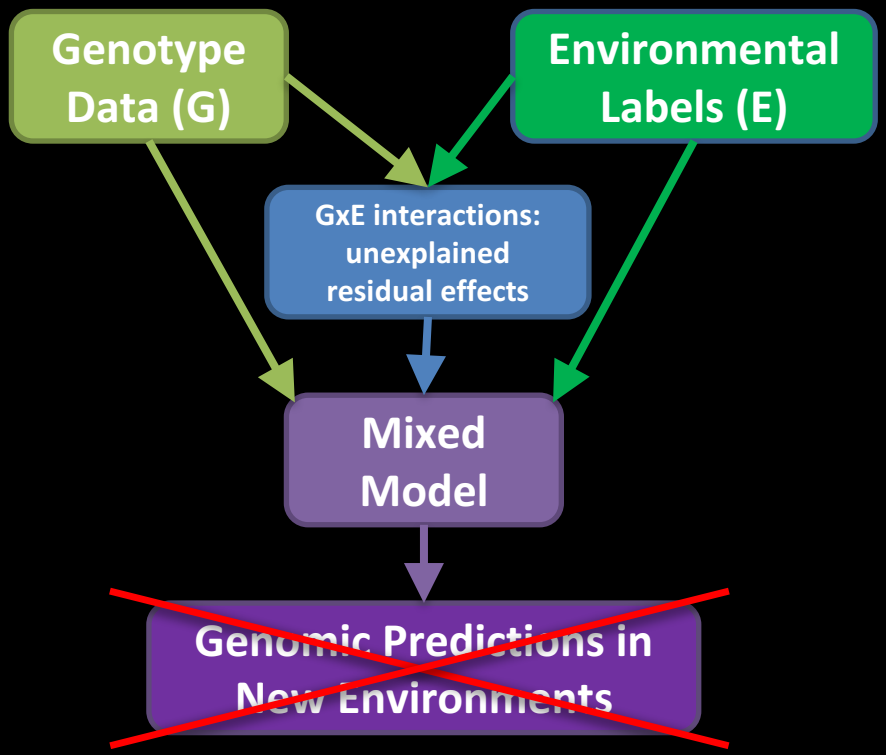
Hybrid Prediction with Marker and Climate Data

Anna R. Rogers
Jim Holland

Modeling GxE

- Traditional Approach



Genomes to Fields (G2F) data (2014-2016)

- 2118 maize hybrids tested over 23-32 locations, locations vary by year
  - 21,122 markers filtered and imputed from inbred parents, hybrid genotypes generated in TASSEL
- Weather data gathered at each location
  - 30 minute increments
  - Temperature, humidity, solar radiation, rainfall, wind speed
- Photoperiods calculated from United States Naval Observatory Data

Genomes to Fields (G2F) Initiative, 2017.

Data Cleaning –Genotype Data

- Genotype Data: 1221 parental lines of G2F hybrids
  - Filter out markers with <90% data present, monomorphic markers (22008 markers remain)
  - Impute missing values using LD KNNi method
    - Filter out duplicate lines (1221 unique parental lines)
    - Highest heterozygosity within individual: 31.4% → Some parent lines are actually hybrids?
    - Sites without any heterozygotes: 982
  - Filter out sites with high (>5%) heterozygosity
    - 21,122 sites remain
    - Highest heterozygosity with individual: 29.9%
  - "Coin Flip" in R to remove heterozygous sites
    - TASSEL "create hybrid genotypes" function does not tolerate heterozygous calls → sets marker to missing if either or both parents are heterozygous
  - Create hybrid genotypes in TASSEL
    - Now have both Inbred and Hybrid Genotype sets to perform PCA
  - Output genotypes as expected counts of the minor allele
    - Impute major allele for any missing values to prep for PCA
  - Standardize each markers to mean = 0, variance = 1

Data Cleaning –Weather Data

- Weather Data:
  - 30 minute increments transformed first into daily data in R:
    - Means, High-Lows, Cumulative Rainfall
  - Daily Data imputed for missing values, and all wind speed values webscraping from Weather Underground
  - Imputation of calculatable results (T, T_d, RH) using formulas from Lawrence, M. (2005)
  - Photoperiod calculated using Sunrise Sunset Tables from USNO
  - Growing Degree days calculated for daily data
  - 30 day period means, rain accumulation, cumulative growing degree days calculated in R

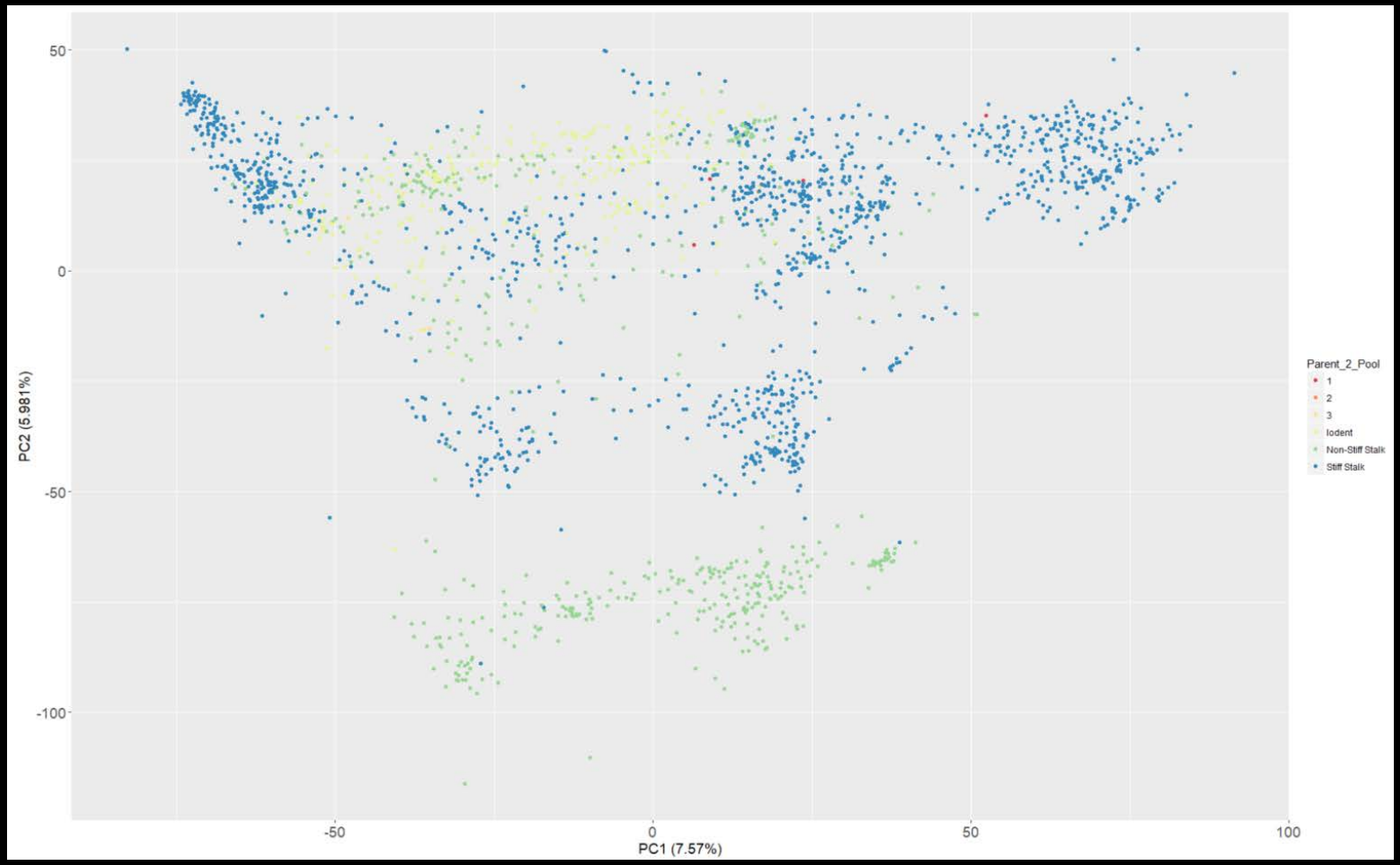
Data Cleaning Phenotype -Data

- Phenotype Data
  - G2F clean data came filtered to remove data for erroneous values
    - Ear Height < 20 cm
    - Days to Pollen Shed < 20 days
      Days to Pollen Shed >100 days
    - Days to Silking > 100 days
    - Weight < 1.0 lbs
      Sets weight, grain yield, grain moisture, test weight to missing
  - More filtering was necessary:
    - Removed all values with comments column indicating errors in planting, destroyed plots, concerning commentary
    - Filtering on stand count/area to remove plots < 15,000 plants/acre
    - Removal of Local Checks and other lines with no genotype information
    - Name matching for genotype lines to phenotype lines (inconsistent naming year to year)

PC Analysis of Hybrid Genetic Data

- First 10 PC's account for 37% of variance among hybrids
- Appear to be separating out groupings of hybrids – but hybrids are admixtures by design, so groups not clearly defined

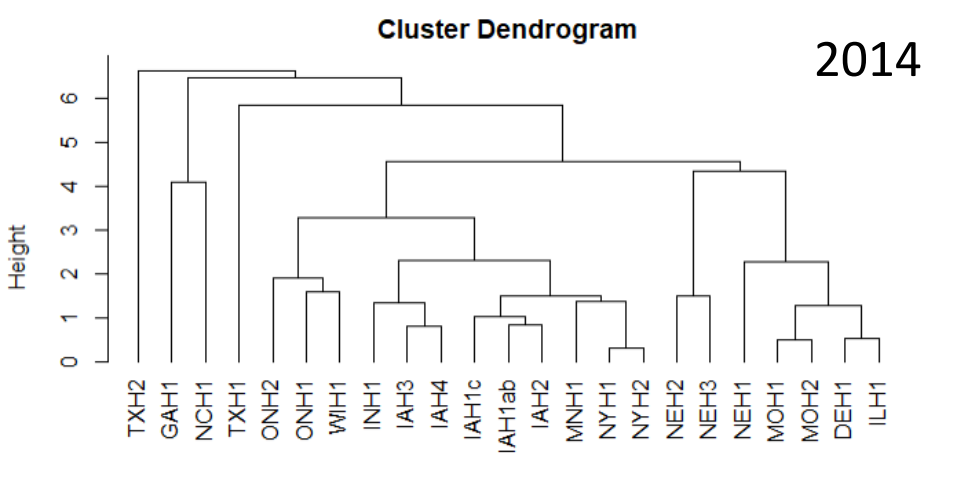
Hybrid genetic similarities based on PCA of marker data

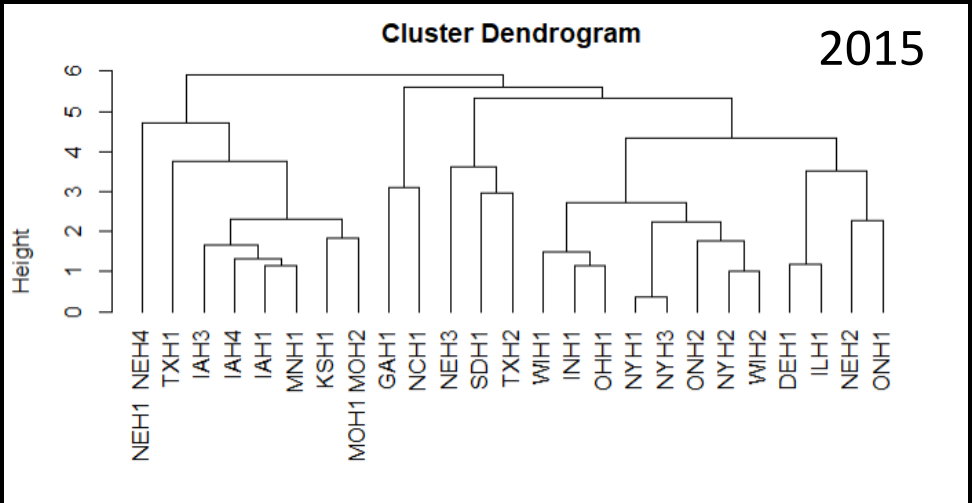


We will attempt to define hybrid clusters with these

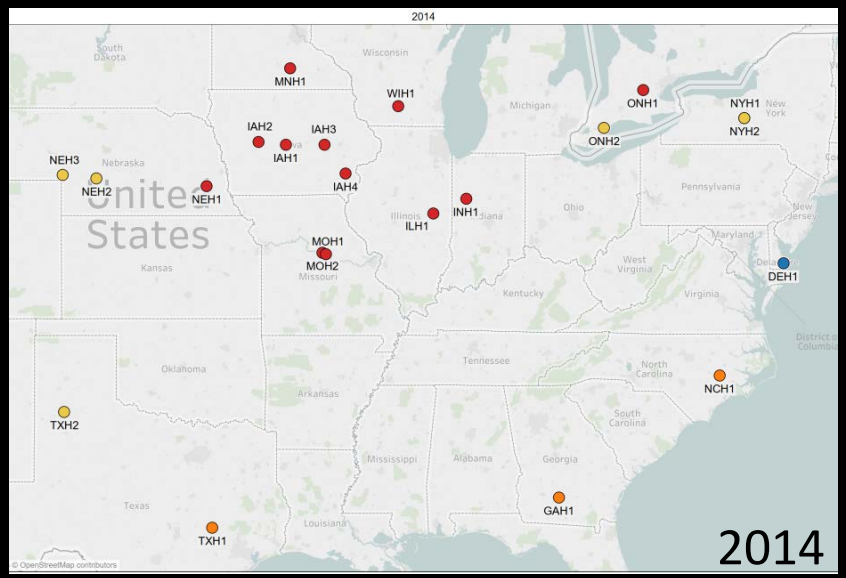
Factor Analysis of Weather data

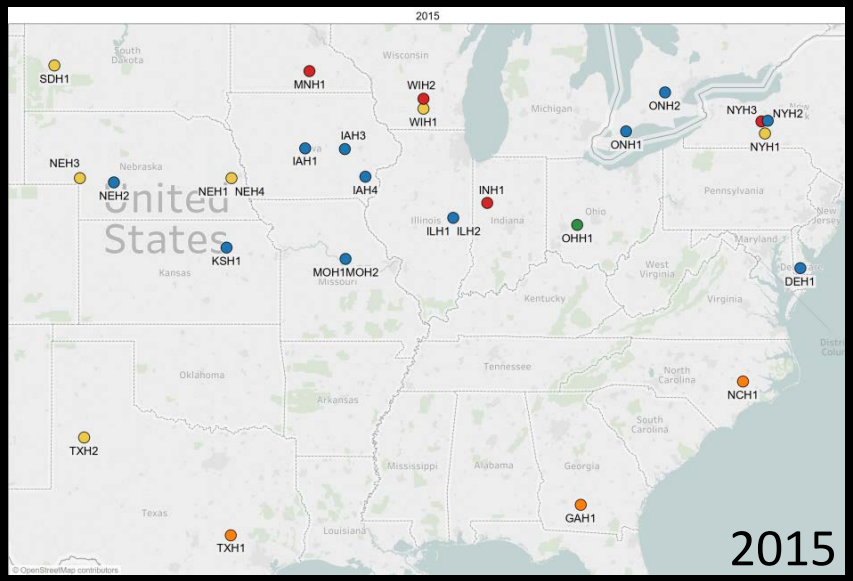
- 2014 and 2015 Data:
  - Scree plotting suggests 5-6 factors, using 6 as of now
  - Hierarchical Clustering done to group environments → after imputation of wind values, environments fall out in relatively good geographical patterns

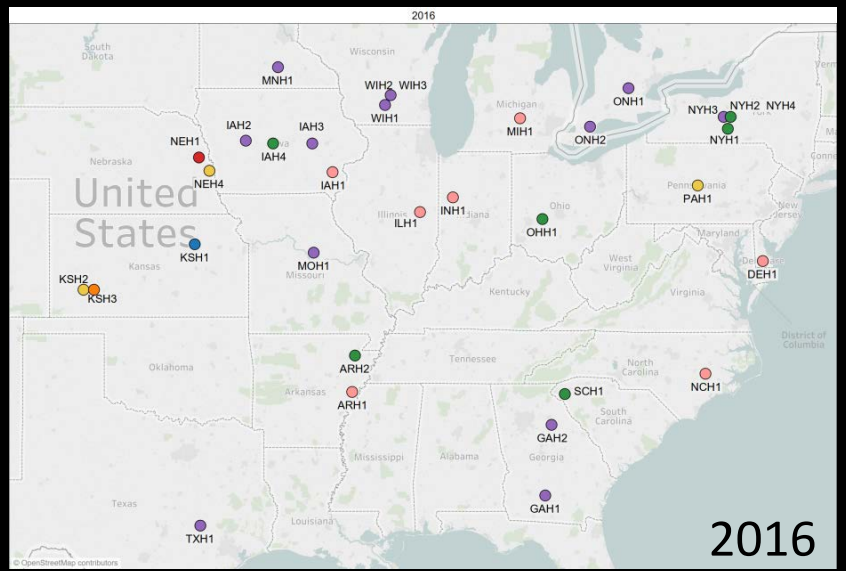




Weather factor-based clusters relate to geography





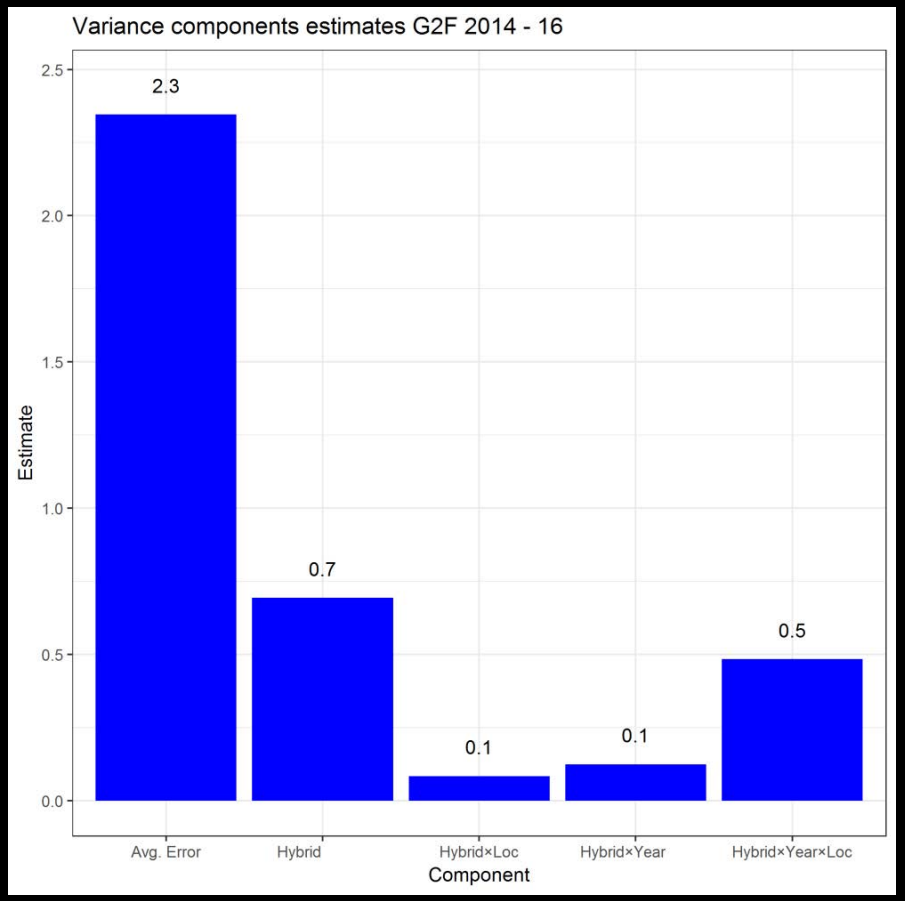


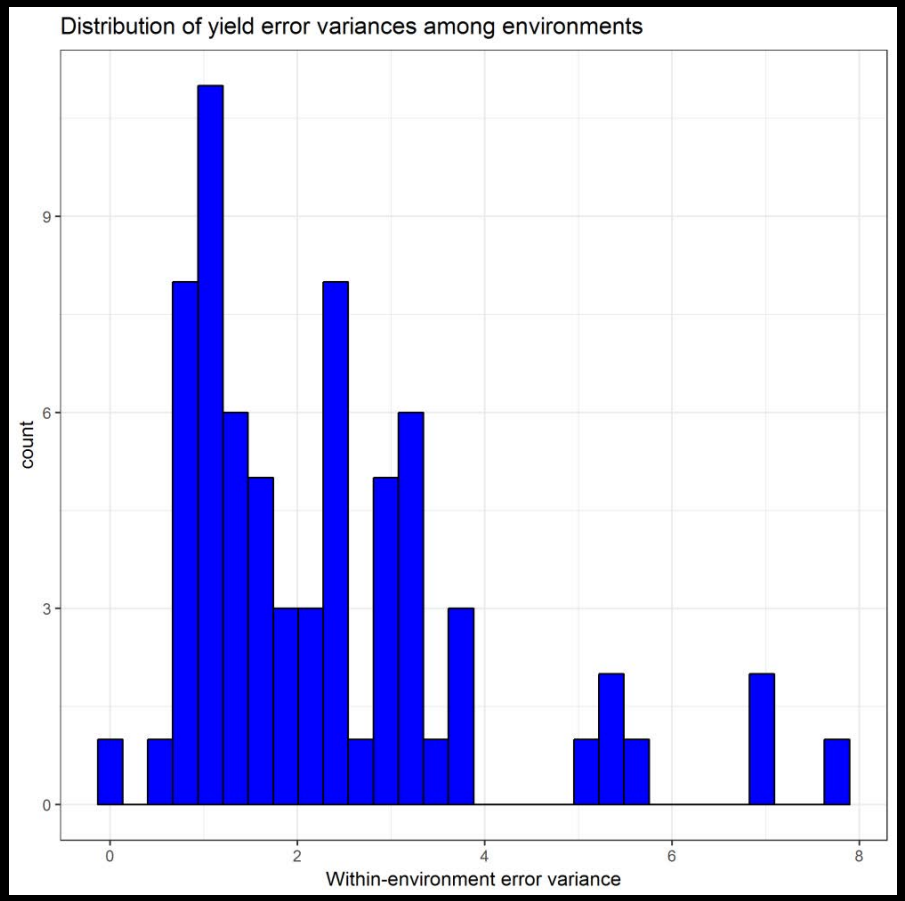
Ward Clusters
1
2
3
4
5
6
7

Next Steps for Factor Analysis

- Cluster 2016 Data alone
- Factor Analysis on data from all three years together
  - Hypothesis: Areas will cluster with themselves from year to year (i.e. NCH1_2014 clusters with NCH1_2015 and NCH1_2016
- Some locations not present in every year → will be interesting to see where they cluster
- Output Factor Analytic scores for use in predictive modeling

Variance components from ANOVA for yield





Total G×E variance ≈ Genetic variance

Huge range of variation among environs

Traditional ANOVA model for G×E

- $Y_{ijk} = \mu + E_i + G_j + GE_{ij} + \varepsilon_{ijk}$

Assume:

- one common variance for genotypes (V_G)
- one common variance for G × E (V_{GE})
- one common variance for residuals
- Covariance of a genotype's performance between

Genotype nested in environment model

$$\begin{array}{c c} & \begin{array}{ccc} E_1 & E_2 & E_3 \end{array} \\ \begin{array}{c} E_1 \\ E_2 \\ E_3 \end{array} & \begin{bmatrix} V_{G(E)} & C_{GEjj'} & C_{GEjj'} \\ C_{GEjj'} & V_{G(E)} & C_{GEjj'} \\ C_{GEjj'} & C_{GEjj'} & V_{G(E)} \end{bmatrix} \end{array}$$

- $Y_{ijk} = \mu + E_i + G(E)_{ij} + \varepsilon_{ijk}$

- We can make this model identical to traditional ANOVA by assuming that:
- Cov($G(E)_{ij}, G(E)_{i'j}$) is same for all environment pairs, this will give same value as V_G in ANOVA
- Var($G(E)_{ij}$) will give same value as $V_G + V_{GE}$ in ANOVA

Genotype nested in environment: unstructured covariance

- $Y_{ijk} = \mu + E_i + G(E)_{ij} + \varepsilon_{ijk}$

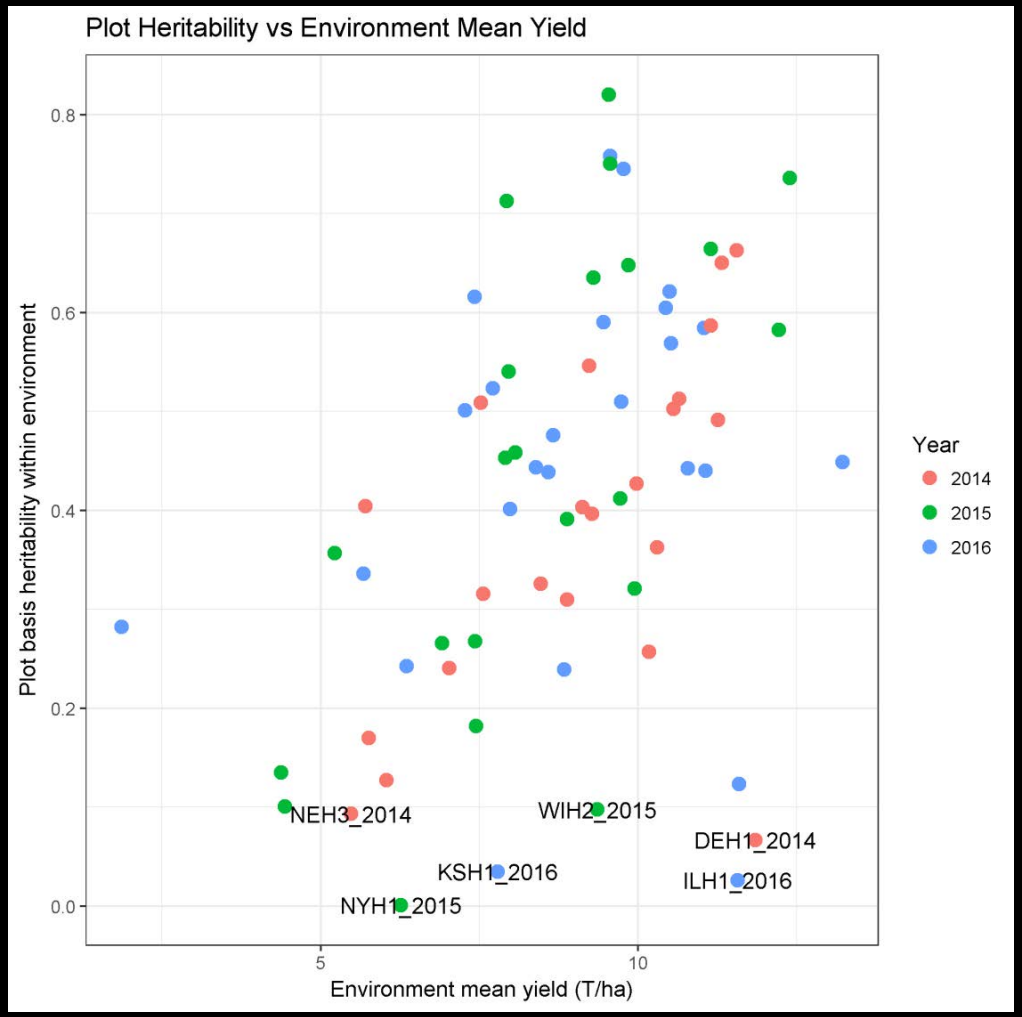
$$\begin{array}{c|ccc} & E_1 & E_2 & E_3 \\ \hline E_1 & V_{G(E1)} & C_{GE12} & C_{GE13} \\ E_2 & C_{GE12} & V_{G(E2)} & C_{GE23} \\ E_3 & C_{GE13} & C_{GE23} & V_{G(E3)} \end{array}$$

- With a mixed model, we can estimate unstructured variance/covariance matrix among environments
- This allows each environment to have a specific genetic variance and each pair of environments to have a specific genetic covariance/correlation
- Also we can allow error variances to be unique for each

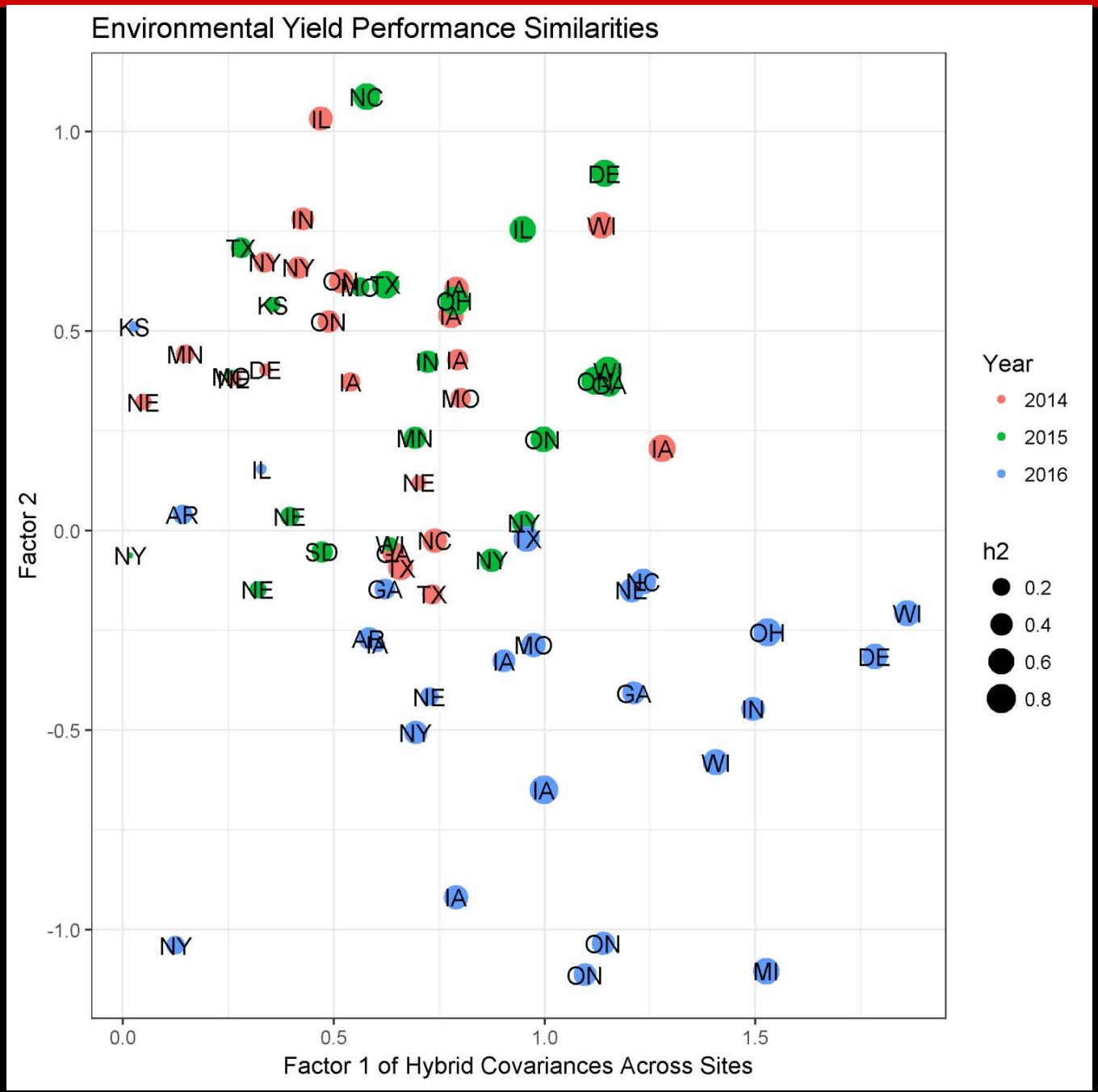
Too many parameters!

- With 70 environments in G2F, we have 70 different $V_{G(E)}$ + 2415 unique pairwise covariances + 70 residual error variances to estimate.
- We can impose a factor analytic structure on the covariance matrix:
  - Similar to PCA, we represent the covariance matrix with fewer dimensions (factors)
  - The covariance matrix can be approximated as the outer product of one factor…or the sum of outer products of a few factors
  - Plus an additional site-specific genetic variance

Factor analytic models with 1 or 2 factors fit G2F yield data well



Lots of variation among sites for heritability and mean yield



Environments closer together have higher genetic covariance.

What we need for prediction

- $f(\boldsymbol{W}) \approx \boldsymbol{GE}$ Some function of weather data matrix that can predict missing values in the **GE** variance-covariance matrix
- Can do this by predicting the factors of **GE** with various statistical learning models
- $f(\boldsymbol{M}) \approx \boldsymbol{G}$ Some function of marker data matrix that can predict missing values in the **G** matrix. vanRaden realized genomic relationship does this well for polygenic

Yield GE Factor – Weather Variable Correlations

Weather Variable	Factor 1	Factor 2
Mean high temp period 1	0.40	-0.33
Mean low temp period 2	-0.25	0.45
Mean precip period 5	0.27	0.39
Mean wind period 5	-0.23	0.14
% days no rain period 3	0.43	-0.46
Mean humidity period 4	0.07	0.27

We can inspect these for causal relationships, but for prediction, best to include all environmental covariates

Genetic prediction in ‘new’ environments

- ‘new’ environment is in our data set, but held out from model training
- ‘new’ genotypes also in our data set, but held out from model training
- Prediction accuracy measured on new genotypes in new environments from various models

Cross-validation schemes

- 1. Hold out random 10% of hybrids
- 2. K-means clustering to get 10 groups based on marker similarity, hold out one group at a time
- A. Hold out one environment at a time
- B. Cluster environments based on weather data, hold out one group at a time
- Do all combinations of the above.

Alternative Approach

- Directly fit marker and marker*weather covariates
- Estimate marker and marker*weather effects using regularization or Bayesian models
- Predict held out observations
- Problem here is heavy parameterization:
- 20k marker effects + 800k* marker*weather interactions + non-genetic effects on 35k phenotypic records!
- BGLR is pretty good at this, but initial tests show that memory demands are big and cross-validation is going to be SLOW…