PURDUE UNIVERSITY | PHENOMICS

Analytical Methods for Temporal Data and Phenotyping

Alencar Xavier

Ph.D. in Plant Breeding and Statistical Genetics
Quantitative Geneticist at Dow AgroSciences

March 14th, 2017

Outline

1. Introduction
2. Time-Space & Treatments
3. Structured Data
4. Phenomic-enable prediction

INTRODUCTION

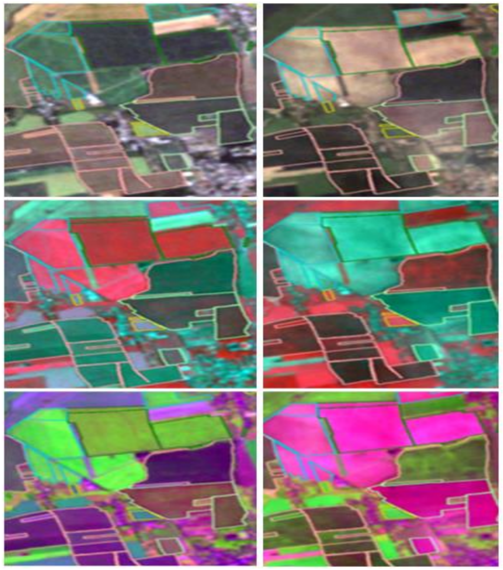
INTRODUCTION & RATIONALE

High-dimensionally inter-correlated data

Space

Time

Spectra



INTRODUCTION & RATIONALE

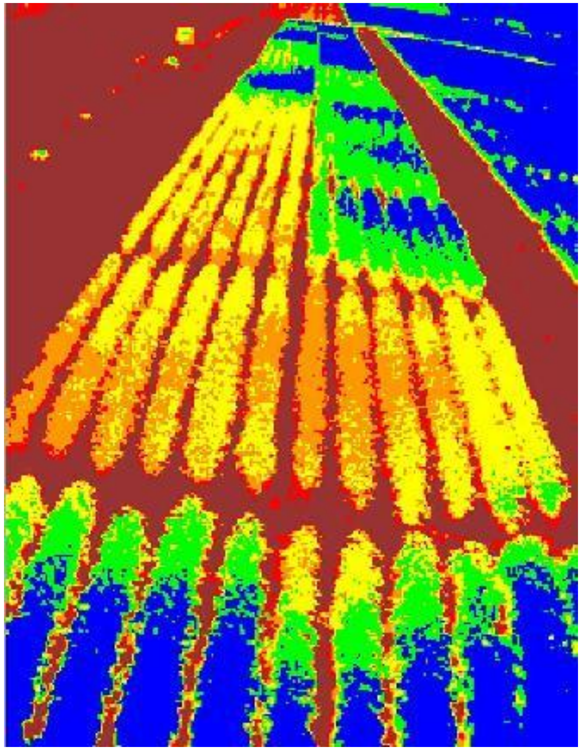
- Complex phenome characterization

1. Phenotype from pictures
   - Each bandwidth is a trait
   - Each combination of bandwidth is a trait

2. Data from various time points

3. Different treatments across the field
   - Cultivars
   - Management practices

4. Field heterogeneity
   - Physical - Porosity, density & texture
   - Chemical - Organic matter & nutrients
   - Biological - Nematodes & pathogens



INTRODUCTION & RATIONALE

Ultimate goal from quantitative analytical techniques:
DISTINGUISH **SIGNAL** FROM **NOISE**

- Promising genetic
- Management practice

Environmental variation

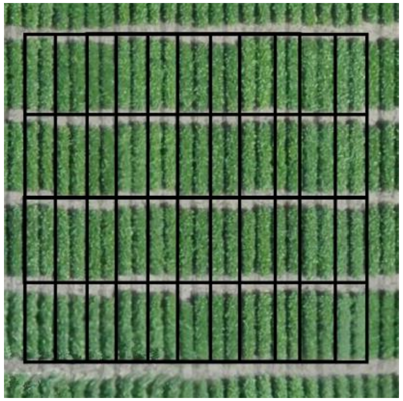
TIME-SPACE & TREATMENTS

TIME-SPACE & TREATMENT

Space

Time

Treatment



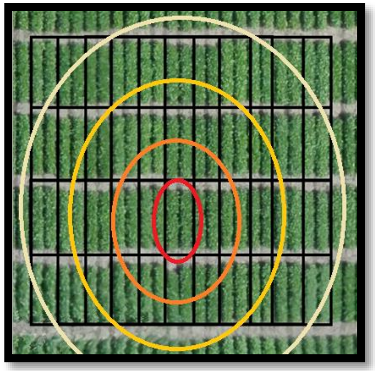
AND multiple phenotypes

TIME-SPACE & TREAT.: STRUCTURED DATA

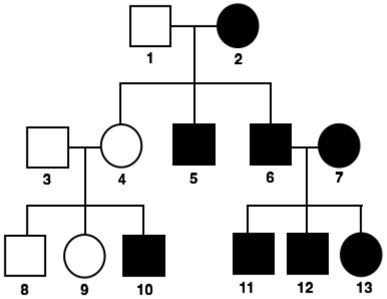
- **Time structure**
  - **Markov chain**: Observation 2 is related to observation 1

Phenology ~ $V_E \rightarrow V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow V_n \rightarrow R_1 \rightarrow R_2 \rightarrow R_n$

- **Spatial structure**
  - Field plots have an **autocorrelation**
  - Neighbor plots trend to
    1. Be exposed to a more similar environment
    2. Compete for the same resources (light & nutrient)



- **Treatment structure**
  - Genetics: Pedigree or genomic data → “Kinship”
  - Management: DOES NOT APPLY



STRUCTURED DATA

STRUCTURED DATA

- Ways to handle auto-correlated data:

1. **Covariate**
   - Fixed effect vector in the model
   - “Quick and dirty” solution
2. **Covariance**
   - Random effect term in the model
   - We state how the levels are correlated
3. **Multivariate**
   - Modeling multiple traits at once
   - Handle correlated signal & correlated noise

STRUCTURED DATA

- **Mixed linear model**

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad \mathbf{u} \sim \mathbf{N}\left(\mathbf{0}, \mathbf{K}\sigma_{\mathbf{u}}^{2}\right) \quad \mathbf{e} \sim \mathbf{N}\left(\mathbf{0}, \mathbf{I}\sigma_{\mathbf{e}}^{2}\right)$$

- y – Phenotype(s)

- Xb – Fixed effects
  - Covariates (adjustment)
  - Management

- Zu – Random effects
  - **Structured terms** (genetics, space, time)
  - **K** – correlation structure among levels

- e – Residuals

SPACE

Covariance

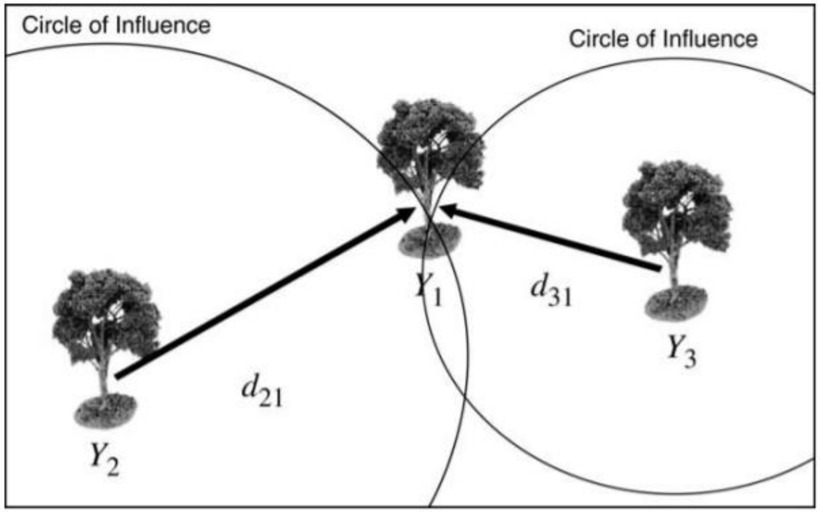


FIGURE 1.—Relationship of three trees in a field.

Muir, W. M. (2005). Incorporation of competitive effects in forest tree or animal breeding programs. *Genetics*, *170*(3), 1247-1259.

Covariate

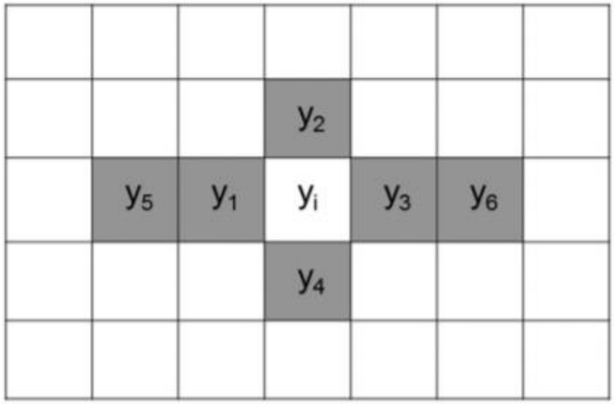
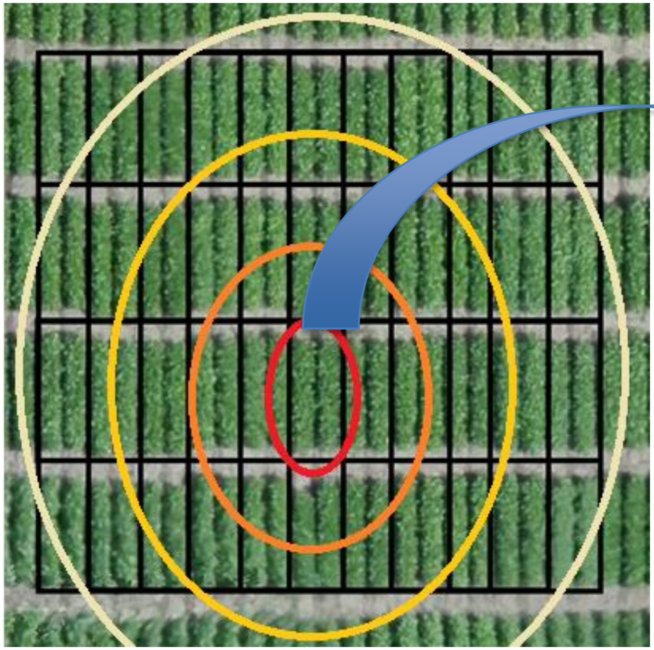


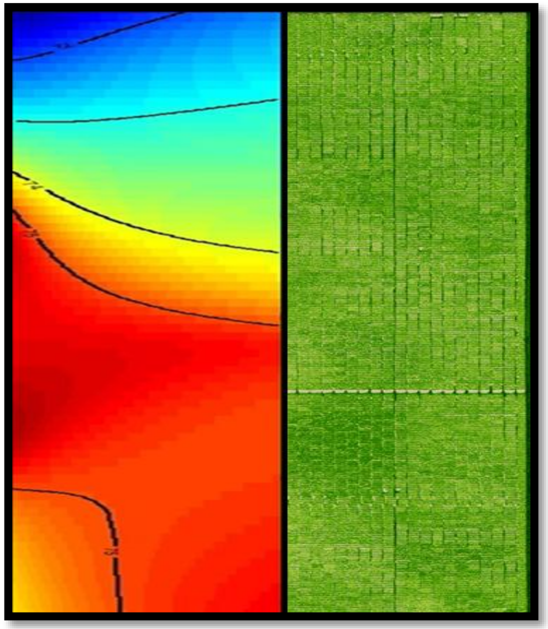
Figure 1 Diagram to calculate the covariable x_i. Y_i is the phenotypic value in the plot. The neighboring plots are indicated with gray color.

Lado, Bettina, et al. "Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data." *G3: Genes| Genomes| Genetics* 3.12 (2013): 2105-2114.

SPACE



0.19	0.21	0.23	0.24	0.23	0.21	0.19
0.33	0.4	0.46	0.49	0.46	0.4	0.33
0.42	0.56	0.75	1	0.75	0.56	0.42
0.33	0.4	0.46	0.49	0.46	0.4	0.33
0.19	0.21	0.23	0.24	0.23	0.21	0.19



Covariance – smooth decay (*Kriging*)

Kriging to control local noise

Table: Correlation between two years of SoyNAM phenotypic data (2013 and 2014, Indiana) and narrow-sense heritability before (BK) and after kriging (AK) for six soybean traits: plant height (Height), days to flowering (Flowering), days to maturity (Maturity), number of reproductive nodes (Nodes), and average canopy closure (Canopy).

		Height	**Flowering**	**Maturity**	**Nodes**	**Canopy**
Correlation	BK	0.67	0.20	0.54	0.22	0.21
	AK	0.71	0.20	0.55	0.26	0.35
Heritability	BK	0.90	0.49	0.82	0.74	0.74
	AK	0.94	0.56	0.88	0.76	0.79

Approaching space

- Covariates ROW & COLUMN
  - Interaction with block / environment

- Covariate of neighbor plots
  - Raw observations (pre-modeling)
  - Residuals (post-modeling)

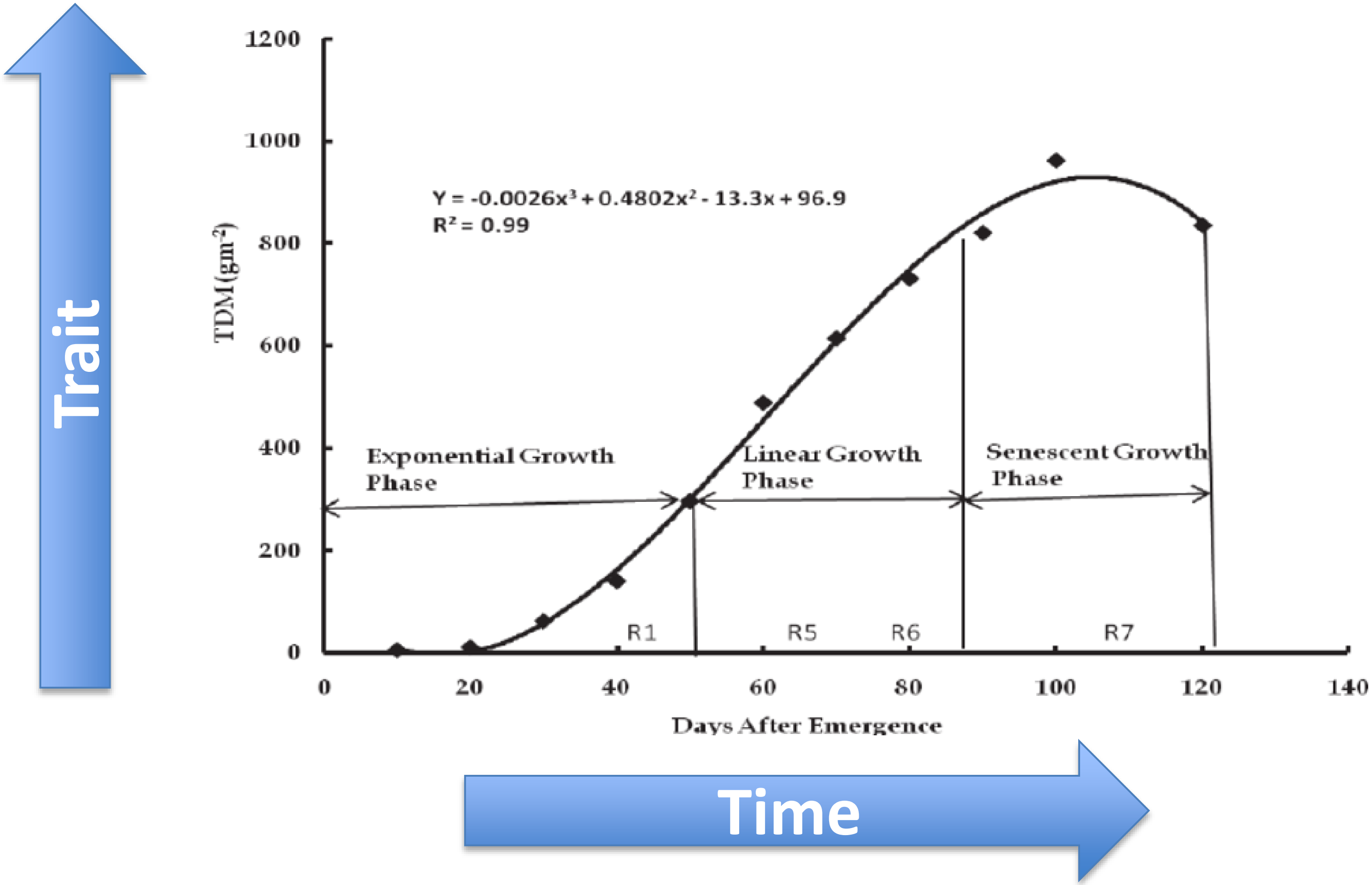
- Random effect with covariance

- Residuals with correlation structure

TIME

- Same field has data points collected from different stages
- Time can be represented as
  - Days after planting
  - Developmental stage
  - Heating degrees day

Modeling the TRAIT using TIME to get the intervals



Soybean total dry matter (TDM) curve from Carpenter & Board (1997)

Getting measures between true observations: Example below is a **logistic curve**

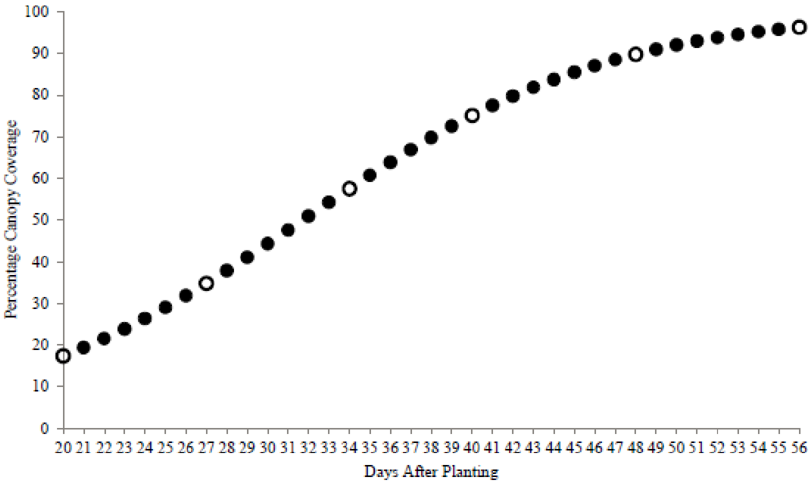
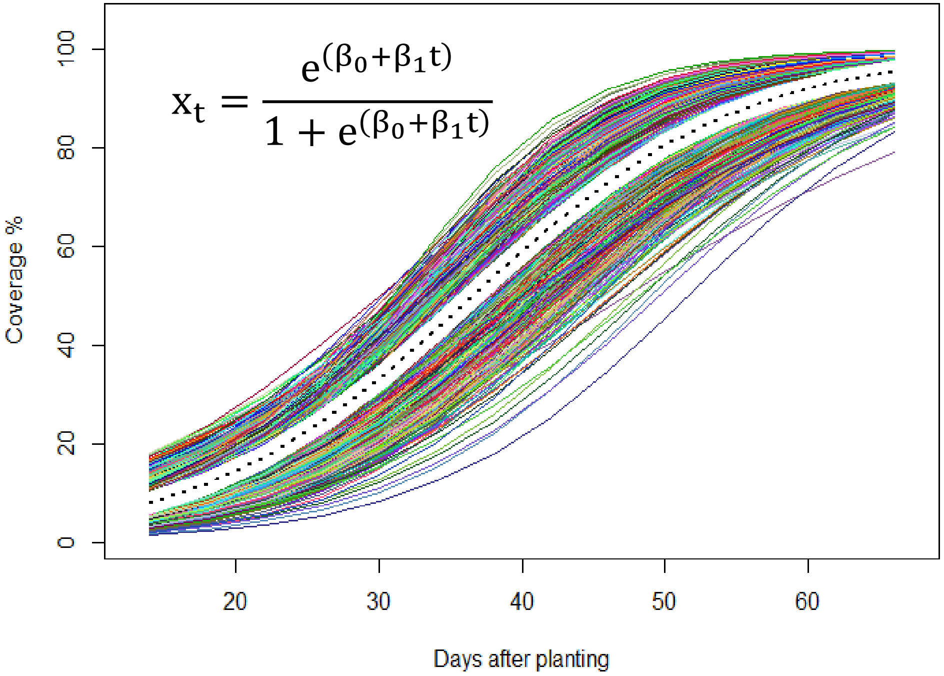


Figure A-5. Asymptotic logistic fit of seasonal percentage canopy coverage for one genotype during the 2014 growing season. Open circles represent actual sampling dates, filled circles denote estimates from logistic model.

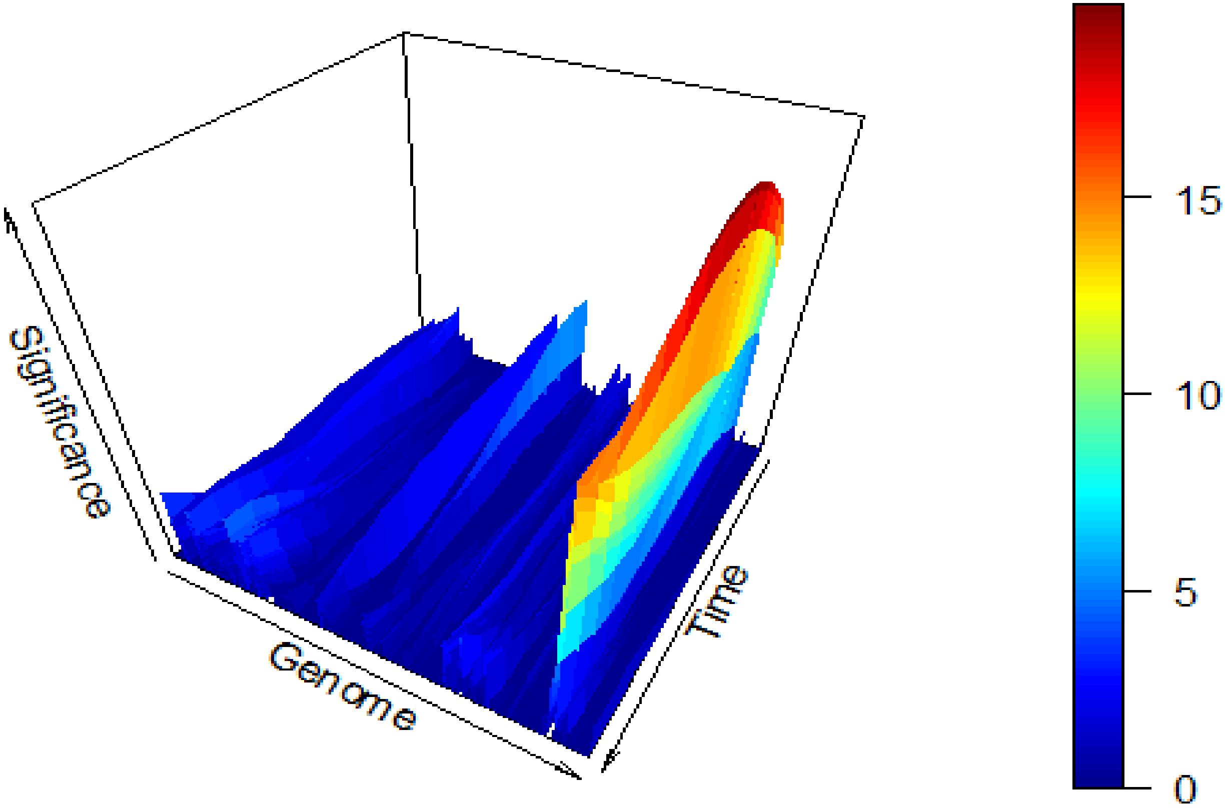
(Hall 2015)



$$x_t = \frac{e^{(\beta_0+\beta_1 t)}}{1+e^{(\beta_0+\beta_1 t)}}$$

TIME

Inference performed by time point



Approaching time

1. Indexation
   - Reduce various time-points to a single phenotype
   - Example: Regression parameters of a growth curve

2. Model stages separately and combine the outcome

3. Covariate: Continuous variable that states the stage

4. Structured random effect (aka. random regression)
   - Add a factor that indicates the stage/date/age
   - Tell the model that the how stages are correlated

5. Multivariate: Treat each stage is a different trait

MULTIPLE TRAITS

THOUGHTS ON DATA & MULTIPLE TRAIT ANALYSIS

1. Data storage
   - Collect & store only all data I can?
   - How much resolution to get what I am looking for?

2. Mean or median
   - What is the level of dispersion of my data?
   - Aerial imagery data trend to display lots of OUTLIERS

3. Computation burden & analysis
   - With big data, comes the need for big computers
   - The headache is proportional to the volume of the dataset

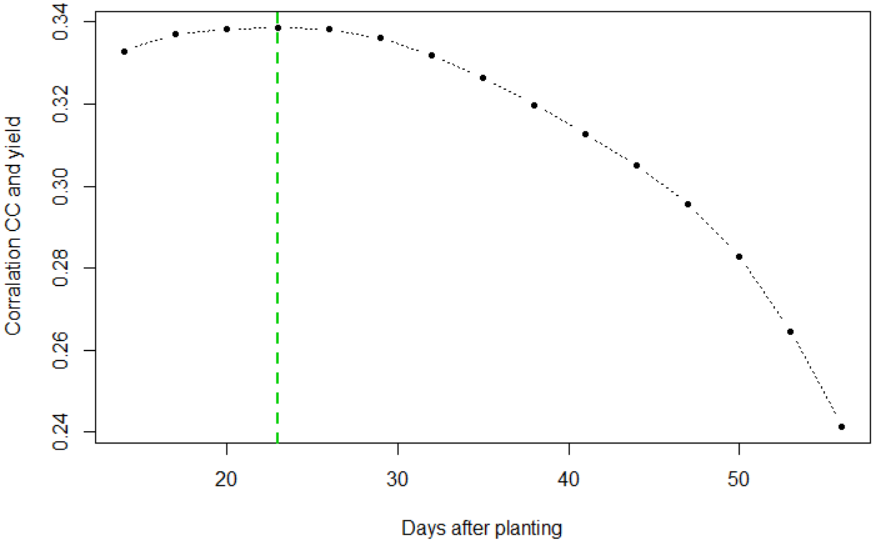
4. Research question
   - **CLARITY IS ESSENTIAL**: What are we looking?

MULTIPLE TRAIT – RESEARCH QUESTION

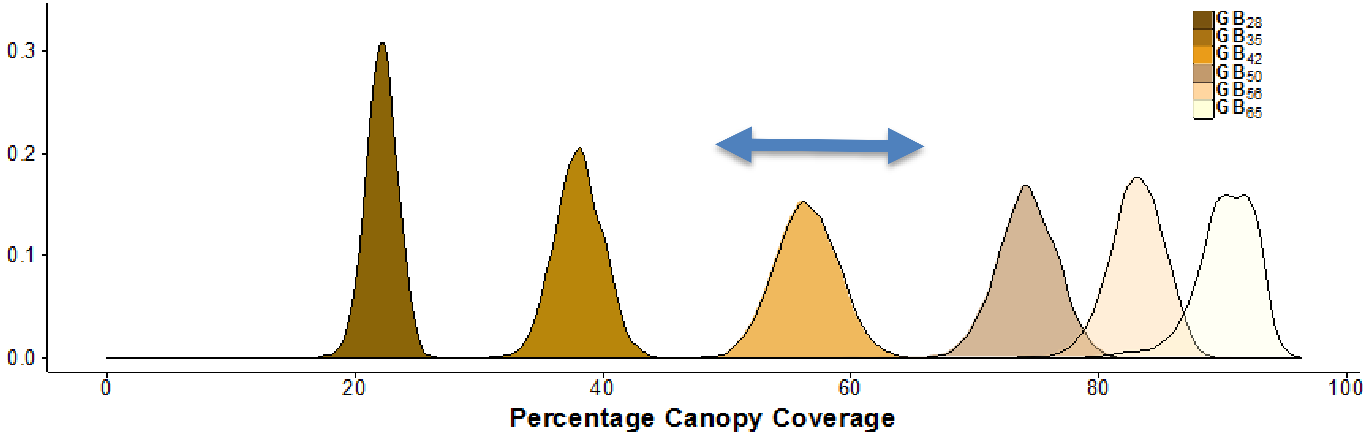
- Are all spectra of useful for the purpose of the experiment?
- Are there just a few very important bands?
- Are there spectral combinations or ratios of importance?
- Are there spectra that are purely environmental and do not depend on the treatment of interest?
- Is reduction of dimensionality (ie. PCs) something to consider?
- Is vegetation indexes / selection indexes something to consider?
- Are there stages more important than other to collect data?
- Are there important combinations of spectrum-stage?

MULTIPLE TRAIT – Avoidance

Is there a best single trait or time point?



Is there a point of maximum variability?

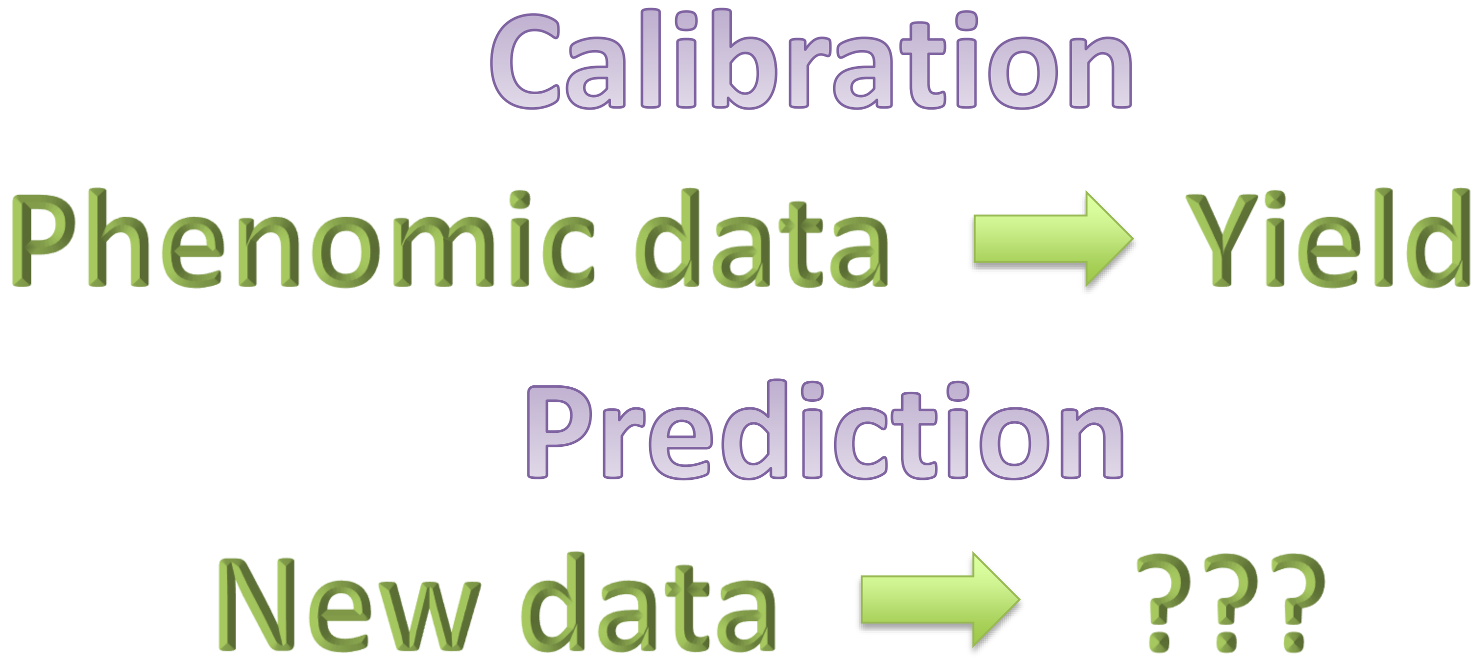


IMPLEMENTATIONS

- Software that enable complex modeling
  - BLUPF90
  - ASREML
  - SAS (proc mixed)
  - R packages:
    - lme4, nlme, MCMCglmm, asreml, BGLR, EMMREML, varComp, NAM, MTM (github), MixedModel (github)
  - Matlab
  - MTG2

PHENOMIC-ENABLE PREDICTION

PREDICTION



Calibration

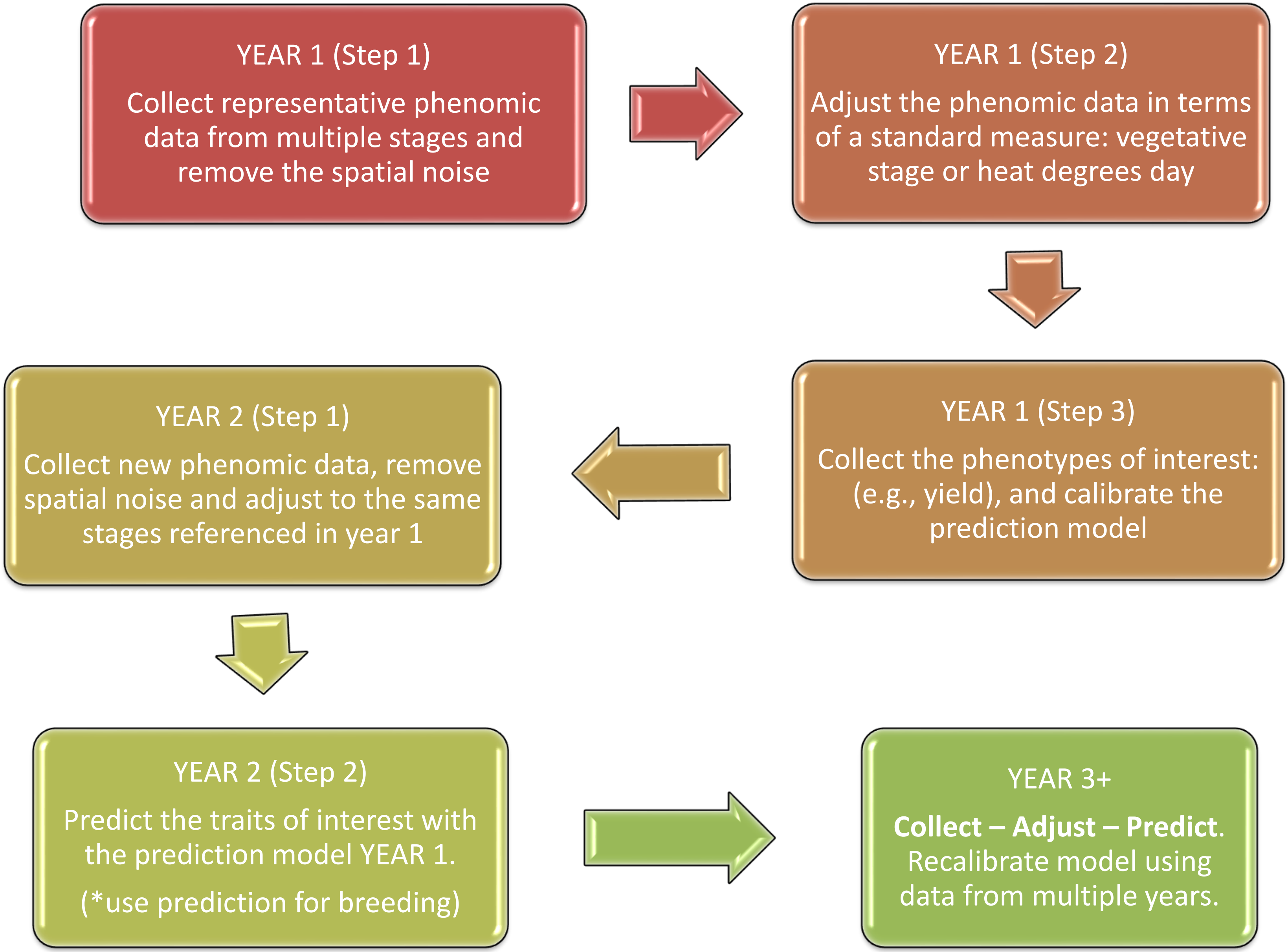
Phenomic data → Yield

Prediction

New data → ???

- High-throughput phenotypes to predict the phenotypes of meaningful agronomic traits (yield, maturity and height)
- *No guarantees* that the predictions hold accurate across environments (or across generations)
- No need to know what are the important parameters.

PREDICTION



YEAR 1 (Step 1)

Collect representative phenomic data from multiple stages and remove the spatial noise

YEAR 1 (Step 2)

Adjust the phenomic data in terms of a standard measure: vegetative stage or heat degrees day

YEAR 1 (Step 3)

Collect the phenotypes of interest: (e.g., yield), and calibrate the prediction model

YEAR 2 (Step 1)

Collect new phenomic data, remove spatial noise and adjust to the same stages referenced in year 1

YEAR 2 (Step 2)

Predict the traits of interest with the prediction model YEAR 1.

(*use prediction for breeding)

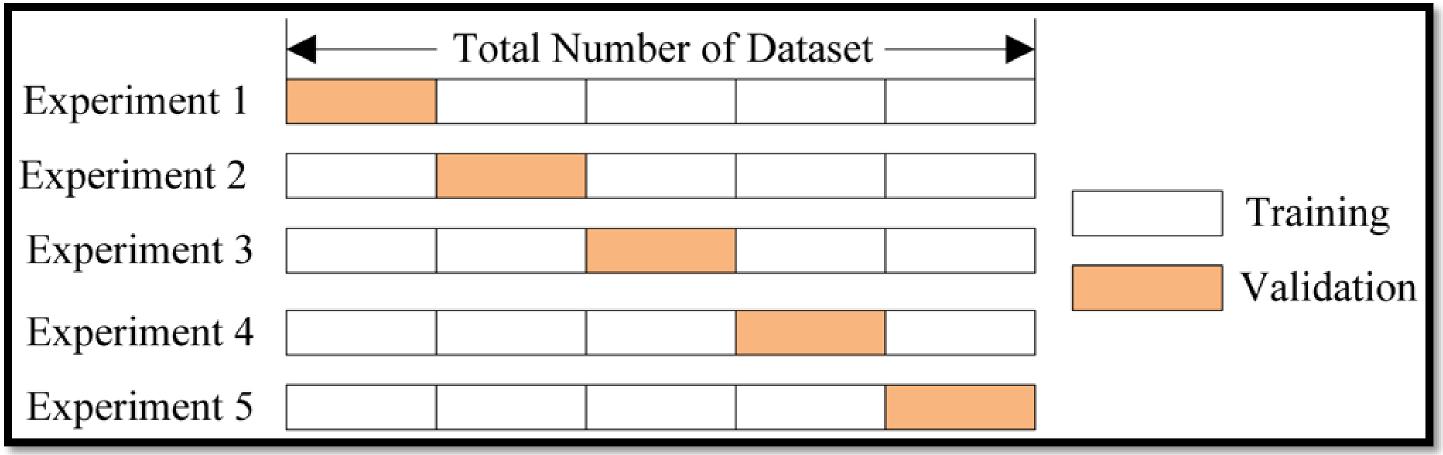
YEAR 3+

Collect – Adjust – Predict. Recalibrate model using data from multiple years.

CROSS-VALIDATION

- **Cross-validation:** define the model or method of prediction

- **How does it work?**
  1. Split the data into *k* subsets (at random or not)
  2. Calibrate the proposed model with all but 1 subset
  3. Predict the subset left out
  4. Repeat the procedure to all subsets
  5. Average the accuracy (ie. correlation or prediction error)
  - Repeat 1-5 for different methods or models under evaluation



IMPLEMENTATIONS IN R packages

- **Bayesian methods**: BGLR
- **Random Forest**: ranger
- **Ridge, Lasso and Elastic-net**: glmnet
- **Mixed models**: rrBLUP, BGLR, NAM
- **Support Vector Regression**: kernlab
- **Partial least square**: pls
- **Neural Network**: nnet, neuralnet
- **Boosting**: gbm
- **k-Nearest Neighbors**: kknn

That’s all!

QUESTIONS??

http://alenxav.wixsite.com/home