliers for higher-order raw moments in the training set, specifically concentrating around Re = 90. The two outliers from the predicted raw moments also have a Re of 90, suggesting that more future work should be done to investigate why observations with a Reynolds number of 90 yield different raw moments from other observations. Moreover, all four predicted raw moments are right skewed, which is similar to the raw moments in the training data.



Conclusion

This case study explores the effect of flow intensity (Reynolds number), gravitational acceleration (Fr), and particle characteristics (Stokes number) on various properties of particle clusters, including average particle cluster size, variance of cluster distribution, and the extent of extremity of the distribution (kurtosis).

The average size of particle clusters shrink with more intensive flow and larger gravitational acceleration, the two of which have an even larger impact on cluster size jointly. While flow intensity and gravitational acceleration are easily quantifiable, the influence of Stokes number on the average size vary under different conditions because it is a dimensionless number characterizing particle behavior. In this case study, we use particle inertia to represent particle behavior. When particle motion is closely related to fluid motion (small Stokes number), the average cluster size increases as particle inertia increases. However, when particle motion is not influenced by the fluid, that relationship is reversed.

Aside from the average size of particle cluster, variance, skewness and the extent of extremity of the distribution of clusters are also important for turbulence. We found that that particle clusters tend to be smaller, more concentrated, and have moderate extremities for more turbulent flow. When the fluid speed is below a certain threshold (Fr < 1), stronger gravitation acceleration is associated with more concentrated, normally distributed cluster with moderate extremities. However, when the fluid flows faster, the cluster distribution starts to have more extreme outliers as gravitation acceleration increases. In this case, it can be harder to predict an accurate cluster distribution. Additionally, when fluid motion affect particle motion minimally, higher particle inertia is associated with less variable clusters; however, as the impact of fluid motion increases, the clusters become more variable.

The above insights into the physical characteristics of particle clusters has allowed us to make predictions of unknown distributions using GAM model. Although the GAM model might not be able to explain changes in particle clusters when flow intensity is at a specific level (Reynolds number = 90), the properties of predicted distribtion generally match the model's trend. Nonetheless, future work should investigate the behavior of turbulence at levels of certain flow intensity.