Catalog

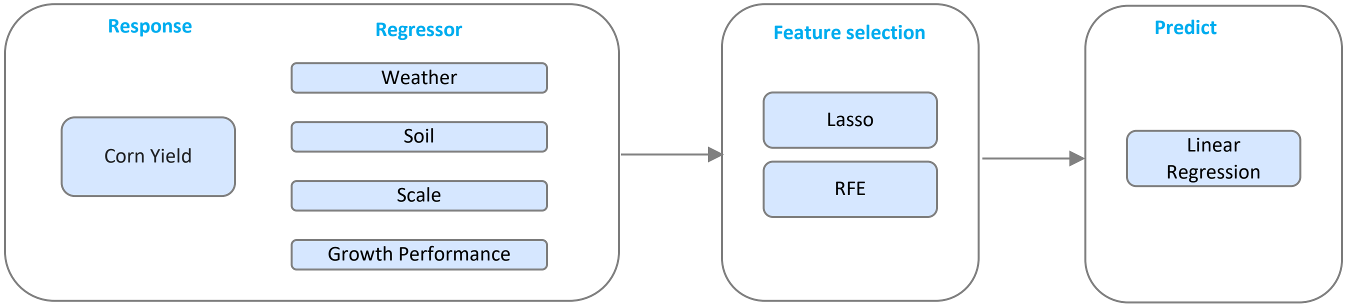
Using Linear Regression Method to Predict Corn Yield

Background and Method

Background

- Crop yield prediction is essential for global food security but is complex, we are focusing **on corn yield.**
- Our regressors represent four different aspects: **weather(weekly), soil, scale, and growth performance(weekly)**, and the response is **corn yield[1] (BU/A)**.
- We collected data from 1990 to 2018, with a total of **8,352 observations and 688 features**.
- Our goal is **to forecast corn yield performance** for the Corn Belt states Illinois, Iowa, and Indiana.

Method



Group Member: Yaopeng Ma, Changhe Ji, John Zhang

1: Bushel per Acre: Mass of bushels per area of an acre. Using a bushel of wheat having 60 pounds.

Using Historical Data as Potential Data Solution

Experimental Setup

2018

Weather and Growth Performance Information at Week t

Year	Observed			Unobserved	
2018	Week1	Week2	...	Week t	...

+

Year	Historical Data				
1990	Week1	Week2	...	...	Week40
1991	Week1	Week2	...	...	Week40
...	...				
2017	Week1	Week2	...	...	Week40

Weather and Growth Performance Information at Week t

Scenarios	Observed			Historical Data		
1	Week1	Week2	...	Week t	...	Week 40
2	Week1	Week2	...	Week t	...	Week 40
...	...					
28	Week1	Week2	...	Week t	...	Week 40

Response
y1
y2
...
y28

Summarize
Max
Median
Min

Key Points

- As crops grow, we can collect more detailed information throughout the year, so we can **improve our forecasts on a weekly basis**, which could be beneficial to policy makers and farmers.
- We run all data from previous years as the **underlying data to construct prediction intervals**.
- We summarize the **maximum, median, and minimum** as significant result for our prediction.

Model I : Linear Regression Using Lasso as Feature Selection Method

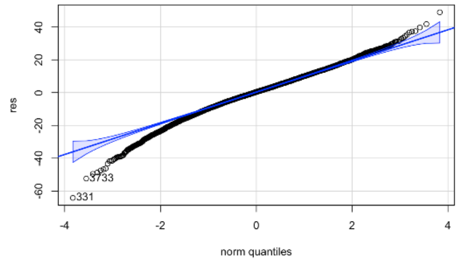
Function :

$$y = \beta_0 + \beta_w X_w + \beta_{soil} X_{soil} + \beta_{scale} X_{scale} + \beta_g X_g + \mathcal{E}$$

Model fitting Performance

2017-2018

QQ-Plot



Result

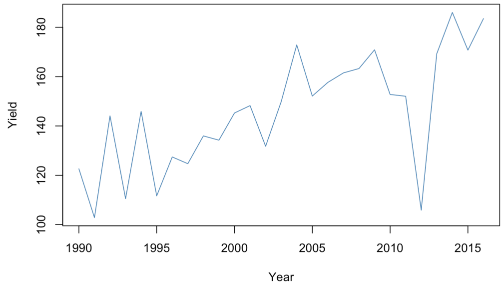
```
Residual standard error: 10.82 on 7247 degrees of freedom
Multiple R-squared:  0.8872,    Adjusted R-squared:  0.8792
F-statistic: 110.1 on 518 and 7247 DF,  p-value: < 2.2e-16
```

Explain

- The idea of the lasso is adding a penalty term to constrain the equation. The consequence of imposing such a penalty is to shrink the coefficient values towards zero, this would **set the less contributive regressor to have a zero coefficient** which achieves the purpose of variable elimination.
- We successfully reduce the variables **from 688 to 519 with 88% R-squared**.
- We select the data from 1990-2016 to train the model and apply the model to predict the corn yield in 2017. And train model based on the data from 1990-2017 to 2018.

Corn Yield Trend

1990-2016



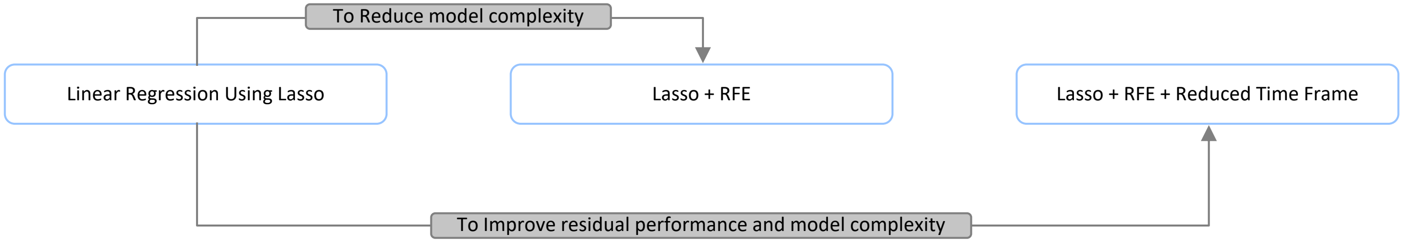
Bias and Improvements

- Model is **complex** with 519 variables: **RFE**
- **Residual not fit perfectly** : applied interaction terms; get ride of Genetic Modified Seed effects **by reduce time frame**

Improvements for Model I Using Different Strategies

Model Progress

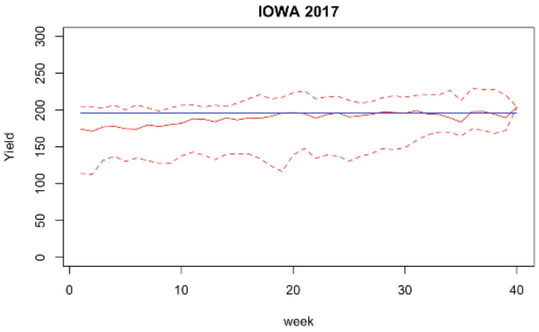
Improvement needed | Good | Great

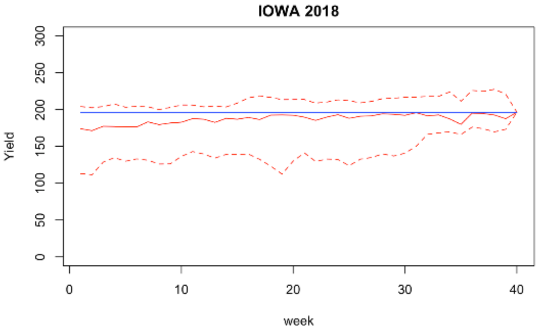


	Model I	Model II	Model III
Model Complexity	▪ 519 variables	▪ 250 variables	▪ 109 variable
Residual Analysis	▪ Distribution is skewed left that concentrated on the right side of the distribution	▪ Limited improvement	▪ Limited improvement
Converge & Explain	▪ After **week 20**, we have a predict result that close to the real result. ▪ Data after **week 35 would still affect** the model significantly.	▪ After **week 28**, predict result getting closer to real result.	▪ After **week 20,** we have a predict result that close to the real result ▪ Furfure information added to the model have **marginal effect.**
Accuracy	▪ Most outcomes are **accurate** based on RRMSE	▪ Most outcomes are accurate based on RRMSE, **slightly decrease** from Model I.	▪ Most outcomes are accurate based on RRMSE, **slightly decrease** from Model I.

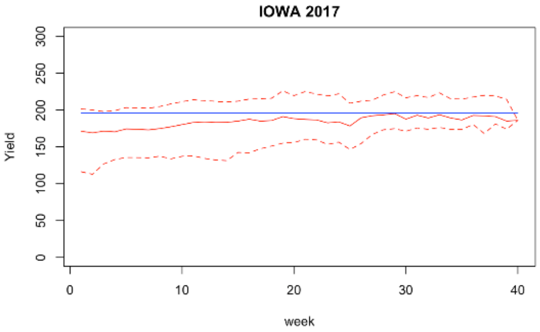
Iowa: Prediction Interval based on Weekly Updated Data under Different Model

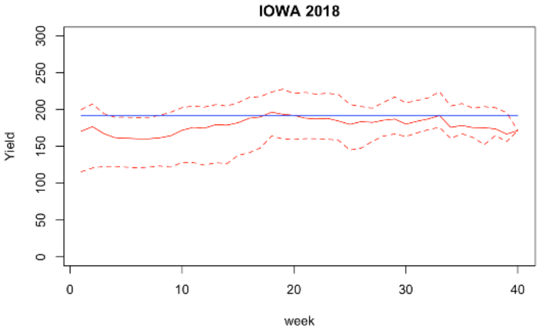
Mode I
Yield[1]:BU/A ;2017 - 18



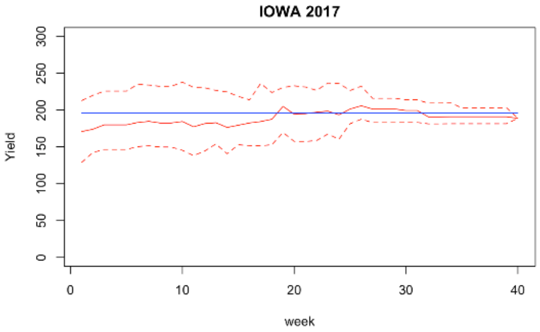


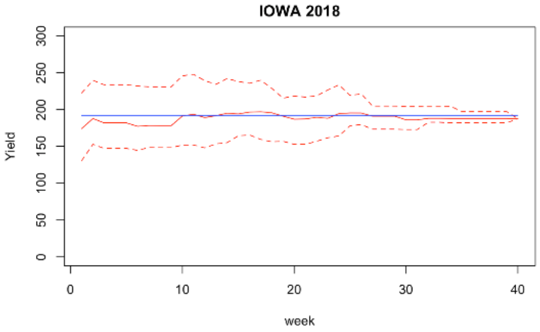
Model II
Yield:BU/A ; 2017 - 18





Model III
Yield:BU/A ; 2017 - 18





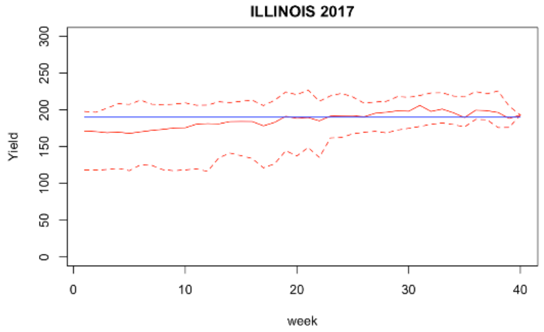
Group Member: Yaopeng Ma, Changhe Ji, John Zhang

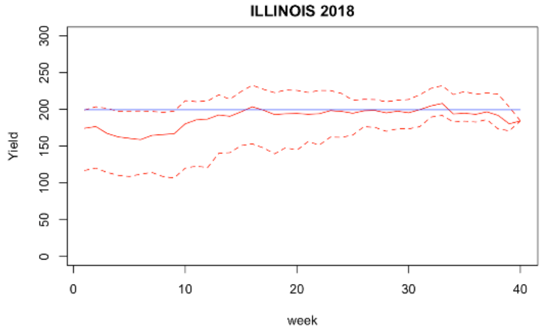
1: Volume per Acre: 56lb = 1 BU

Illinois: Prediction Interval based on Weekly Updated Data under Different Model

Mode I

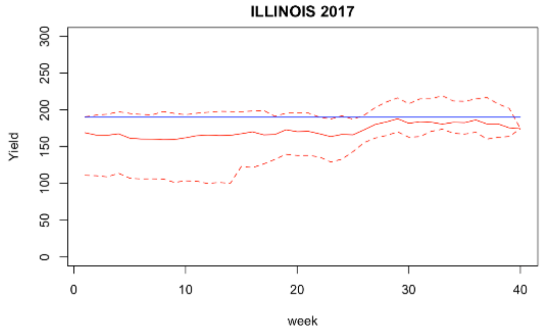
Yield:BU/A ;2017 - 18

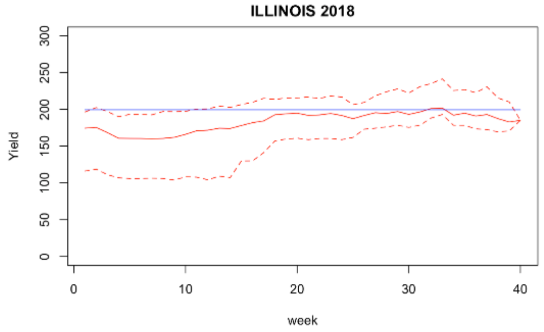




Model II

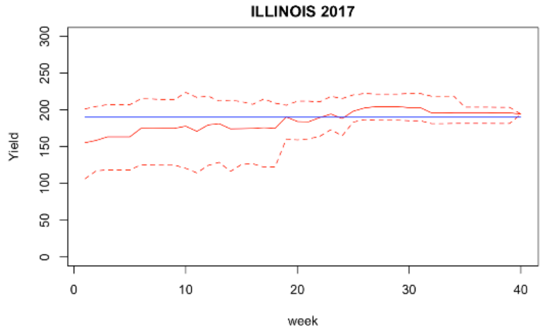
Yield:BU/A ; 2017 - 18

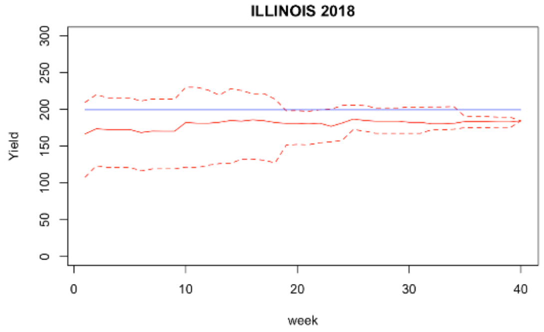




Model III

Yield:BU/A ; 2017 - 18

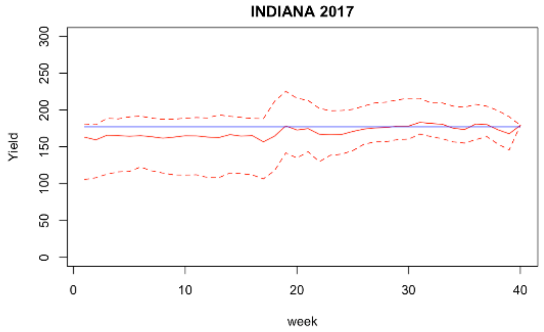


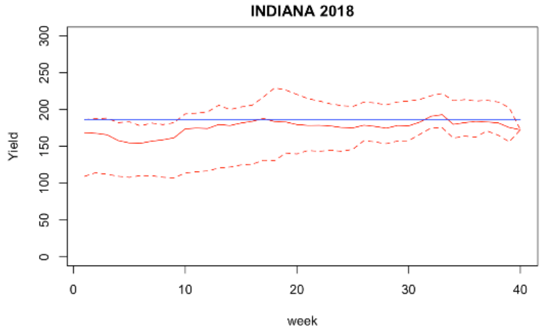


Indiana: Prediction Interval based on Weekly Updated Data under Different Model

Mode I

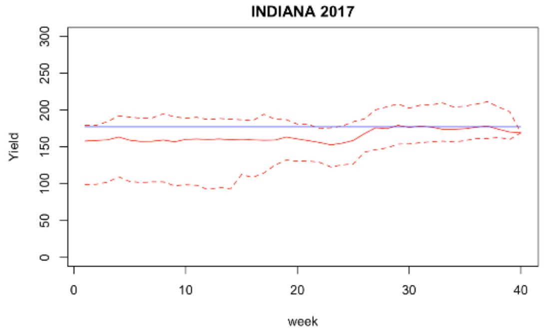
Yield:BU/A ;2017 - 18

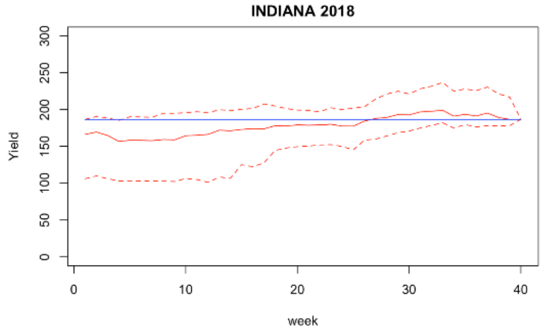




Model II

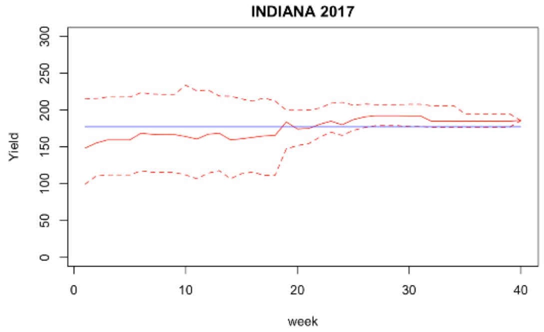
Yield:BU/A ; 2017 - 18

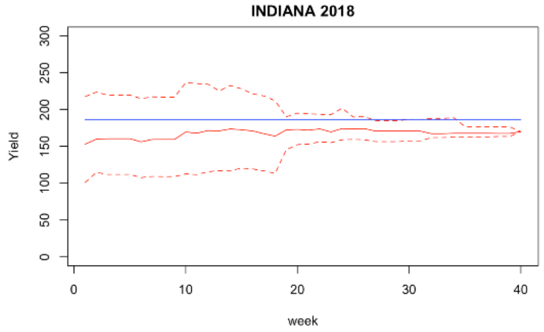




Model III

Yield:BU/A ; 2017 - 18



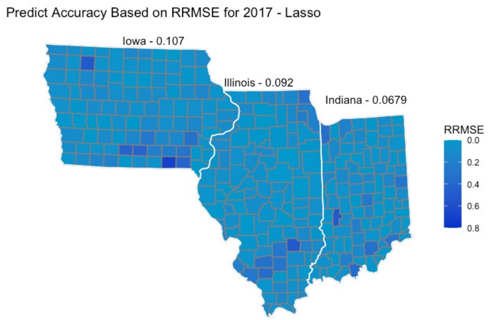


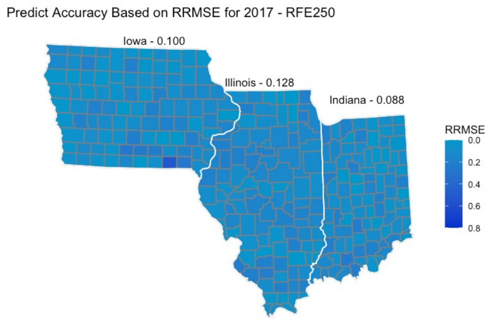
Prediction Accuracy at County Level for Different Models

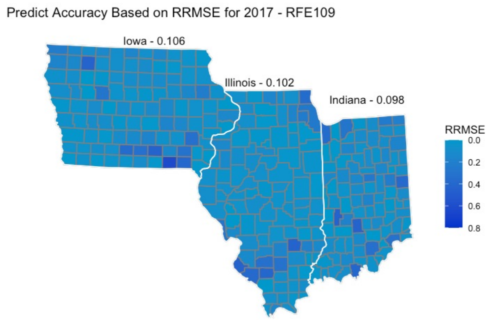
Mode I
2017 - 18

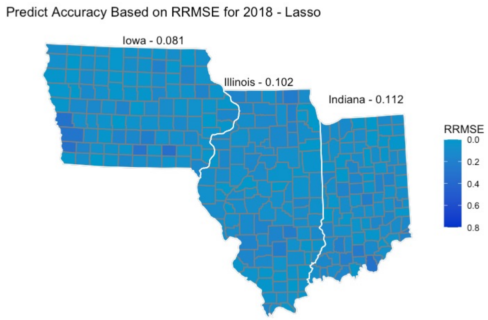
Model II
2017 - 18

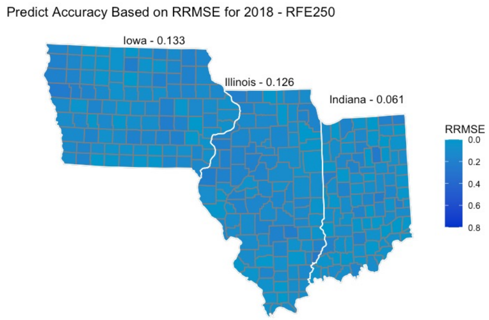
Model III
2017 - 18

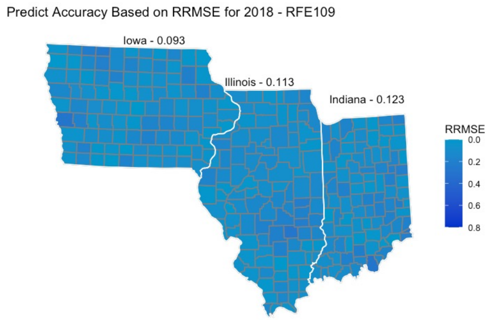












Conclusion: Model III Would Be the Optimize Solution for Prediction

Model Comparison

	Model Fitting			RRMSE		
Model	**Variables**	**R-squared**	**Adj R-squared**	**Illinois**	**Indiana**	**Iowa**
I- 17	519	89%	88%	9%	7%	11%
II- 17	250	84%	83%	13%	9%	10%
III - 17	109	75%	74%	10%	10%	11%
I- 18	519	89%	88%	10%	11%	8%
II- 18	250	84%	84%	13%	6%	13%
III - 18	109	76%	75%	11%	12%	9%

- Model I performed the **best in both model fitting and prediction accuracy**, but the **complexity** needs to be improved.
- The complexity of Model II is lower, but the prediction **accuracy dropped** based on RRMSE.
- Model III performed **best in terms of model complexity,** and prediction result **outperforming Model II** but slightly weaker than Model I.
- After comparison, all models have their own biases and strengths. We choose **Model III as the optimal solution** for our project.

Thank you