Keep Calm and Yield Corn:

Using Cluster Analysis to Identify "Super Corn"

SMU SINGAPORE MANAGEMENT UNIVERSITY

ISSS602 Data Analytics Lab

Ong Hui Lin | Soh Hui Shan | Teo Yaling

INTRODUCTION

From the husks to the cobs to the kernels, almost every part of corn can be used to improve our lives today. Most of the corn produced are used for ethanol fuel production, animal feed and food such as our favourite popcorn. Although it is grown mainly in wet and hot climates, it can also thrive in cold and dry climates, making it an extremely versatile crop. However, in 2012, the harshest drought disaster in U.S. has reduced corn yield by 26% and caused $30 billion agricultural losses. Many farmers' cornfields shrivelled, producing small to non-existent yields. Scientists and researchers are therefore motivated to create resilient hybrids that fight drought. So the million dollar question is –

Which are the Super Corns?

OBJECTIVE

(1) Identify stressed vs non-stressed environments by grouping aggregated weather data into meaningful clusters based on weather indicators for heat and drought – Maximum Temperature (TMAX) and Precipitation (PREC).

(2) Identify "Super Corn" hybrids by comparing differences in selected hybrids' mean yield across all clusters.

ANALYSIS FLOWCHART

Data Source

Datasets from Syngenta Crop Challenge in Analytics 2019:

(1) *performance_data* – Contains the yields of 2452 corn hybrids in 1560 environments from 2008 to 2017.

(2) *weather_data* – Contains daily weather at each planting environment (ENV_ID), which is uniquely identified by latitude, longitude, and year.

Data Preparation

(1) Imported datasets into SAS JMP Pro 14.

(2) Cleaned *performance_data* to eliminate discrepancies in PLANT_DATE and HARVEST_DATE.

(2) Aggregated yields of each HYBRID_ID + ENV_ID combination as mean(YIELD).

Data Integration

(1) Joined *weather_data* to *performance_data* to bring in daily weather data between PLANT_DATE to HARVEST_DATE using JMP's Query Builder.

(2) Aggregated daily weather data to their mean values at each HYBRID_ID+ENV_ID combination.

Interactive Exploratory Data Analysis

Analysis Methods

(1) Performed variable selection using multivariate analysis to examine correlation of weather variables.

(2) Performed cluster analysis of selected weather variables (TMAX, PREC) using JMP's Normal Mixtures Clustering Platform.

(3) Performed ANOVA of mean yields of HYBRID_ID across all clusters.

Insights Discovery

METHODOLOGY & RESULTS

(1) EDA

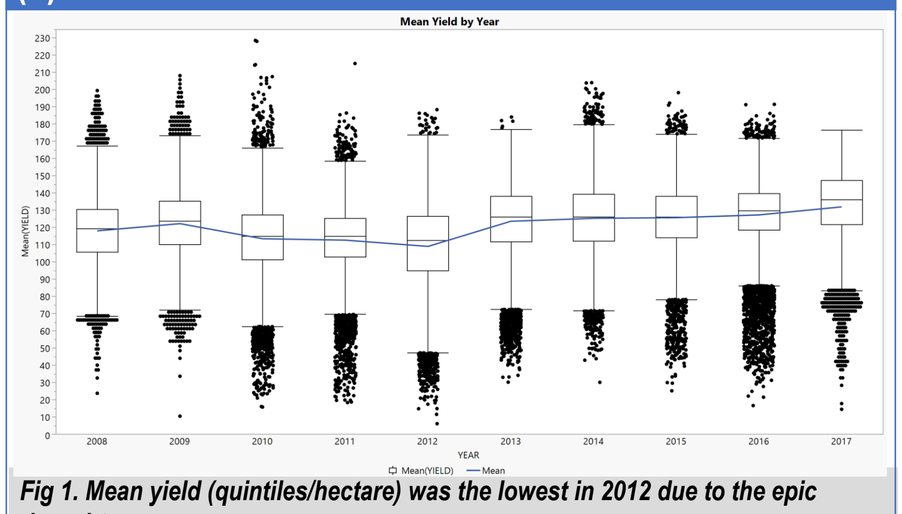
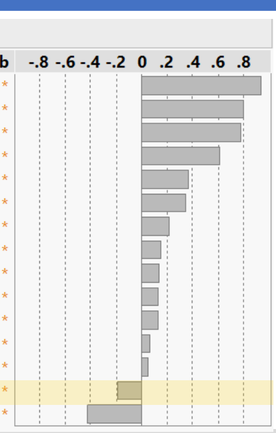


Fig 1. Mean yield (quintiles/hectare) was the lowest in 2012 due to the epic drought.

(2) MULTIVARIATE ANALYSIS - CORRELATION

Pairwise Correlations

Variable	by Variable	Correlation	Signif Prob
Mean(VP)	Mean(TMIN)	0.9354	<.0001*
Mean(TMAX)	Mean(SRAD)	0.7990	<.0001*
Mean(TMIN)	Mean(TMAX)	0.7804	<.0001*
Mean(VP)	Mean(TMAX)	0.6156	<.0001*
Mean(TMIN)	Mean(SRAD)	0.3651	<.0001*
Mean(VP)	Mean(PREC)	0.3453	<.0001*
Mean(TMIN)	Mean(PREC)	0.2165	<.0001*
Mean(PREC)	Mean(DAYL)	0.1488	<.0001*
Mean(VP)	Mean(SRAD)	0.1371	<.0001*
Mean(VP)	Mean(DAYL)	0.1350	<.0001*
Mean(TMIN)	Mean(DAYL)	0.1281	<.0001*
Mean(SRAD)	Mean(DAYL)	0.0674	<.0001*
Mean(TMAX)	Mean(DAYL)	0.0558	<.0001*
Mean(TMAX)	Mean(PREC)	-0.1864	<.0001*
Mean(SRAD)	Mean(PREC)	-0.4270	<.0001*



Weather Measures: Day Length (DAYL), Solar Radiation (SRAD), Minimum Temp (TMIN), Vapour Pressure (VP), PREC, TMAX.

Fig 2. Pairwise Correlations of all weather variables.

Key environment measures – Mean(TMAX) and Mean(PREC) have low correlation of -0.19. Both were subsequently used in clustering.

(3) CLUSTER ANALYSIS USING NORMAL MIXTURES

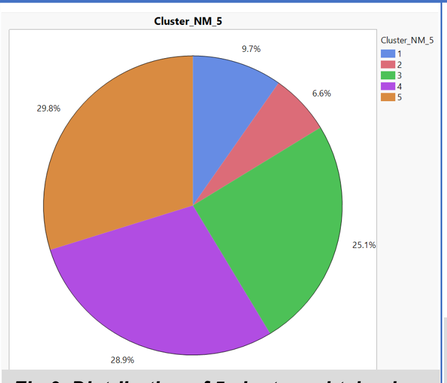


Fig 3. Distribution of 5 clusters obtained from JMP's Normal Mixtures Clustering Platform.

1. Optimal Northern Zone
TMAX ☀ ☀
TMIN ❄ ❄ ❄ ❄
PREC 🌧 🌧 🌧 🌧 🌧
YIELD 🌽 🌽 🌽 🌽

2. Heat-stressed Southern Zone
TMAX ☀ ☀ ☀ ☀ ☀
TMIN ❄
PREC 🌧 🌧 🌧
YIELD 🌽

3. Drought-stressed Corn Belt
TMAX ☀ ☀ ☀ ☀
TMIN ❄ ❄ ❄
PREC 🌧
YIELD 🌽 🌽 🌽

4. Optimal Corn Belt
TMAX ☀ ☀ ☀
TMIN ❄ ❄
PREC 🌧 🌧 🌧 🌧
YIELD 🌽 🌽 🌽 🌽 🌽

5. Cold Great Lake
TMAX ☀
TMIN ❄ ❄ ❄ ❄ ❄
PREC 🌧 🌧
YIELD 🌽 🌽

Fig 10. Cluster Summary

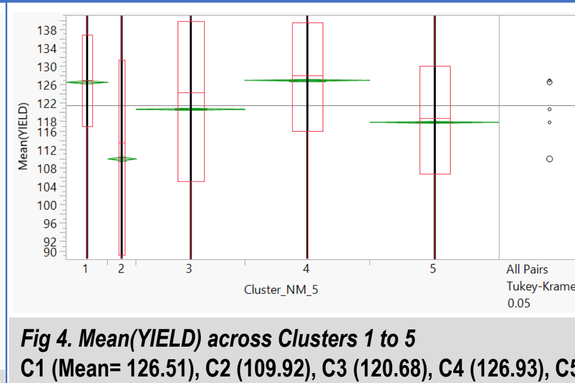


Fig 4. Mean(YIELD) across Clusters 1 to 5

C1 (Mean= 126.51), C2 (109.92), C3 (120.68), C4 (126.93), C5 (117.84)

Pairwise comparisons: All p-values <0.0001, except for C1 vs C4

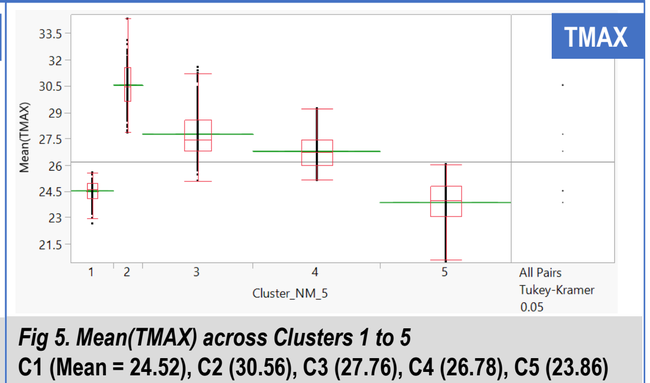


Fig 5. Mean(TMAX) across Clusters 1 to 5

C1 (Mean = 24.52), C2 (30.56), C3 (27.76), C4 (26.78), C5 (23.86)

Pairwise comparisons: All p-values <0.0001

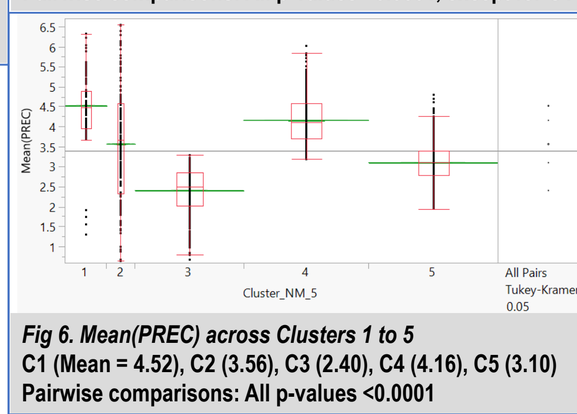


Fig 6. Mean(PREC) across Clusters 1 to 5

C1 (Mean = 4.52), C2 (3.56), C3 (2.40), C4 (4.16), C5 (3.10)

Pairwise comparisons: All p-values <0.0001

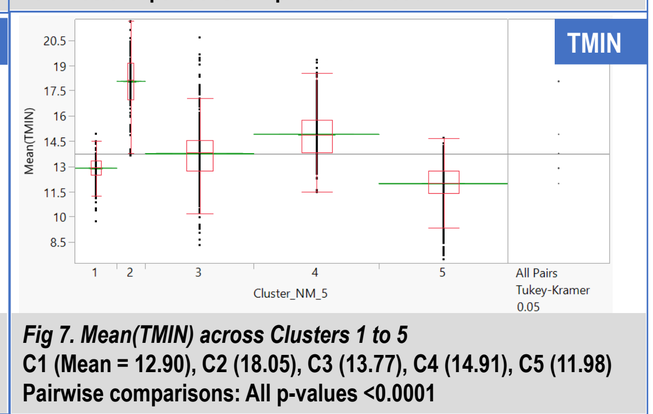


Fig 7. Mean(TMIN) across Clusters 1 to 5

C1 (Mean = 12.90), C2 (18.05), C3 (13.77), C4 (14.91), C5 (11.98)

Pairwise comparisons: All p-values <0.0001

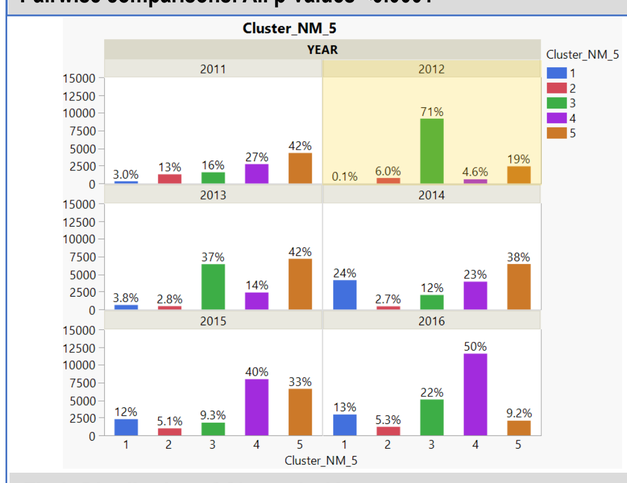


Fig 8. Distribution of Cluster by Year

71% of Hybrids in 2012 (worst drought in US) was located in C3.

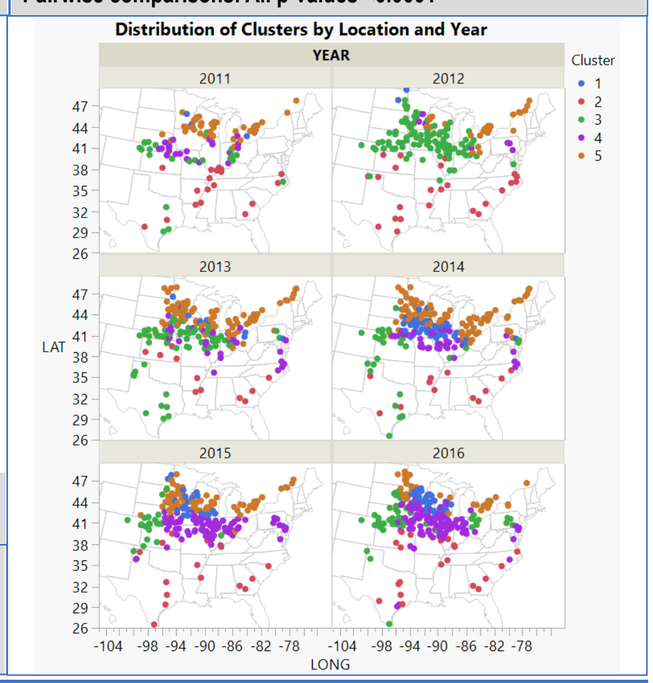


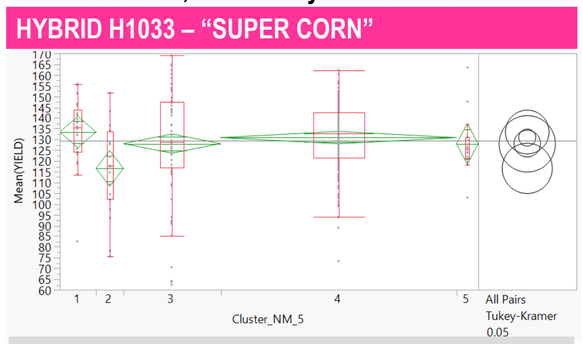
Fig 9. (right) Distribution of Cluster by Geographical Location and Year

Clusters were relatively well separated across United States corn harvesting regions.

(4) ANOVA – IDENTIFICATION OF "SUPER CORN"

ANOVA was performed for selected hybrids planted in all 5 clusters. For heat and drought-resistant "super corn", there was no significant difference in mean yield across C1 to C5 (at significance level of 0.01). Amongst the heat and drought-susceptible "normal corn", there was significant difference in mean yield across C1 to C5, and mean yield is observed to be lower in C2 and/or C3 (at significance level of 0.01).

HYBRID H1033 – "SUPER CORN"

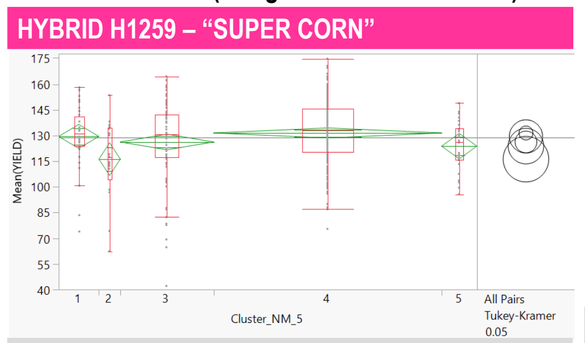


C1 (Mean = 133.51), C2 (116.58), C3 (128.12), C4 (130.91), 5 (128.00)

ANOVA: p-value 0.0208 > 0.01

Pairwise comparisons: All p-values > 0.01

HYBRID H1259 – "SUPER CORN"

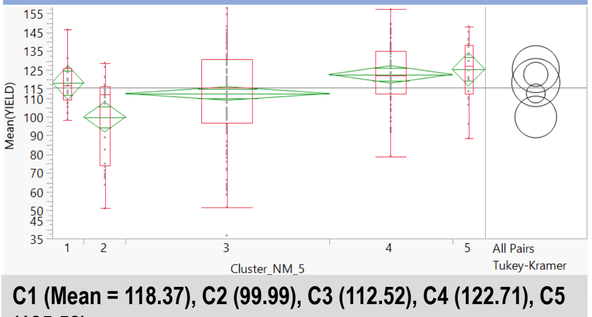


C1 (Mean = 129.42), C2 (116.18), C3 (126.19), C4 (131.44), C5 (123.76)

ANOVA: p-value 0.0125 > 0.01

Pairwise comparisons: All p-values > 0.01

HYBRID H1202 – "NORMAL CORN"



C1 (Mean = 118.37), C2 (99.99), C3 (112.52), C4 (122.71), C5 (125.52)

ANOVA: p-value <0.0001

Pairwise comparisons: C2 vs C5, C2 vs C4, C3 vs C4 (p-value < 0.01)

CONCLUSION & RECOMMENDATIONS

- ✓ Optimal Northern Zone (Cluster 1) and Optimal Corn Belt (Cluster 4) were ideal environments; Heat-stressed Southern Zone (Cluster 2) and Drought-stressed Corn Belt (Cluster 3) were stressed environments while Cold Great Lake (Cluster 5) defined the cold and dry environments.
- ✓ Effects of weather conditions on yield were shown to be statistically significant with the mean yield differing across the 5 environment clusters, except for Optimal Northern Zone (Cluster 1) and Optimal Corn Belt (Cluster 4).
- ✓ ANOVA was performed to identify the "super corns" which could thrive in all types of environments. With increasing extreme climate changes, corn producers can focus on cultivating these "super corns" to ensure continuity of yield amidst unfavorable weather conditions.