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, unpredictable, making it difficult to understand. 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.

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 $Re \in \{90, 224, 398\}$) in the training data. 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. This relationship between St and R moments appeared logarithmic, so we log-transformed St. *are we going to deal with infinity?* Also, while treating the Fr variable as categorical, we observed that there is a linear and decreasing relationship between Fr (gravitational acceleration) and R moments. Lastly, there is also a linear and decreasing relationship between Re (Reynolds number) and R moments. We also explored multicollinearity through VIFs for each model, which were very low.

Our final models for each of the moments for inference is a cubic polynomical with interaction terms between Re and Fr:

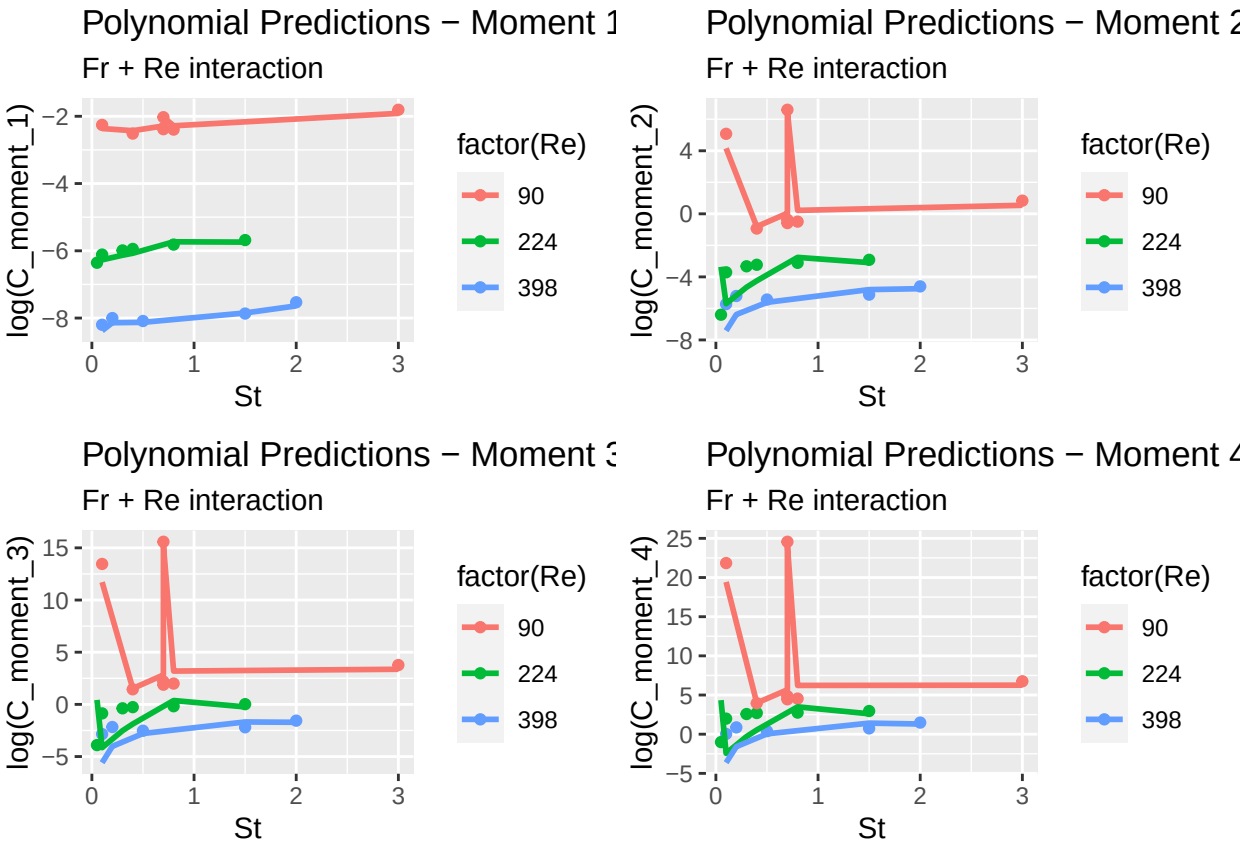
$$\begin{aligned} log(R_{moment1}) = &-2.048 + 2.002(log(St)) + 0.161(log(St))^2 + 0.446(log(St))^3 \\ &-3.824(Re_{224}) - 6.015(Re_{398}) - 0.286(Fr_{0.3}) - 0.317(Fr_{\infty}) \\ &+0.238(Re_{224} * Fr_{0.3}) + 0.389(Re_{224} * Fr_{\infty}) + 0.504(Re_{398} * Fr_{\infty}) \end{aligned}$$

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

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

$$log(R_{moment1}) = 23.168 + 18.0035(log(St)) - 22.202(log(St))^2 + 8.469(log(St))^3$$
$$-15.075(Re_{224}) - 23.325(Re_{398}) - 19.018(Fr_{0.3}) - 18.619(Fr_{\infty})$$
$$+12.755(Re_{224} * Fr_{0.3}) + 12.576(Re_{224} * Fr_{\infty}) + 19.746(Re_{398} * Fr_{\infty})$$

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.



	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

DISCUSS REAL-WORLD INTERPRETATION

Prediction 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. To address the limitations of our current models, we provide the following related models using splines that can be used for extrapolation. In these models, Re is nuemric, and we transformed Fr to get rid of the infinity value that it takes on.

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.

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 only 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.

DISCUSS PREDICTION SPLINE MODEL W INTERACTIONS

Results

Conclusion