

----- GENERAL INFORMATION -----

DATA TITLE: University extension service hybrid maize yield trials, 1934-2014

DATA ABSTRACT: The original publications of hybrid maize yield trials in the US for 1934-2014 were collected for Illinois, Iowa, Kansas, and Nebraska, digitized, and tabulated. The dataset includes brand and hybrid names, trial location, and yield (bushels per acre) for all trials and years. Additional data on agronomic phenotypes, soil type, and average weather is included when reported in the source publications.

AUTHORS:

Author: Aaron Kusmec

ORCID: 0000-0003-2295-385X

Institution: Iowa State University

Email: amkusmec@iastate.edu

Author: Lakshmi Attigala

ORCID: 0000-0002-7124-5617

Institution: Iowa State University

Email: lakshmi@iastate.edu

Author: Srikant Srinivasan

ORCID: 0000-0001-8605-3952

Institution: Plaksha University

Email: srikant.srinivasan@plaksha.edu.in

Author: Cheng-Ting "Eddy" Yeh

ORCID: 0000-0002-1392-2018

Institution: Iowa State University

Email: eddyeh@iastate.edu

Author: Patrick S. Schnable

ORCID: 0000-0001-9169-5204

Institution: Iowa State University

Email: schnable@iastate.edu

Corresponding author: Patrick S. Schnable

----- FILE DIRECTORY -----

All file names use the four-digit year (YYYY) format and date ranges are formatted with a dash between years.

- DataShare_Codebook.csv
 - Contains descriptions of the common variables found across yield reports for all four states.
 - Note that there are many columns that are occasional (e.g., damage caused by a specific pest) or inconsistently reported and named (e.g., lodging). These are not defined, and users should consult the original publications for more information.
 - Missing Data Codes: “-”, “*”, “NA”, “N/A”, “---”, “”, “M”
- **Curated Yield Data - CSV [folder]**
 - Additional Metadata [folder]
 - Tabular data files containing location information for each trial. File names indicate the state in which the trials were conducted and correlate to the data in the folder with the same name.
 - Illinois [folder]
 - File names are formatted as “Illinois_Corn_Trials_YYYY.csv”.
 - Note: 1962 through 1969 are combined into one file due to changes in the reporting structure.
 - Additional notes about the data, such as “no brand” are appended to the end of a few file names.
 - Iowa [folder]
 - File names are formatted as “Iowa_Corn_Trials_YYYY.csv”
 - “Iowa_Corn_Trials_Rain_YYYY.csv” contains additional information on precipitation and temperature when available.
 - “Iowa_Corn_Trials_Soil_YYYY.csv” contains additional information on soil type when available.
 - Kansas [folder]
 - File names are formatted as “Kansas_Corn_Trials_YYYY.csv”
 - “Kansas_Corn_Trials_Rain_YYYY.csv” contains additional information on precipitation and temperature when available.
 - “Kansas_Corn_Trials_Soil_YYYY.csv” contains additional information on soil type when available.
 - “Kansas_Corn_Trials_Rain_Soil_YYYY.csv” contains additional information on precipitation, temperature, and soil type when available.
 - Nebraska [folder]
 - File names are formatted as “Nebraska_Corn_Trials_YYYY.csv”
- **Curated Yield Data - Excel [folder]**
 - Contains original Excel formatted data files.
 - Note: Has the same folder structure as the CSV files but the number of years in each file varies.

- **Historical Input Data [folder]**

- These files contain the curated input data. They serve as input to scripts beginning in “03.model temp” of the associated GitHub repository (see below).
- “all_trials_distributions.csv” contains season total precipitation and exposure time to different temperatures for all curated trials.
- “DataShare Codebook pt2.xlsx” contains descriptions of the variables in each of the curated input files.
- “munged_trials_all.csv” contains the curated yield records from all trials.
- “trials_unique.csv” contains curated trial data including dates and locations (administrative units and latitude/longitude). If trial results from multiple locations were reported as a single, combined value (e.g., yield in “District 1”), dates and locations for all individual trials are separated by semicolons (“;”).

----- METHODS AND MATERIALS -----

----- DATA COLLECTION METHODS -----

The data were collected from print and online publications produced by extension services at the University of Illinois at Urbana-Champaign, Iowa State University, Kansas State University, and the University of Nebraska-Lincoln from the earliest available records through 2014. PDFs of online publications and scans of print publications from Illinois, Iowa, and Kansas were converted to text files using ABBYY Fine Reader v12.1.3 optical character recognition (OCR) software and then entered into Microsoft Excel spreadsheets.

Publications from Nebraska were converted with the assistance of the Amazon Mechanical Turk (AMT) service because the scans were too low quality for OCR. In the first campaign, AMT contractors entered data from the scans into Microsoft Excel spreadsheets manually. In a second campaign, contractors were asked to check the accuracy of data entry against the source PDFs. As a quality control measure, 10% of the entries in a spreadsheet were randomly selected and subtly changed. A spreadsheet was considered checked if at least 70% of the introduced errors were identified.

----- DATA PROCESSING METHODS -----

All checked spreadsheets were manually curated for consistent formatting, spelling, and capitalization of brand (seed company) and hybrid names. Further processing and standardization of hybrid and location names and geo-tagging information was performed on the original Excel files using custom R scripts, which can be accessed at https://github.com/amkusmec/heat_selection/tree/main/01.munge_yield.

----- DATA CURATION -----

The ISU Library exported the Excel files to CSV format using Excel Archival Tool (<https://github.com/mcgrory/ExcelArchivalTool>), standardized file names, corrected minor typos in the csv files, and updated the readme.

----- SOFTWARE -----

Name: ABBYY Fine Reader

Version: 12.1.3

URL: <https://pdf.abbyy.com/>

Developer: ABBYY

Name: R

Version: 4.2.2

URL: <https://cran.r-project.org/>

Developer: R Core Team

Additional notes: R packages required for data processing include “tidyverse”, “readxl”, and “RJSONIO”.

----- LICENSING -----

This work is licensed under the Creative Commons Attribution (CC-BY) 4.0 International License. For more information visit: <https://creativecommons.org/licenses/by/4.0>