Predicting Turbulence: Cluster Analysis via Voronoi Tessellation

Dav King

2023-10-28

Introduction

Turbulence has been described as the greatest unsolved problem in physics. Its understanding is key across a number of fields, including meteorology, astronomy, and engineering. However, at this point in time, we lack both understanding of how turbulence comes about and the ability to predict when it will. The purpose of this project is to attempt to provide insights into both of those dimensions.

Although it is too resource-intensive to work as a general method for prediction, direct numeric simulation of particle flow can generate a prediction of fluid turbulence. At any point in this simulation, we can partition the particles into regions of local influence in a Voronoi diagram, and calculate a probability distribution of Voronoi volumes that allows us to better understand turbulence (on average, smaller Voronoi volumes means lots of particles close to each other, implying clustering of particles and therefore increased turbulence). We can use the parameters (Reynolds number Re for fluid turbulence, Froud number Fr for gravitational acceleration, and Stokes number St for particle characteristics) at which the simulations were run and the probability distributions of Voronoi volumes that they generated in homogenous isotropic turbulence in order to build models that predict this distribution at new parameters, emulating the simulation with less computational demand.

The goal of this project is to build a model that effectively predicts the first four moments of the distribution of Voronoi volumes, given St, Re, and Fr. Specifically, we want to A) better understand the factors affecting turbulence, and B) predict the formation of turbulence going forward. This modeling, which uses fairly inflexible and therefore more interpretable techniques, allows us to try to interpret how these variables affect particle clustering and make predictions over new data without overlearning the random noise in this small dataset.

Methodology

Preliminary exploration of the data revealed several important considerations. First, the entire training set contained only 89 observations and three predictors, and two of the three predictors only had 3 separate values. Because of the relative lack of size in the training set, only fairly inflexible models were considered to minimize the risk of overfitting to random noise in the data, and 10-fold cross-validation was employed for model selection rather than a hold-out set approach. Second, analysis of both St and all four raw moments of the distribution revealed heavy right-skew. Visual inspection and a Box-Cox analysis suggested that a log-transformation of the data would be appropriate for each variable. Thus, each of these variables was only considered when log-transformed. Third, some of the values of Fr were infinite. Regression is mathematically impossible on infinite input variables; thus, two approaches were considered: replacing the infinite values with an arbitrarily large integer (which would have approximately the same effect as an infinite value, and would be acceptable given that Fr cannot actually take on the value of infinity), or splitting the data into a categorical variable (infinite vs finite). Because there was not much difference in the other two values of Fr (0.052 and 0.3) relative to an arbitrarily large integer (which would be orders of magnitude greater), the decision was made to categorize Fr into Fr_f (any finite value) or Fr_∞ (an infinite value).

In order to calculate a model, for each of the four raw moments, several types of models were considered.

Exploration began with simple linear regression, followed by multiple linear regression and the addition of interaction effects. Shrinkage methods to create a less flexible model, such as Ridge and LASSO regression, were included. To consider the possibility for non-linear predictive capacity in these predictors, polynomial regression models were also considered, and shrinkage methods were evaluated on those polynomial fits as well. All of these model types enabled the examination of different ways these variables might predict particle clustering (especially in tandem with one another), while remaining simple enough that they were still interpretable (serving the goal of learning inferentially from this model) and less likely to be distorted by random noise in the data (a condition of high variance, which would be likely given the small training set; this served the goal of building a model with high predictive reliability). Although more flexible nonlinear (e.g., splines) and nonparametric (e.g., kNN-regression) models were briefly considered, they required a great deal of caution given the size of the data, and initial evaluation did not appear promising.

For each moment, several models were identified which appeared promising for consideration as the final model. Once these models were selected, the training data were divided into 10 folds, and cross-validation was run on each of these folds to directly estimate the test MSE of each model. In the aftermath, an estimate of test MSE standard error was calculated (though it is not strictly accurate, it is still useful). The model with the lowest test MSE was selected, provided its standard error did not make it appear to be a fluke (in most cases, the model with the lowest test MSE also had the lowest or close to the lowest standard error). Finally, the model was calculated on the full set of training data. Although 10-fold cross-validation does introduce some bias, as the estimates are calculated on a subset of the full available training data, it was considered as the optimal solution here because it did not take many observations out of the training set and enabled a direct estimation of test error that was comparable across the different types of model specification.

The plan in this study was to calculate a different model for predicting each of the four raw moments of the Voronoi volume distribution. However, when optimal models were selected, moments 1 and 2 utilized the same model specification, and moments 3 and 4 shared one as well (albeit, in both cases, with different coefficients). Moments 1 and 2 relied on a linear regression model with a quadratic polynomial expansion of Re and an interaction between this expansion of Re and Fr, while moments 3 and 4 relied on a LASSO shrinkage of this same specification (with a small, but meaningful, λ).

Results

Moment 1

$R\_mo\hat{m}ent\_1 = e^{-4.943} \times 0.194 \cdot St \times e^{-20.661 \cdot Re} \times e^{5.455 \cdot Re^2} \times e^{0.012 \cdot Fr_\infty} \times e^{1.275 \cdot Re \times Fr_\infty} \times e^{-0.600 \cdot Re^2 \times Fr_\infty}$

This model is able to predict the distribution of R_moment_1, or the mean of the probability distribution, with a very high level of accuracy. The model has a training R^2 of 0.9972, suggesting that we can explain 99.72% of the variance in the mean of the cluster probability distribution with this model, and the calculated test MSE of this model from the cross-validation testing was 0.018, suggesting that we estimate this model will mispredict values of R_moment_1 calculated from new testing data by $e^{0.018} \approx 1.018$, on average. The model is overall significant, F(6,82) = 4829, $p < .001$. One variable, Fr_∞, is not by itself significant, suggesting that after controlling for all other variables, Fr by itself being infinite does not significantly change our predictions for R_moment_1. However, since the interactions $Re \times Fr_{infty}$ and $Re^2 \times Fr_\infty$ are significant, we know that Fr is still an important variable in this analysis. While there are slight trends in the scale-location of residuals for this model, in general the assumptions for a linear model appear met and are not cause for much concern.

This model tells us several things about the distribution, though it does stray towards the side of prediction at the cost of interpretability. First, all three of the predictors affect the mean of the distribution, and each appears to do so in a non-linear fashion. Holding Re and Fr constant, increases in St predict dramatic increases in R_moment_1 (an increase of almost 0.2 in our prediction for one additional unit of St). We know that higher values of St make it harder to solve the equation for modeling turbulence. With a limited understanding of physics, this may then be interpreted as “when it is harder to solve the equation for this turbulence, we predict higher average volume and therefore less turbulent clustering”, which may even be a byproduct of limitations in simulation. More in-depth explorations of what exactly this relationship between St and R_moment_1 reveals should be left to physicists more comfortable in their understanding of this value. It is worth noting that both of these values are on the same (log) scale.

When Fr is finite, Re predicts R_moment_1 in a quadratic fashion. Re quantifies the intensity of turbulent flow. The fact that it is highly significant even at low quantities suggests that it is very useful as a predictor of the mean size of clusters. To interpret this finding very loosely in the logarithmic context, it suggests that at a baseline of low values of Re, when increasing Re we might predict a decrease in R_moment_1 (multiplication by $e^{-20.661 \cdot Re}$ will decrease our overall prediction faster), but as our values of Re increase substantially we would predict increases in R_moment_1 (as the Re^2 term comes to dominate). Re also predicts R_moment_1 differently depending on whether Fr is finite or infinite. If Fr is infinite (or arbitrarily large), there is some “insulation” against these effects of Re - increases at low values of Re would predict less of a decrease in R_moment_1, and increases at high values of Re would predict less of an increase in R_moment_1. This, in context, suggests that higher gravitational acceleration reduces the intensity of turbulent flow, which makes sense (it would take more energy to fight against high gravitational acceleration). In general, all three numbers seem important in the prediction, although Fr is not significant in its own right.

Moment 2

$$R\_mo\hat{m}ent\_2 = e^{-0.722} \times 0.856 \cdot St \times e^{-30.048 \cdot Re} \times e^{8.335 \cdot Re^2} \times e^{-1.579 \cdot Fr_\infty} \times e^{14.405 \cdot Re \times Fr_\infty} \times e^{-3.820 \cdot R^2 \times Fr_\infty}$$

The second raw moment of a probability distribution is its variance. The prediction for R_moment_2 uses the same model specification as the prediction for R_moment_1, but the coefficients are very different. This model is slightly less effective in making predictions than the model for moment 1. It is still highly significant, F(6,82) = 36.32, $p < .001$. However, it has a reduced $R^2 = 0.7266$, suggesting that we can only explain about 72.66% of the variance in the second raw moment of the cluster probability distribution with this model (though this is still a very high amount). Additionally, the estimate for test MSE on this model was 4.345, suggesting that on average the model will mispredict R_moment_2 for new input variables by $e^{4.345} \approx 77.092$. This is a large value, which stems from some of the shortcomings of this (or any) model in being able to predict such a skewed response variable - it is hard to predict such a wide range of values with so little input data. While there are some slightly messy trends in the model assumption plots for this model, overall the assumptions for linear regression appear to be met well enough.

In this model, all predictors are significant except for the interaction $Re^2 \times Fr_\infty$. There are some notable changes in the coefficients for this model. As might have been expected, the effect of St is inflated here. We know that an increase in St makes it harder to solve the Navier-Stokes equation for particle clustering, and the working assumption in the previous section was that St may make computation harder and decrease average Voronoi volumes as a result. That flows directly into an interpretation here, which is that higher values of St make calculation of the equation harder and therefore dramatically (a ratio of 1:.856) increase the variance in the probability distribution. The effects of Re by itself are again quadratic and in the same directions as in the previous model, referring to the same differential explanation. However, in this model Fr_∞ is significant, and the interaction between $Re \times Fr_\infty$ is much larger, suggesting that infinite gravitational acceleration predicts a substantial decrease in the variance of this distribution. This, again, makes sense - with more gravitational acceleration at play, a system would seem to be more “tempered”, and thus vary less overall.

Moment 3

$$R\_mo\hat{m}ent\_3 = e^{3.759} \times 1.226 \cdot St \times e^{-40.168 \cdot Re} \times e^{11.415 \cdot Re^2} \times e^{-3.136 \times Fr_\infty} \times e^{26.375 \cdot Re \times Fr_\infty} \times e^{-6.686 \cdot Re^2 \times Fr_\infty}$$

The third raw moment of a probability distribution is skewness. This model uses the same specification of variables as models 1 and 2, but relies on a LASSO shrinkage model with $\lambda = 0.0088$ (a very small value, but still one that resulted in a model with decreased test MSE). Because λ is so small, we can assume it reasonable to check model assumptions as if it were not present, and this model generally meets the assumptions for linear regression. This model continues to provide evidence to support the interpretations of these input variables from the previous sections - increased St means harder calculations and therefore more uncertainty

& skew in the probability distribution, increased Re means more turbulent flow and again more skew in the distribution of particle cluster volumes, and infinite gravitational acceleration brings things back towards a more normal and less chaotic distribution. This model is much worse at predicting moment 3 than our models for 1 and 2 were - it can only explain about 58.08% of the variance in moment 3, and the calculated test MSE for $log(R\_moment\_3)$ was 15.496.

Moment 4

$$R\_mo\hat{m}ent\_4 = e^{8.255} \times 1.543 \cdot St \times e^{-50.672 \cdot Re} \times e^{14.732 \cdot Re^2} \times e^{-4.694 \cdot Fr_\infty} \times e^{38.427 \cdot Re \times Fr_\infty} \times e^{-9.831 \cdot Re^2 \times Fr_\infty}$$

Like model 3, this model employs the same framework of variables as all models thus far, but also uses the LASSO framework with $\lambda = 0.0107$ to penalize variables with large coefficients. Once again, we use the (closely related) linear model to check assumptions, and determine that this model is acceptable for linear/LASSO regression. This model is, once again, worse at predicting R_moment_4 - with $R^2 = 0.5161$, it explains barely more than half of the variance in the Kurtosis of the probability distribution (though this is still a substantial amount), and it has a test MSE on $log(R\_moment\_4)$ of 32.962, which is very high. Again, these stem from the incredible skewness of the distribution of R_moment_4. This model continues the trend from previous models with respect to the direction and strength of input variables St, Re, and Fr, suggesting that (regardless of whether this interpretation of them is correct), these variables tend to predict the probability distribution in a similar fashion regardless of which moment we consider, even if the effects constantly intensify (but get worse at prediction) as the raw moment increases.

Conclusion

While I will not surmise to be a physicist to any degree, this modeling project can give some key insights into our understanding of turbulence. First and foremost, this project has identified four models which are very effective at predicting the first four moments of the probability distribution of cluster volume. This saves us from having to run as many computationally expensive simulation models. Although they become less effective predictors as the raw moments increase, all of the models are capable of explaining over half of the variance in the response, and the model which predicts the mean of the distribution explains almost 100% of said variance. While these models would certainly benefit from additional training data, they are already quite effective emulators.

All of the variables tend to predict their respective moments in the same fashion, although this effect gets stronger as the moments increase. St is on the same log scale as every moment, and in every moment, increases in St increase the response variable (suggesting that increased computational difficulty leads to more uncertainty and non-normalcy in the probability distribution of cluster volumes). Re is quadratically related to the log of the response variables, suggesting that with enough turbulent flow, we would expect more uncertainty and more turbulence in general. Re also interacts with Fr - when Fr is infinite, it seems to "temper" the effects of Re, as both the linear and the quadratic term of Re cancel each other out to some extent (increasingly in higher moments). Fr itself tells us a lot - while it does not say much about the mean (so it may not predict turbulence in and of itself), infinite gravitational acceleration seems to predict less variance, skewness, and kurtosis in the probability distribution, suggesting that with more gravity, turbulent effects may have a harder time coming about and thus be tempered down.

This modeling process does come with some caveats. This model was trained on only 89 data points, so very non-flexible models were tested and selected. However, it is very possible that a more flexible model would be able to more accurately capture some aspects of this probability distribution. Additional data on which to train would be very useful in enabling us to determine the true nature of some of these relationships. Also, while this model technically allows us to predict at any continuous value of St or Re, very restricted input values were used for training. I would not recommend that this model is used for values that are much different from what it has been trained on. Finally, Fr is used in a very broad fashion with finite/infinite values, but it is possible that the threshold for the difference between these groups is actually much lower than infinity - once again, more training data would be useful, and until then, increased caution when predicting at values far from what the model was trained on is necessary.