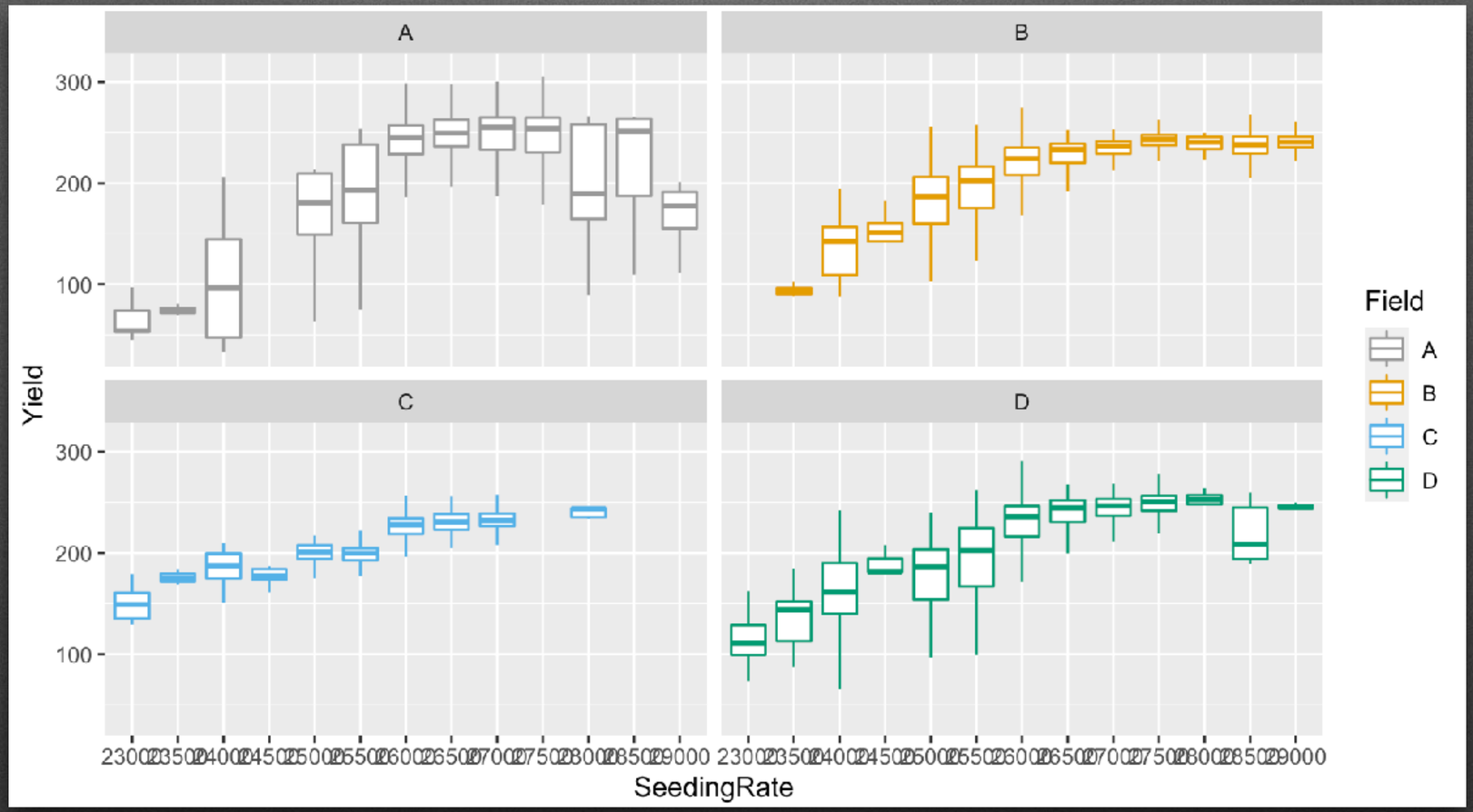
Causal Inference from Large Scale Yield Data

Peter Claussen
GDM Solutions

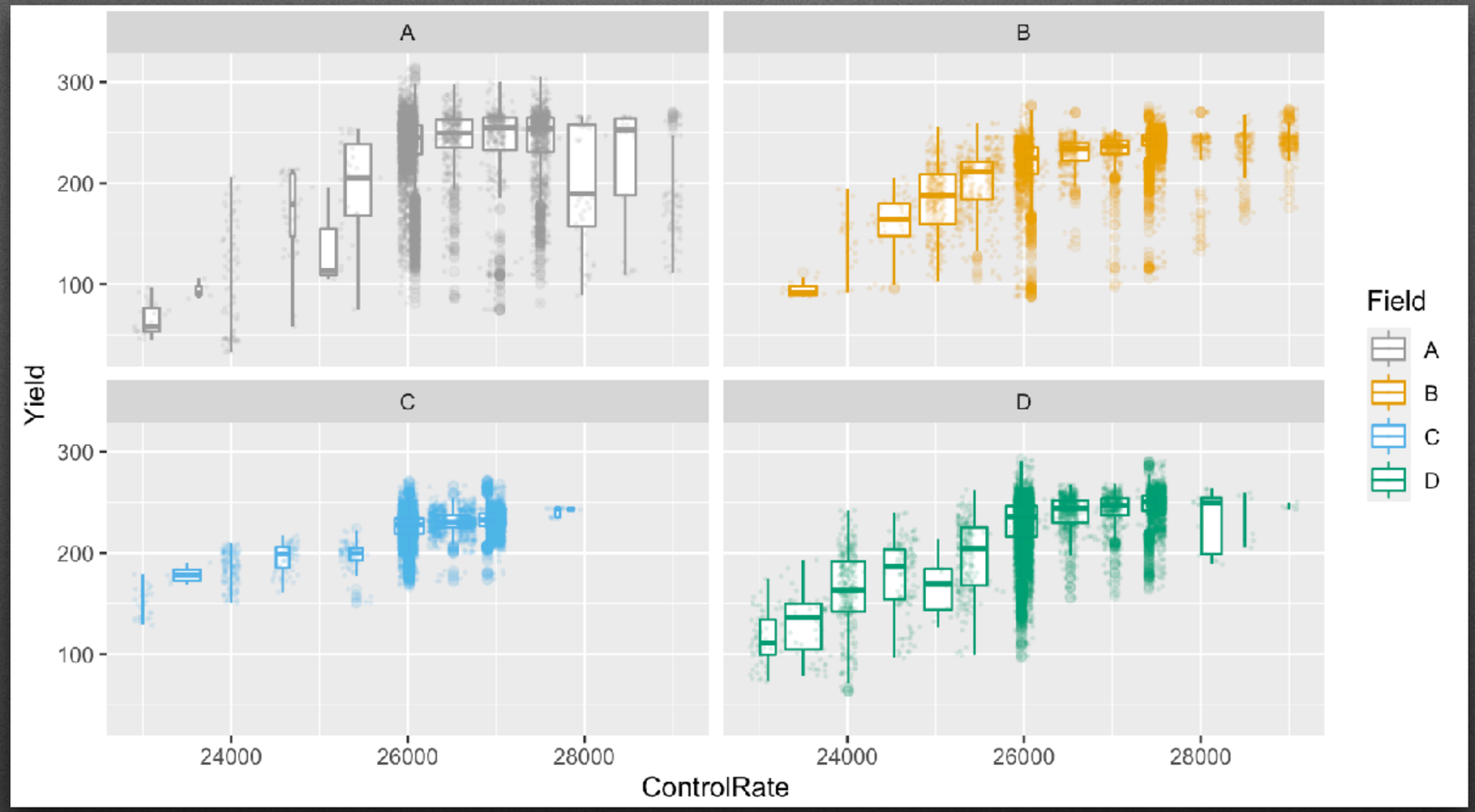
A simple problem

- We have 4 data sets, each representing a single corn field.
- The data contain two columns of interest. Yield is measured in bushels per acre, while ControlRate is measured in seeds per acre.
- Each row in the file represents a geo-spatially tagged point located by Easting and Northing, measured in meters from the southwest corner of the field.
- Aggregate the data by ControlRate, calculate an average Yield at each ControlRate for each field.
- What is the relationship between ControlRate and Yield?



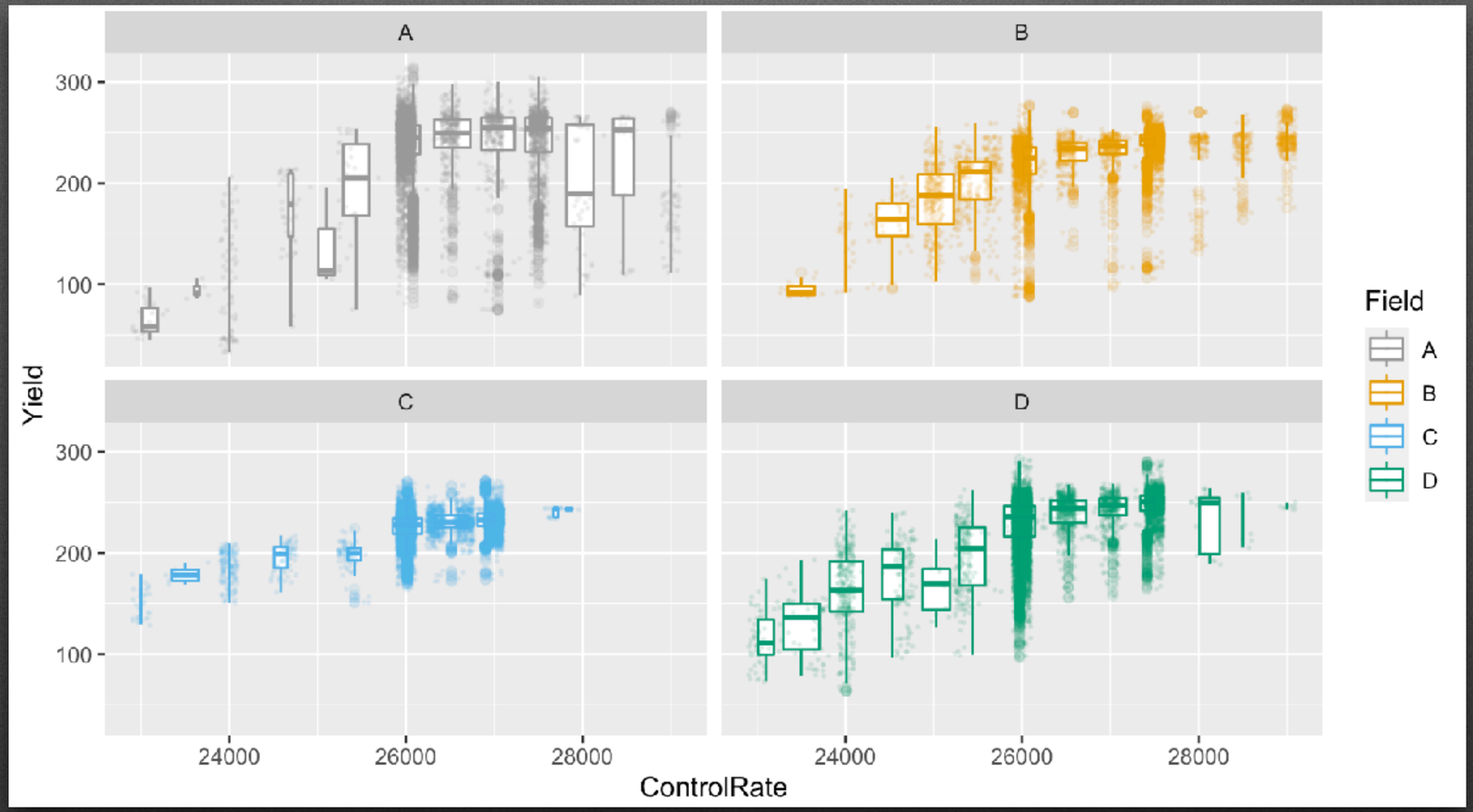
Regression Analysis?

Could we regress yield on seeding rate and infer an optimum?



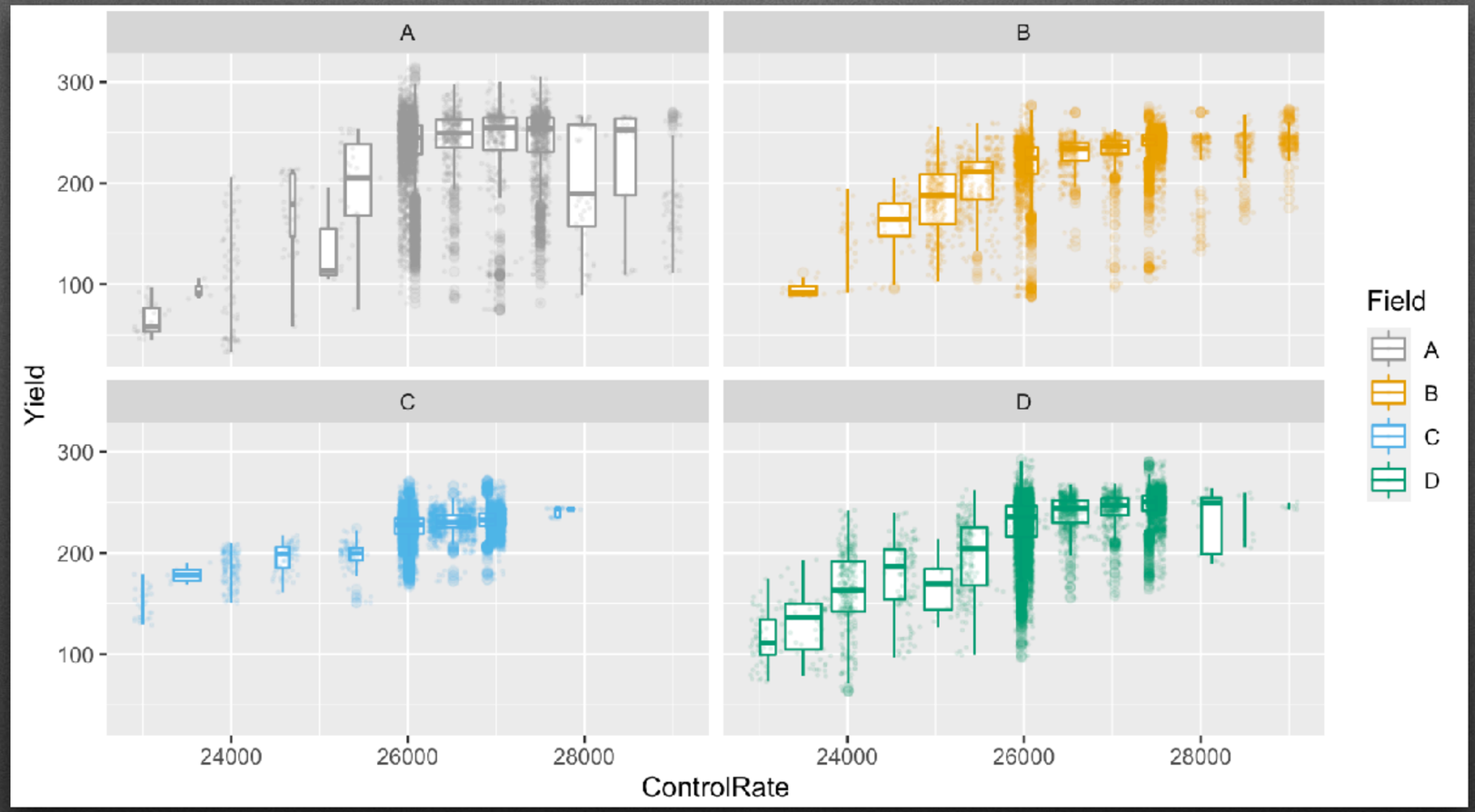
Regression Analysis?

Looks are deceiving. In this case, different Control Rates were not randomized equally over exchangeable units. Instead, we have a seeding rate prescription, and the resulting yield data.



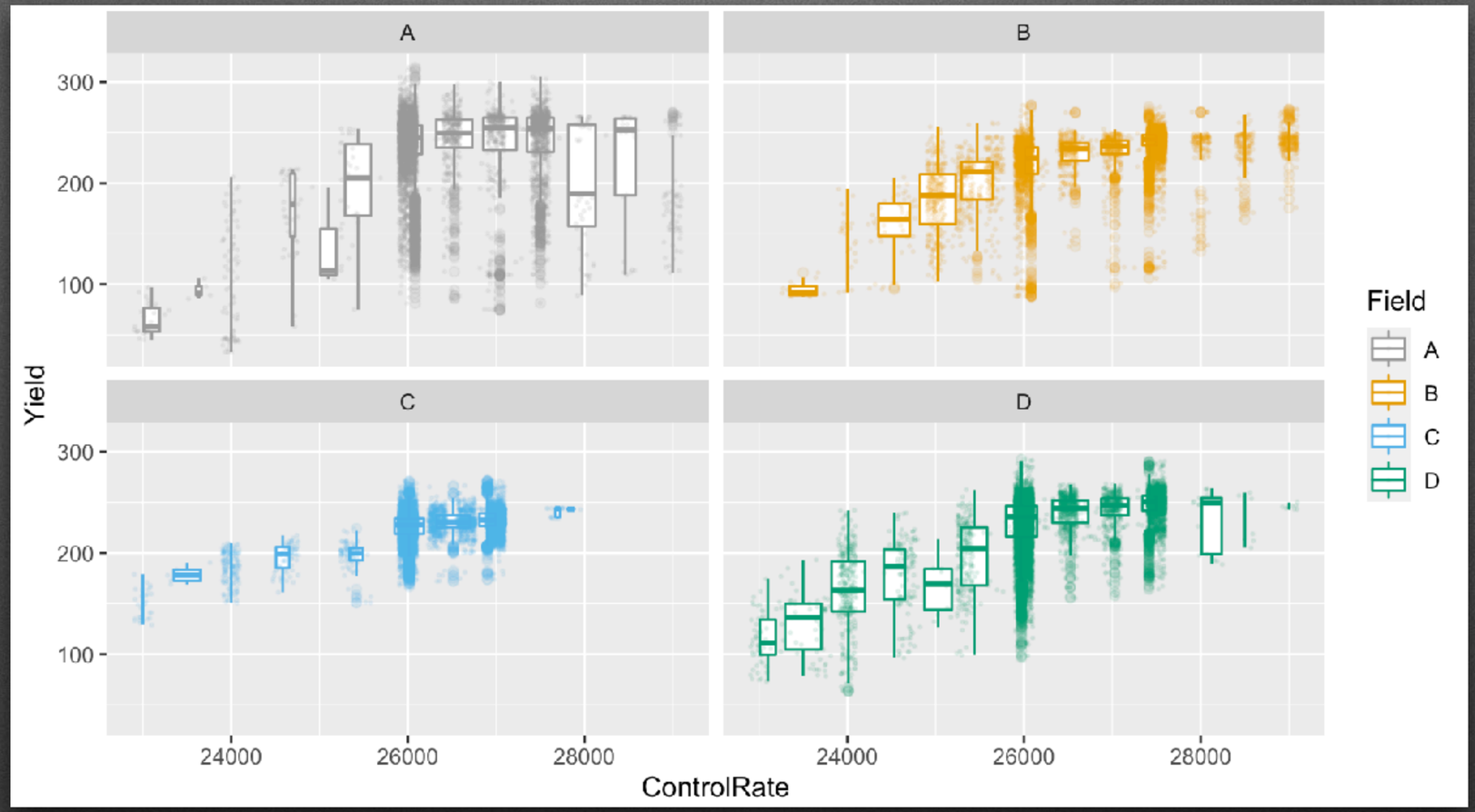
Regression Analysis?

In the absence of randomization over exchangeable units, this simply correlation.



Regression Analysis?

Correlation may be a hint about causal relationships, but we need to identify or control for other confounding variables.

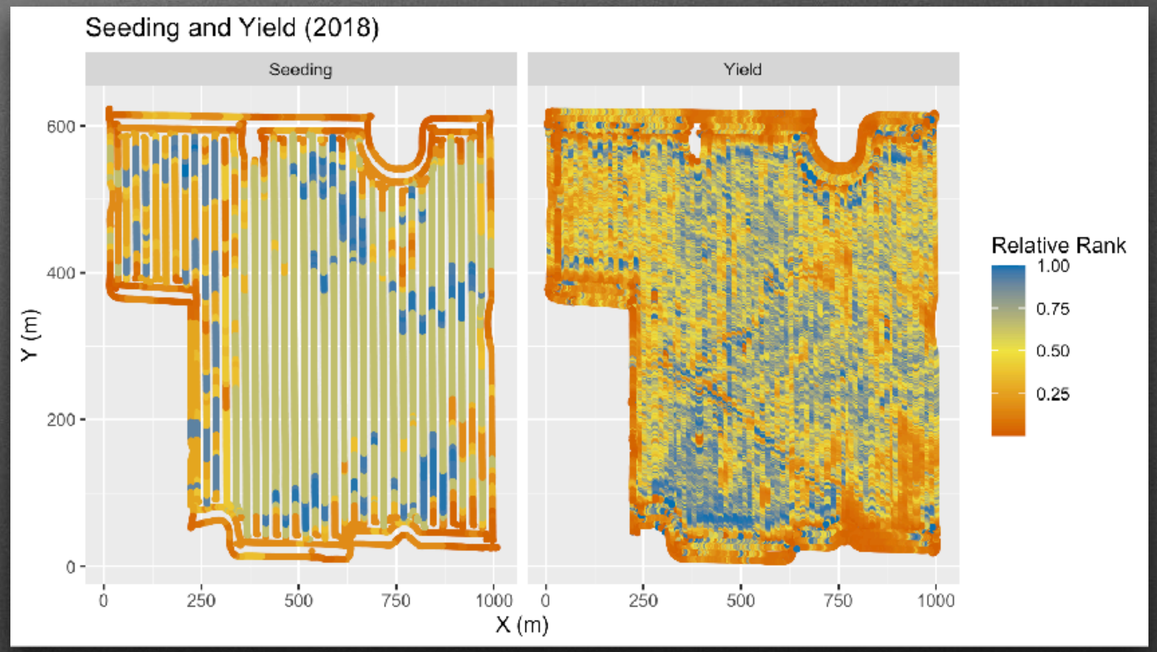


Regression Analysis?

We'll focus on just one field, and consider additional data.

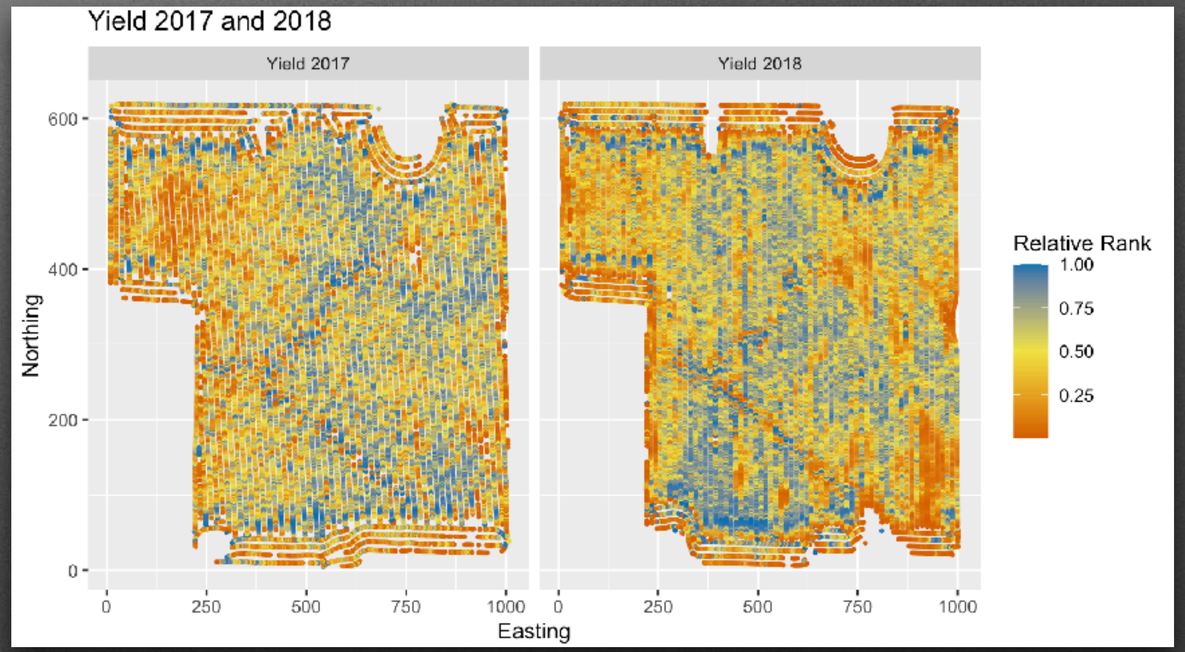
Simple Regression

- $Y18_i = \beta_0 + \beta_1 R18_i + e_i$
  - Is Seeding Rate for 2018 a predictor of Yield for 2018?



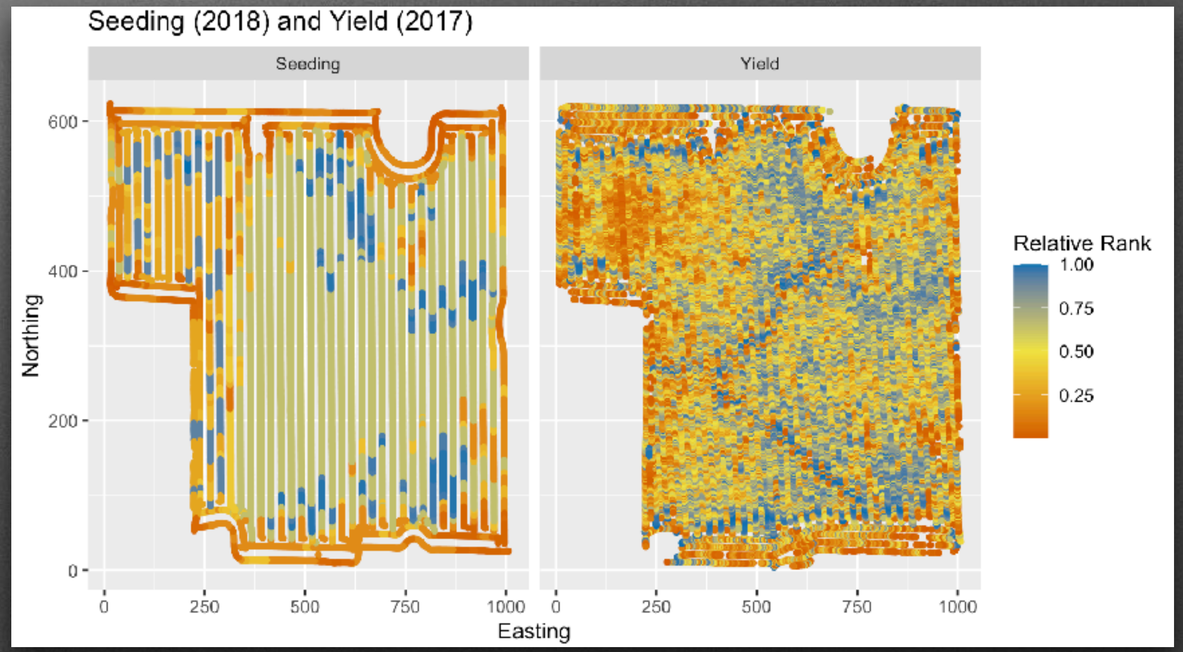
Counterfactual?

- Can we ask what might have happened if there had been no seeding map?
  - Can we estimate what yield would have been, based on a prior season estimate, at uniform rates?
  - Does Yield from 2017 predict Yield for 2018?



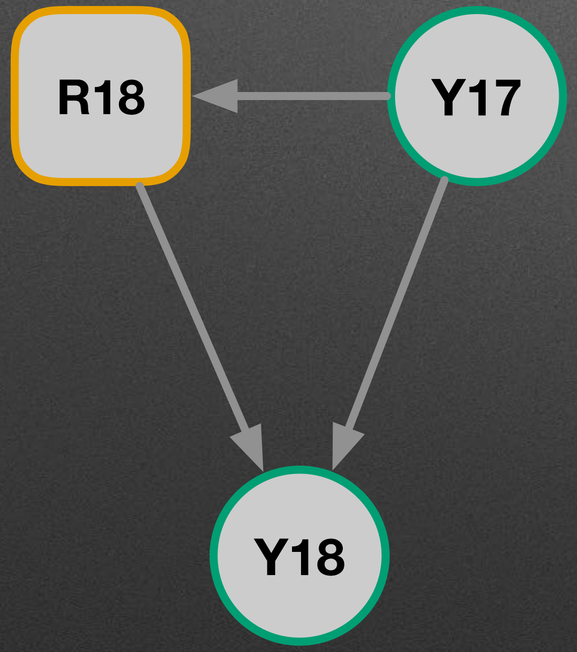
Confounding Interaction?

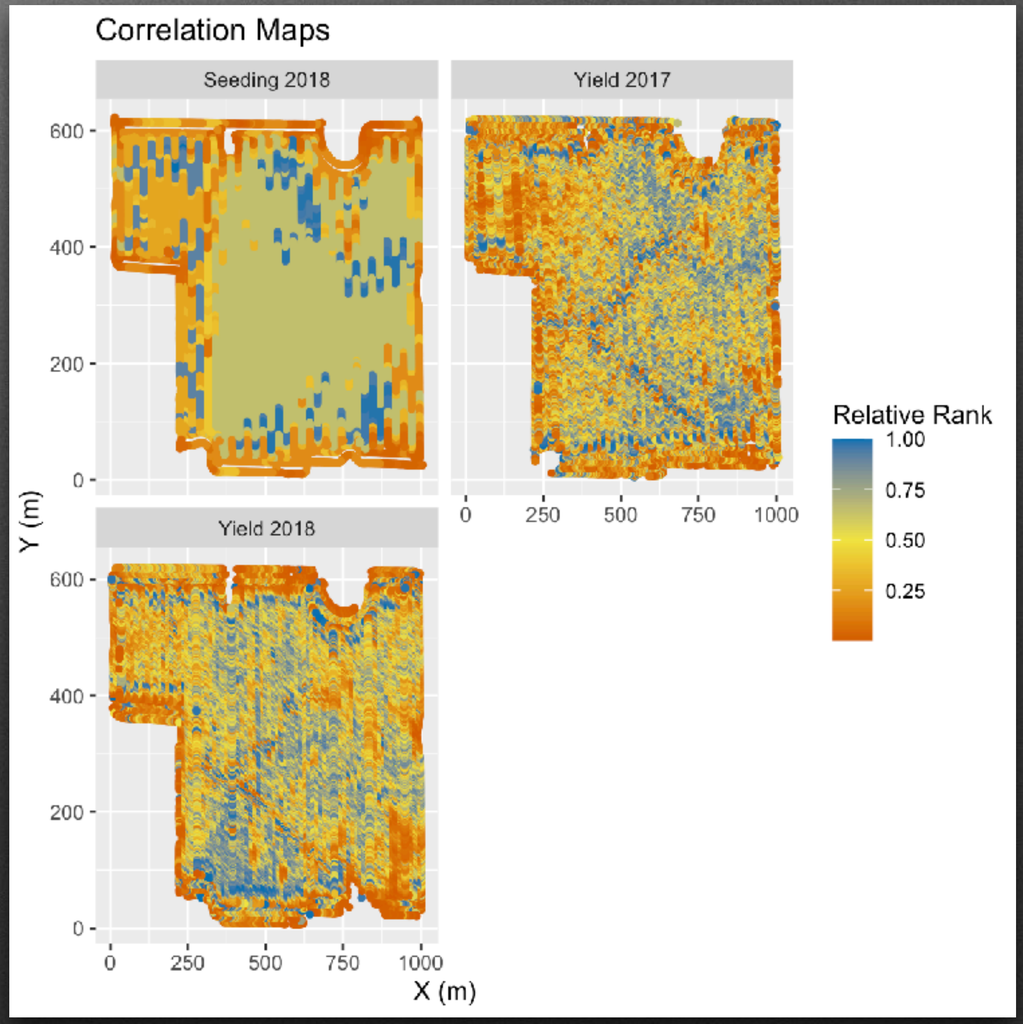
- Could one predictor have influence on another predictor?
  - Are two correlated variables both correlated with a third?
  - Did the 2017 Yield map determine 2018 Seeding map?



Exploring Causal Relationships

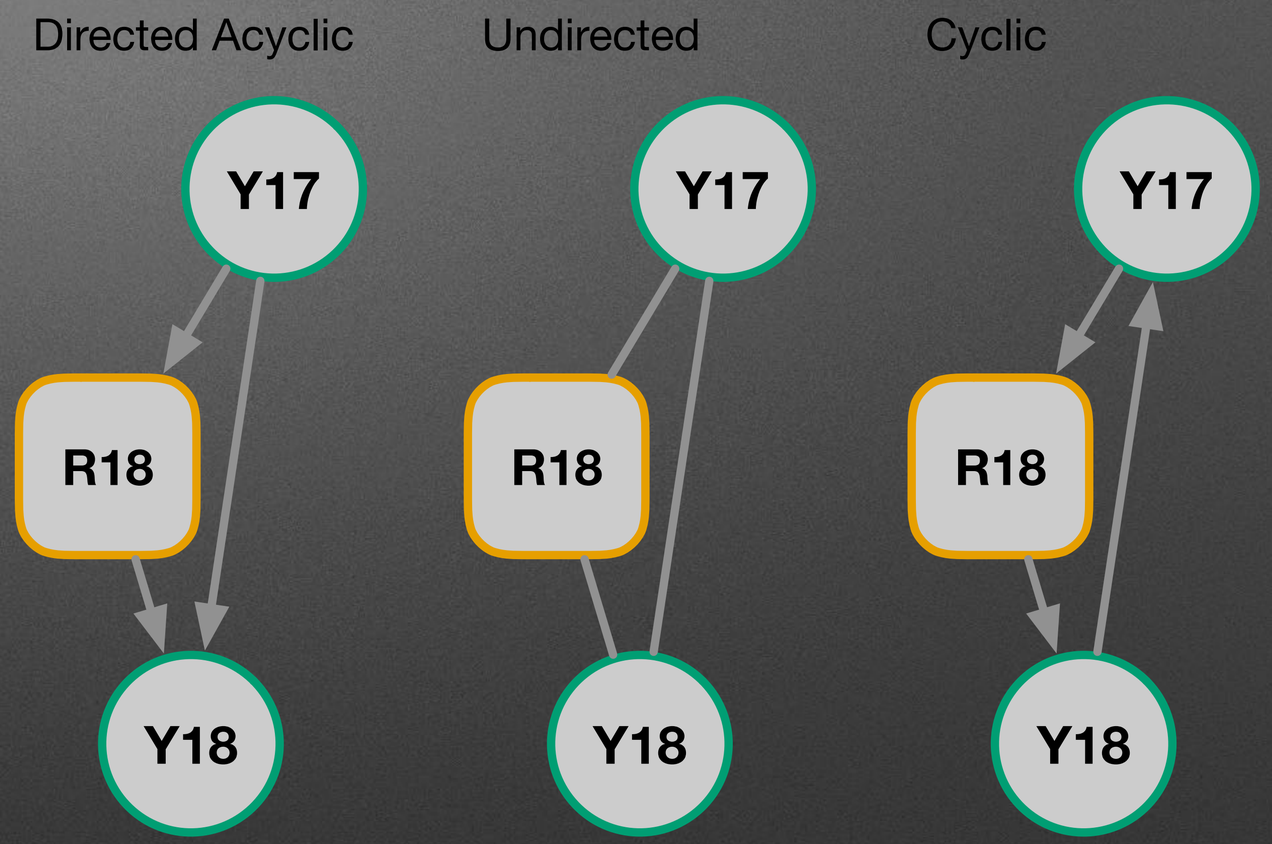
- The relative merit of possible causal relationships can be explored using directed acyclic graphs.





Directed Acyclic Graphs

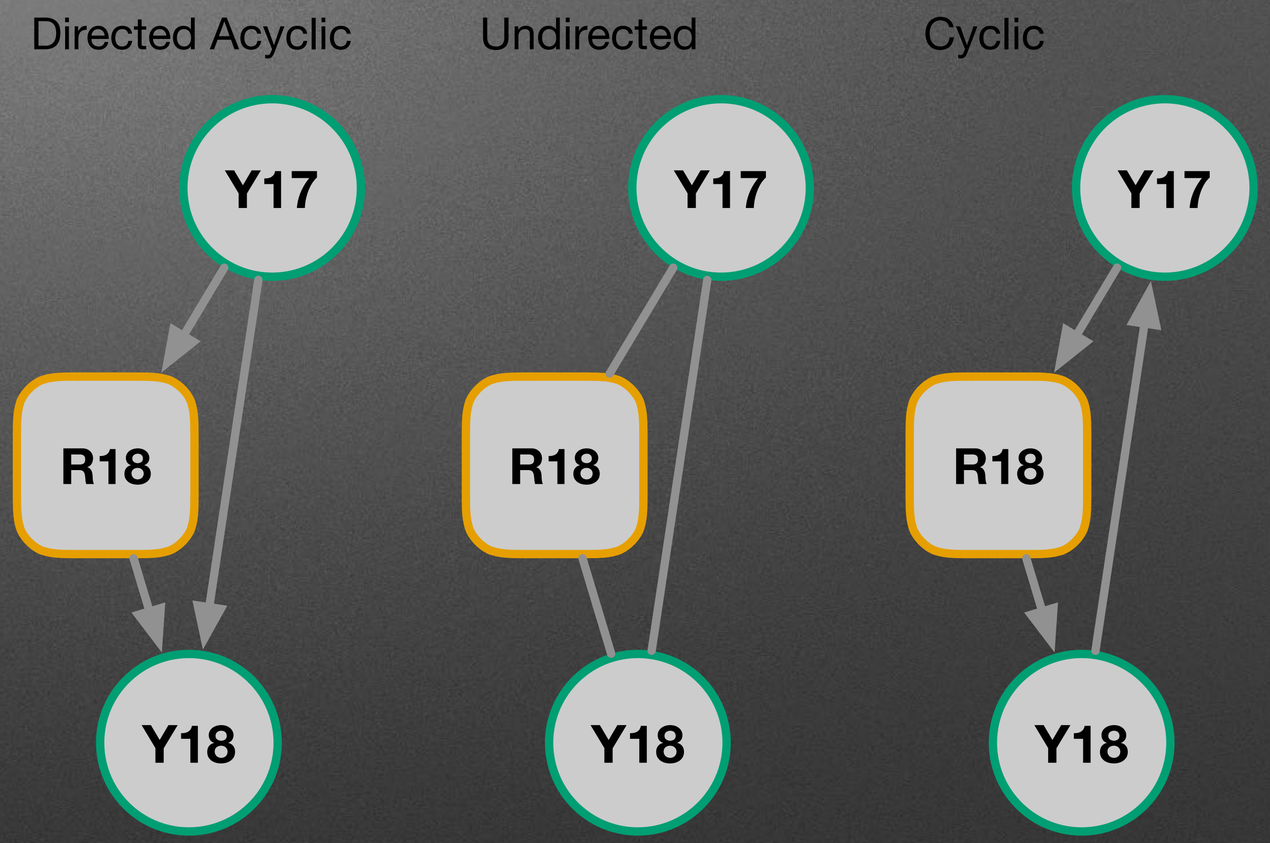
- Nodes
  - (Measurable) Variable
- Arcs
  - Relationships
- Arrows connect parents to children
  - Direct relationship
- Paths
  - Sequence of arcs leading from one node to another.
  - Indirect relationships



Directed Acyclic Graphs

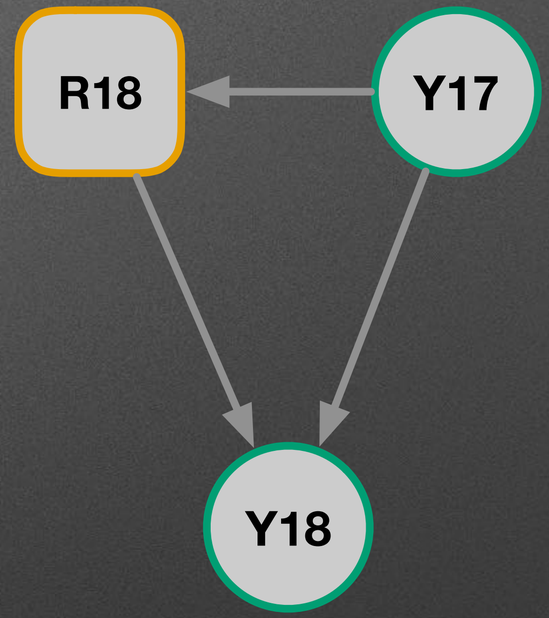
- Bayesian Network
  - Probabilistic model with
    - set of variables
    - conditional dependencies
  - Represented as a Directed Acyclic Graph

- Compare with a linear model
  - response variable
  - one or more predictor variables



Model Comparison

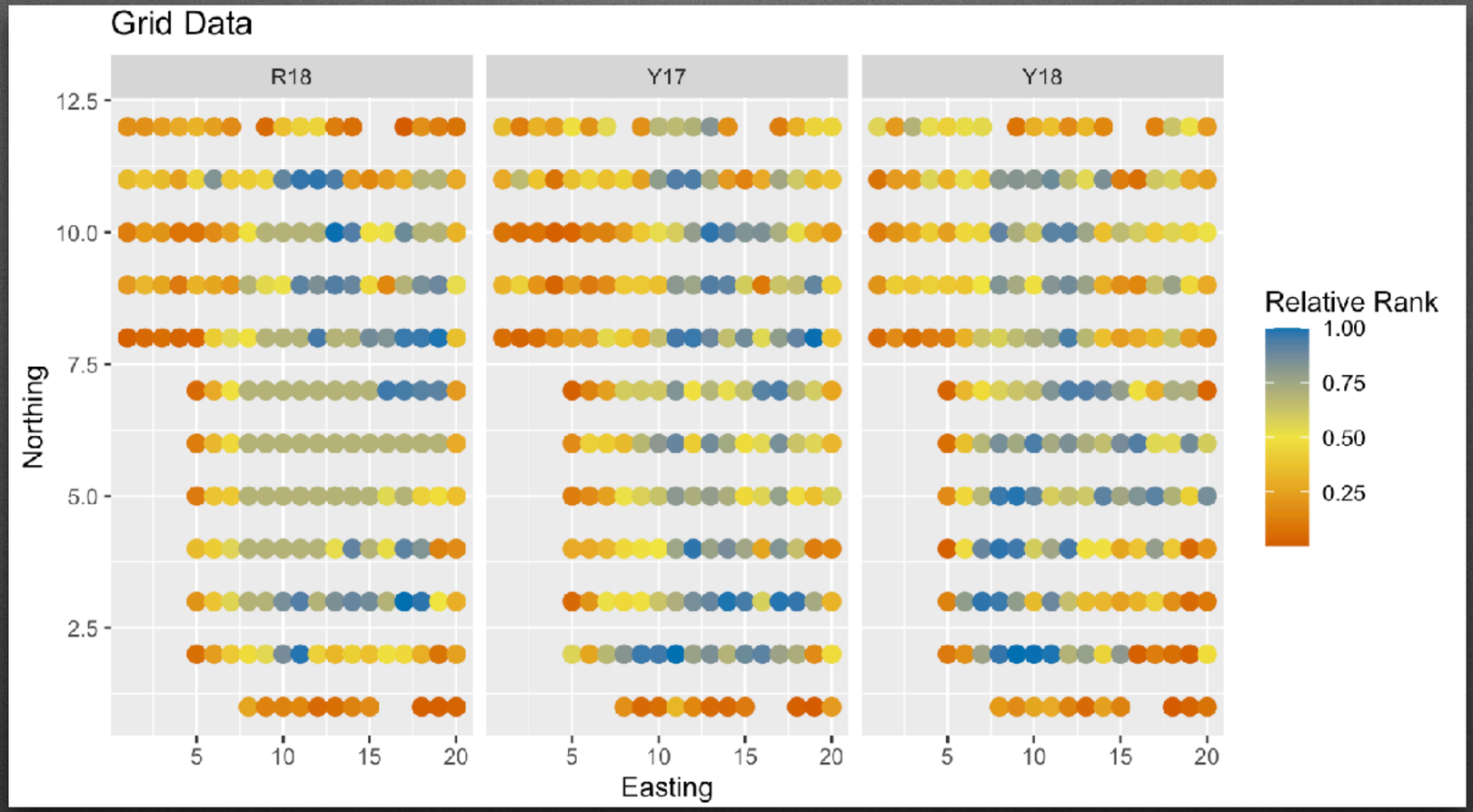
- Multiple Regression : Linear model
  - $Y18_i = \beta_0 + \beta_1 Y17_i + \beta_2 R18_i + \beta_3 Y17_i \times R18_i + e_i$
  - $e_i \sim \mathcal{N}(0, \sigma^2)$
- Bayesian Networks : Probabilistic models
  - $Y17 \sim \mathcal{N}(\mu_{Y17}, \sigma^2_{Y17})$
  - $R18 \mid Y17 = y \sim \mathcal{N}(\mu_{R18} + \beta_0 Y17, \sigma^2_{R18})$
  - $Y18 \mid R18 = r, Y17 = y \sim \mathcal{N}(\mu_{Y18} + \beta_1 R18 + \beta_2 Y17, \sigma^2_{Y18})$



Computational Details

- Multiple Regression
  - R stats
  - `model3.lm <- lm(Y18 ~ R18 + Y17 + R18*Y17, data=model3.dat)`
- Bayesian Network
  - bnlearn
  - `model3.dag <- model2network("[Y17][R18|Y17][Y18|R18:Y17]")`
  - `fit3 = bn.fit(model3.dag, model3.dat)`

https://github.com/PeterClaussenSDSU/ManagementZoneML



Data

For simplicity, we analyze yield response on a 50 x 50m grid.

Model Fitting

- $Y18_i = \beta_0 + \beta_1 Y17_i + \beta_2 R18_i + \beta_3 Y17_i \times R18_i + e_i$

- $Y17 \sim \mathcal{N}(\mu_{Y17}, \sigma^2_{Y17})$
- $R18 \mid Y17 = y \sim \mathcal{N}(\mu_{R18} + \beta_0 Y17, \sigma^2_{R18})$
- $Y18 \mid R18 = r, Y17 = y \sim \mathcal{N}(\mu_{Y18} + \beta_1 R18 + \beta_2 Y17, \sigma^2_{Y18})$

Regression Coefficients

TERM	ESTIMATE	STD ERR	T	P(>T)
(INTERCEPT)	-2497	0.0522	-4.778	<0.0001
Y17	42.39	0.9099	4.659	<0.0001
R18	0.1017	0.0196	5.196	<0.0001
Y17:R18	-0.0016	0.0003	-4.647	<0.0001

Conditional Densities

Model	Intercept	R18	Y17	SD
Y17	57.869			2.526
R18 | Y17	13849		227.784	436.155
Y18 | AR18 + Y17	-71.980	0.011	0.160	10.889

Testing Models

- Importance of Effects
  - Linear Models
    - Coefficient t-tests
    - Analysis of Variance
  - Bayesian Network
    - Arc Strength
- Model Comparison
  - AIC
  - BIC
  - BGE

Testing Models

- Arc Strength
  - Change in score when an arc is removed from the graph, retaining the rest of the graph structure
  - AIC, BIC, etc.
  - may be expressed as a probability (likelihood)

- Bayes Factor
  - marginal likelihood ratio of two competing hypothesis

Model Comparison

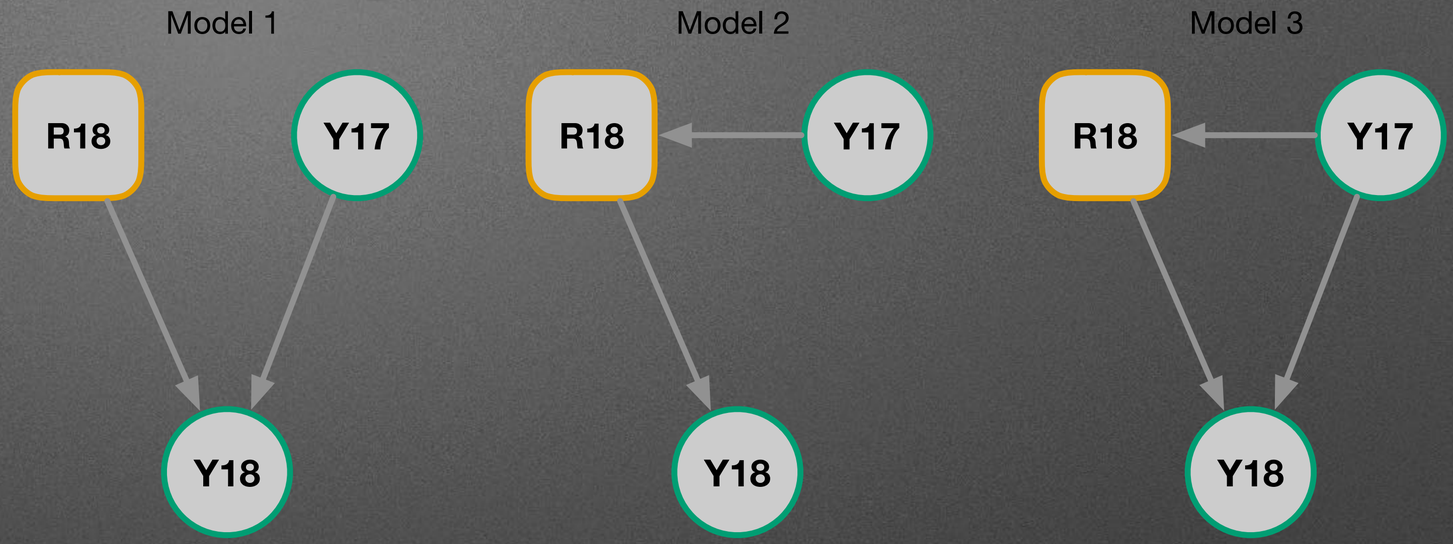
- Regression Models

- ```
  Model1.lm <- lm(Y18 ~ R18 + Y17, data=model1.dat)
  Model2.lm <- lm(Y18 ~ R18 + R18:Y17, data=model1.dat)
  Model3.lm <- lm(Y18 ~ R18 + Y17 + R18*Y17, data=model1.dat)
  ```

- Bayesian Networks

- ```
  model1.dag <- model2network("[Y17][R18][Y18|R18:Y17]")
  model2.dag <- model2network("[Y17][R18|Y17][Y18|R18]")
  model3.dag <- model2network("[Y17][R18|Y17][Y18|R18:Y17]")
  ```

Model Comparison



	LINEAR MODELS		BAYESIAN NET		
MODEL	AIC	BIC	AIC	BIC	BGE
1	1558.127	1571.399	-2892	-2951	-2905
2	1558.223	1571.495	-2789	-2851	-2802
3	1539.206	1555.797	-2790	-2858	-2805

Importance of Effects

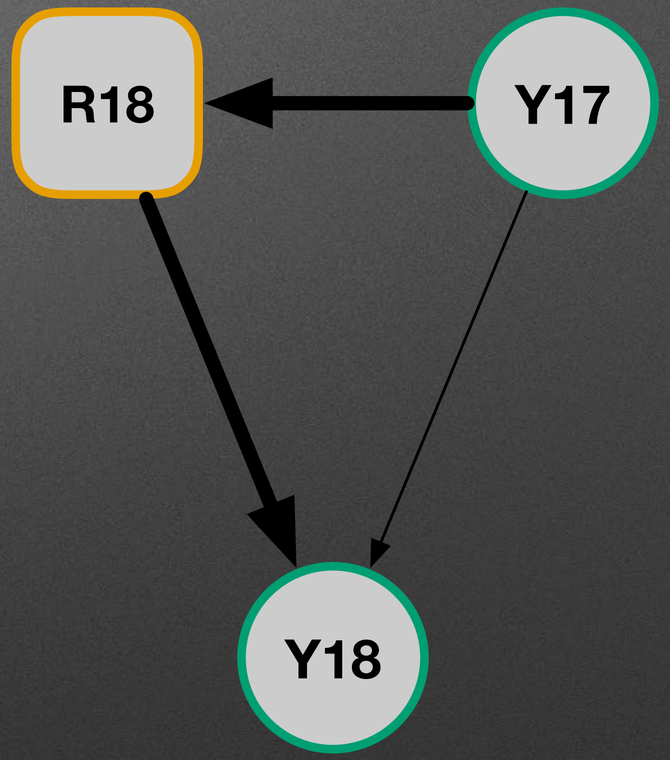
ANOVA

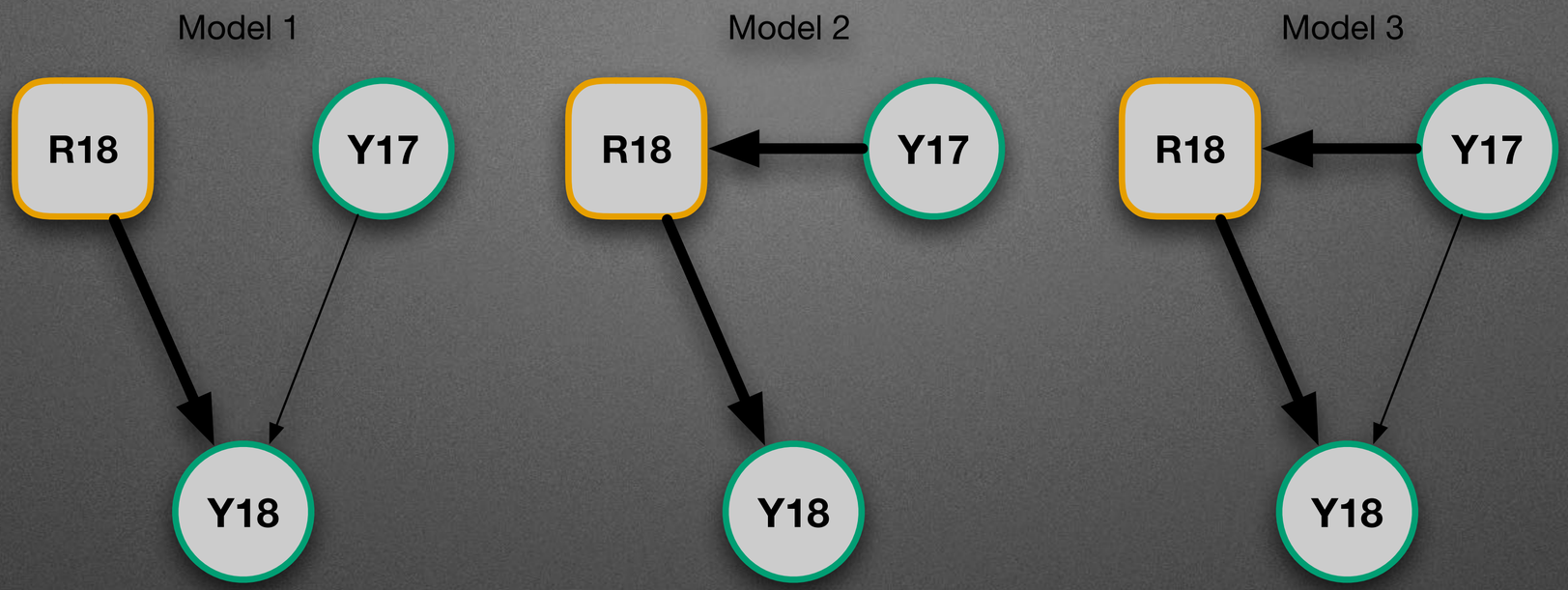
SOURCE	DF	SS	MS	F	P
R18		11429431	11429431	95851	<0.0001
Y17		11422986	11422986	95797	<0.0001
R18 X Y17		551	551	4.62	0.03
R18 (TYPE III)		6582	6582	54	<0.0001
Y17 (TYPE III)		137	137	1.13	0.29
RESIDUAL	201	23967	119		

Arc Strength

FROM	TO	LOGLIK	AIC	BIC	BGE
Y17	R18	-103.153	-102.154	-100.494	1.0000
R18	Y18	-18.388	-17.389	-15.730	0.9991
Y17	Y18	-0.045	0.954	2.613	0.0012

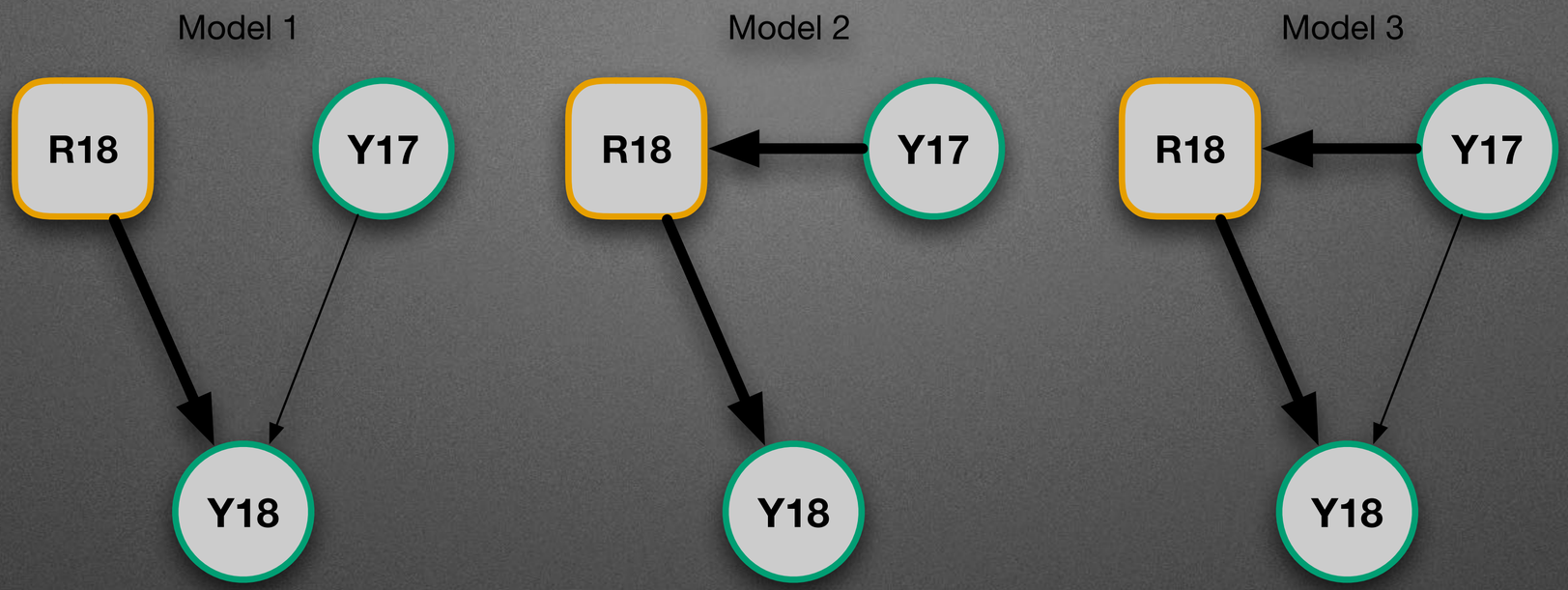
Model 3





Model Comparison

2017 Yield adds little information to this network. Does this suggest the yield map was effective?

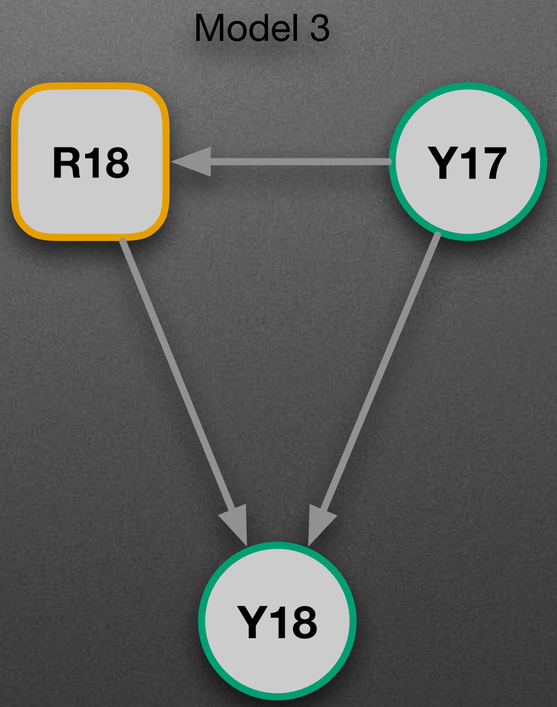


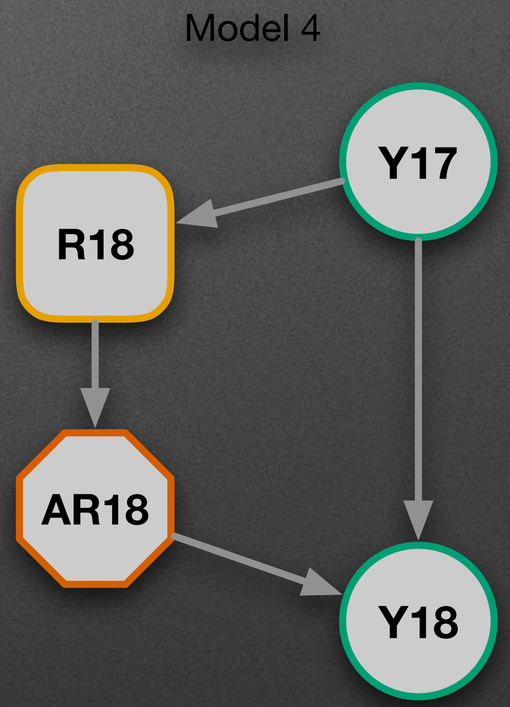
Model Comparison

2017 Yield adds little information to this network. Does this suggest the yield map was effective?

Intermediate Variables

- Intermediate variables block the direct path between variables of interest.
- This may confound a direct relationship, and make an alternate path more plausible.
  - Possible back-door path

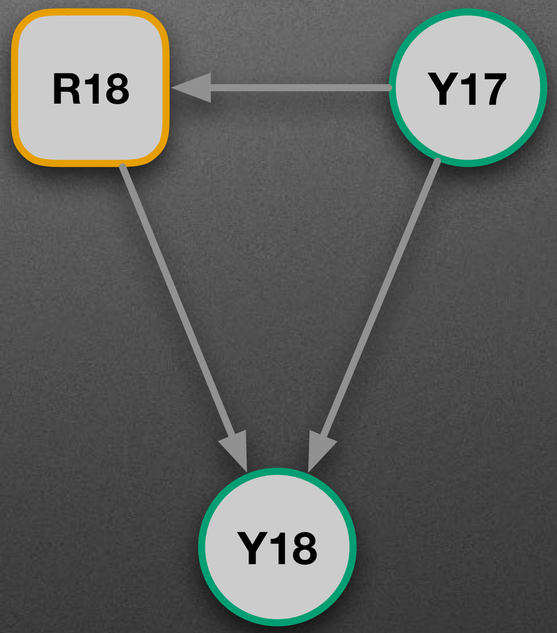




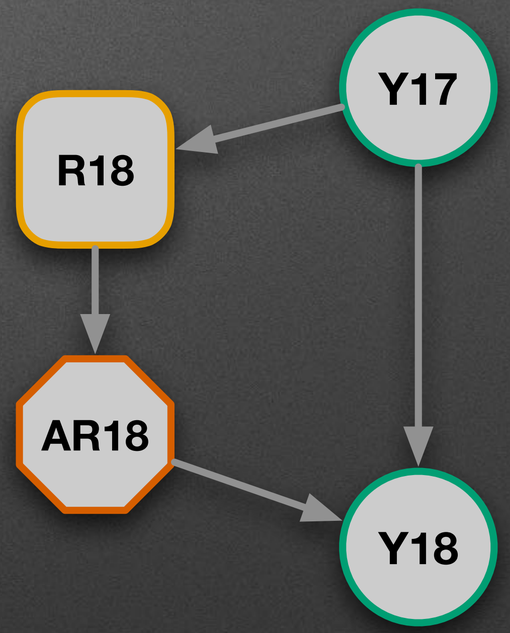
Intermediate Variables

- In this case, the ControlRate was a machine setting controlled by the operator (or seeding map prescription)
- AppliedRate (AR) was measured at the seeder, during planting.
- Thus, Control Rate is not a direct effect on Yield

Model 3



Model 4

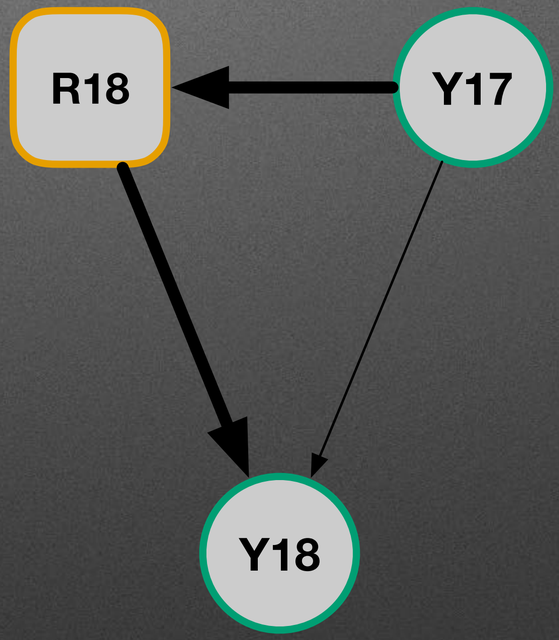


Intermediate Variables

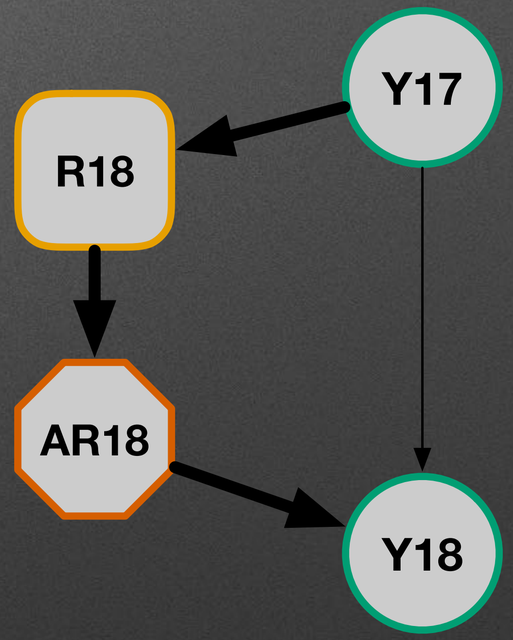
		MODEL 3
FROM	TO	STRENGTH
Y17	R18	1.0000
R18	Y18	0.9991
Y17	Y18	0.0012

		MODEL 4
FROM	TO	STRENGTH
Y17	R18	1.0000
R18	AR18	1.0000
AR18	Y18	0.9990
Y17	Y18	0.0013

Model 3



Model 4

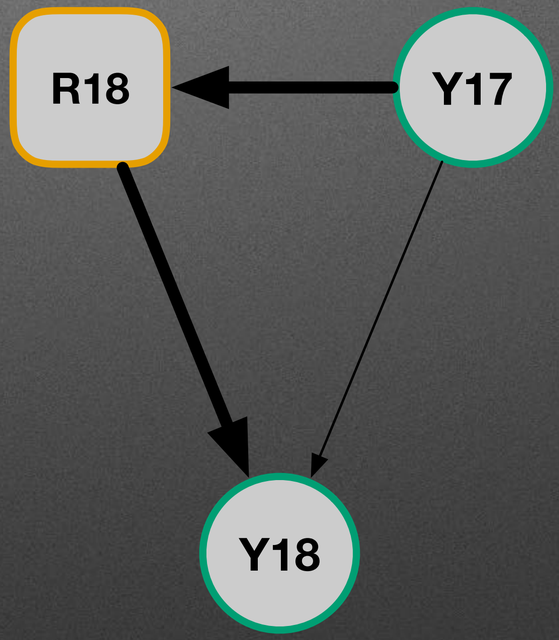


Intermediate Variables

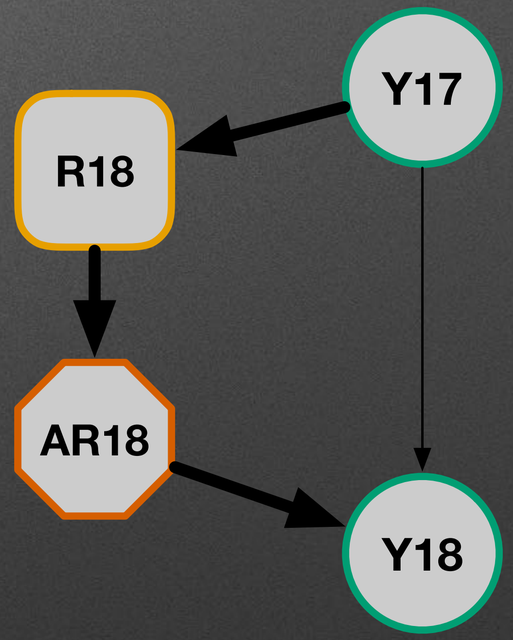
		MODEL 3
FROM	TO	STRENGTH
Y17	R18	1.0000
R18	Y18	0.9991
Y17	Y18	**0.0012**

		MODEL 4
FROM	TO	STRENGTH
Y17	R18	1.0000
R18	AR18	1.0000
AR18	Y18	0.9990
Y17	Y18	**0.0013**

Model 3

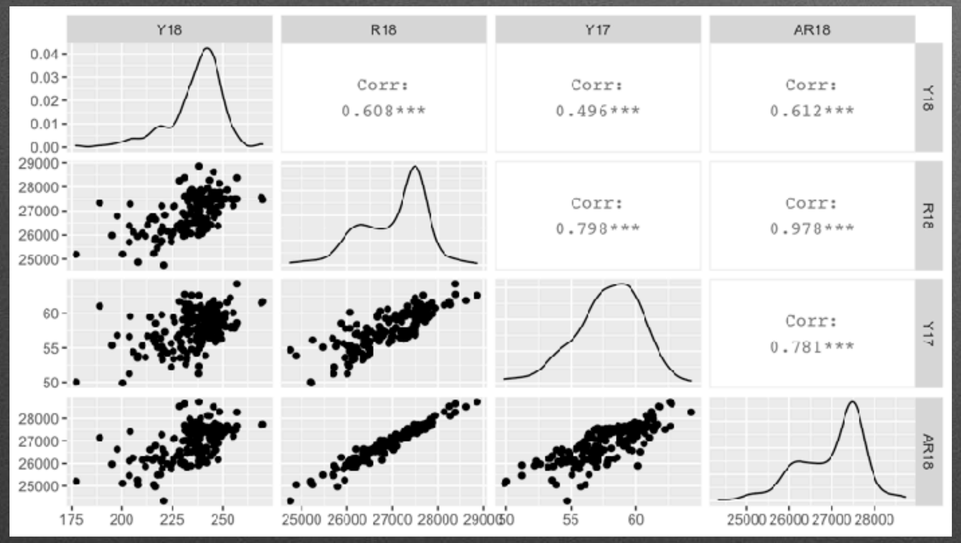


Model 4

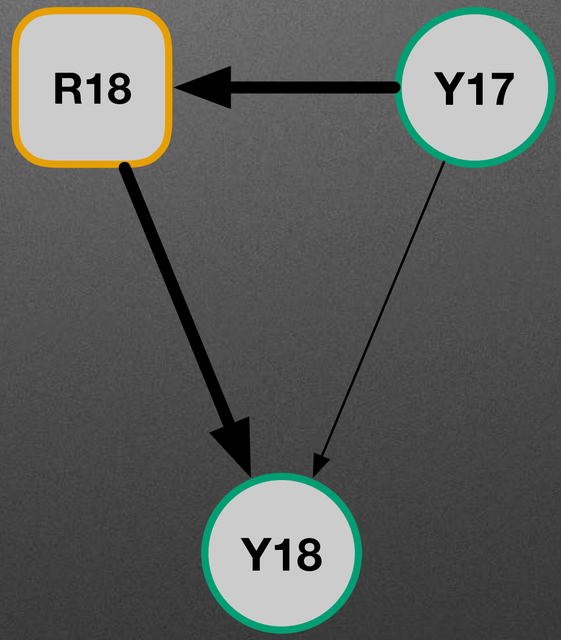


- We see a **very** slight increase in the plausibility of an alternate path, when we add an additional node in a direct path

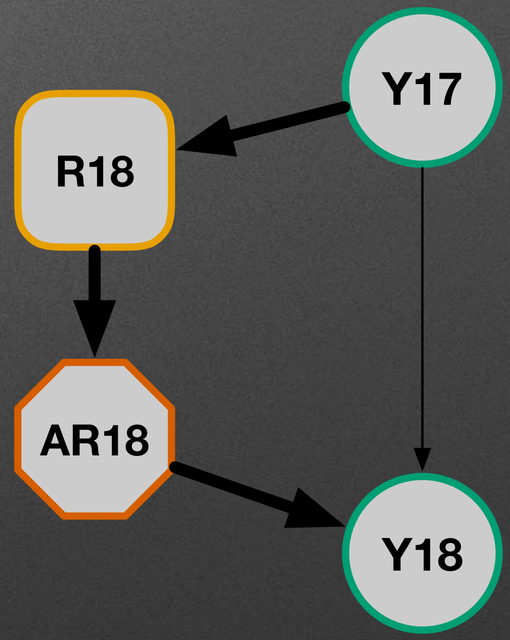
Intermediate Variables



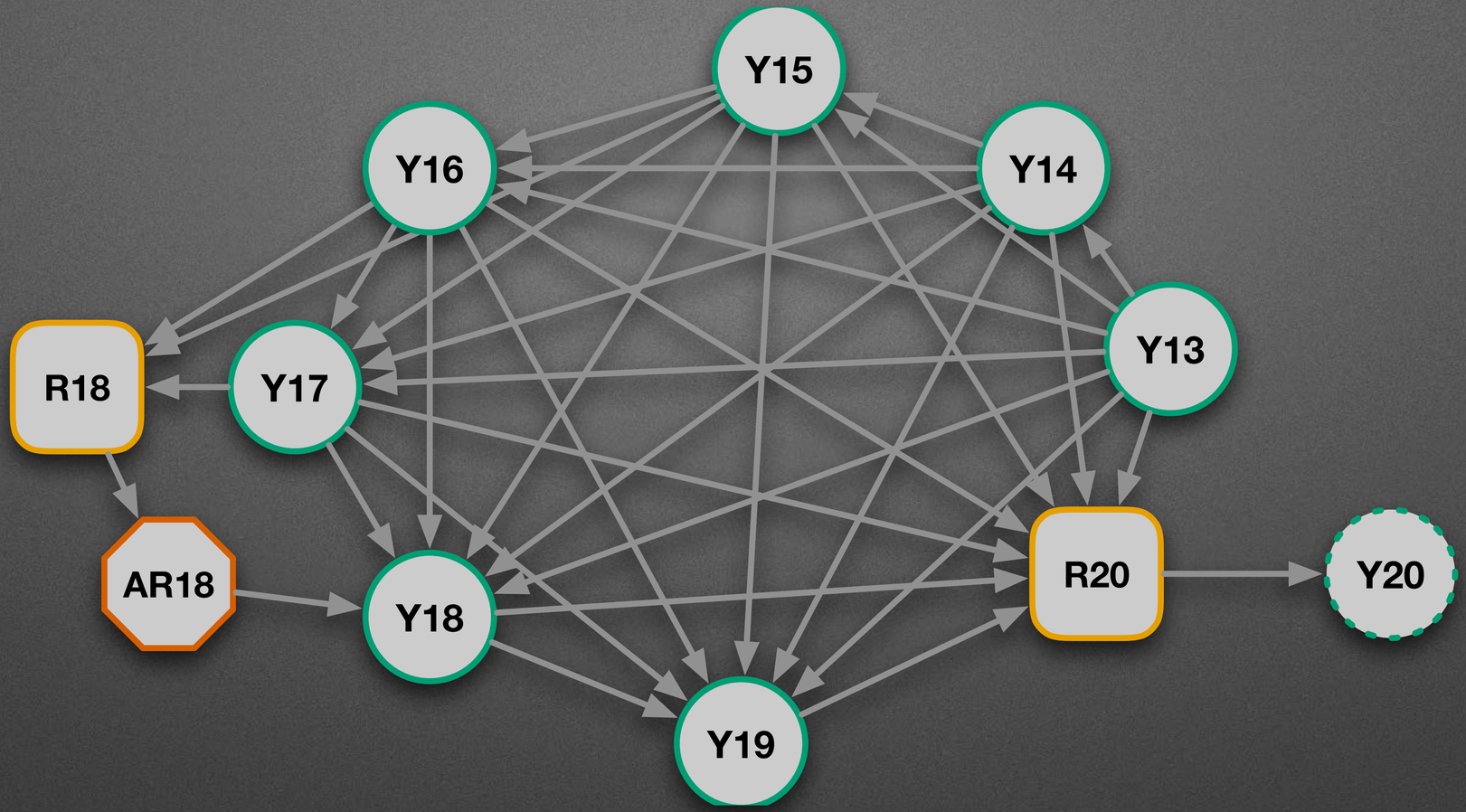
Model 3



Model 4

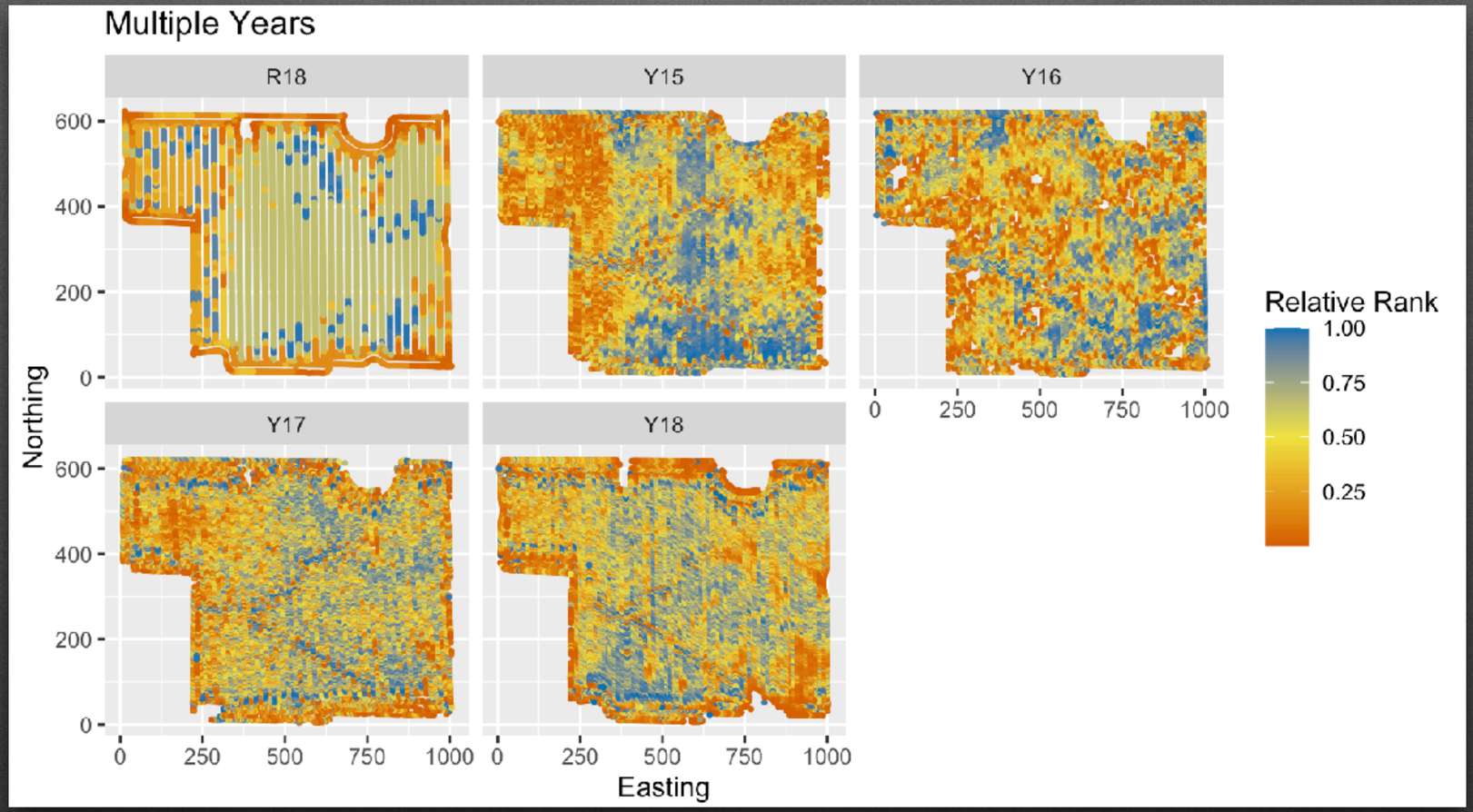


- In this case the change was small because Control Rate and Applied Rate are very highly correlated.



More Interesting Models

Up to now, we've worked with a simple network, to get an understanding or intuition of Bayesian networks. We can continue with these data, building more detailed models.

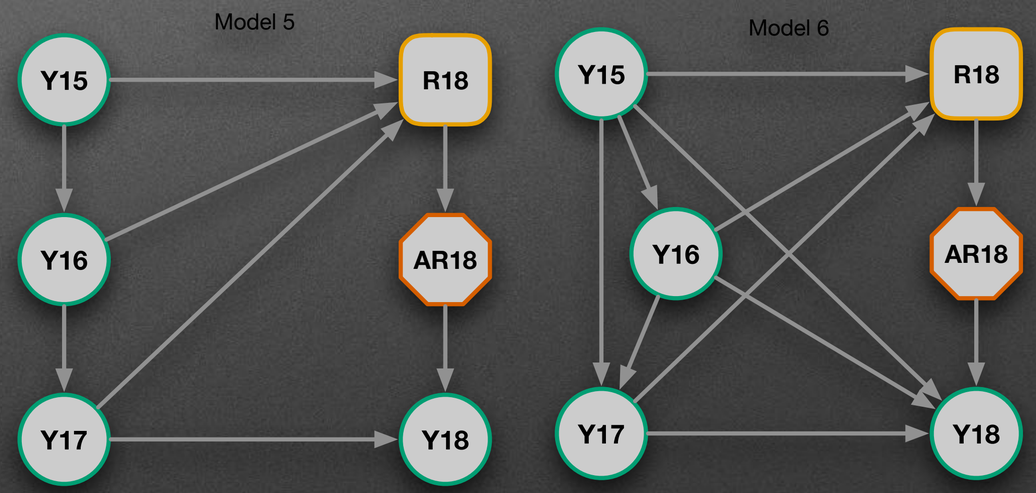


Historical Yields

How do prior years' maps influence seeding rate prescriptions and/or actual yield?

Historical Yields

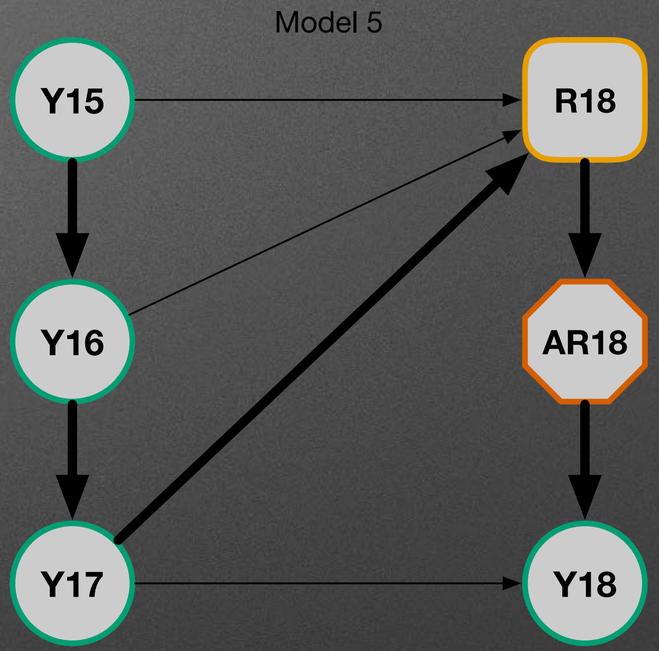
- Model 5
  - Only prior yield directly affects current yield.
  - Prior yields indirectly affect yield by determining variable rate.
- Model 6
  - All prior years directly affects all yield.
  - Prior yields indirectly affect yield by determining variable rate.



Model 5

Model	Intercept	R18	Y15	Y16	Y17	AR18	SD
AR18 | R18	1077.361	1.039					161.078
R18 | Y15 + Y16 + Y17	13724.494		-1.052	-4.270	239.757		435.182
Y15	35.382						7.868
Y16 | Y15	99.311		0.708				11.962
Y17 | Y16	46.463			0.092			2.224
Y18 | AR18 + Y17	-54.675				0.258	0.010	10.852

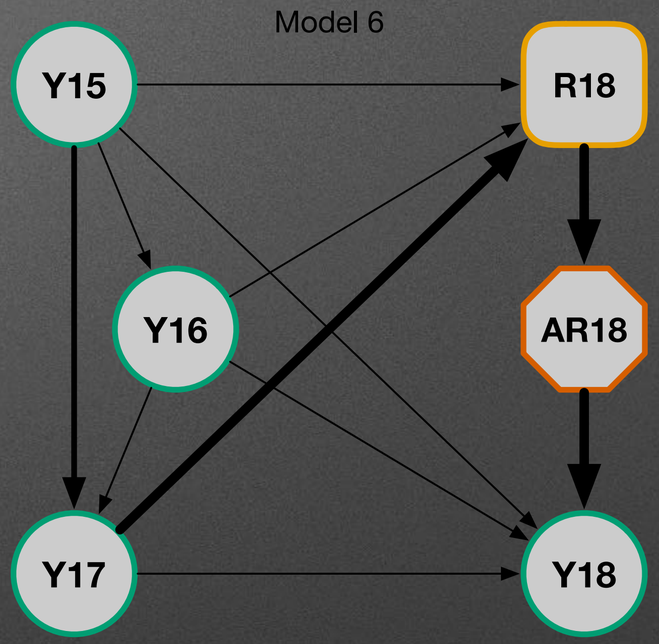
Edge	From	To	Strength (AIC)	Strength (BF)
1	Y15	Y16	-19.1010	0.999
2	Y16	Y17	-25.4679	1.000
3	Y17	R18	-89.0156	1.000
4	Y16	R18	-0.2090	<0.0001
5	Y15	R18	0.9800	0.0200
6	R18	AR18	-317.6883	1.0000
7	AR18	Y18	-18.0892	0.9990
8	Y17	Y18	0.8612	0.0013



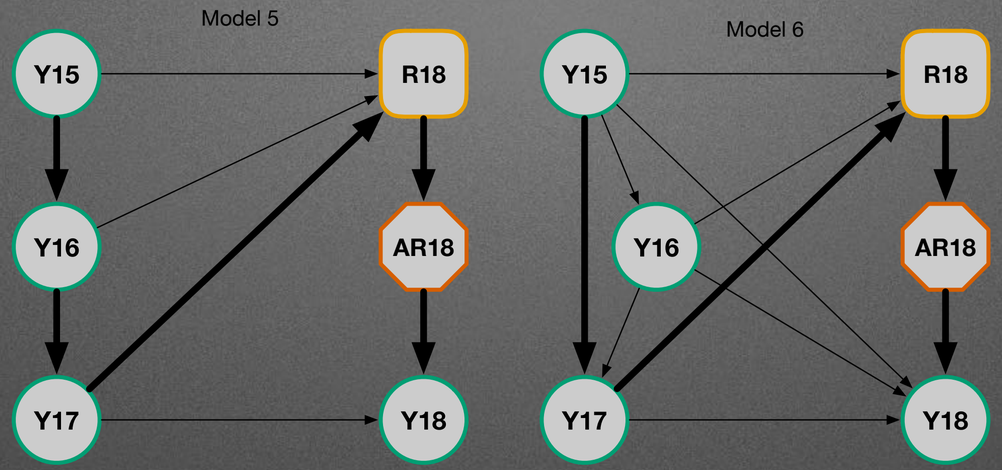
Model 6

Model	Intercept	R18	Y15	Y16	Y17	AR18	SD
AR18 | R18	1077.361	1.039					161.078
R18 | Y15 + Y16 + Y17	13724.494		-1.052	-4.270	239.757		435.182
Y15	35.382						7.868
Y16 | Y15	99.311		0.708				11.962
Y17 | Y15 + Y16	46.148		0.080	0.072			2.155
Y18 | AR18 + Y15 + Y16 + Y17	-61.380		-0.328	0.039	0.663	0.010	10.664

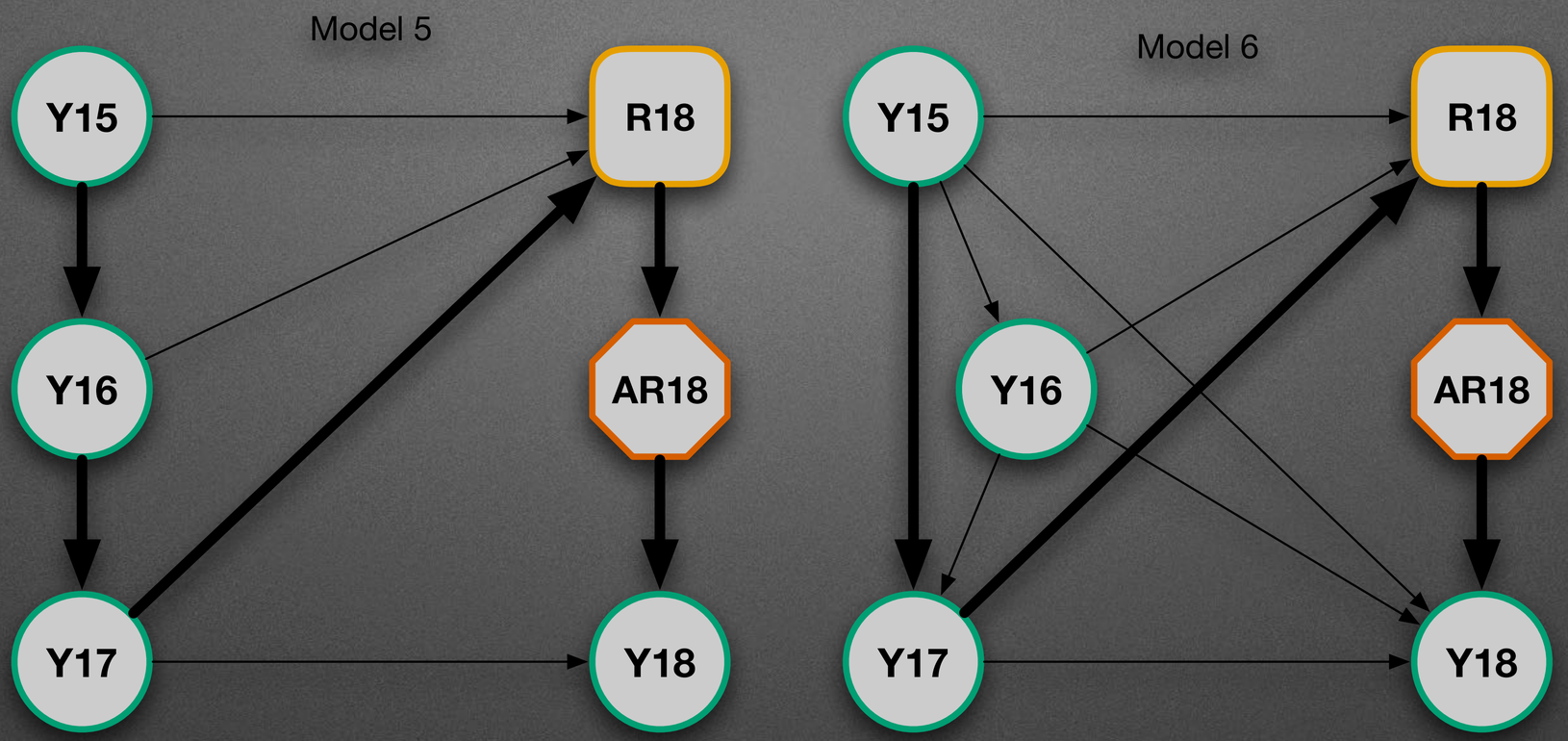
Edge	From	To	Strength (AIC)	Strength (BF)
1	Y15	Y16	-19.101	0.997
2	Y16	Y17	-14.005	0.999
3	Y17	R18	-89.016	1.000
4	Y16	R18	-0.209	<0.001
5	Y15	R18	0.980	<0.001
6	R18	AR18	-317.688	1.000
7	AR18	Y18	-17.117	0.997
8	Y17	Y18	0.229	0.002
9	Y15	Y17	-5.920	0.589
10	Y16	Y18	0.841	<0.001
11	Y15	Y18	-3.547	0.022



Model Summary

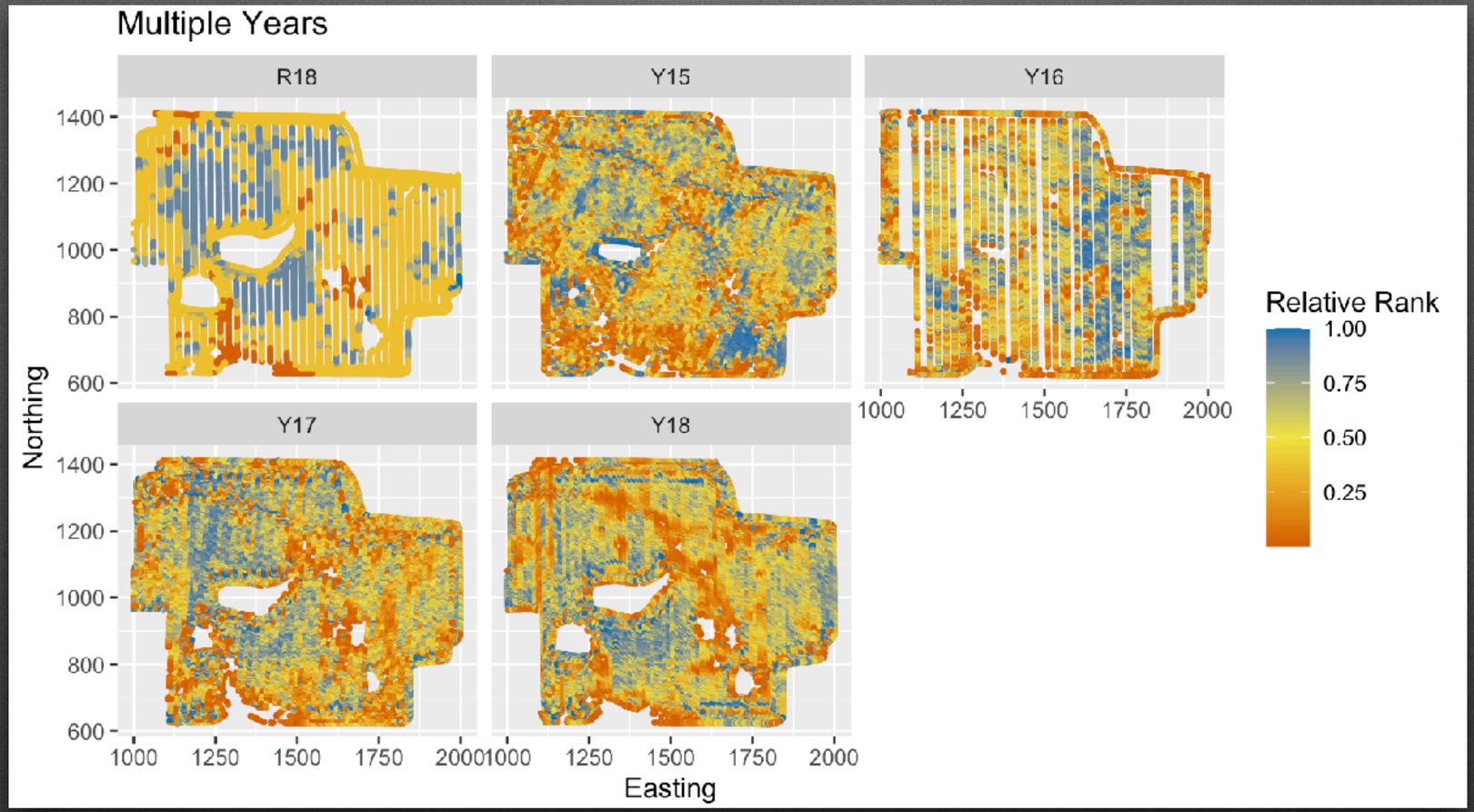


			MODEL 5		MODEL 6	
Edge	From	To	Strength (AIC)	Strength (BF)	Strength (AIC)	Strength (BF)
1	Y15	Y16	-19.101	0.999	-19.101	0.997
2	Y16	Y17	-25.468	1.000	-14.005	0.999
3	Y17	R18	-89.016	1.000	-89.016	1.000
4	Y16	R18	-0.209	<0.0001	-0.209	<0.001
5	Y15	R18	0.980	0.020	0.980	<0.001
6	R18	AR18	-317.688	1.000	-317.688	1.000
7	AR18	Y18	-18.089	0.999	-17.117	0.997
8	Y17	Y18	0.861	0.001	0.229	0.002
9	Y15	Y17			-5.920	0.589
10	Y16	Y18			0.841	<0.001
11	Y15	Y18			-3.547	0.022



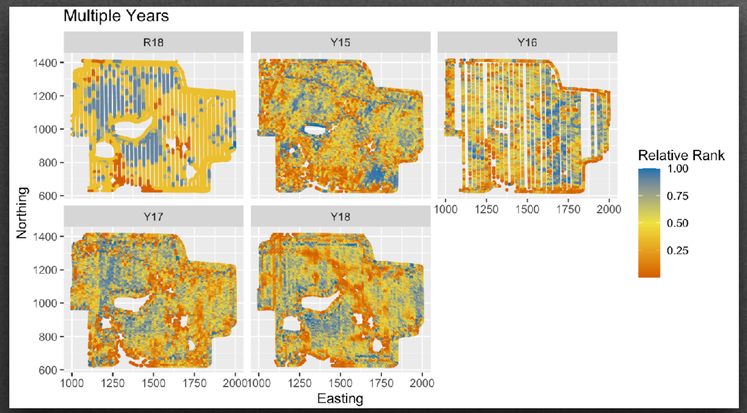
Causal Inference?

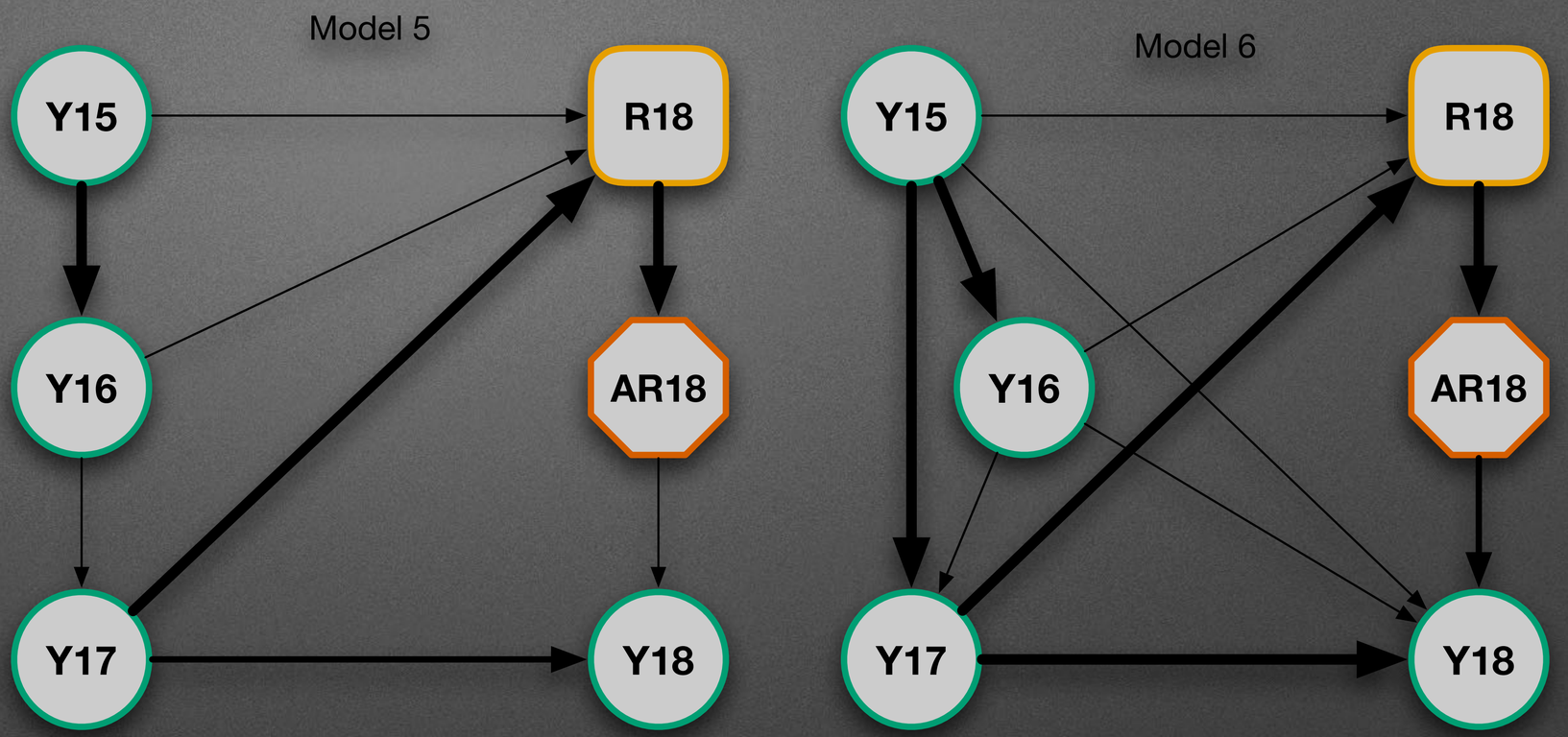
Did the Seeding Map have an effect?



Repeatability

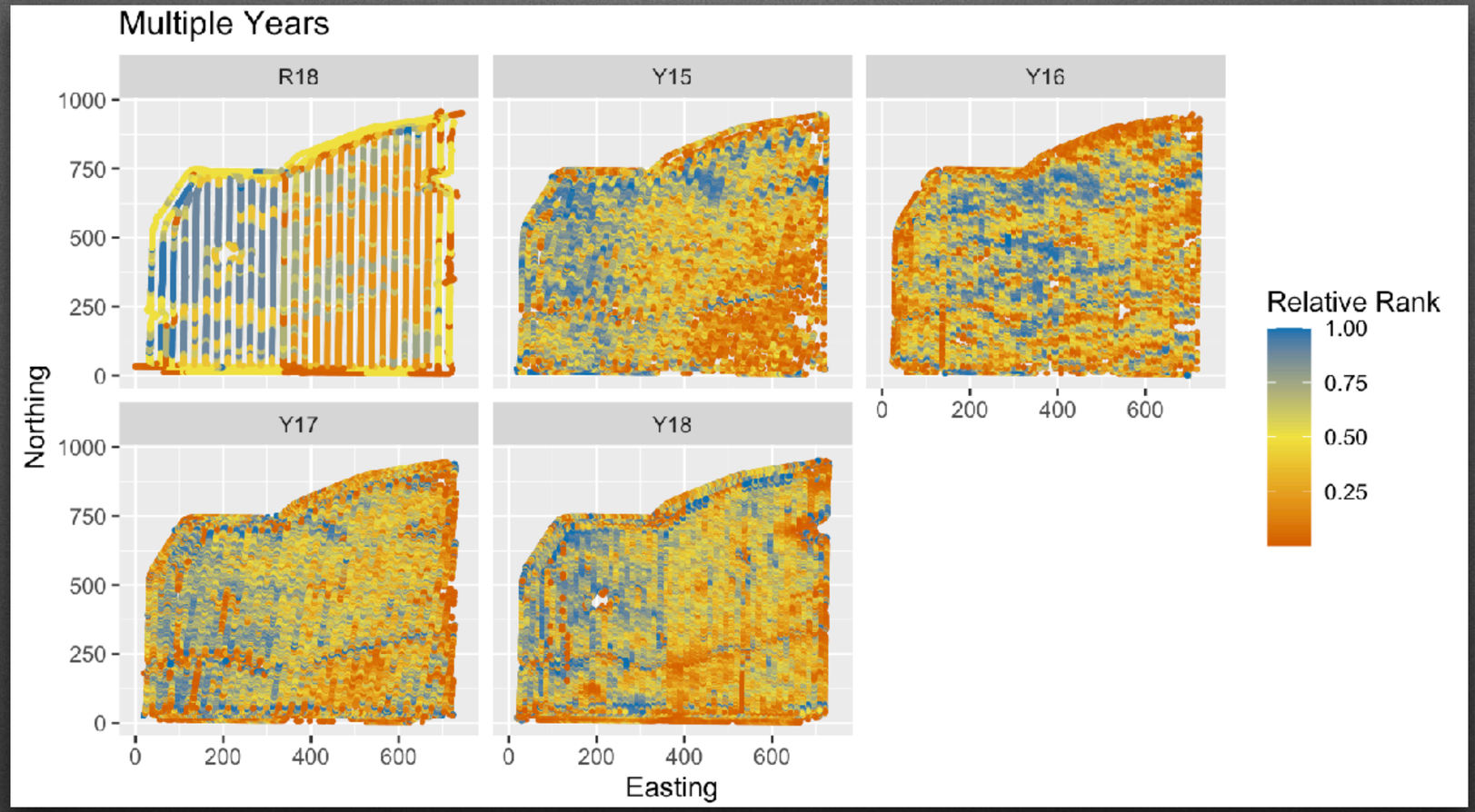
Field B : Northeast corner adjacent field, similar rotation





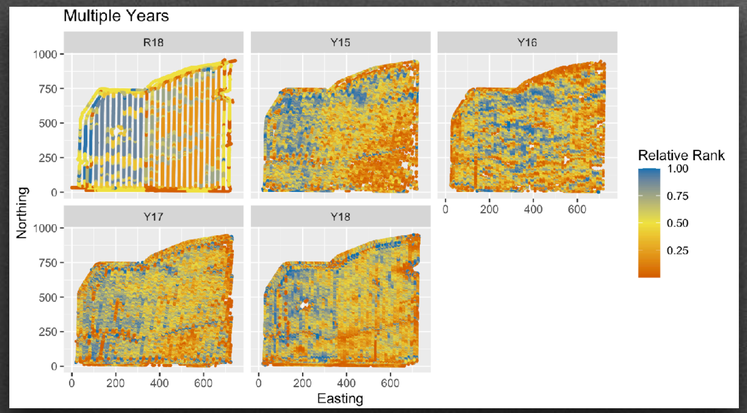
Repeatability

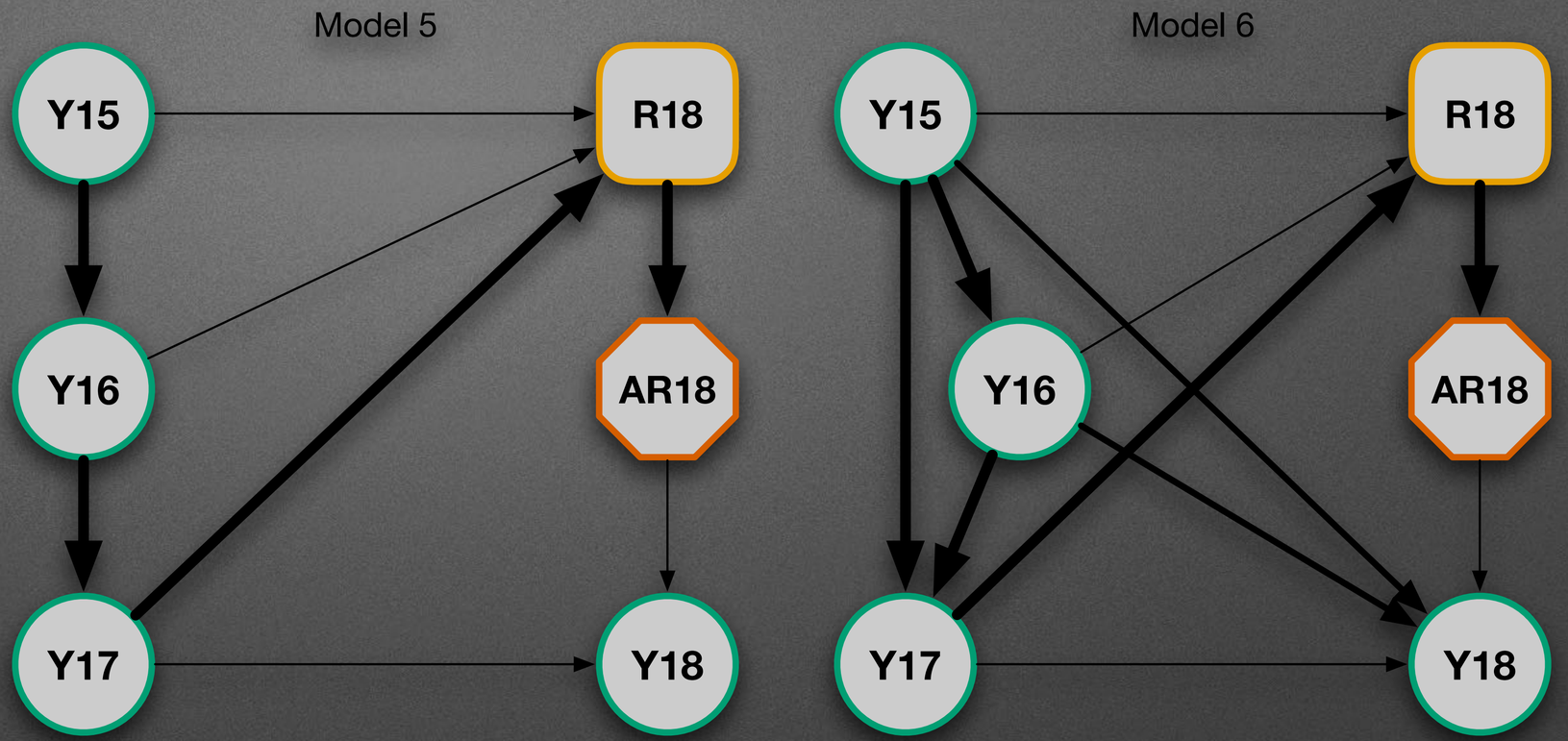
Field B : Northeast corner adjacent field, similar rotation
Seeding map was less effective



Repeatability

Field E : 800 Northwest, different soil map, similar rotation



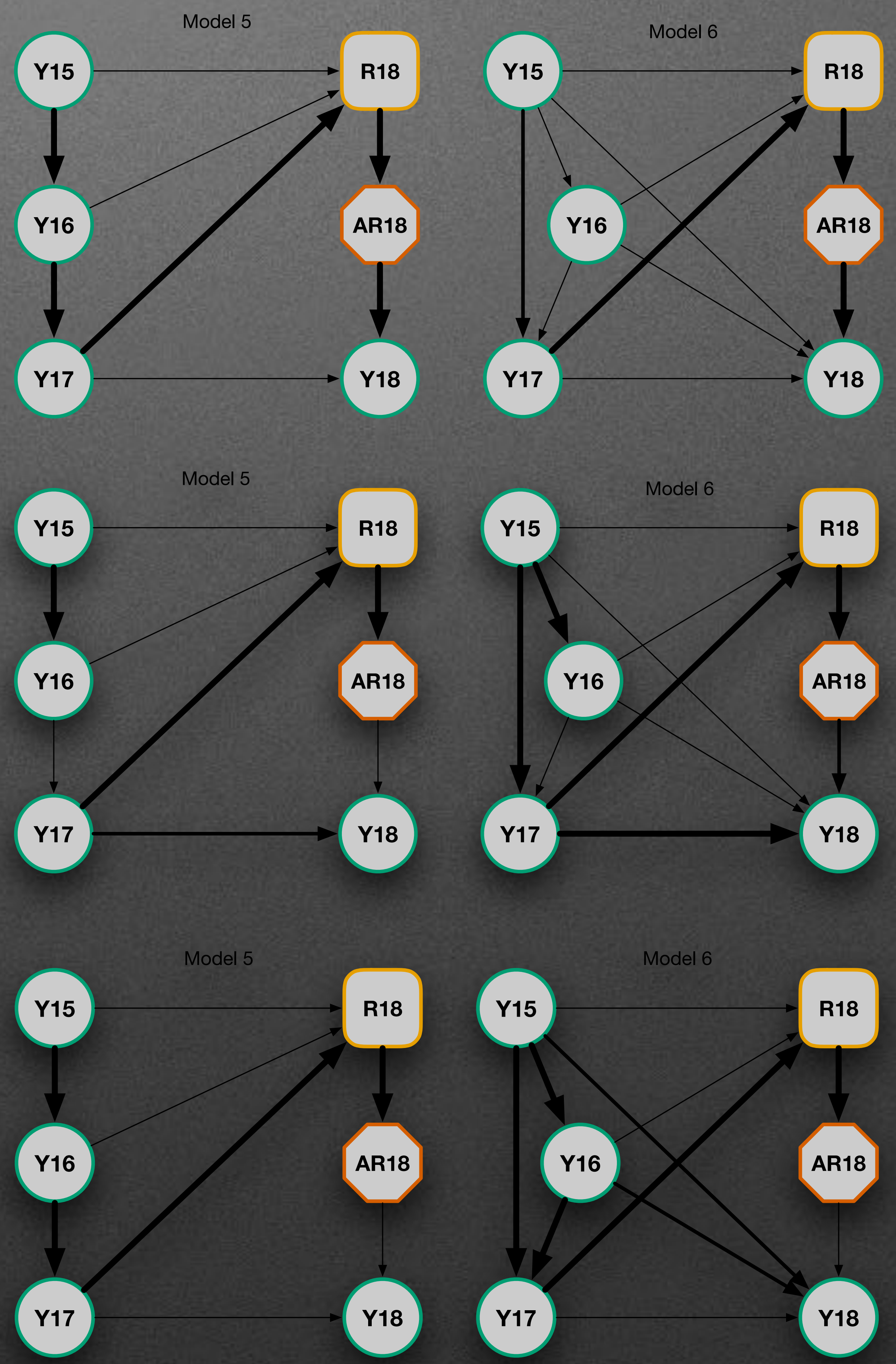


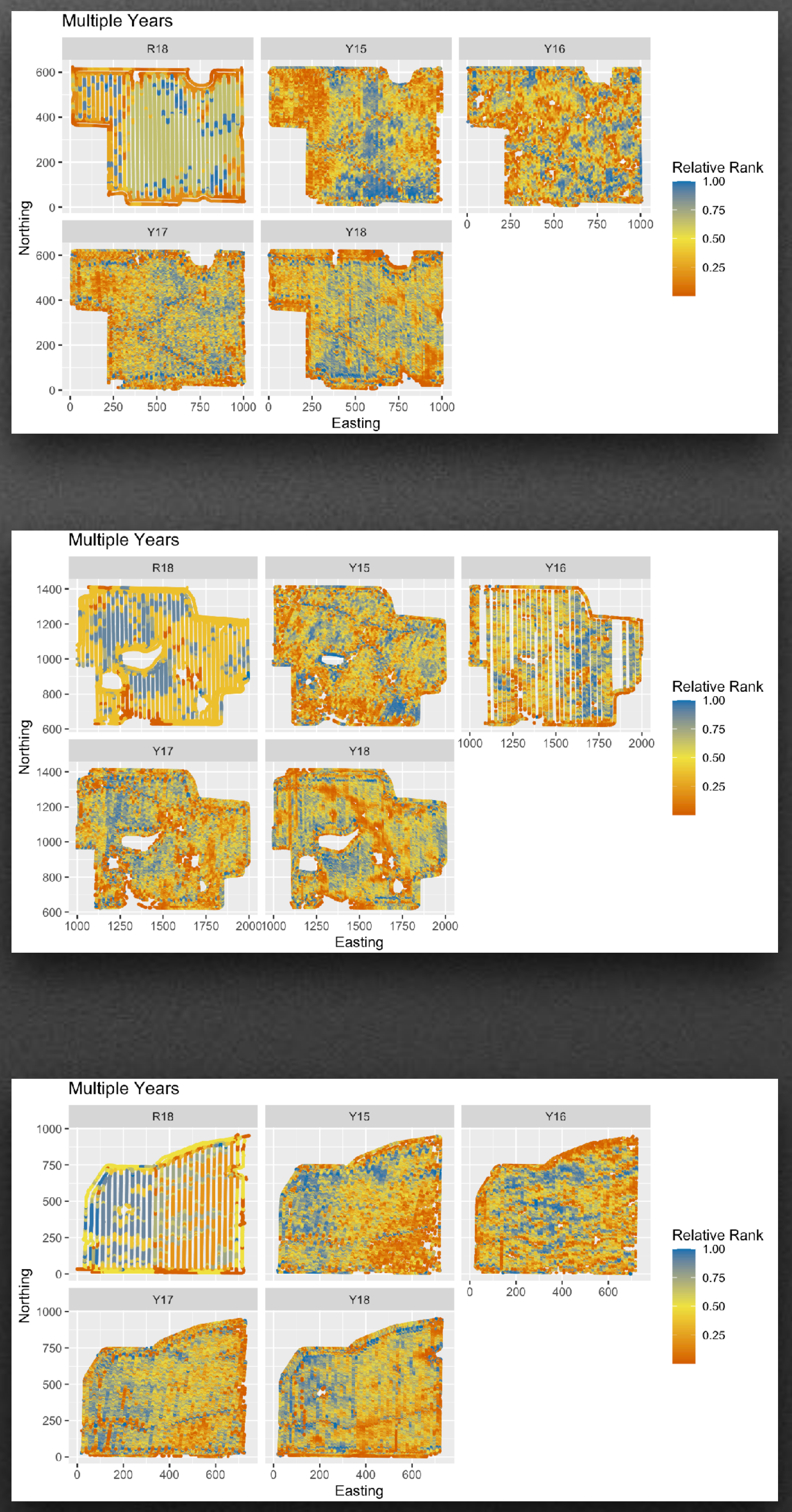
Repeatability

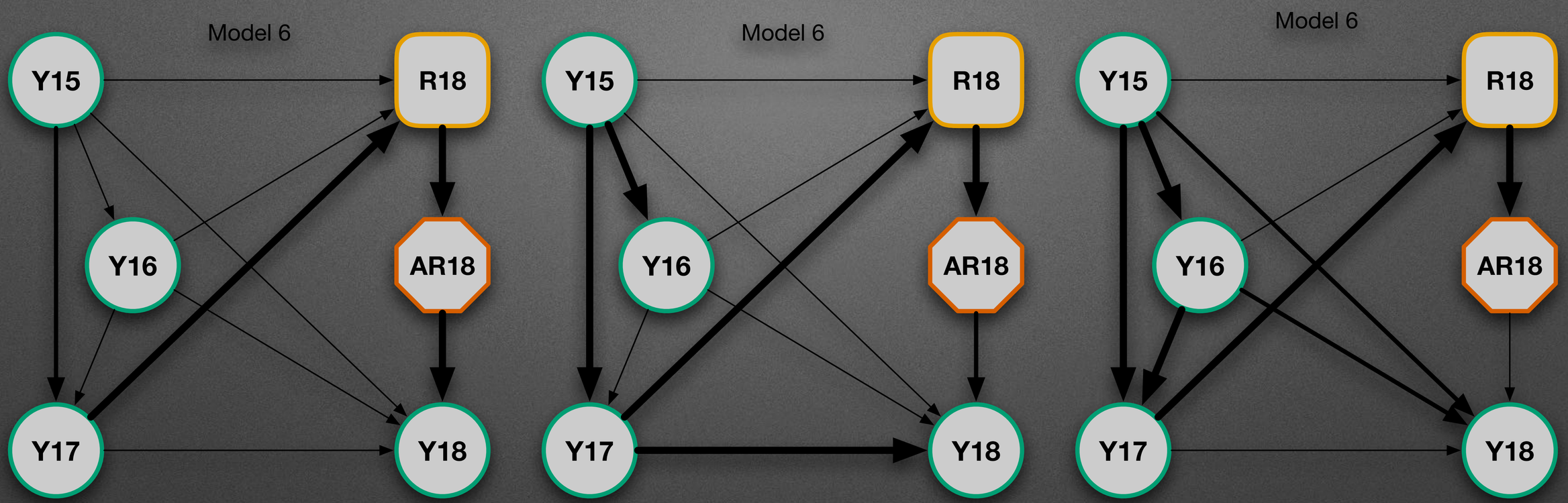
Field E : 800 Northwest, alternate rotation
Seeding map was not effective

Models 5 and 6

- Summary
    - Seeding maps influenced only by 2017 yields.
    - Prescriptions were not consistently effective at influencing yield.

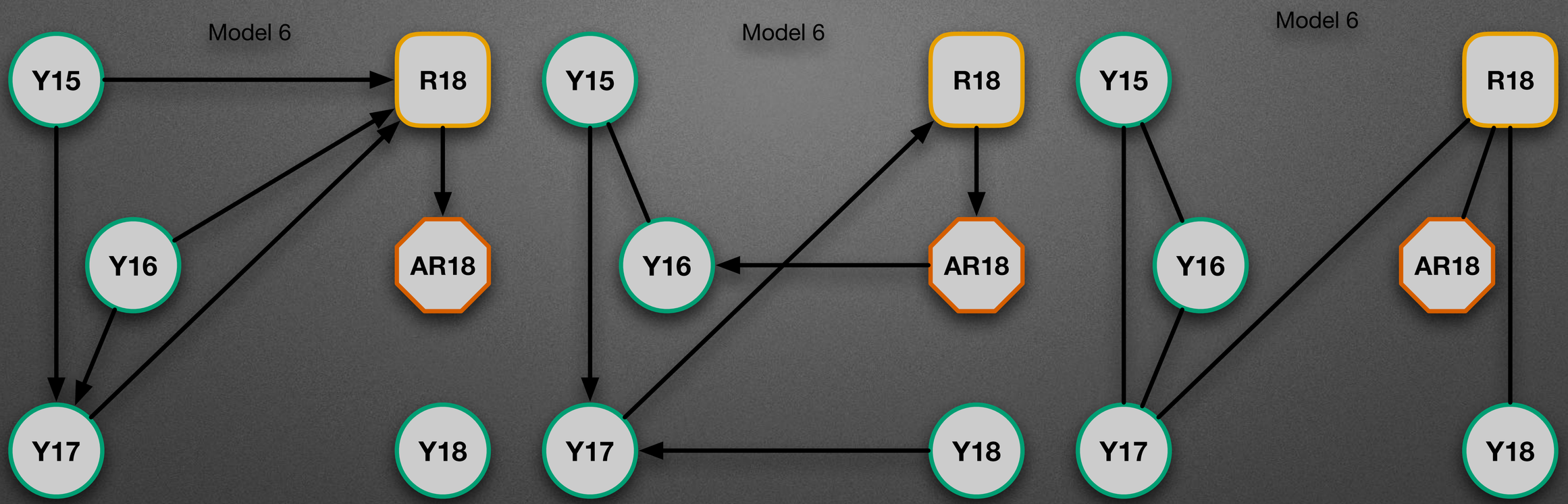






Cautionary Note

So far, we've used DAG to study plausible, predefined causal relationships.
Sometimes DAG are used to 'learn' (probabilistic) causal relationships.

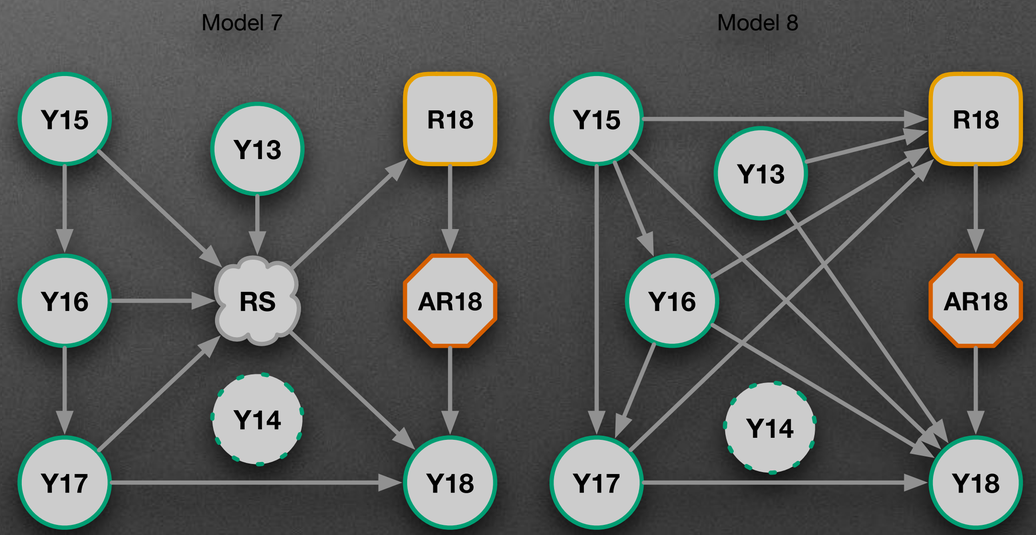


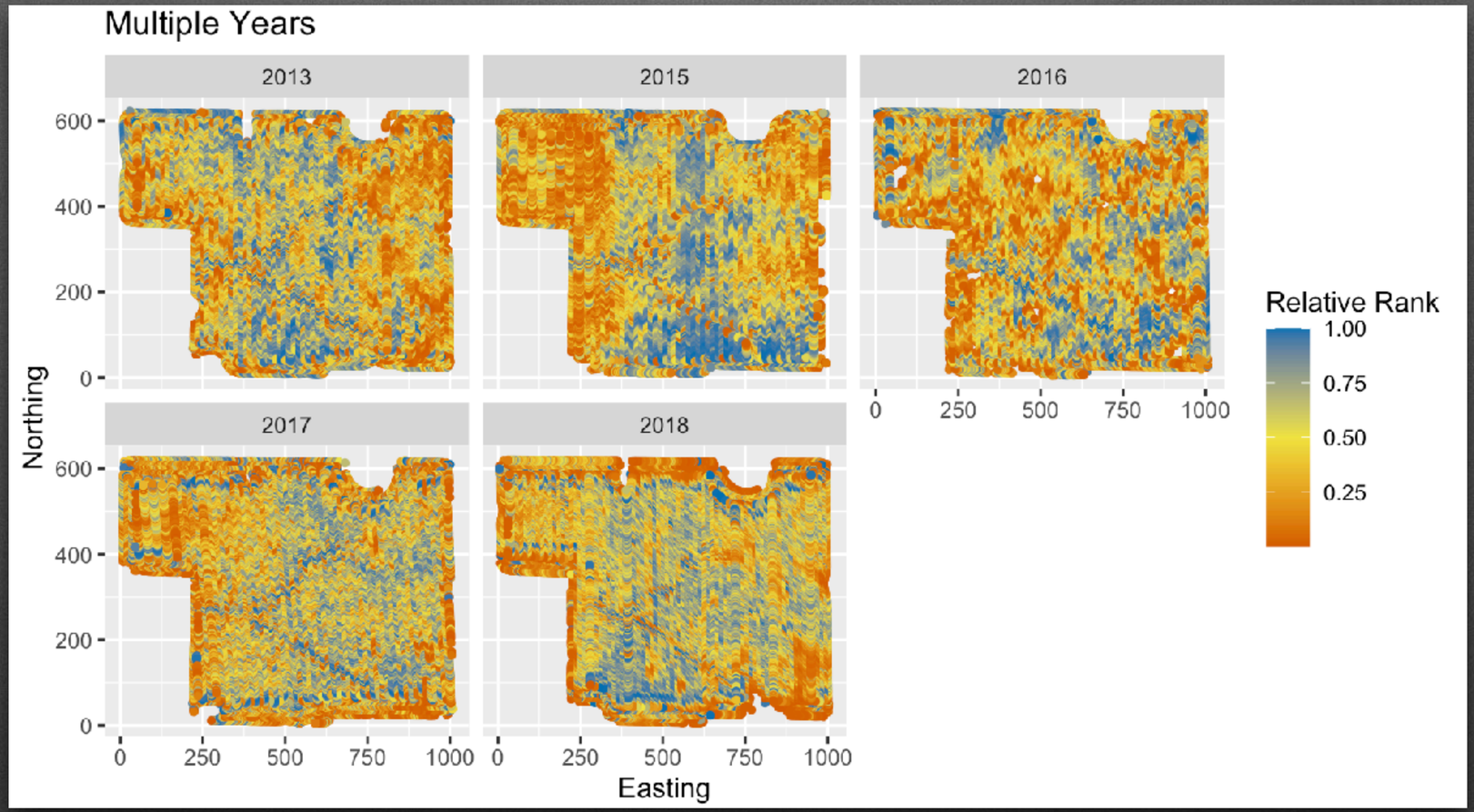
Cautionary Note

It doesn’t always go well.

Response Surface

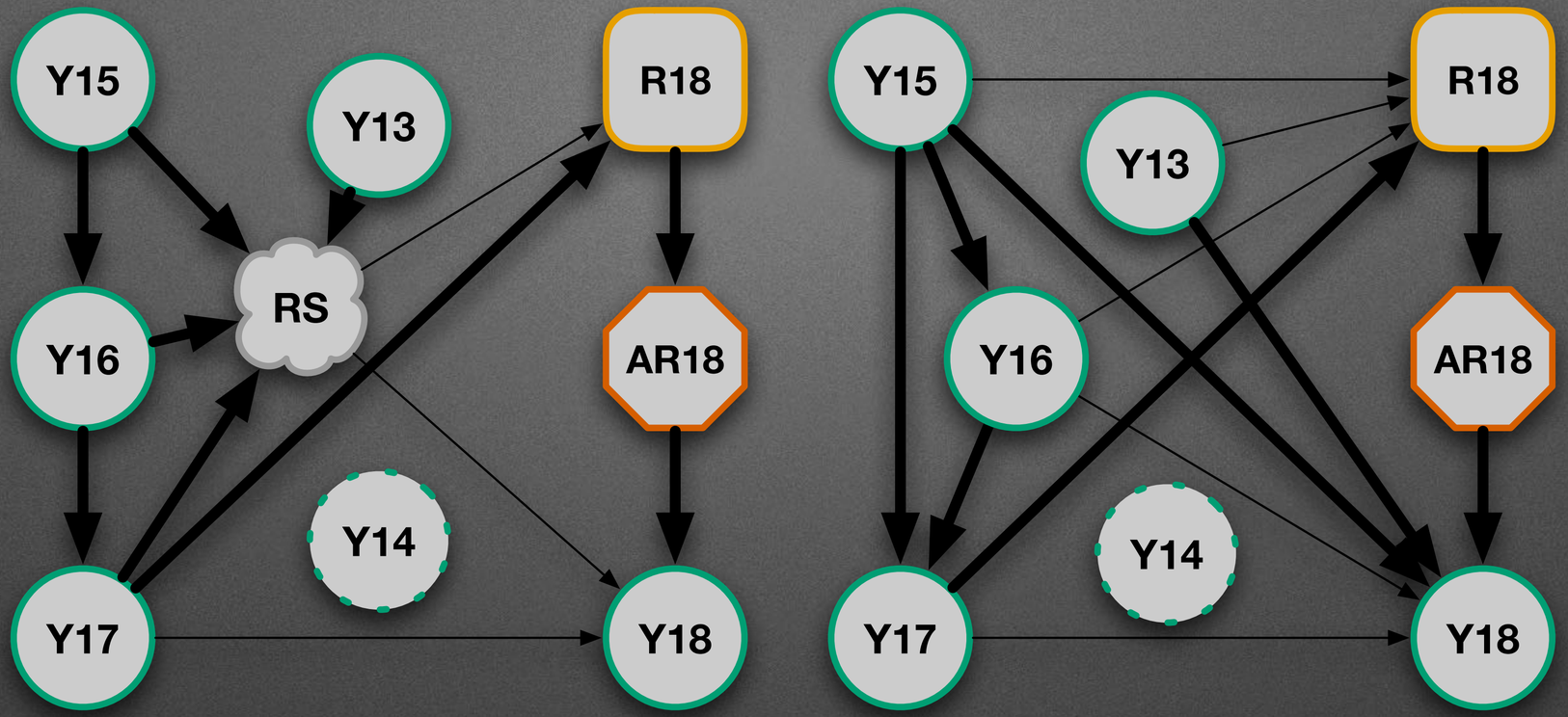
- Model 7
  - Only prior yield directly affects current yield.
  - Pooled information (yield rank) from prior years estimate a response surface. Did the response surface determine 2018 seeding rate?
  - Does the response surface determine 2018 yield?
- Model 8
  - Prior yield from each year directly influences 2018 yield.
  - Prior yield from each year directly influence 2018 seeding map.





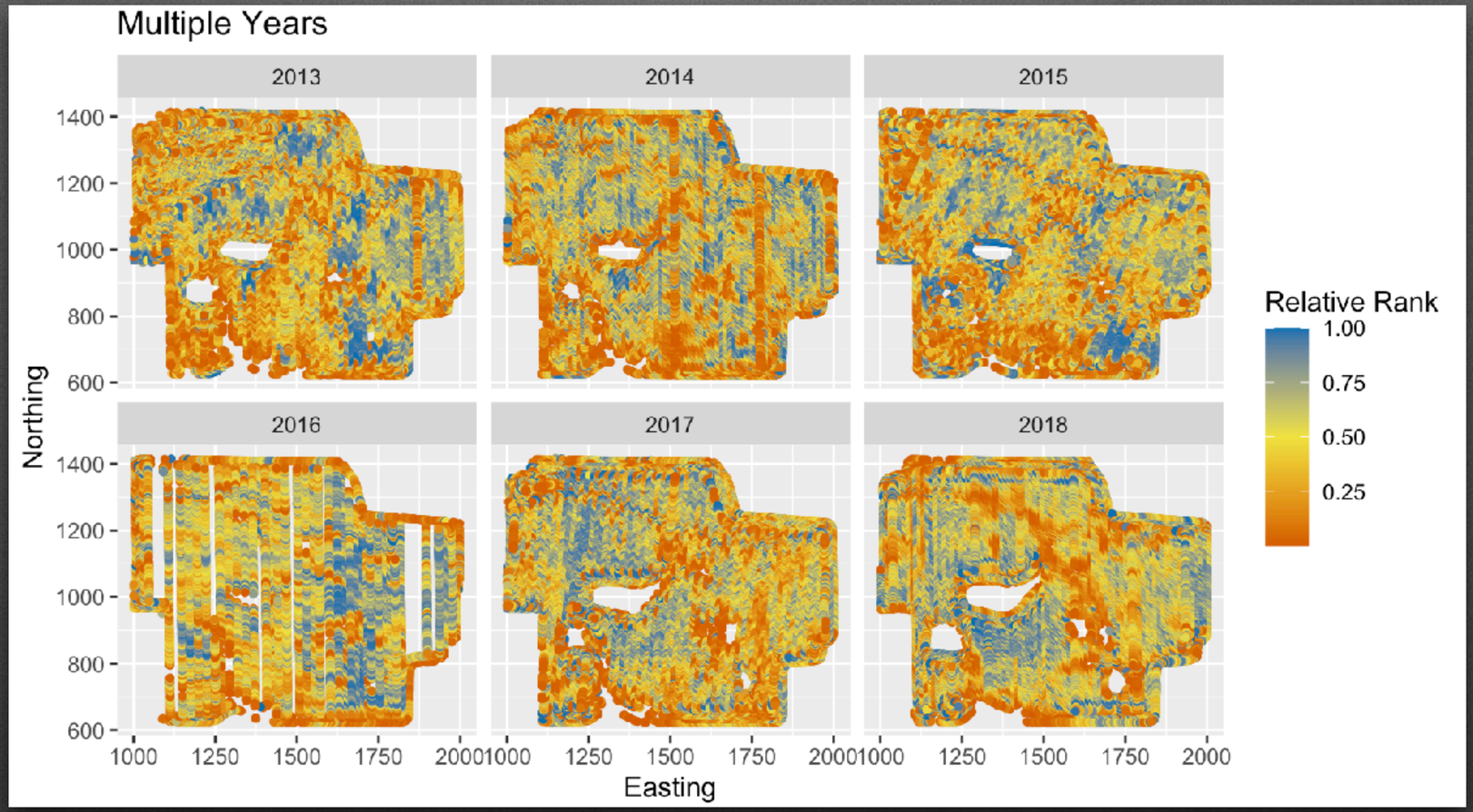
Response Surface

Field A - One additional year. Data are missing for 2014.



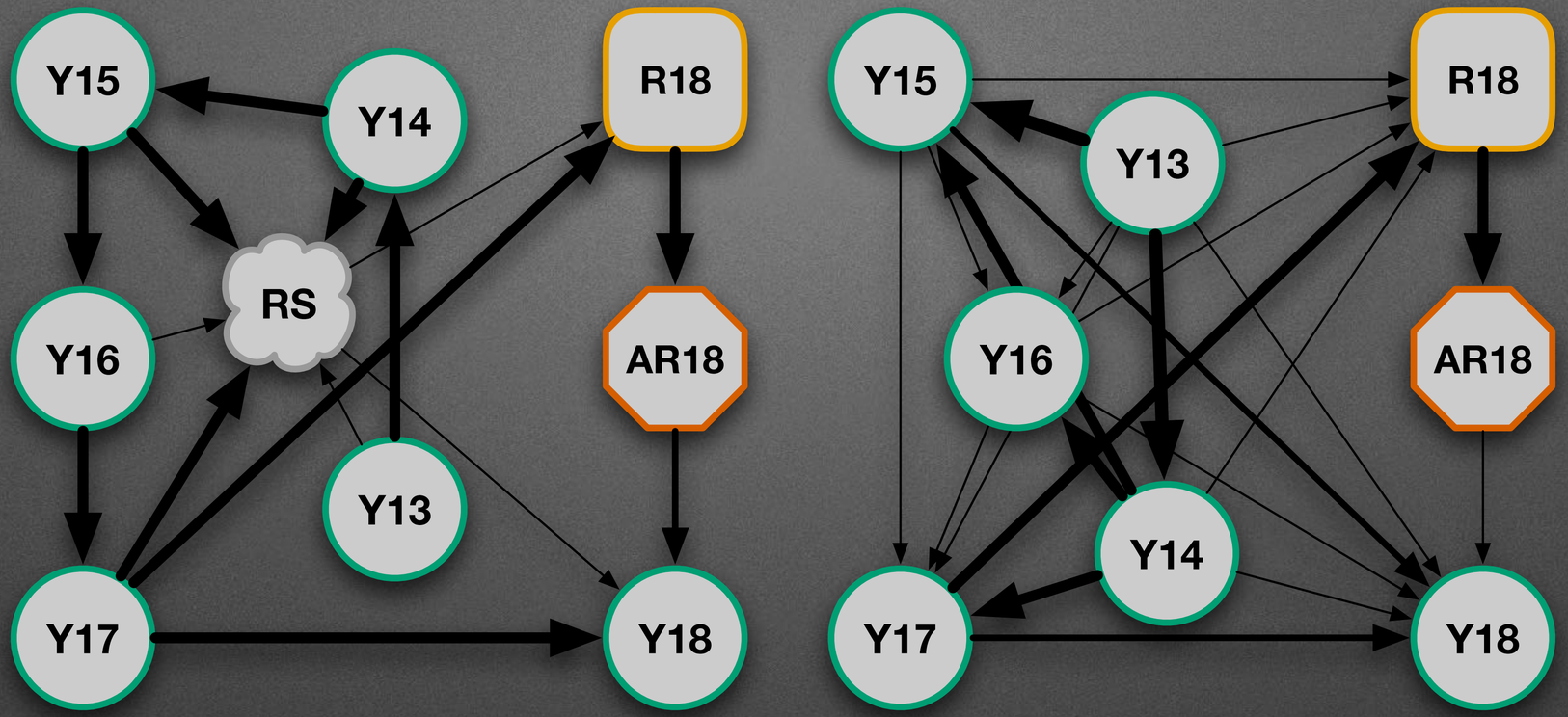
Response Surface

Field A - One additional year. Data are missing for 2014.



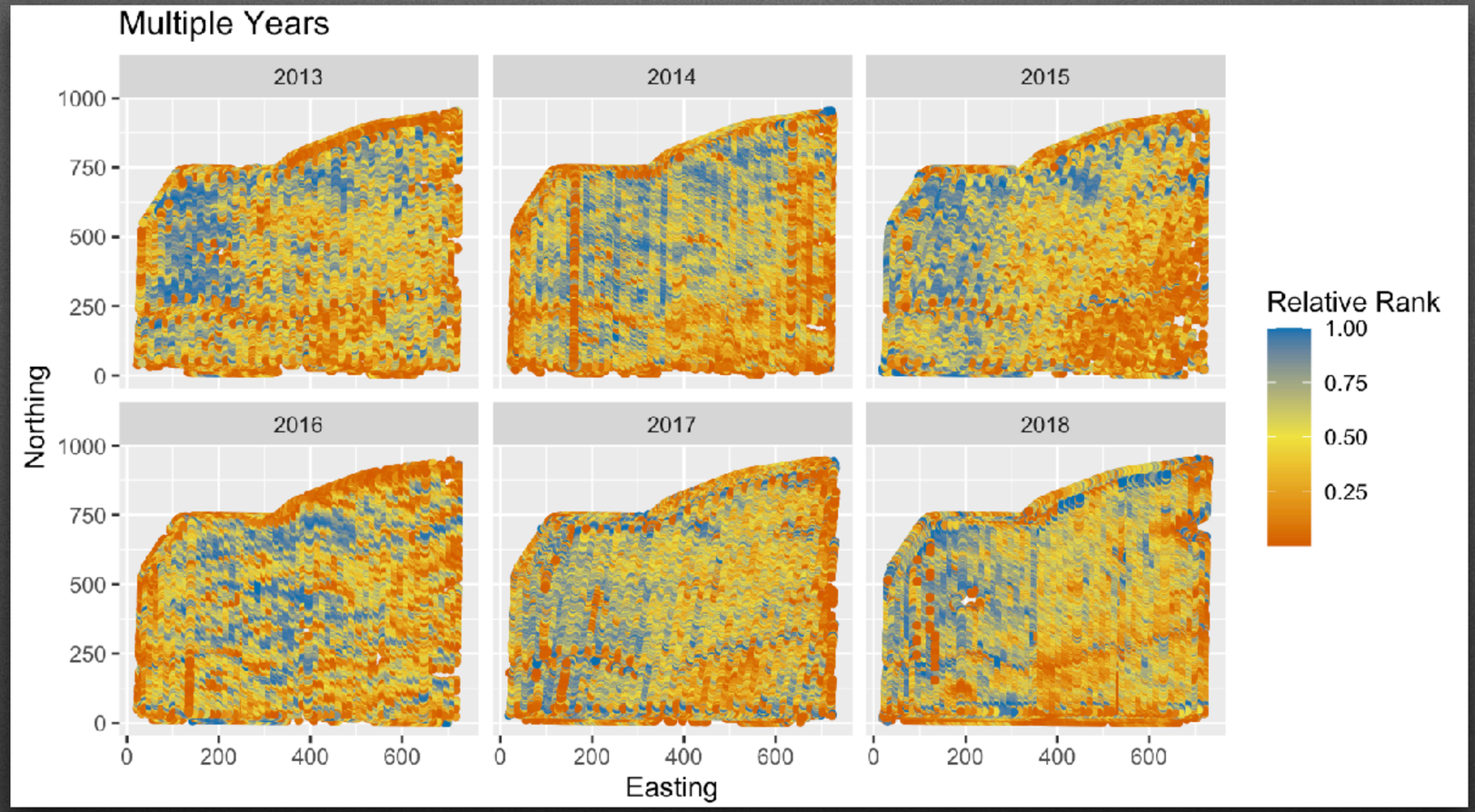
Response Surface

Field B - Two additional years.



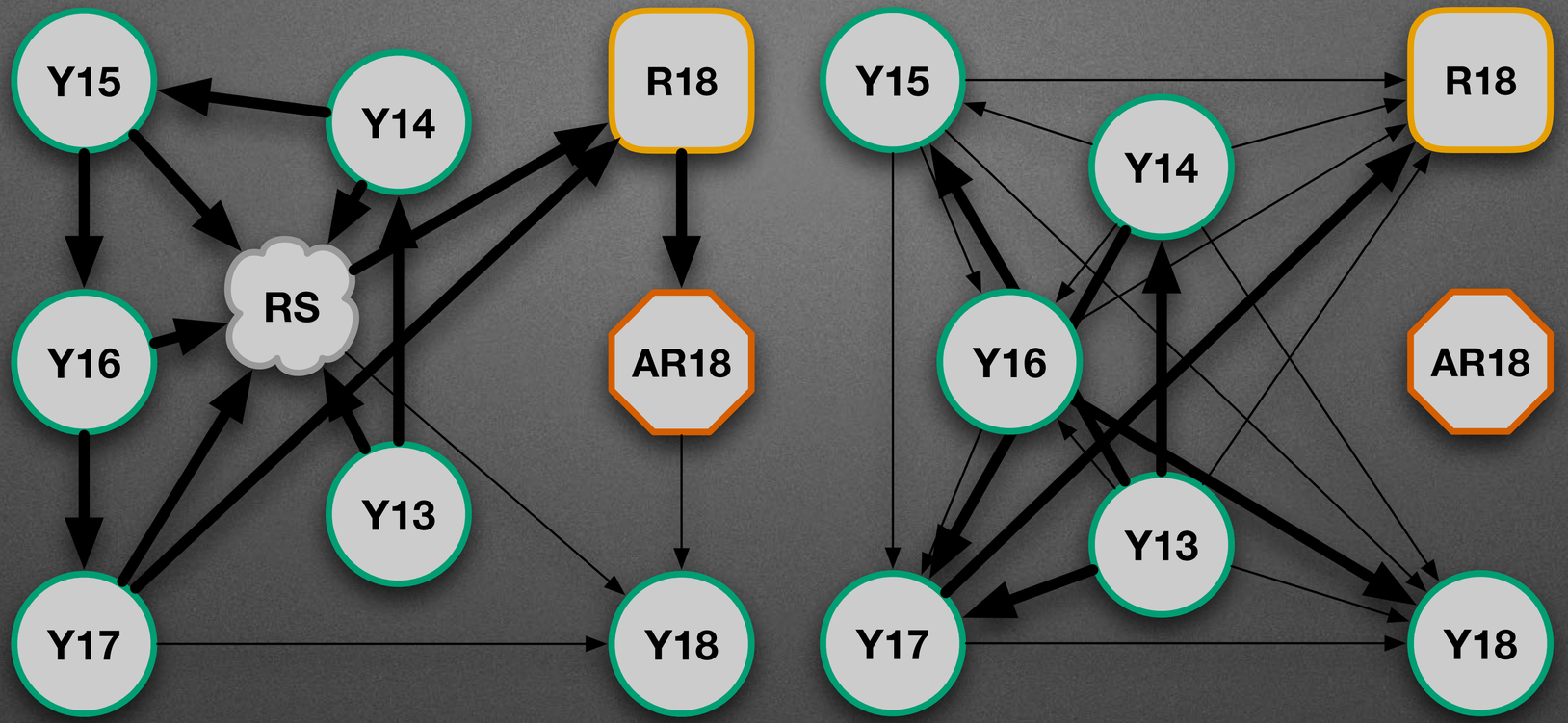
Response Surface

Field B - Two additional years.



Response Surface

Field E - Two additional years.

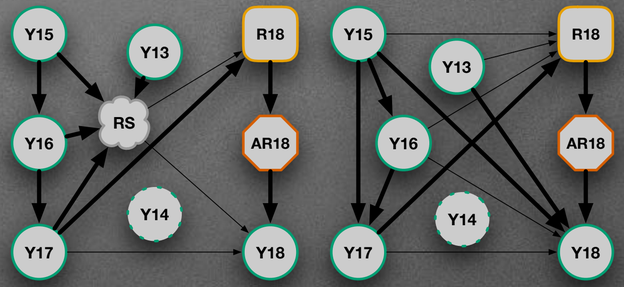


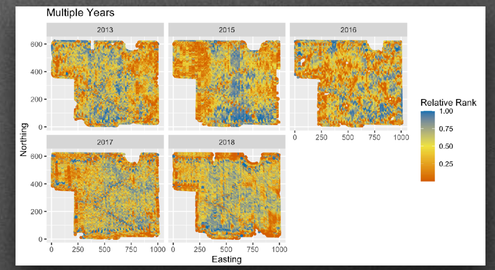
Response Surface

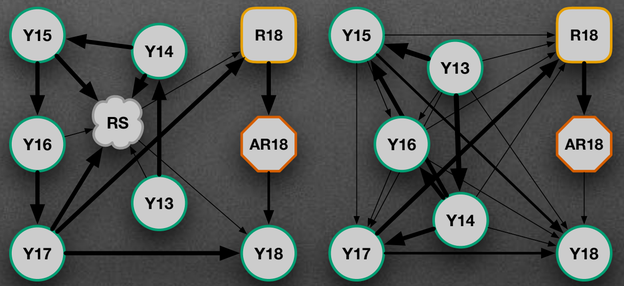
Field E - Two additional years.

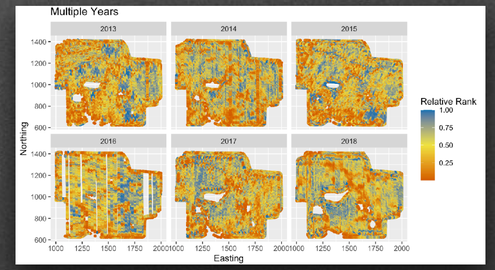
Models 7 and 8

- Summary
  - Response surface (yield map) does not appear to influence 2018 seeding map.
  - Response surface (yield map) does not appear to influence 2018 yield.



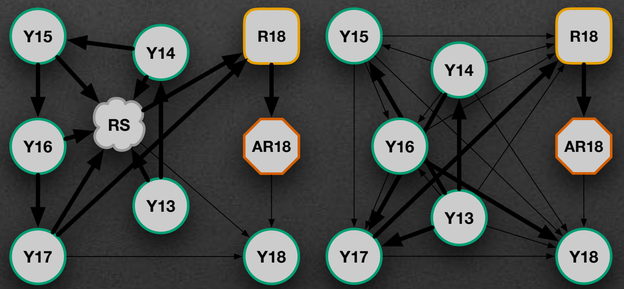


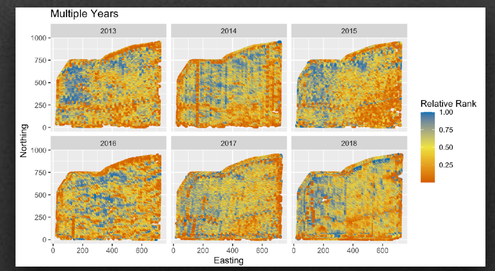




Model 7

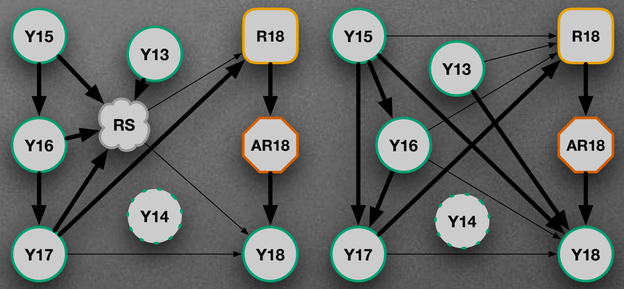
Model 8

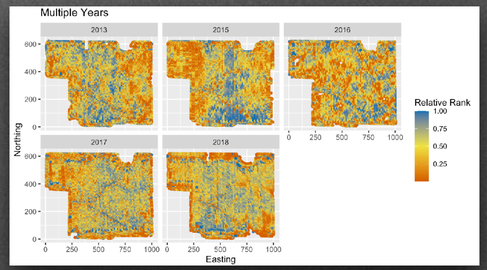


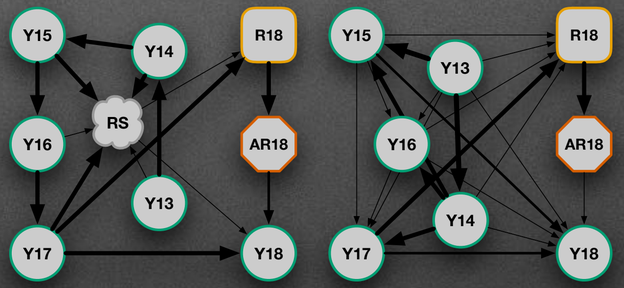


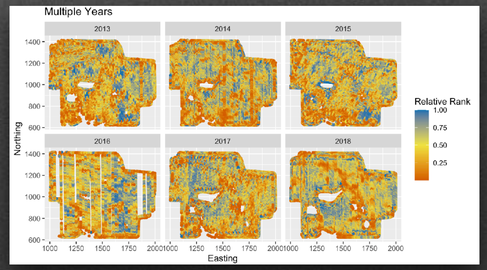
Models 7 and 8

- Summary
  - Response surface (yield map) does not appear to influence 2018 seeding map.
    - Can we further test the effectiveness of seeding prescription based on single year yields?
  - Response surface (yield map) does not appear to influence 2018 yield.
    - What do we learn from an average yield map?



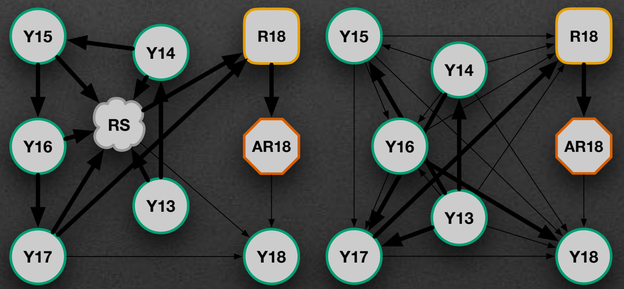


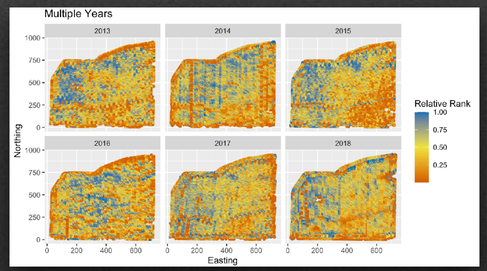


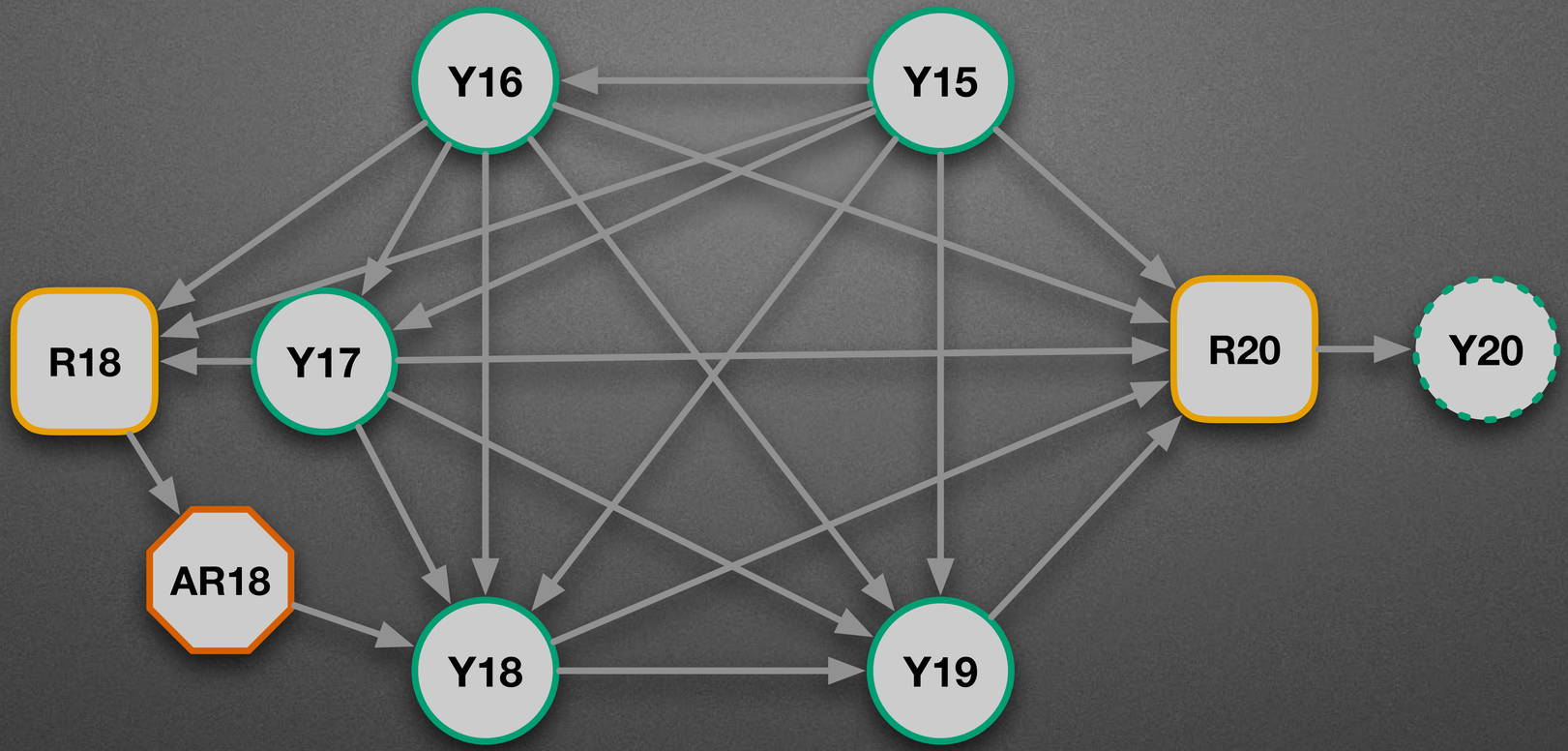


Model 7

Model 8



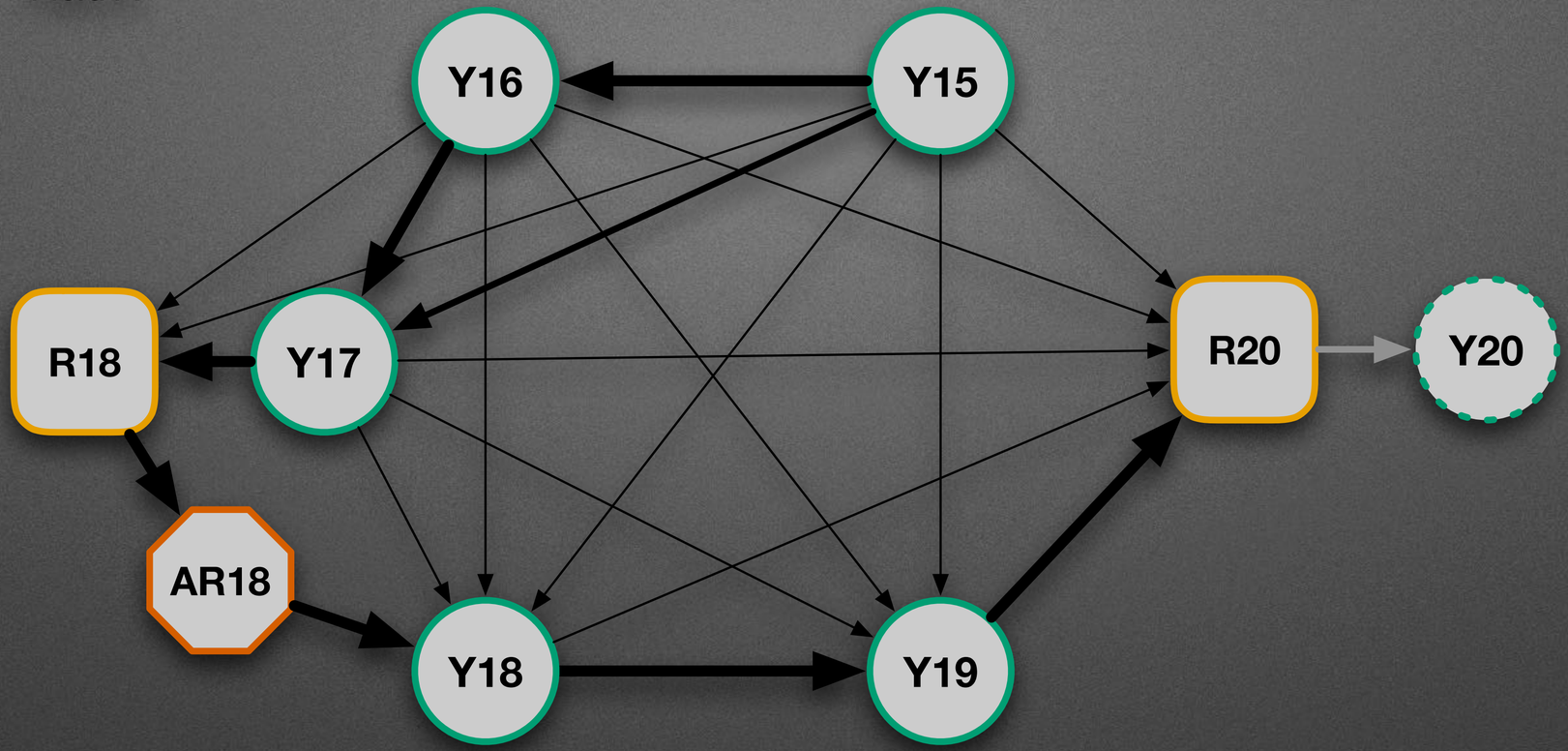




Model 9

The best predictor for seeding maps for 2018 was the single prior year yield. Yields 2 or more years prior tended to have little influence. Did these trends continue into 2019 and 2020?

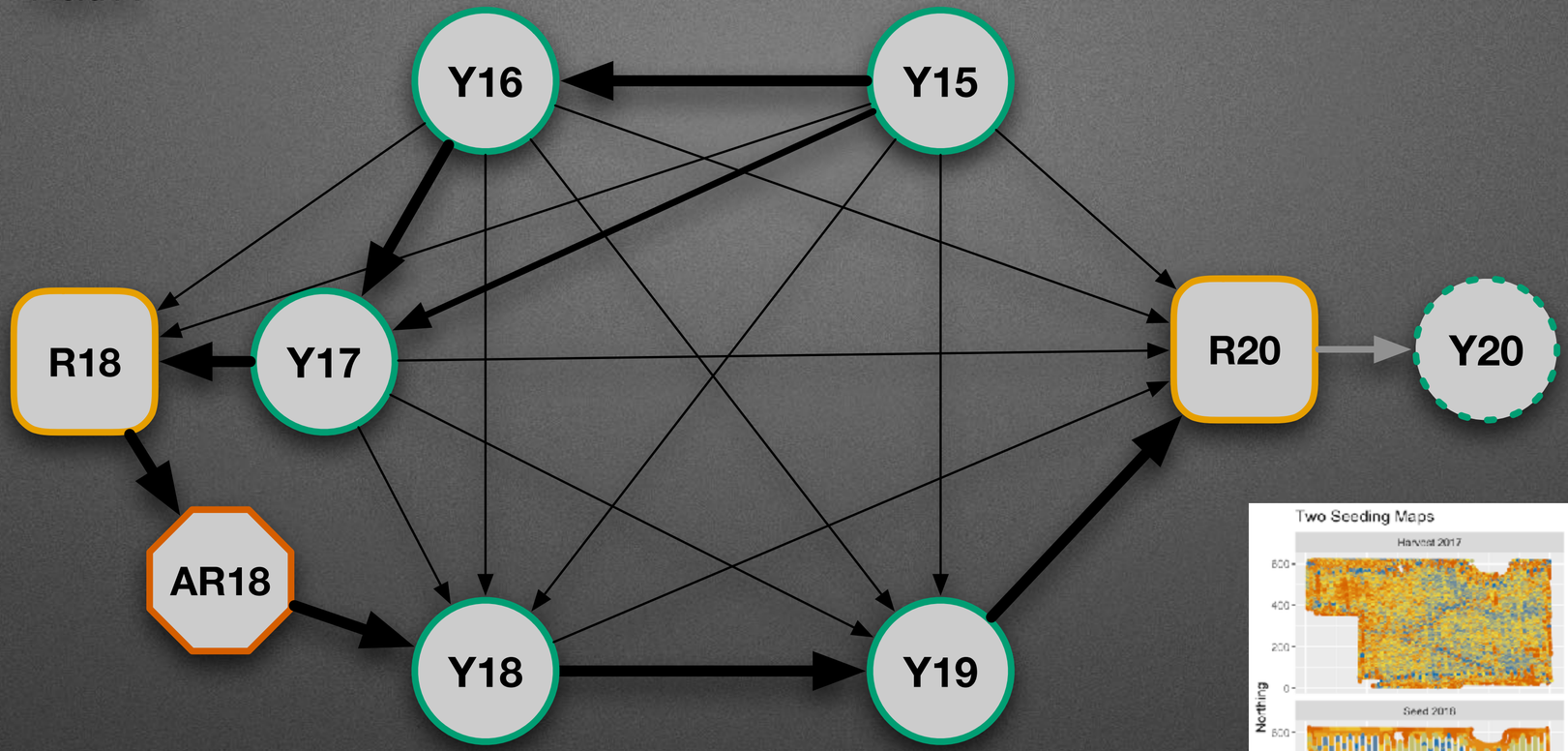
Field A

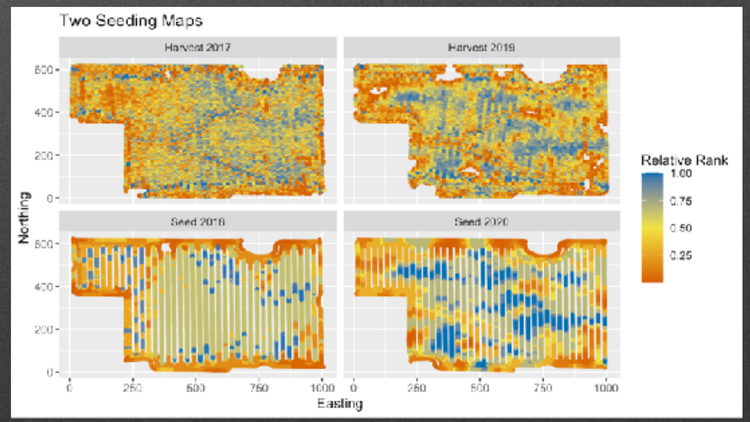


Field A

There is only one significant path from 2015 yield to the 2020 seeding map.

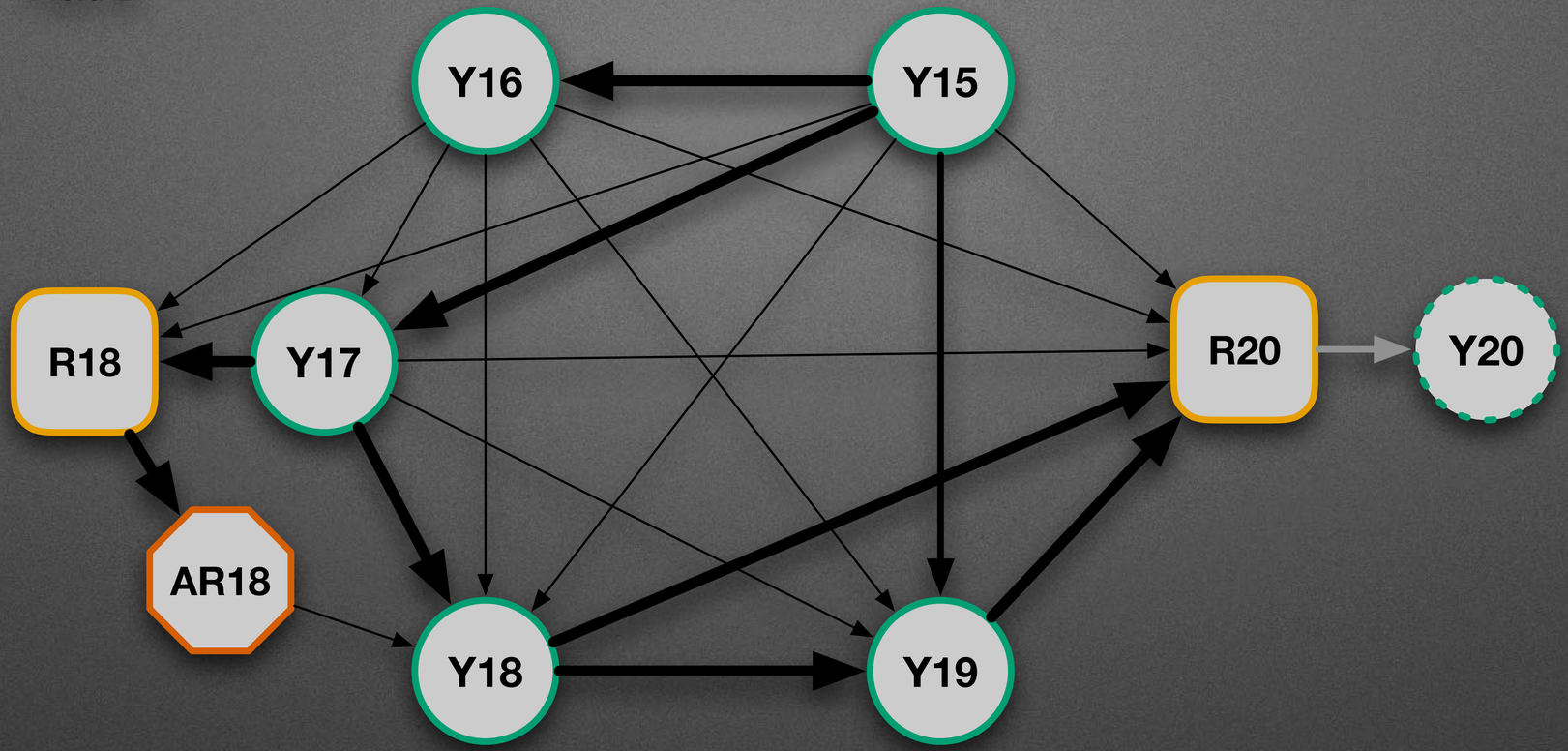
Field A





Field A

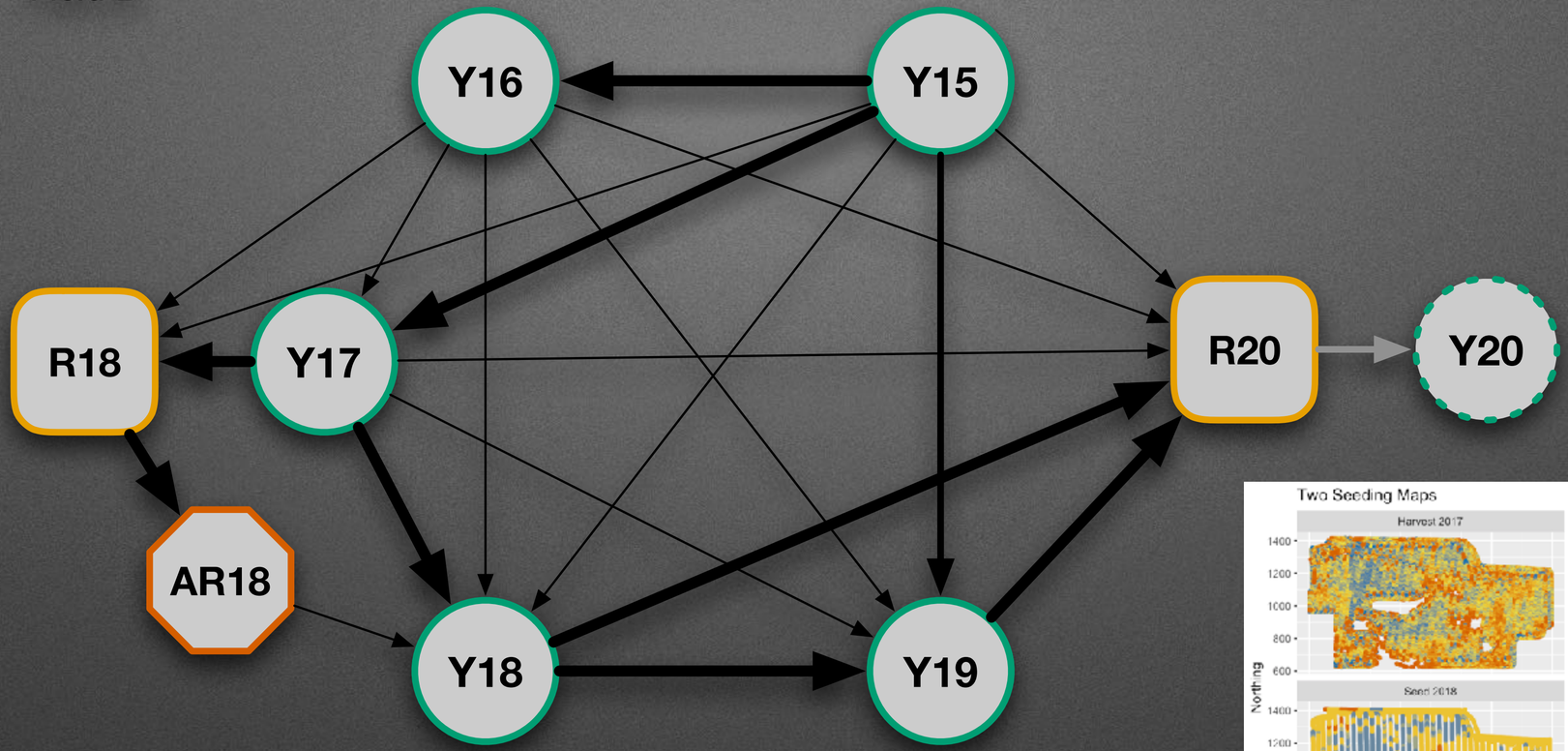
2020 seeding map more closely follows from 2019 yield.

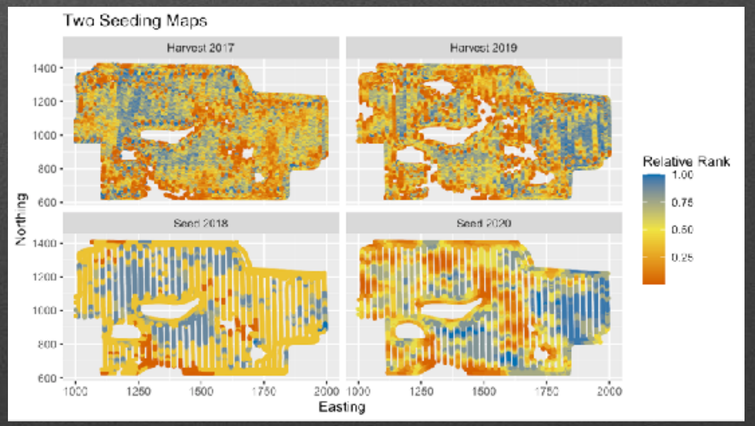


Field B

Yield tended to be influenced by two years prior rotation.

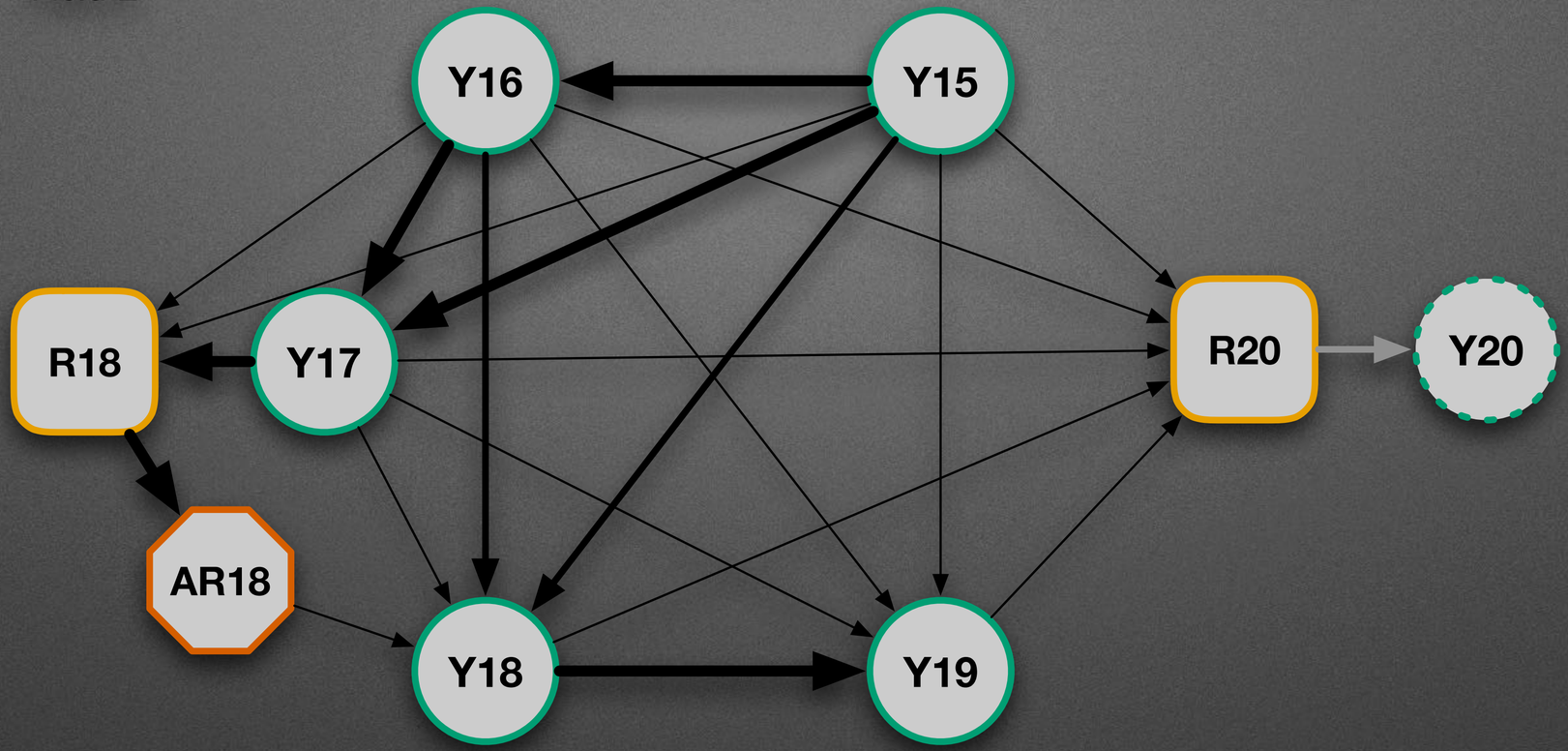
Field B





Field B

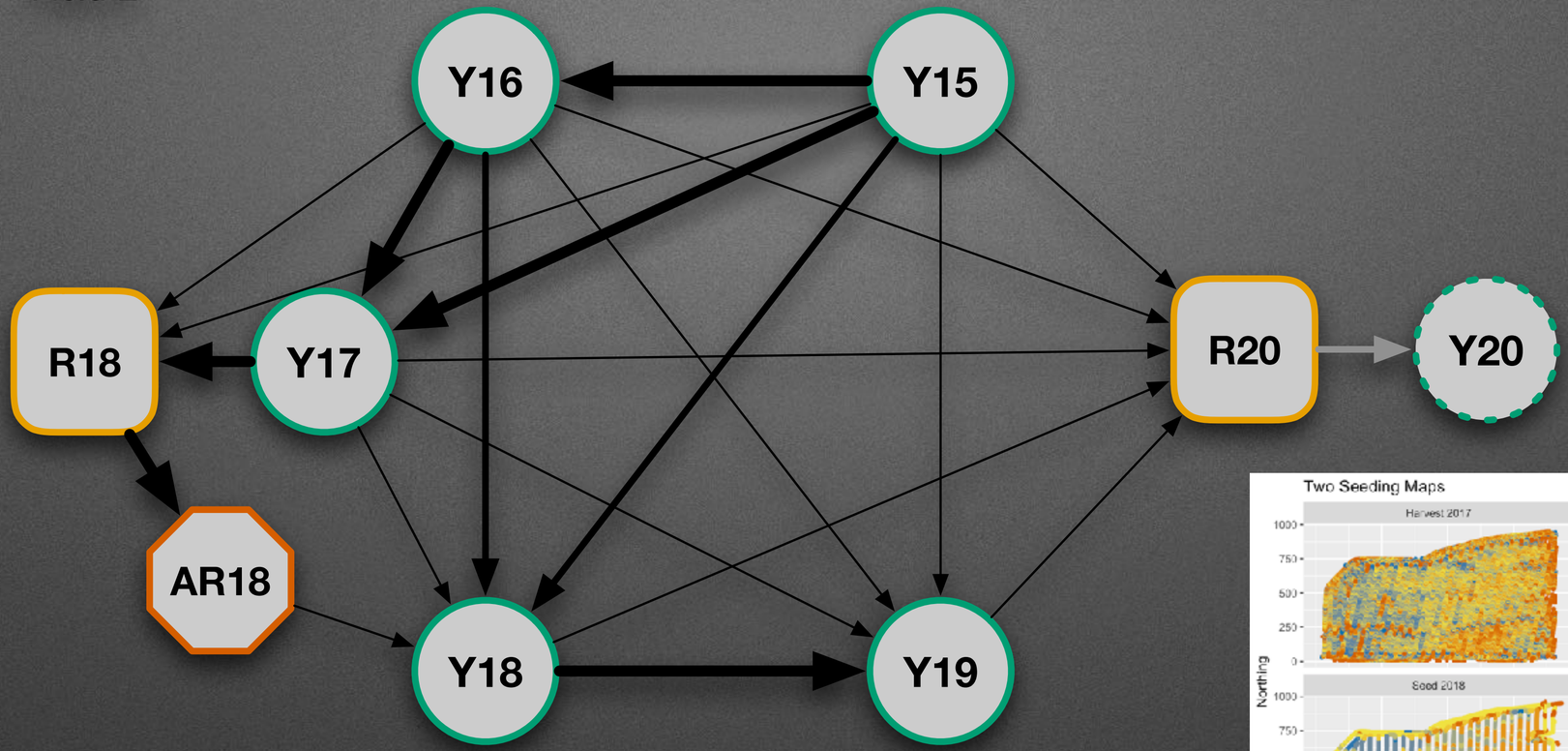
Yield tended to be influenced by two years prior rotation.
Yield map primarily influenced by prior yield.

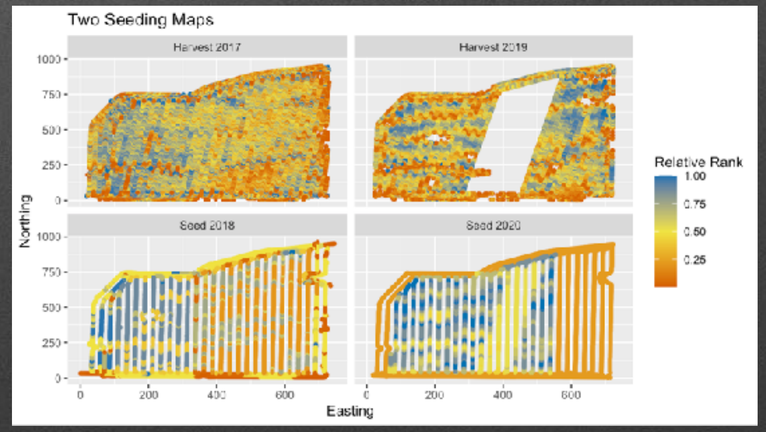


Field E

Sequential correlation was sporadic, but similarly limited to two years rotation.

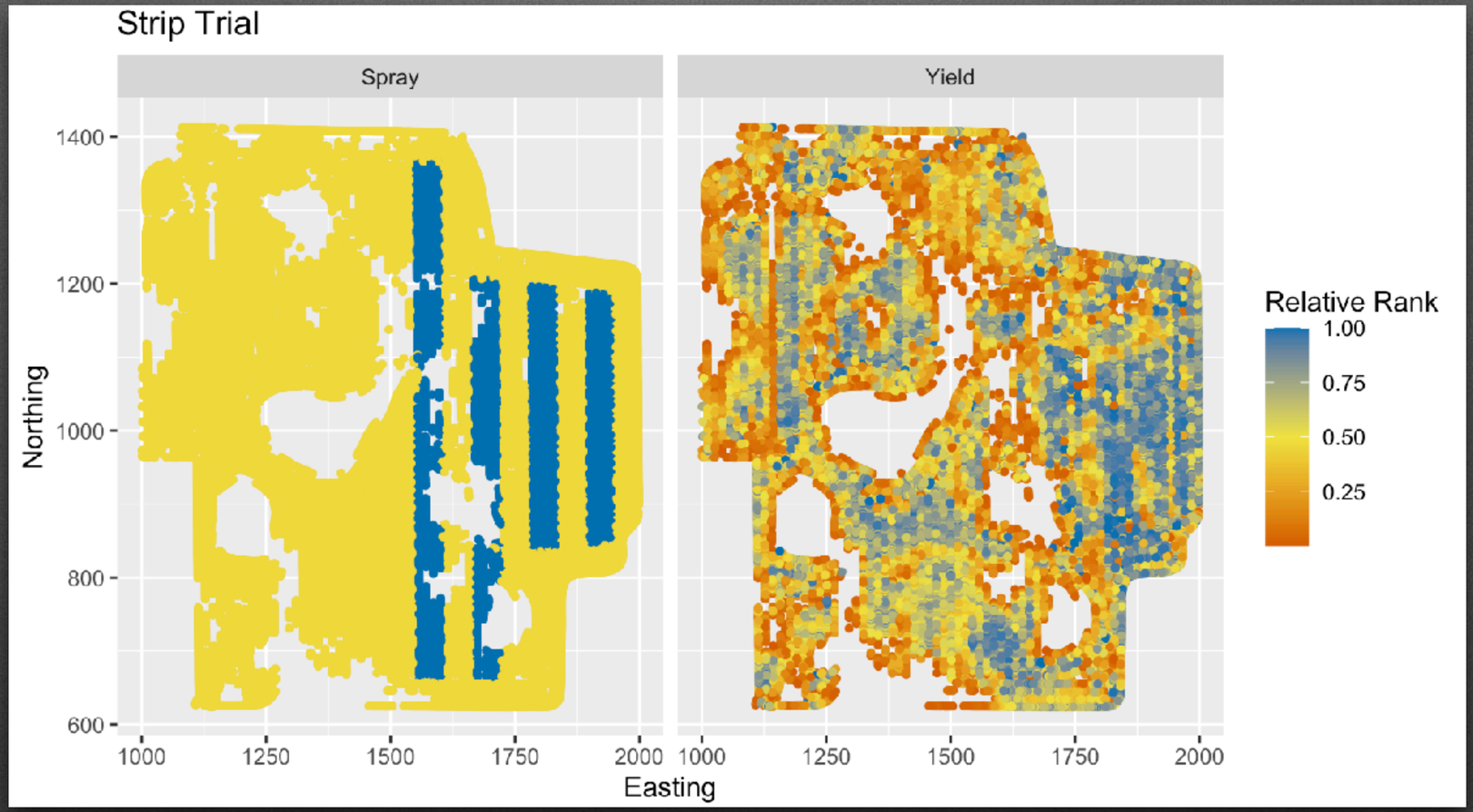
Field E





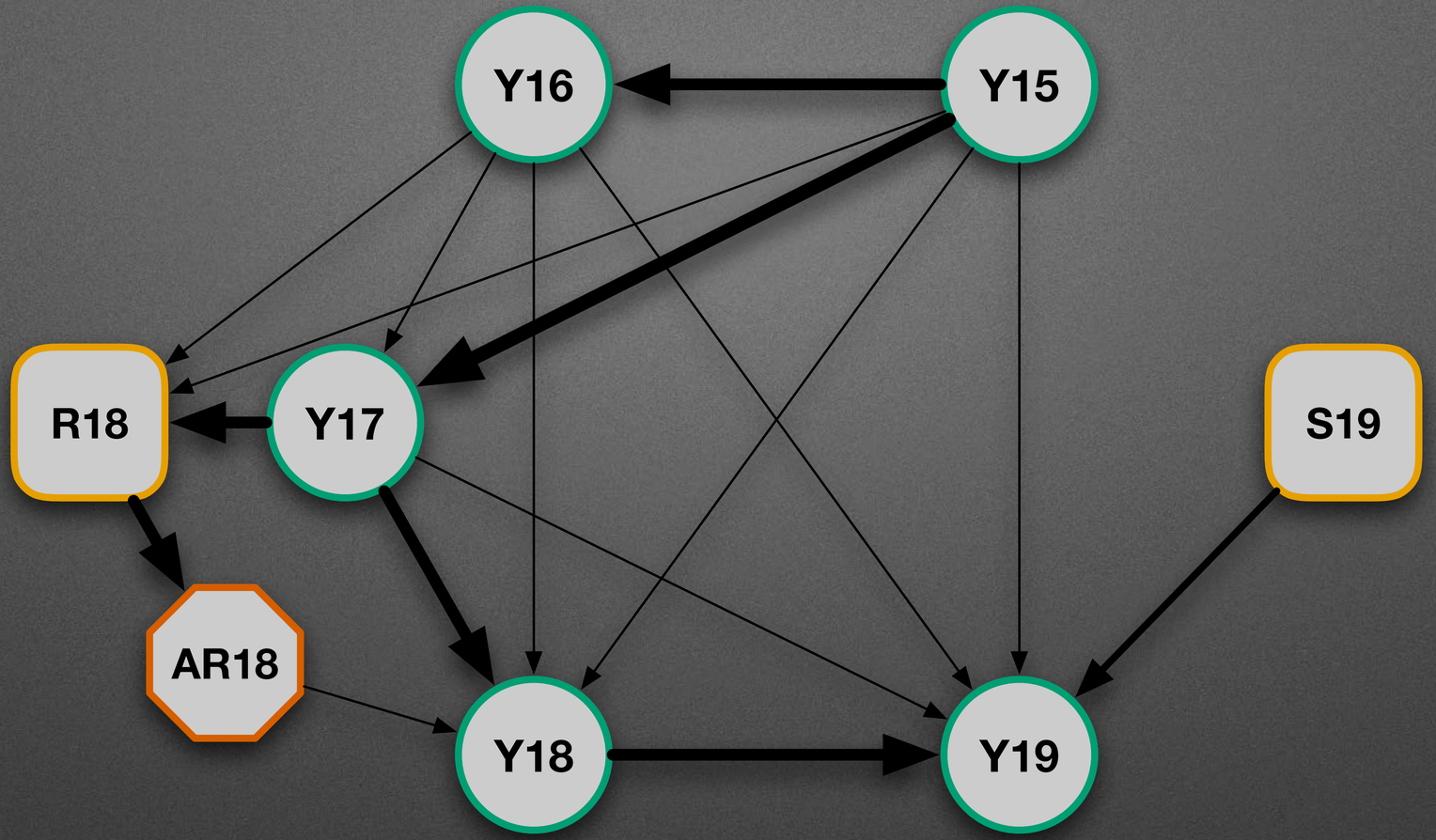
Field E

Sequential correlation was sporadic, but similarly limited to two years rotation.
But 2019 data was incomplete and only partially used to determine seeding map.



Strip Trial

Field B, Soybean 2019. Strips of fungicide were sprayed in the eastern half of the field.

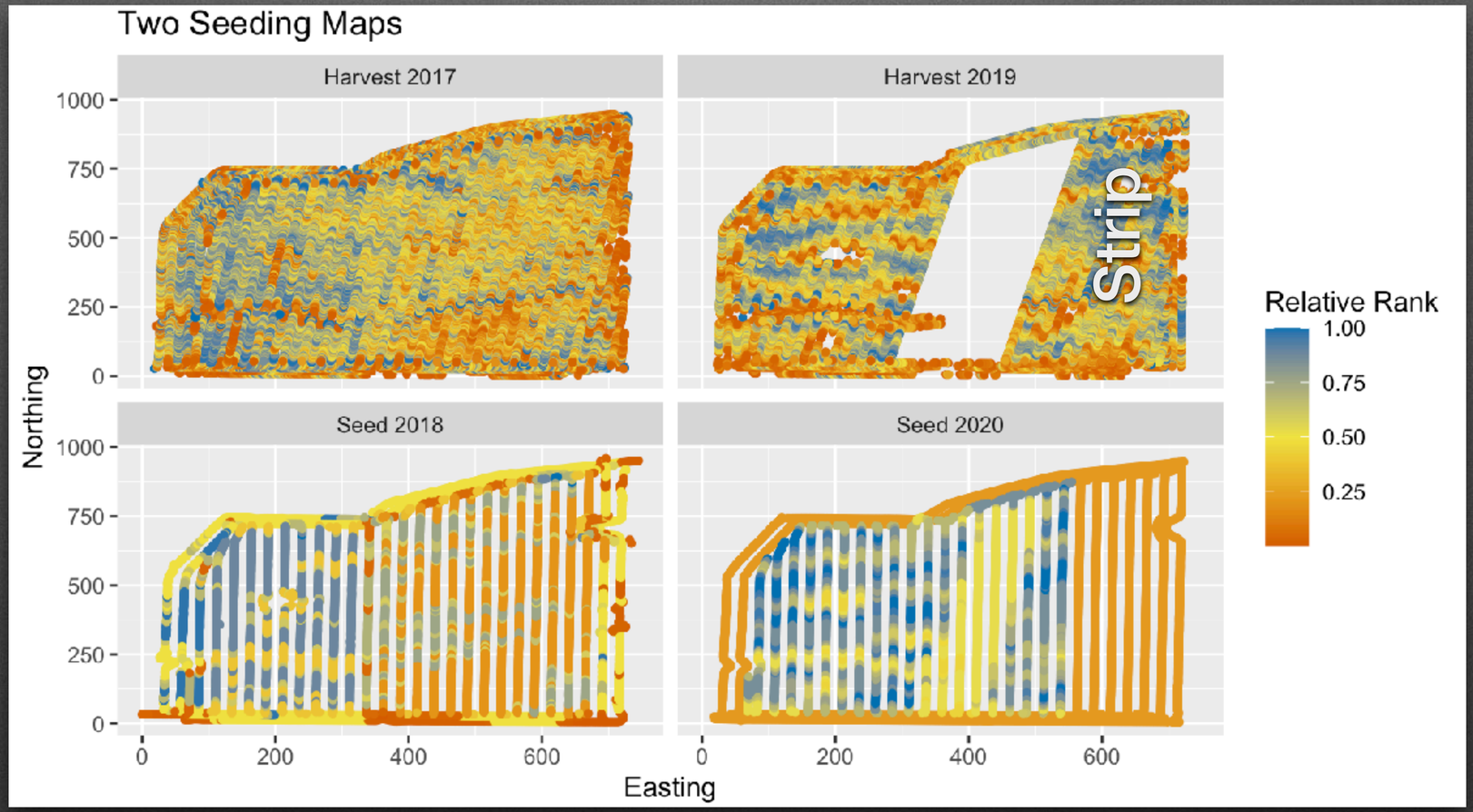


Model 10

Fungicide spray had less influence on 2019 yield than did 2018 yield, but may have relatively more influence on yield than the 2018 seeding map

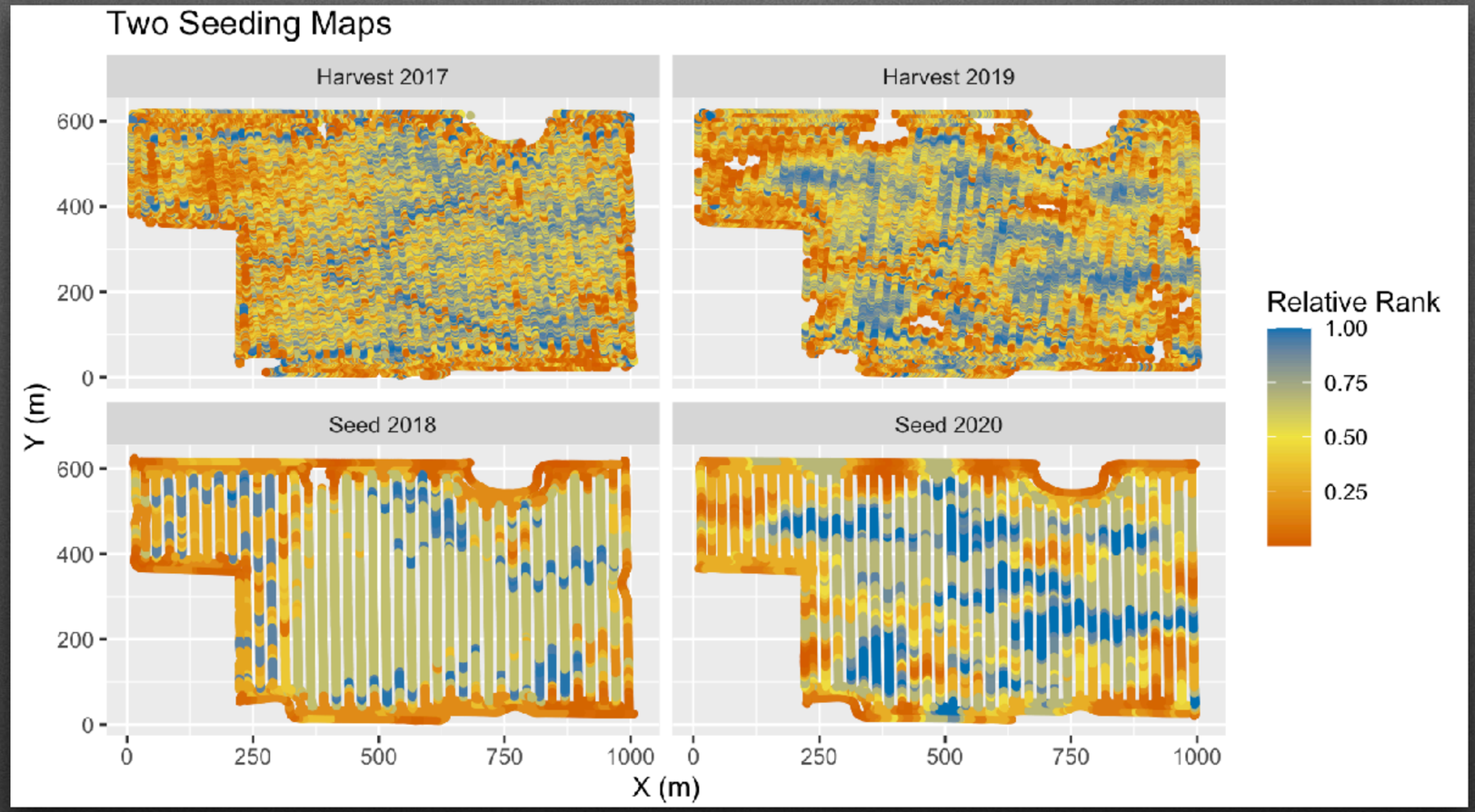
Final Thoughts

- Bayesian nets are an appealing tool for exploring causal relationships.
- Can we design experiments to test the predictive capabilities of Bayesian nets?
- How can we combine multiple realizations of the same Bayesian net?
  - Model averaging?
- The data here were aggregated over simple grids. Finer scale spatial resolution may be possible, but present other analytic issues.

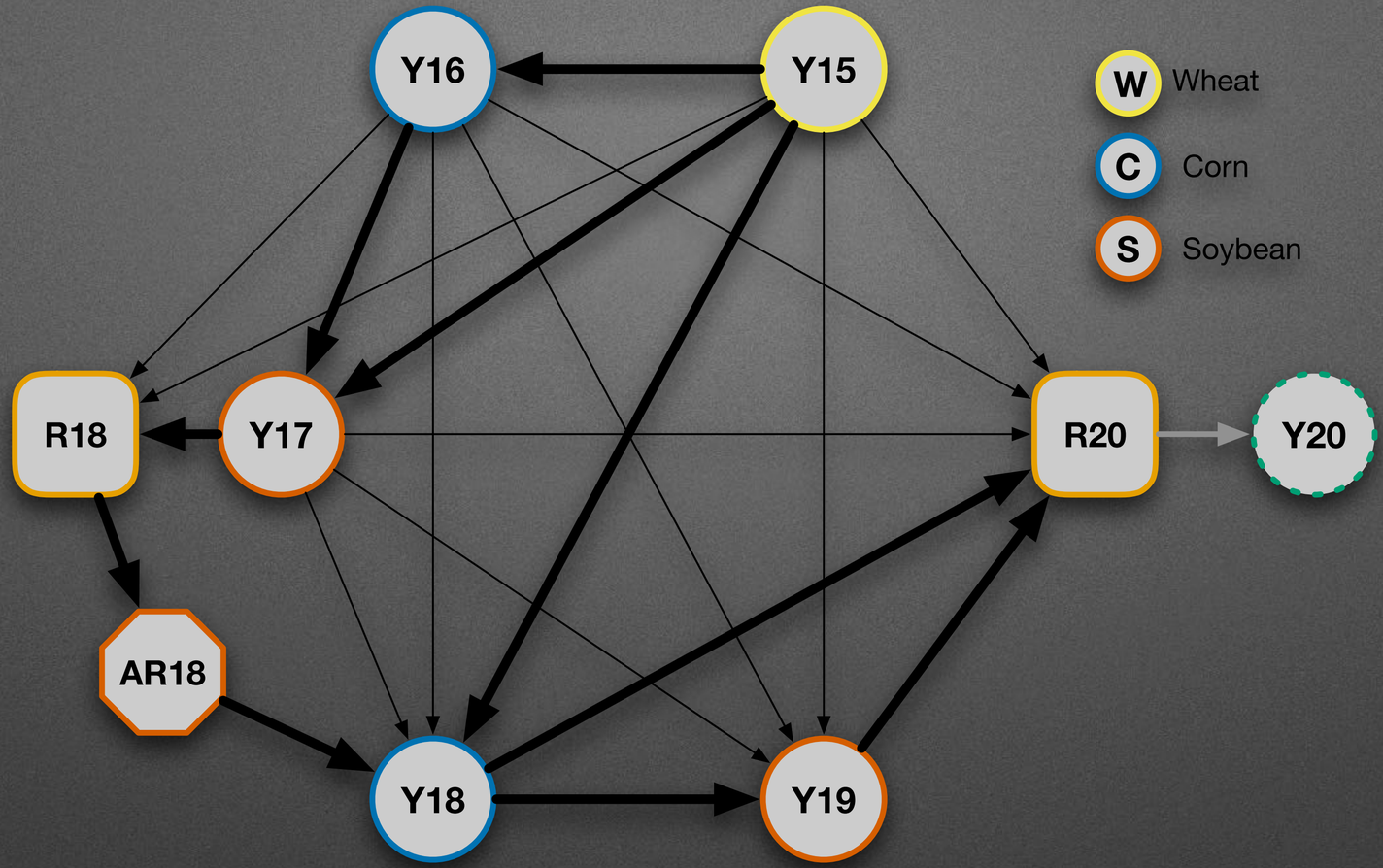


Final Thoughts

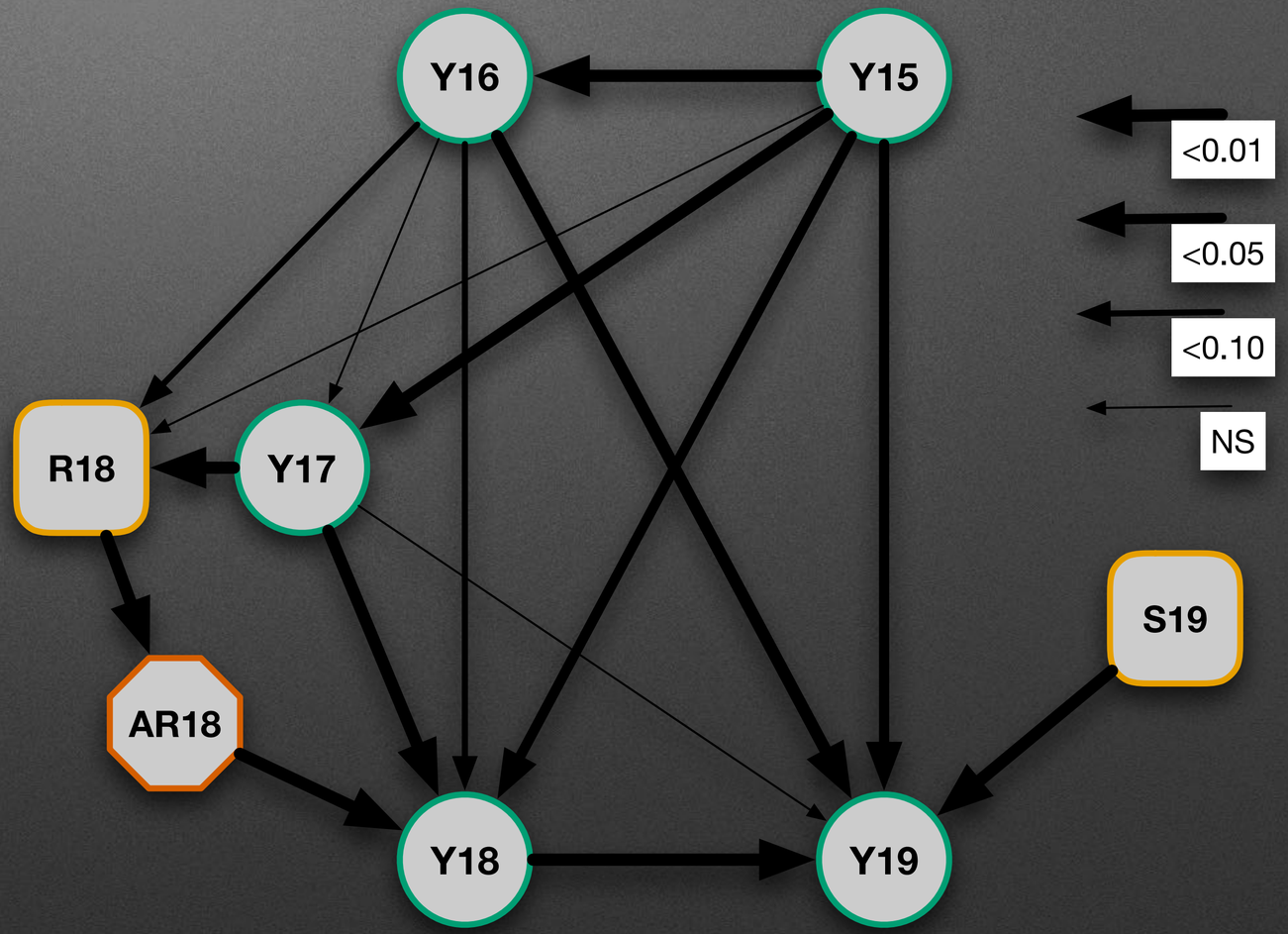
Can we combine prior history in planned experiments using Bayesian nets?


Two Seeding Maps
Harvest 2017
Harvest 2019
Seed 2018
Seed 2020
Y (m)
X (m)
0
200
400
600
0
250
500
750
1000
Relative Rank
1.00
0.75
0.50
0.25

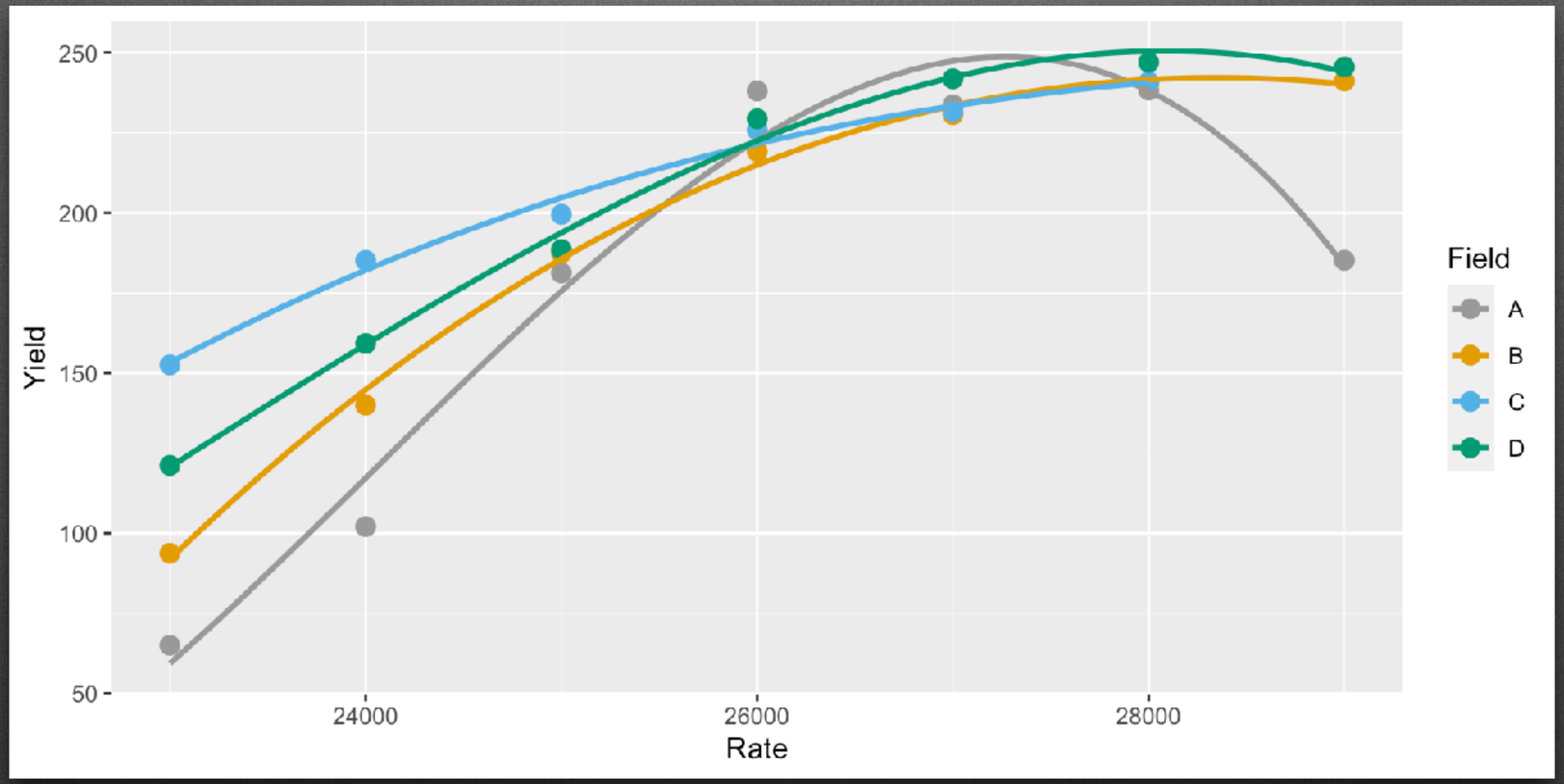
Refinement


Y16
Y15
W
Wheat
C
Corn
S
Soybean
R18
Y17
R20
Y20
AR18
Y18
Y19

EDGE	FROM	TO	STRENGTH
1	Y15	Y16	<0.0001
2	Y16	Y17	0.8863
3	Y15	Y17	<0.0001
4	Y17	R18	<0.0001
5	Y16	R18	0.0556
6	Y15	R18	0.7981
7	R18	AR18	<0.0001
8	AR18	Y18	<0.0001
9	Y17	Y18	<0.0001
10	Y16	Y18	0.0844
11	Y15	Y18	0.0286
12	S18	Y19	0.0099
13	Y18	Y19	<0.0001
14	Y17	Y19	0.2192
15	Y16	Y19	0.0053
16	Y15	Y19	0.0132



Learning Models


250
200
150
100
50
Yield
24000
26000
28000
Rate
Field
A
B
C
D

Estimation

- Relatively more corn was harvested in regions with higher seeding rates, up to certain rate. Above that rate, yield did not correlate with seeding rate.

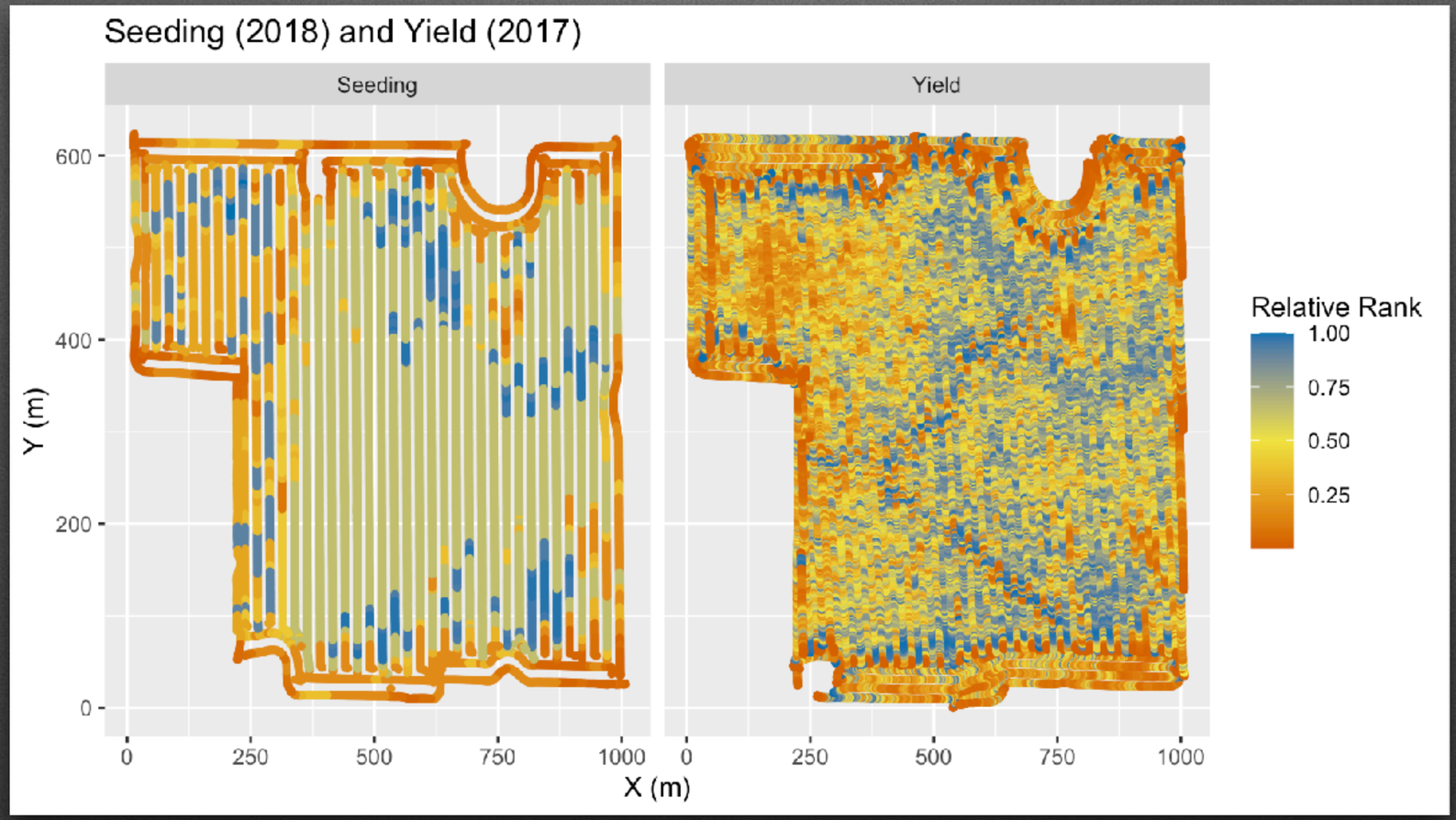
- This is a correlation.

Inference

- Seeding rate increases yield, up to an optimal rate. Above that rate, seeding rate does not affect yield.
- This is a statement about causation.

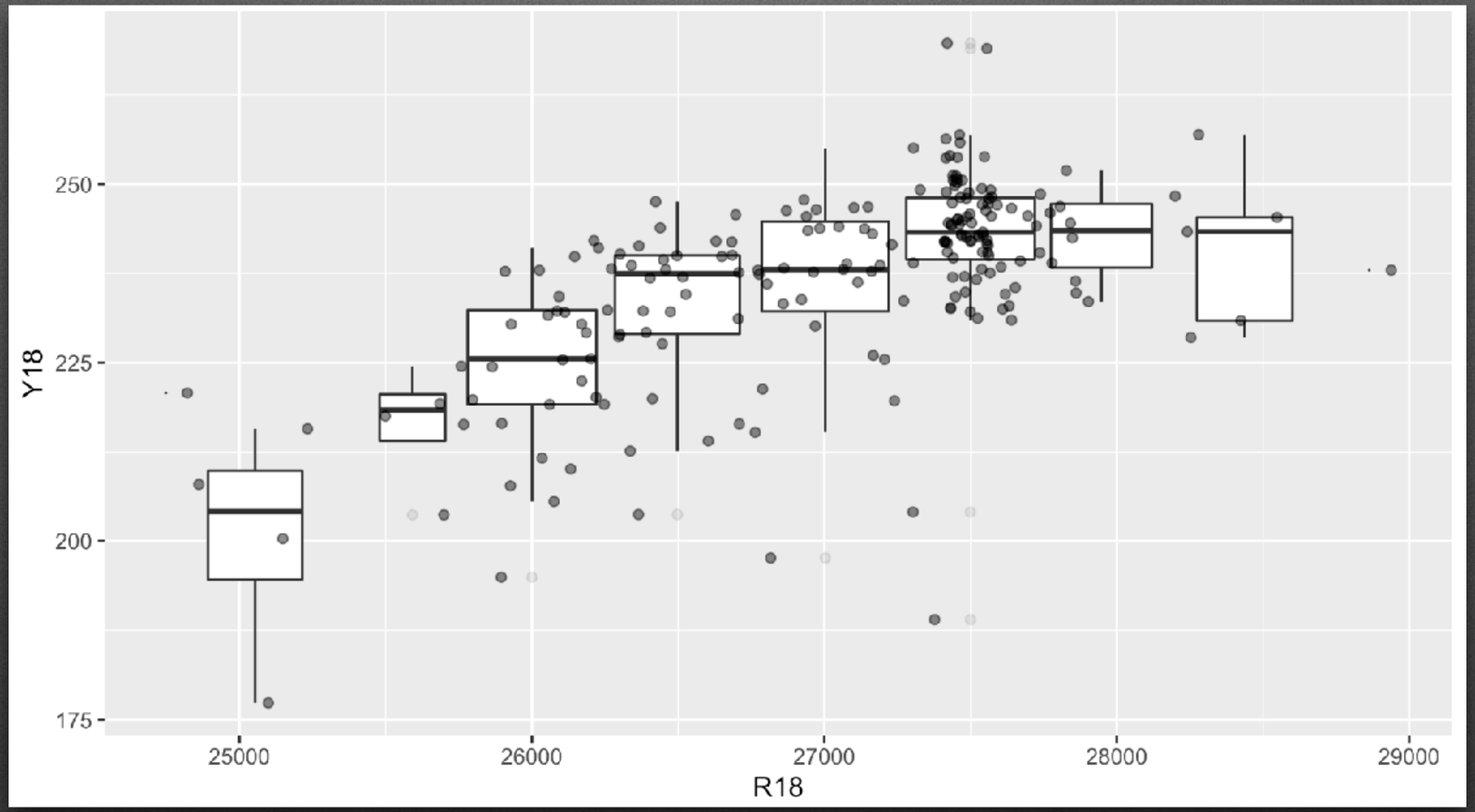
Correlation vs Regression

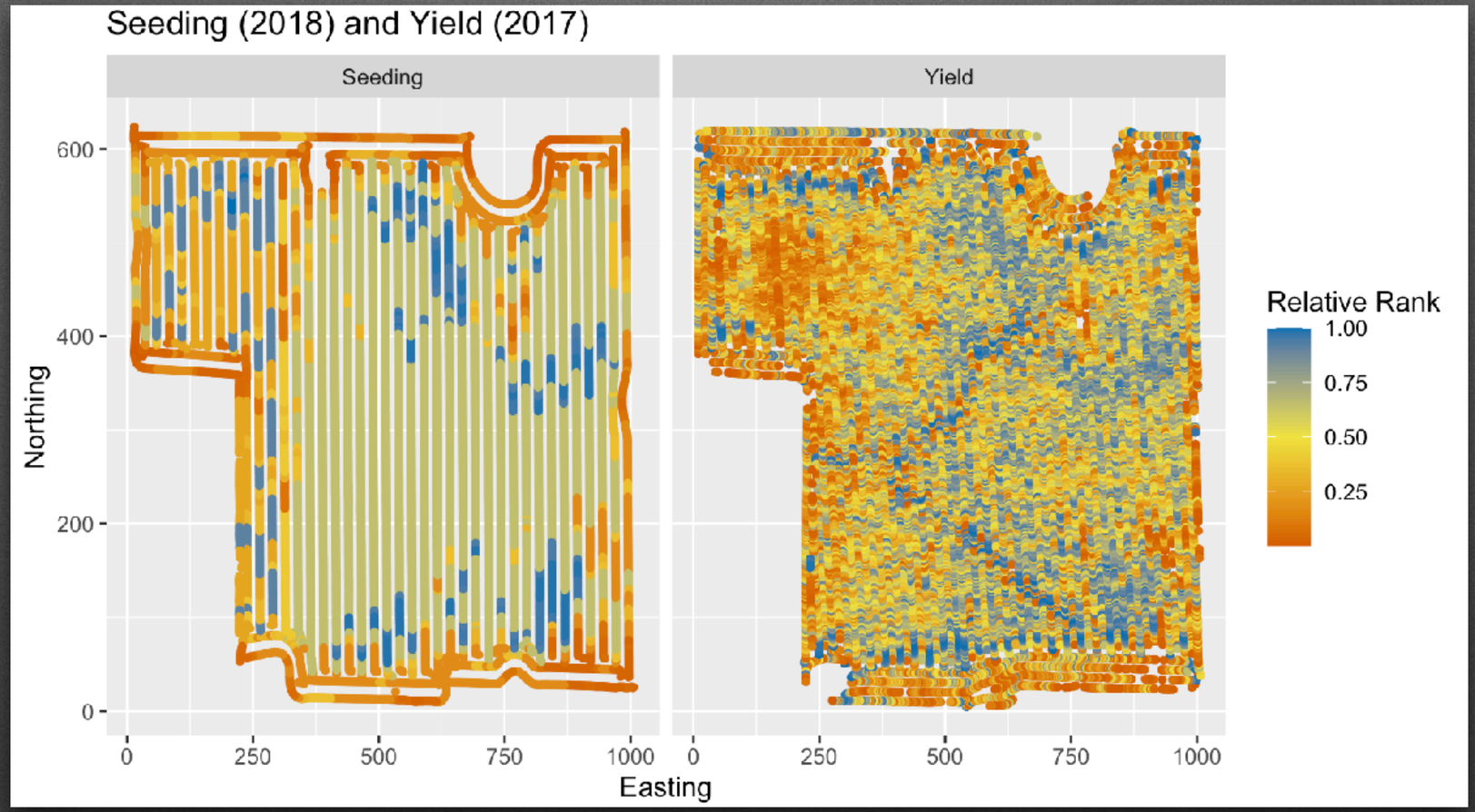
- Correlation
- $$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
- Regression
- $$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Not a good model

R18 and Y17 are highly correlated. In the event, seeding rate for 2018 was largely determined by yield from 2017.


250
225
200
175
Y18
25000
26000
27000
28000
29000
R18



Regression Analysis?

There is an additional variable, not included in these data, that both may be dependent on.

Correlation is not causation

– everyone

Correlation is not causation but it sure is a hint.

– Edward Tufte

Data Aggregation

Briefly, we produce a common basis for analysis by dividing each field into a grid (50x50 meters) and averaging Yield or ControRate values within each cell.

Analysis of Covariance

- Since the seeding rates were not randomized, we want to account for spatial variability. One simple method is to include the previous yield in the linear model; this can be consider multiple regression or an analysis of covariance.
- That is, we regress Yield from 2018 on Control Rate 2018, controlling for Yield 2017.
  - $Y18_i = \beta_0 + \beta_1 Y17_i + \beta_2 R18_i + e_i$

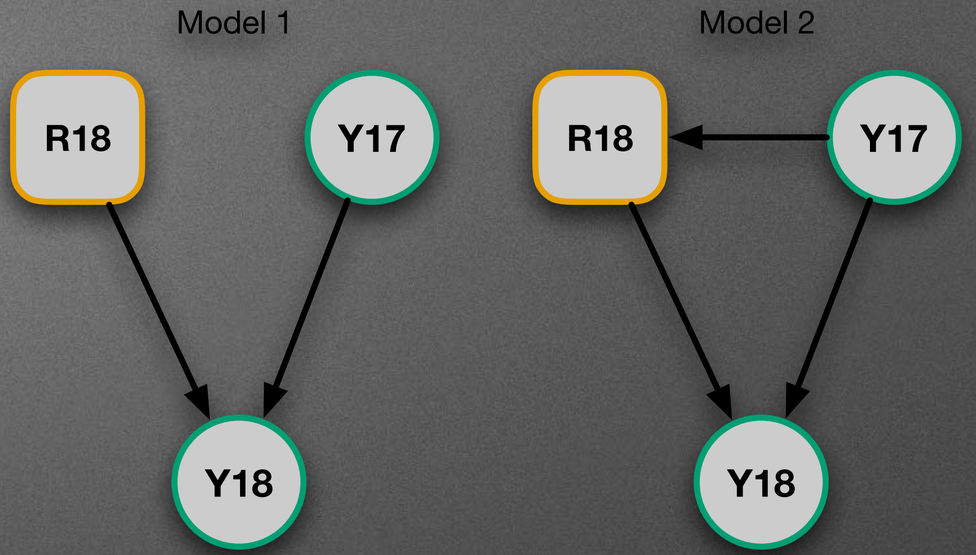
SOURCE	DF	SS	MS	F	P
R18		11429431	11429431	94164	<0.0001
Y17		11422986	11422986	94111	<0.0001
R18 (TYPE III)		6582	6582	54	<0.0001
Y17 (TYPE III)		137	137	1.13	0.29
RESIDUAL	202	24518	121		

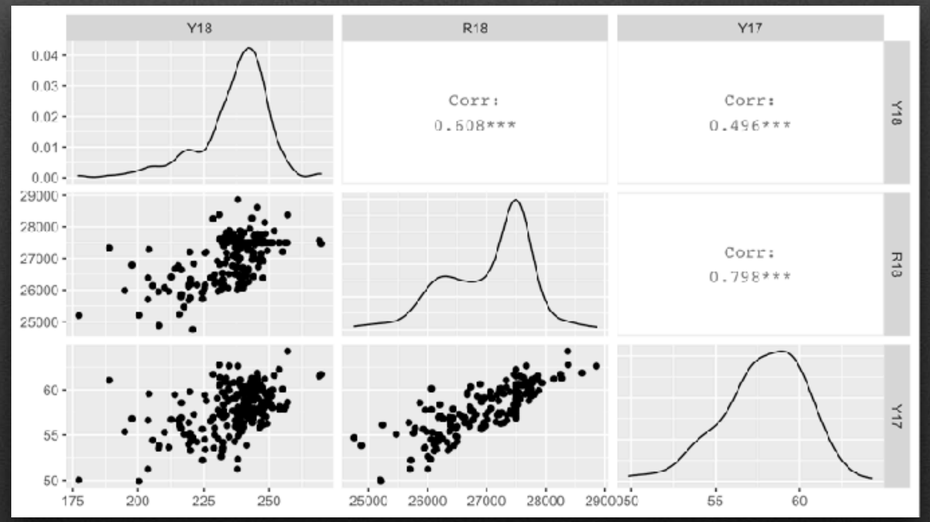
Analysis of Covariance

Not a good model. R18 and Y17 are highly correlated.

Testing Models

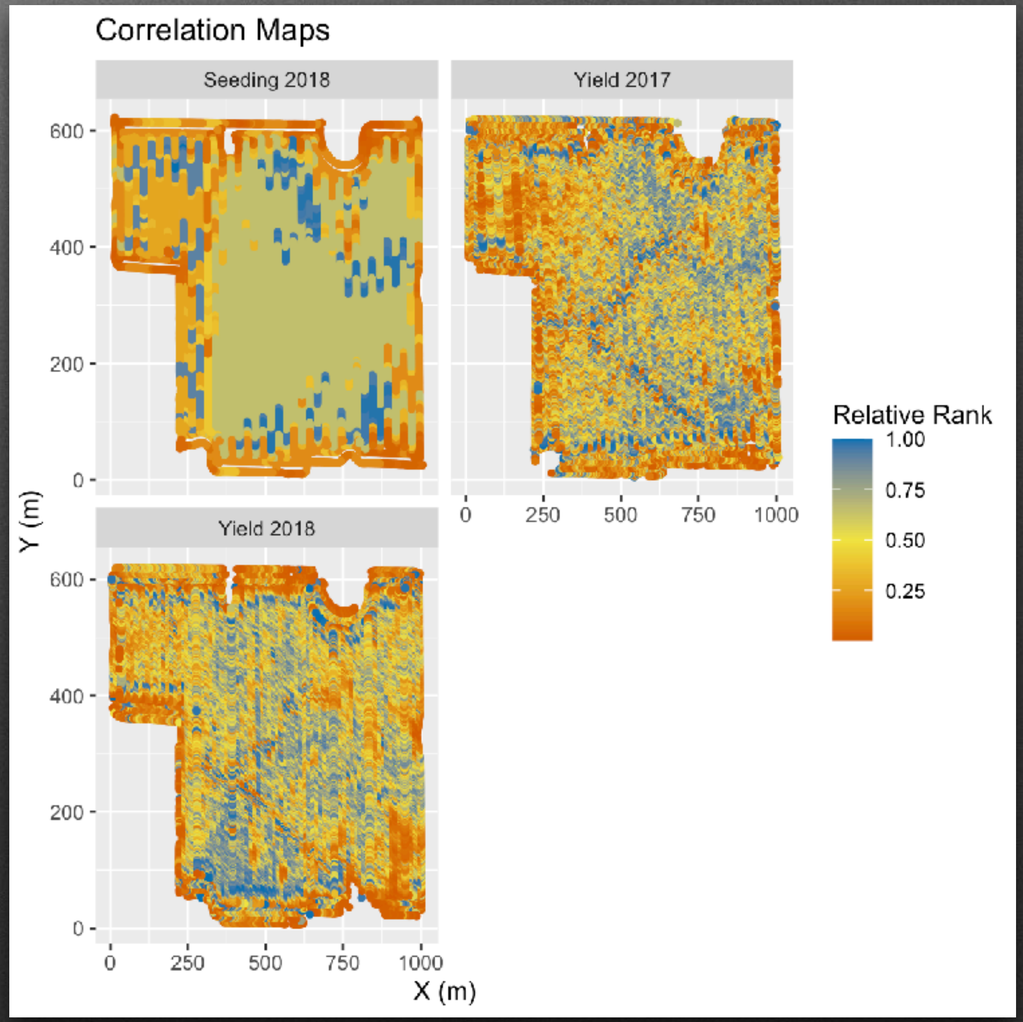
- Model Comparison
  - Information criteria
    - AIC common for linear models, computed from log-likelihood
  - BGE
    - Bayesian Gaussian equivalent score
    - Unique to DAG
    - Similar to BIC


Model 1
R18
Y17
Y18
Model 2
R18
Y17
Y18


Y18
R18
Y17
0.04
0.03
0.02
0.01
0.00
Corr:
0.508***
Corr:
0.496***
29000
28000
27000
26000
25000
Corr:
0.798***
60
55
50
175
200
225
250
25000
26000
27000
28000
29000
55
60
Y18
R13
Y17

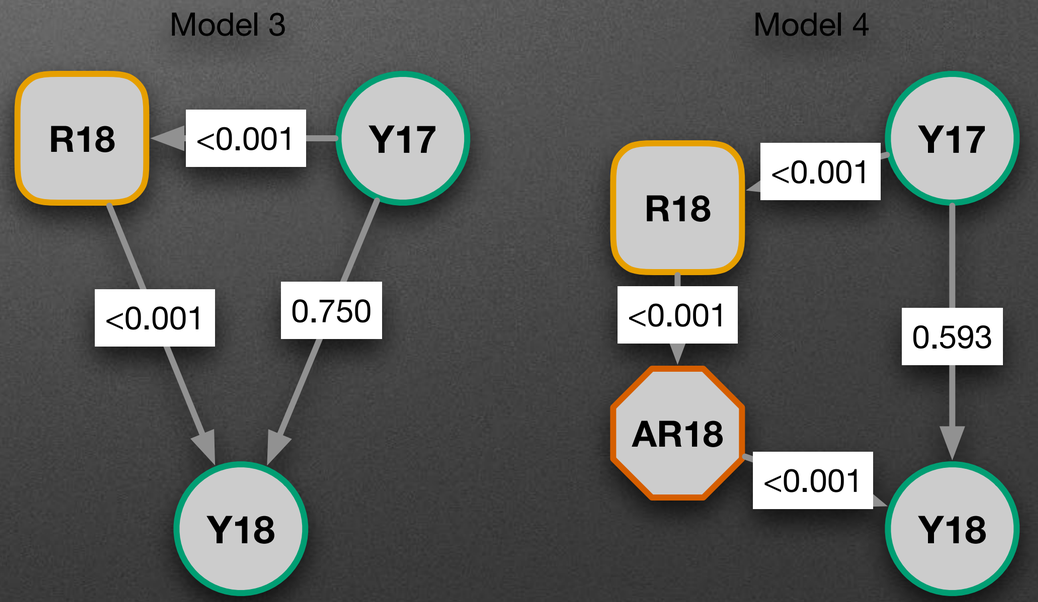
Plausible Causal Relationships

- Did the seeding plan for 2018 determine the yield map for 2018?
- Did the yield from 2017 determine the seeding map for 2018?
- Did yield from 2017 have an impact on the yield for 2018?
  - Soybeans (2017) -> Corn (2018)?



Intermediate Variables

Model	Intercept	R18	Y17	Y18	AR18	SD
AR18 | R18	3273	0.874				195.974
R18 | Y17	19817		102.756			314.277
Y17	37.691					3.931
Y18 | AR18 + Y17	-12.940		1.512		0.006	12.085



SOURCE	DF	SS	MS	F	P
R18		11429431	11429431	95851	<0.0001
Y17		11422986	11422986	95797	<0.0001
R18 X Y17		551	551	4.62	0.03
R18 (TYPE III)		6582	6582	54	<0.0001
Y17 (TYPE III)		137	137	1.13	0.29
RESIDUAL	201	23967	119		

Analysis of Covariance

We can account for some of the confounding with an interaction term. Can we discount the possible influence of prior year's yield?