P2: Data Acquisition And Cleaning

Ram Pangaluri, Fabian Leon, Ashok Shanker

February 13, 2022

Data Sources & Extraction

Texas A&M AgriLife Research (TXAR)

Texas A&M AgriLife Research conducts the grain sorghum performance tests each year to provide growers in Texas with accurate and unbiased information on hybrid performance at locations across the state. For producers, selection of superior hybrids that are well adapted for a given region is essential for maximizing yield and profit. The TAES dataset contains data from multi-environment trials of commercially released sorghum hybrids grown from 1970-2021 across different counties/cities around Texas.

Characteristics/Attributes: This dataset is extensive and has 50+ columns describing hybrid sorghum performance and the management practices used to grow each trial. Some important variables include:

- Year - Year survey data point was obtained
- County - Texas County Name
- Hybrid - Specific sorghum genotype
- Irrigation - Irrigation level/amount
- Brand - Company brand of sorghum plant/seed

Response variables that we are interested in visualizing include many agronomically relevant traits included in the dataset:

- Grain Yield
- Plant height
- Days to Flowering
- Testweight

These data were extracted from both TXAR publications by the TXAR sorghum breeding program and more recently, provided directly by the variety testing program from their non-public database.

United States Department of Agriculture (USDA)

The USDA's National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year and prepares reports covering virtually every aspect of U.S. agriculture. Production of grain sorghum, prices paid and received by farmers, farm labor and wages, farm finances, chemical use, and changes in the demographics of U.S. producers are available and can downloaded as a csv tables. The main variable that we are interested in visualizing with this dataset is grain yield with years, locations, and agronomic practices also contributing to this metric.

Transformation

Texas A&M AgriLife Research (TXAR)

The TXAR dataset includes fields where notes were taken in paragraph form and do not numerically quantify much useful information. These need to be excluded from the final dataset we'll use. Similarly, we must delete certain columns that have too much missing data to be useful. These data columns include: Avg % Bird Damage, Avg % Midge Damage, Avg Iron Chlorosis, Avg Weathering, Avg Smut, Avg Stand, Avg Plant Population, Avg HeadsPerAcre, Sulfur (lb/ac), Zinc (lb/ac). Additionally, the variables in all remaining columns must be converted to units in the metric system for consistency in analysis and visualization. Finally, the dataset includes some observations from Curry County, New Mexico and since we are only interested in Texas, we wil exclude these data.

United States Department of Agriculture (USDA)

The csv file of the USDA NASS dataset also comes with unnecessary variables that contain some organizational or metadata that aid the USDA programs but mean very little to our analysis. These variables include: Week Ending, State ANSI, Geo Level, and Domain Category. Other variables are entirely empty despite being included in the dataset so these must also be excluded: Zip Code, Region, watershed code, Watershed.

Final Data

Texas A&M AgriLife Research (30,330 observations of 30 variables)

Year	Location	County	AgriLife-Region	Region	Irrigation	Hybrid	Brand	Maturity	DA	PH	EX	MST	GY	LDG	plant-date	plot-length	rainfall
1970	Bushland	Potter	Northern High Plains	1N	NO	429	P-A-G	M	63	76.2	2.54	13.7	1.6	0	15-Jun	30	7.87
1970	Fannin County	Fannin	Blacklands	4	NO	429	P-A-G	M	83	111.76	17.78	12.3	5.3	.	3-Apr	30	11.59
1970	Hillsboro	Hill	Blacklands	4	NO	429	P-A-G	M	78	109.22	10.16	15.3	4	12	27-Mar	30	8.94
1970	Lubbock	Lubbock	Southern High Plains	1S	NO	429	P-A-G	M	62	78.74	5.08	.	1.3	.	21-May	30	8.98
1970	Temple	Bell	Blacklands	4	NO	429	P-A-G	M	81	111.76	17.78	12.2	4	.	25-Mar	30	11.93
1970	Bushland	Potter	Northern High Plains	1N	NO	463	P-A-G	M	64	66.04	.	13.3	1.2	0	15-Jun	30	7.87
1970	Fannin County	Fannin	Blacklands	4	NO	463	P-A-G	ML	84	111.76	15.24	15.7	5.3	.	3-Apr	30	11.59
1970	Gregory	San Patricio	Coastal Bend	8S	NO	463	P-A-G	M	76	119.38	33.02	14.3	3.5	.	20-Mar	30	15
1970	Hillsboro	Hill	Blacklands	4	NO	463	P-A-G	E	76	106.68	10.16	13.2	3.6	6	27-Mar	30	8.94
1970	Lubbock	Lubbock	Southern High Plains	1S	NO	463	P-A-G	ML	67	76.2	5.08	.	1.6	0	21-May	30	8.98

United States Department of Agriculture (7,957 observations of 11 variables)

Year	State	Ag District	Ag District Code	County	County ANSI	Commodity	Data Item	Domain	Value	CV (%)
2020	TEXAS	BLACKLANDS	40	BOSQUE	35	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	36.2	16.2
2020	TEXAS	BLACKLANDS	40	COOKE	97	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	65.5	7.1
2020	TEXAS	BLACKLANDS	40	CORYELL	99	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	64.7	11.2
2020	TEXAS	BLACKLANDS	40	DENTON	121	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	48.9	18.5
2020	TEXAS	BLACKLANDS	40	GRAYSON	181	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	70.6	6.7
2020	TEXAS	BLACKLANDS	40	HILL	217	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	94.7	9.5
2020	TEXAS	BLACKLANDS	40	HUNT	231	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	69.8	20.7
2020	TEXAS	BLACKLANDS	40	JOHNSON	251	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	89.1	11.4
2020	TEXAS	BLACKLANDS	40	MCLENNAN	309	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	66.7	8.5
2020	TEXAS	BLACKLANDS	40	MILAM	331	SORGHUM	SORGHUM, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	88.2	4.9

Both datasets will remain as csv files for easy loading into the R environment.