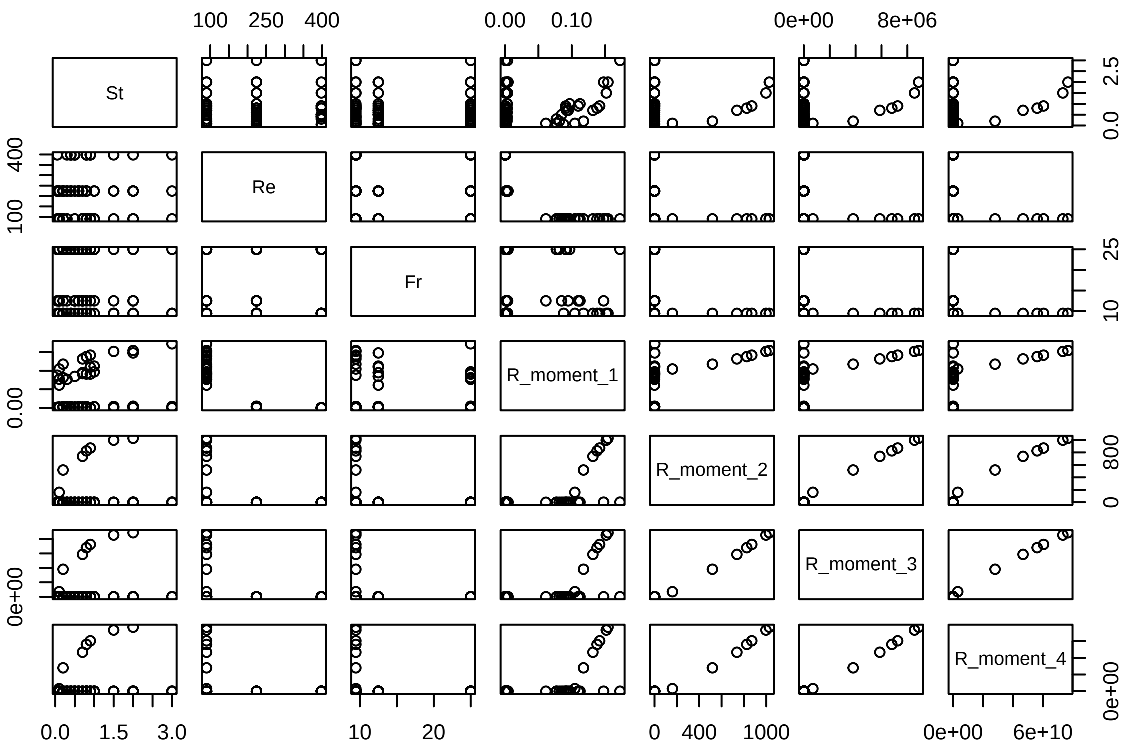
yihan_shi_casestudy

Yihan Shi

10/31/2022

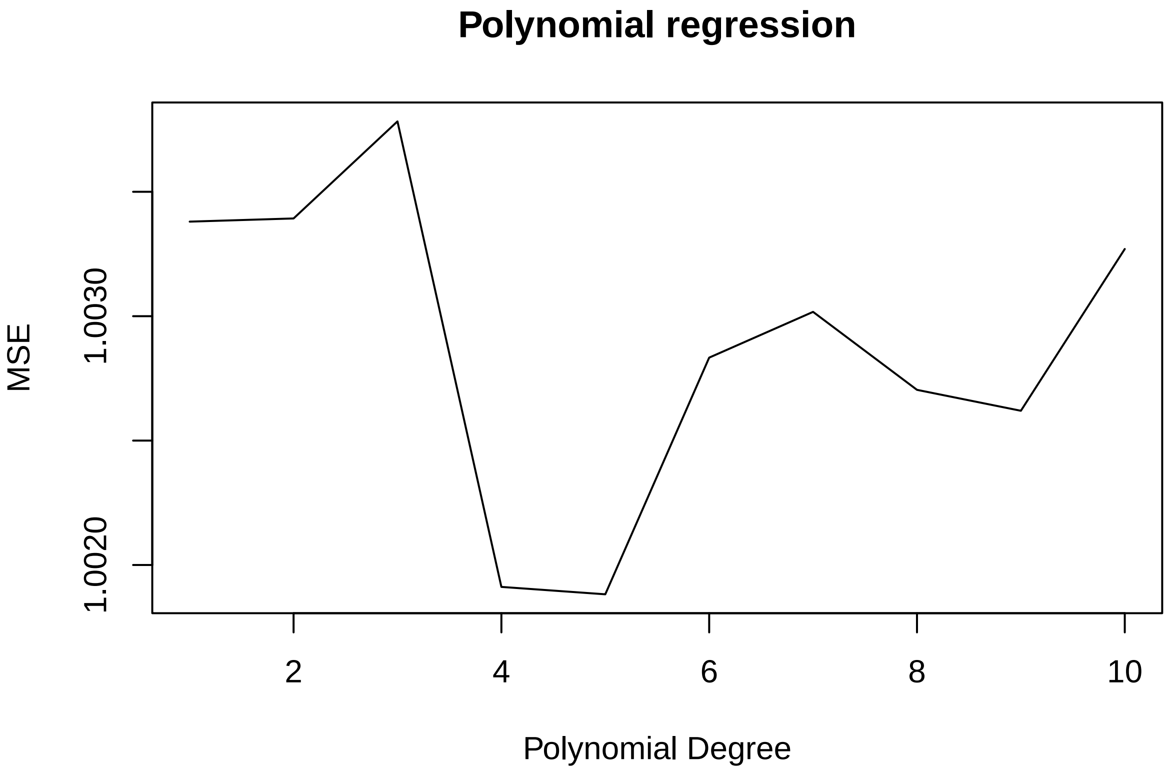
Introduction

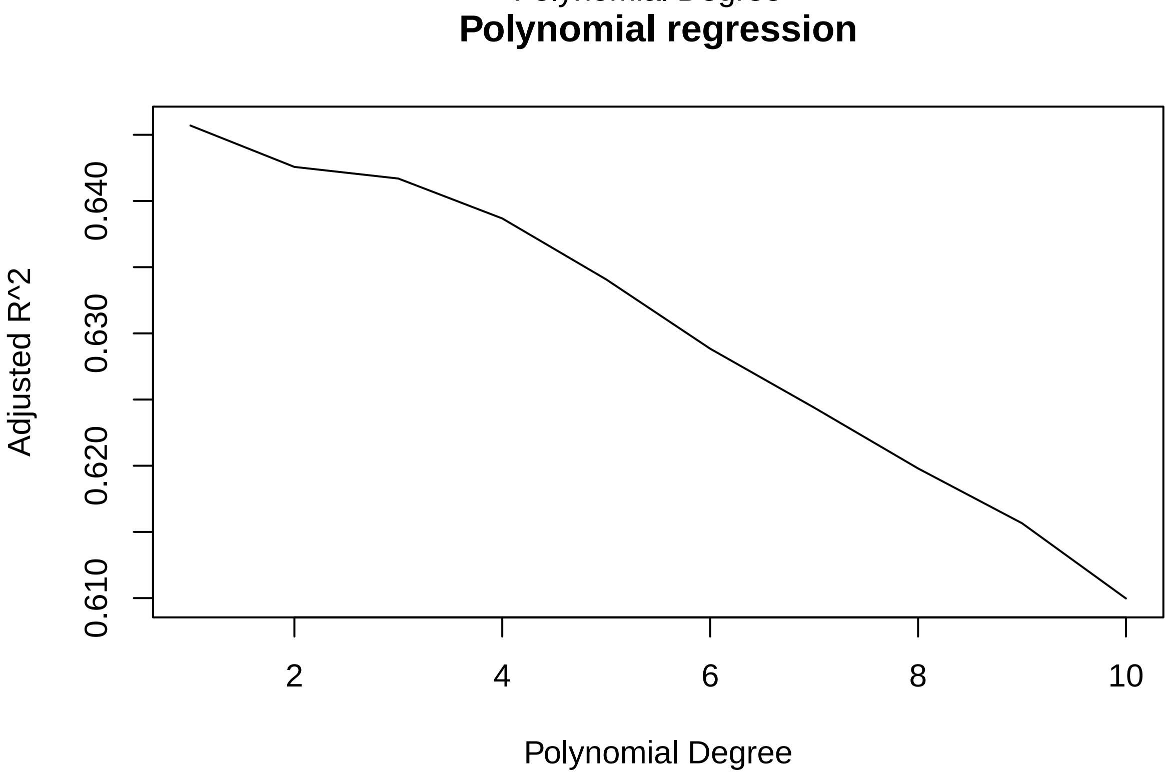
"We are experiencing some turbulence, please fasten your seat belt." Many of us might have heard this radio on the plane and felt bumpy. When we mix paint in water, we can also observe turbulence as the color dissipates. Turbulence is so common and easily observed in daily life, yet its causes and effects 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 hope to use these 3 features to explain changes in particle distribution as well as extrapolate beyond the scope of the known observations.



From the pair plot, there is somewhat linear or polynomial relationship between St and R_moment_1.

Polynomial regression

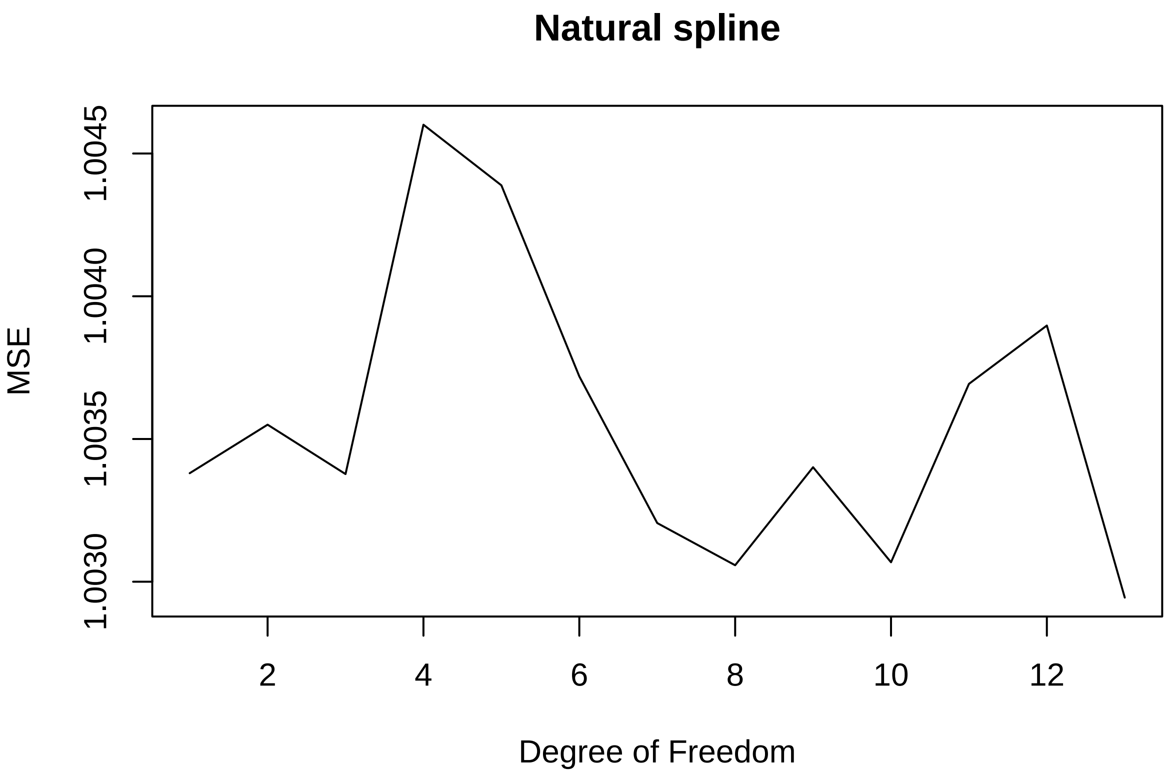


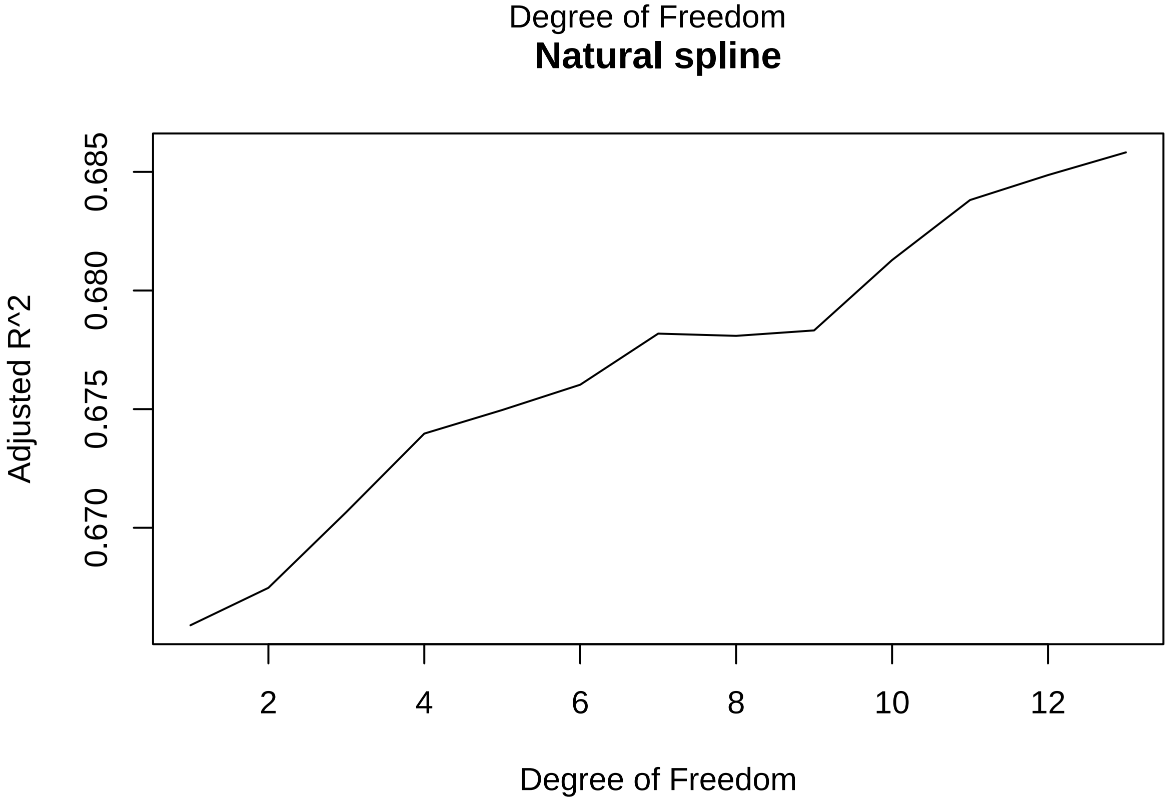


Using cross-validation with 5 folds, we see the adjusted R^2 decreases as polynomial degree increases, and the lowest MSE is achieved at degree of 4. We found that a linear regression is not suitable. Even though it has a relatively high adjusted R^2, the MSE is way higher than other polynomial models. We decided that the optimal polynomial order for St is 4. After degree of 4, the adjusted R^2 starts to decrease quickely.

We tried removing data of both high leverage and residual. However, this didn't change the MSE and adjusted R^2 greatly. Since these observations take up 7% of the full training data, we decided that we don't want to exclude them.

Natural spline

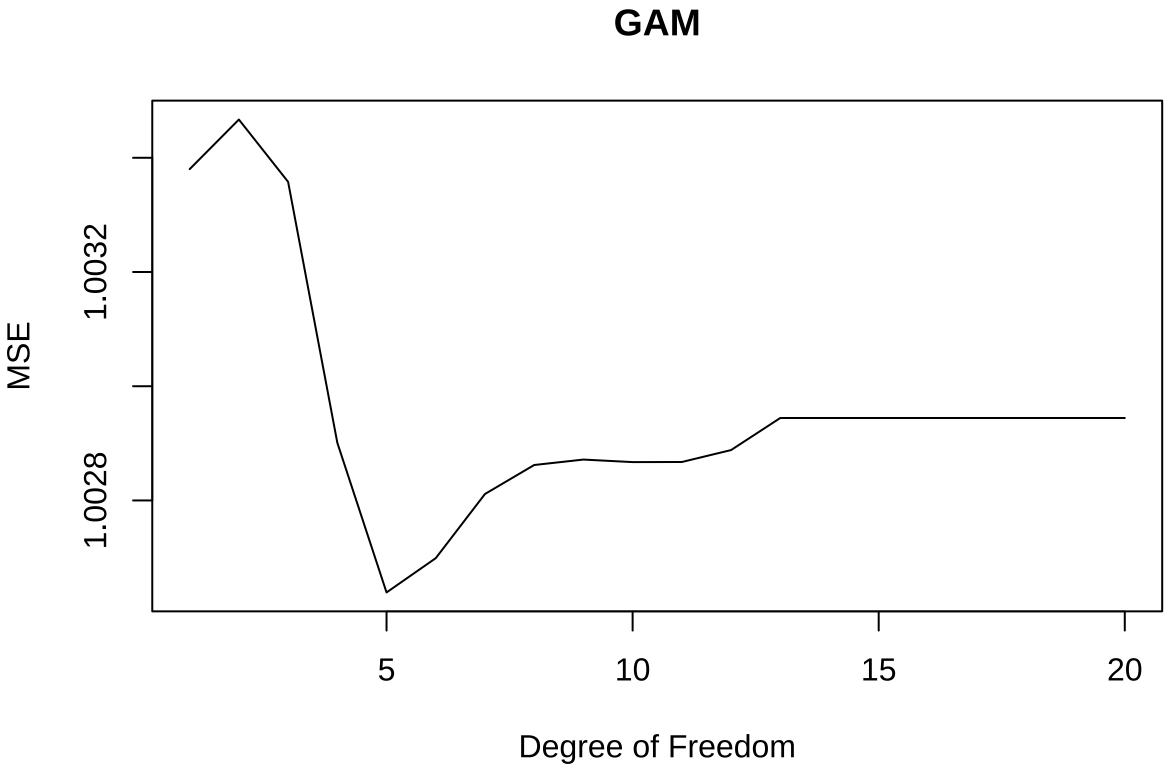


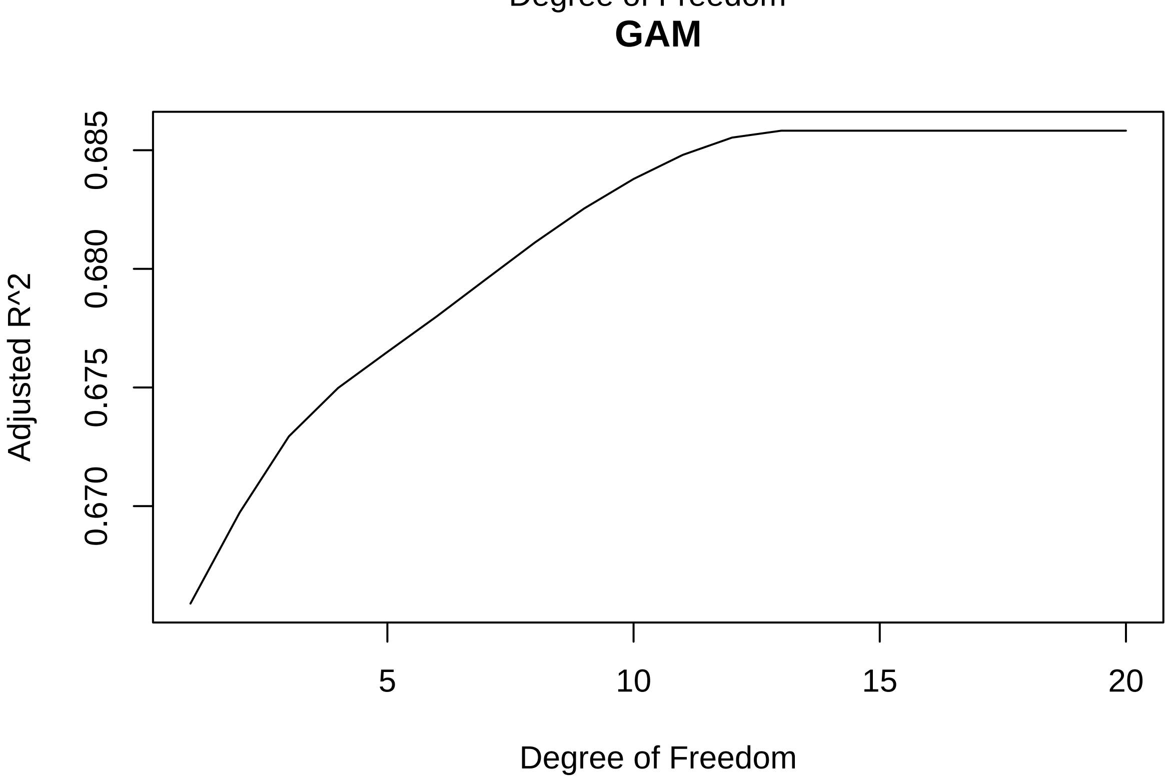


Using

cross-validation with 5 folds, we see the adjusted R^2 increases as the degree of freedom increases, and the lowest MSE is achieved at degree of 13. The optimal polynomial degrees of freedom for St is 13, which achieves both the highest adjusted R^2 and MSE.

GAM





Using

models	formula	mse	adj.R
Polynomial regression	log(R_moment_1) ~ Fr + Re + poly(St, 4) + Fr * Re	0.9722442	0.6850798
Natural spline	log(R_moment_1) ~ ns(St, df = 13) + Fr + Re + Fr * Re	0.9727633	0.7198809
Generalized additive model	log(R_moment_1) ~ s(St, 14) + Re + Fr + Fr * Re	0.9735706	0.7198809

cross-validation with 5 folds, we see the adjusted R^2 increases as the degree of freedom increases. The optimal degree of freedom for GAM model to achieve the lowest MSE is 5. However, we might conculde that a degree of freedom around 14 is optimal because it doesn't have a big increase in MSE compared to degree of freedom at 5. It also achieves the highest adjusted R^2.

For prediction: GAM is the best because it has the highest adjusted R^2. However, since we cannot intrepret the coefficients of each predictor, we would use polynomial regression for inference.