P7 - Final Report

Ram Pangaluri, Ashok Shanker, Fabian Leon

https://fleon.shinyapps.io/STAT689-Project/ (https://fleon.shinyapps.io/STAT689-Project/)
https://github.com/leon-fabian/STAT689-Project (https://github.com/leon-fabian/STAT689-Project)

Introduction

Sorghum [Sorghum *bicolor* (L. Moench)] is the world's fifth most important cereal crop, in terms of both production and area planted (Balakrishna et al. 2019). It is an increasingly relevant grain crop due to its resilience to drier and hotter climates. In the United States, sorghum is typically grown in dryland areas in Kansas and Texas with Texas accounting for ~1.8 million of the United States' ~5.8 million acres of sorghum production ("USDA Nass Quickstats" 2021). The circumstances of where it is grown mean that sorghum usually does not meet its optimal yield potential. To ensure that the best decisions are being made to achieve maximal sorghum yields despite poor growing conditions we attempt to visualize trends in the environmental, genetic, and agronomic factors affecting yield.

Motivation

The idea for this visualization project was to use the Texas A&M Variety Testing ("Texas Agrilife Crop Testing" 2022) and USDA NASS data repositories ("USDA Nass Quickstats" 2021) on grain sorghum production ("Sorghum 101" 2022) to visualize yield and agricultural production trends for the state of Texas. The first dataset is from Texas A&M Agrilife, which contains data from 1970 - 2021 Texas Sorghum Variety Trials. The second is the dataset from the USDA National Agricultural Statistics website that was collected to reflect this same time period. While agronomists attempt to look at the trends through their research, individual experimental attempts seldom span as long a time period. Visualizing these datasets is the best representation of historical trends for Texas production regions as could be compiled.

Usage

The app can be found here https://fleon.shinyapps.io/STAT689-Project/ (https://fleon.shinyapps.io/STAT689-Project/). Best practice is to open the app in browser. Some parts of the app like the map and bar chart race are slower to load, so allow some time for them to load in. If they still don't appear html files of these visualizations can be found here https://drive.google.com/file/d/1wgHxxF2yTMw8LPsy24gXHz5EbZ_qAlLP/view?usp=sharing (https://drive.google.com/file/d/1wgHxxF2yTMw8LPsy24gXHz5EbZ_qAlLP/view?usp=sharing). Downloading and running these html files will allow the ability to see the visualizations without much buffering/loading.

Visualization Design & Implementation

In our project, the Shiny framework ("ShinyR" 2021) is used for creating a set of 5 visualizations as a web application. Shiny lets us create a reactive application with visualization strategies to choose the processed data displayed as per the project vision. Our development cycle involved creating the basic app with our visualization goals, making changes, and experimenting with the results obtained. User interface components provided by the framework are customized in the back-end as per our methodology to visualize the data to answer the research questions that originated during the proposal.

Our app consists of five menu items in the dashboard, which let the user choose the desired visualization to appear on the right side of the application. Two of the five items have UI reactivity to let the user play around to compare and contrast the filtered data plots. On the back-end server, the reactive code is extracted out of the app UI enters the server to perform specific visualization where each item is coded to display the output back on the UI. Our front-end is a ShinyDashboard that allows the user to choose a particular visualization with support for interactive functionality. We have used the following Visualization techniques in our project:

Line Graph - Historical Yield Improvements Here, we visualize how elite hybrids' average sorghum yields have changed over time and what periods saw the most significant increases in yield from the TXAR data. The ggplot2 library (Wickham 2016) was used to implement this visualization technique.

Scatterplot - Traits Relationship An interactive scatterplot window that plots and displays Pearson's R for whichever two variables(yield, plant height, exsertion, days to flowering) the user selects from the drop-down menu is implemented through the ggplot2 library using the TXAR dataset.

Choropleth Map - County-wise yield map
We construct a Choropleth map using Plotly (Sievert 2020) to see the average yields per county from the USDA data. A year slider bar from 1970-2020 selects the desired year and loads the corresponding Choropleth map.

Racing Bar Graph - Company Brands over the years
The dataset is representative of 180 company brands that have produced the ~4000 hybrids in the trials. This animated gif presents which brands were most prolific in releasing these hybrids over the years by allowing us to see the accumulation of hybrids from each brand from 1970 to 2020. This graph is implemented using ddplot library.

LMM Analysis - Statistical Analysis
We implemented a linear mixed model analysis using ggplot2 for output yield, where we attempt to understand how different factors affect the overall yield. The model is plotted as a scatterplot to visualize the effects and interactions. These results will also allow for predictions of future yield trials.

Methodology

The generalized additive model's smoothed line graph of historical yields is a useful tool to see the peaks and valleys of sorghum yields throughout the years. From this plot, we can then retrospectively focus on a certain time period and reflect on what was happening in the industry at that time that might've caused such an increase or decrease in yields.

A biological system like a sorghum plant or a field of sorghum contains many dependencies and in some sense is a zero-sum system. The correlation plots on the "Trait Relationships" tabs between metrics of hybrid performance indicate the tradeoffs within this crop system. Strong associations can be exploited once visualized with a major example being the association between taller plant height and higher grain yield.

The choropleth map is a strong visual for historic progress as it allows for the reader to witness where the highest yielding sorghum production zones were in the state throughout history. Additionally this visual allows us to see the clear decline in the number of Texas counties that grew sorghum beginning in the early 2000s. This loss of sorghum acreage is very clearly reflected with the abscence of more and more counties as the map's slider progresses.

The grain sorghum industry, like the rest of the agricultural industry, has seen consolidation in companies working to produce new hybrid seeds. The racing bar chart in the "Company Brands" tab allows a visual representation of which company brands were most profilic in submitting new hybrid products to these trials throughout history. These bars reflect private seed companies and also public institutions that developed hybrids like Texas A&M or USDA agencies.

The final tab of the data visualization dashboard addresses the environmental variables and cultural practices affecting grain yield. Individual effects are plotted as follows: Yield ~ Year + Effect, with the effect being visualized in the plot with multiple lines.

Evaluation Plan

For evaluation, we compare the initial proposal with the final product of the application produced. Each question from the proposal is addressed and checked to determine whether the final product satisfies our project goals. The evaluation for each visualization technique goes as follows:

Line Graph
Goal: Make a line graph with range slider to see how sorghum yields(GY) changed over time and what time periods saw the greatest increases in yield. Users should be able to interact with data and choose date ranges to examine the data closer.
Evaluation: Overall we were satisfied with the outcome of this attribute of the project. The user interaction objective of this visualization was achieved. Furthermore we also made the line graph display further information on a hover action.

Scatter Plot
Goal: Make a drop down menu of variables that allows the user to choose variables and see scatterplot of the data in a window. Also wanted a pearson R for the relationship between the two variables chosen.
Evaluation: Our final product did have drop down menu which included all the variables. One menu for the x axis and and one for the y axis. So far we have not implemented a way to display the pearson R value for a given plot as we had wanted to, instead we opted for a line of best fit.

Choropleth Map
Goal: Make a chloropleth map of average yields per texas county. Upon hovering we wanted users to see the county name and its average yield.
Evaluation: We were satisfied with our final map. Although it is a map, it also can serve as an animation, by pressing the play button users can see how average yields across Texas counties from 1970-2021. We also accomplished the hover information goal, and added distinct colors to distinguish different yield levels.

Racing Bar
Goal: Make a racing bar graph that visualizes the accumulation of hybrids from each brand over the years.
Evaluation: While we did successfully implement this part, we had trouble imbedding the animation into our shiny app. We worked around this by embedding a video of our animation in action instead.

Linear Model Analysis
Goal: The expressed aim of this analysis is to visualize the contribution of certain factors to the response variable: grain yield.
Evaluation: The individual effect plots are easy to understand because they are reflective of only one variable at a time. A more expansive analysis with a model with more terms would only prove harder to visualize and become less meaningul for anyone who is not a subject-matter expert. The limitation of this approach is that further factors may not have linear effects and would require a more flexible model.

Discussions & Future Work

The final data visualization dashboard encompasses a large breadth of agronomic and historical insights. However, beyond the agronomic practices, further work should be conducted to narrow in on the environmental and genetic components of these trials. This area of study is exceedingly complicated by cross-over interactions between the agronomic practices, environmental conditions, and the plant’s genetics. Typically, these interactions

are very specific to a production practice or geographic area. These type of research questions are best addressed with targeted experimentation and not well suited for investigation via a large retrospective like we present.

Delivering these types of trends into the hands of farmers and decision makers is the ultimate goal of these visualizations. Farmers adopting practices is the best best case scenario, however, they are notoriously difficult stakeholders to influence. For example, despite a wealth of previous knowledge and many of our visualizations suggesting that taller plants achieve higher yields, there is still a preference amongst farmers for shorter sorghum hybrids. There exists a misconception that the tallest plants will blow over and lodge despite our data suggesting that only becomes a concern at plant heights taller than ~160cm. As it stands, the average sorghum hybrid height submitted to these trials is ~125cm. There still exists lots of room for greater plant heights in the Texas grain sorghum industry.

The nature of these data is that they are annually updated with the latest year's production. This work will need to be updated annually to incorporate the most current information into these trends.

Team Member Contributions

Ram Pangaluri:
Contributed to design,planning, and writing in P1-5. Helped brainstorm and finalize project story points. Produced preliminary plots for story points and P6 report, including racing bar animation, chloropleth map, and scatterplot variable relationship interaction. Also created intermediate datasets to implement the previous visualizations mentioned. In P7,contributed to writing of the report namely reference, usage, and evaluation plan sections(except LMA part). Also implemented workarounds for animation embedding in the shiny app. Contributed R app files and data to team github,and very active in team discord.

Fabian Leon:
Design, planning, data collection, writing, and revising (P1-7). Brainstorming and development of project story points as the team member with the subject matter expertise for these data. Created the home page, historical yield improvements, and yield factors tabs in the shiny app. Revised all plots and integrated all code into the shiny app, with the exception of the racing bars animation, with the final dataset. For P7, contributed Introduction, Motivation, Methodology, Discussion, and References.

Ashok Shanker:
Contributed to Design, planning, writing and revising (P1-P7). Helped with Brainstorming and Story points in alignment with project goals. Helped create intermediate datasets to implement visualizations used. Created and integrated the Choropleth and Racebar chart tabs into Shiny app. Contributed to the R app files on team github. For P7, contributed to the Visualization Design and Implementation and Evaluation Plan.

References

Balakrishna, D., R. Vinodh, P. Madhu, S. Avinash, P. V. Rajappa, and B. Venkatesh Bhat. 2019. "Chapter 7 - Tissue Culture and Genetic Transformation in Sorghum Bicolor." Book Section. In *Breeding Sorghum for Diverse End Uses*, edited by C. Aruna, K. B. R. S. Visarada, B. Venkatesh Bhat, and Vilas A. Tonapi, 115–30. Woodhead Publishing. https://doi.org/https://doi.org/10.1016/B978-0-08-101879-8.00007-3 (https://doi.org/https://doi.org/10.1016/B978-0-08-101879-8.00007-3).

"ShinyR." 2021. Generic. https://CRAN.R-project.org/package=shiny (https://CRAN.R-project.org/package=shiny).

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Book. Chapman and Hall/CRC. https://plotly-r.com (https://plotly-r.com).

“Sorghum 101.” 2022. Web Page. https://www.sorghumcheckoff.com/sorghum-101/ (https://www.sorghumcheckoff.com/sorghum-101/).

“Texas Agrilife Crop Testing.” 2022. Web Page. http://varietytesting.tamu.edu/grainsorghum/ (http://varietytesting.tamu.edu/grainsorghum/).

“USDA Nass Quickstats.” 2021. Online Database. https://quickstats.nass.usda.gov/ (https://quickstats.nass.usda.gov/).

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Book. Springer-Verlag New York. https://ggplot2.tidyverse.org (https://ggplot2.tidyverse.org).