From PDF Files to Biological Insights into Soybean Breeding: An Example of How Recovered Historical Yield Data Can be Valuable

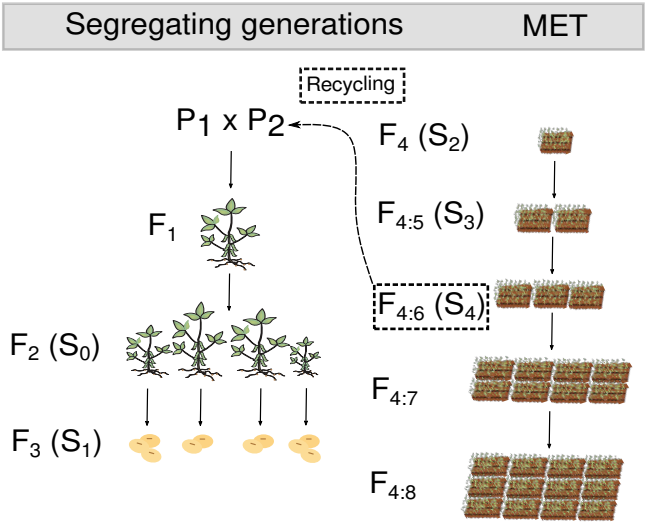
Matheus D. Krause, Kaio O. G. Dias, Asheesh K. Singh, and William D. Beavis

Preprint: https://doi.org/10.1101/2022.04.11.487885

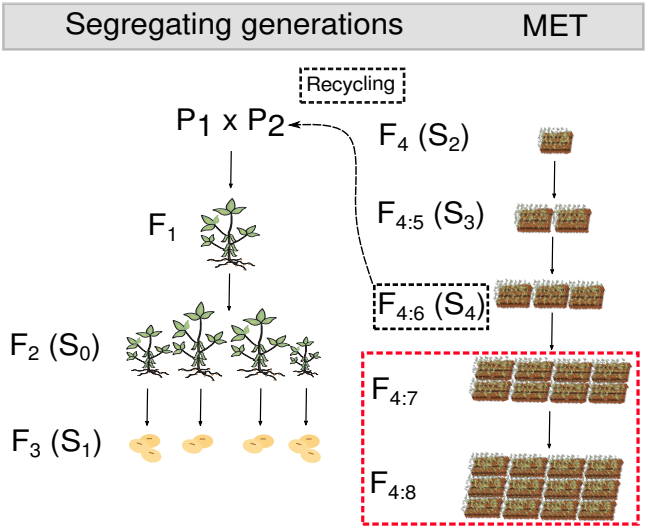
January 15, 2023

Outline

Historical soybean yield data - Backbone of a breeding program



Historical soybean yield data - Backbone of a breeding program



Preliminary and Uniform Regional Trials

Credit to Matthew Carroll (SinghSoybean Group@ISU)

The starting - PDF files from USDA-ARS (1941 - 2022)

UNITED STATES DEPARTMENT OF AGRICULTURE

BUREAU OF PLANT INDUSTRY
cooperating with
BUREAU OF AGRICULTURAL CHEMISTRY AND ENGINEERING
and the
STATE AGRICULTURAL EXPERIMENT STATIONS
OF THE NORTH CENTRAL REGION.

RESULTS OF THE COOPERATIVE UNIFORM
SOYBEAN NURSERIES

1941

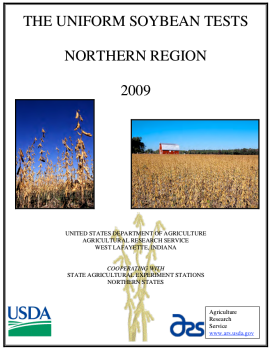
****

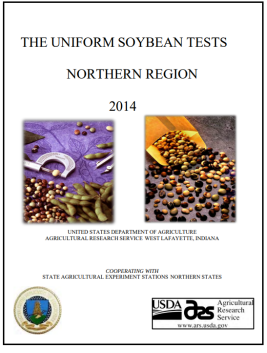
(Not for Publication)

UNITED STATES DEPARTMENT OF AGRICULTURE
AGRICULTURAL RESEARCH ADMINISTRATION

BUREAU OF PLANT INDUSTRY,
SOILS, AND AGRICULTURAL ENGINEERING
cooperating with the
STATE AGRICULTURAL EXPERIMENT STATIONS
of the
NORTH CENTRAL REGION

RESULTS OF THE COOPERATIVE UNIFORM
SOYBEAN TESTS
****
1942
****

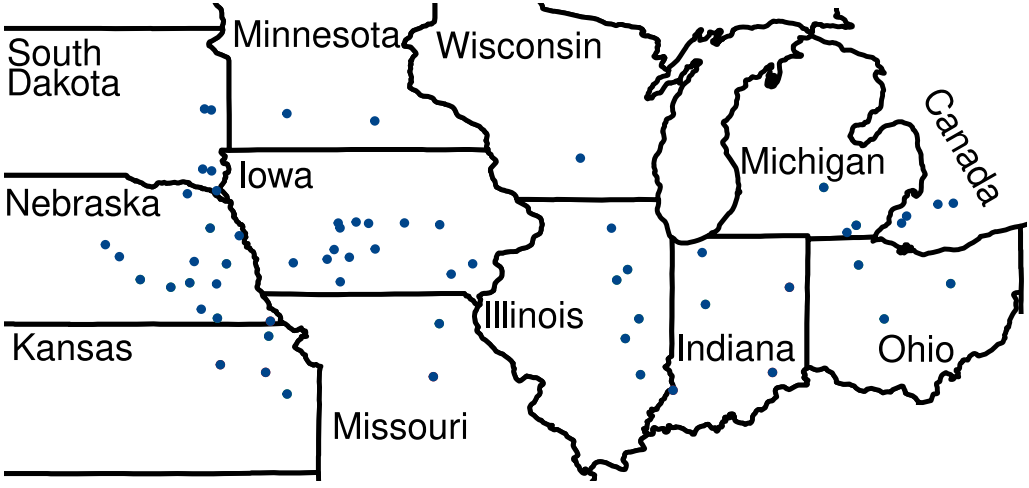




An eye inside the PDF files (West Lafayette, IN)

Strain	Ottawa KS	Albany MO	Novelty MO	Portageville Clay MO	Portageville Loam MO	So Charles-ton OH	Jack-son TN
LD06-7620 (IV)	77.1	79.5	81.6	54.4	73.1	62.0	49.6
LD00-2817P (L)	73.1	81.9	79.6	53.1	73.6	53.8	60.2
LD07-3395bf (SCN)	72.0	79.3	80.7	53.1	77.3	60.2	48.4
LD12-10534	75.3	80.4	65.2	65.2	74.7	58.4	48.8
LG10-3278	62.8	75.1	58.8	66.5	76.3	57.8	55.9
LG11-6759	65.5	72.8	78.8	58.4	75.8	64.9	55.6
LG11-6760	64.4	74.2	76.4	57.0	78.1	61.9	49.1
LG11-6761	65.0	77.7	73.2	60.9	72.8	53.6	47.6
LG13-3925	68.6	80.3	40.1	62.8	81.0	54.6	54.6
LG13-3981	68.5	80.0	68.7	63.6	78.0	57.3	52.7
LG13-3993	70.3	69.9	72.0	64.3	78.1	65.1	52.0
SA10-8471	71.9	72.2	38.2	62.1	77.5	56.6	56.5
SA12-1451	73.3	90.7	72.7	58.2	77.0	74.8	57.9
SA12-1471	66.9	76.9	71.5	59.4	71.0	59.8	54.3
Location Mean	69.6	77.9	68.4	59.9	76.0	60.0	53.1
C.V. (%)	4.4	8.2	12.1	5.6	5.5	9.7	11.0
L.S.D. (5%)	5.1	10.7	14.0	6.8	8.5	11.9	9.8
Row Sp. (In.)	30	30	30	30	30	15	30
Rows/Plot	4	4	4	4	4	6	4
Reps	3	3	3	2	2	3	3

An eye on the map

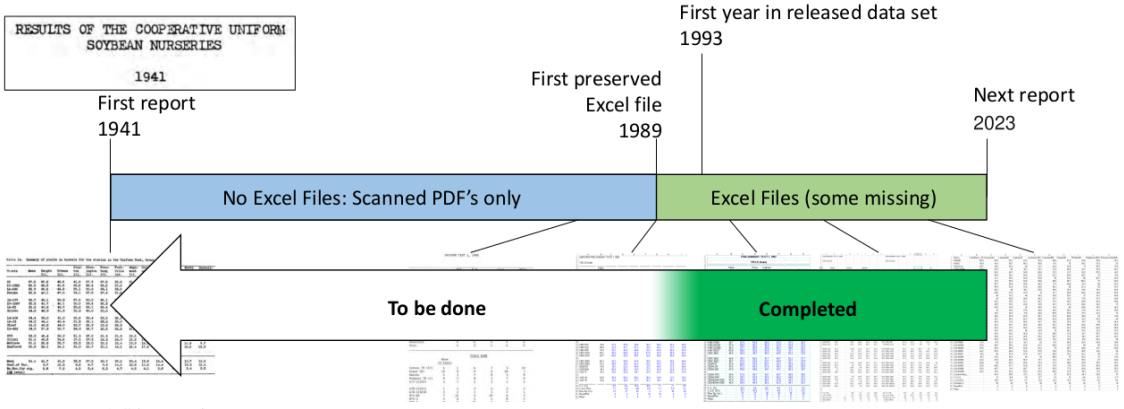


Step-by-step process

1. The data is already there: free; > 80 years of trial information
2. What for? Inputs on public soybean breeding
3. Converting PDF / spreadsheet (when available) files to standard / usable format
4. Curation and cleaning 
5. Biological insights
6. Making the datasets available

Converting PDF and spreadsheet files to standard format

Aaron Lorenz's group: Ben Campbell, Liana Nice, Ana Poets, Peng Zhou, and Cleiton Wartha



From Ben Campbell (PAG 2019)

Converting PDF and spreadsheet files to standard format

Liana Nice

- A marathon of coding ($\sim 10k$):
- R base functions: `paste`, `list.files`, `grep`, `c()`, `strsplit`, `name`, `length`, `gsub` `if(){}else{}`, `for(i in x)`, `lapply`, `sapply`
- Made difficult by non-standard formats across years. Manual?
- Spreadsheet not available: Adobe Acrobat OCR (ZAMZAR)

  "With optical character recognition (OCR) in Adobe Acrobat, you can extract text and convert scanned documents into editable, searchable PDF files instantly. "

- **Main message: there is not an standard way to get this done!**

Converting PDF and spreadsheet files to standard format

Liana Nice

- A marathon of coding ($\sim 10k$):
- R base functions: `paste`, `list.files`, `grep`, `c()`, `strsplit`, `name`, `length`, `gsub` `if(){}else{}`, `for(i in x)`, `lapply`, `sapply`
- Made difficult by non-standard formats across years. Manual?
- Spreadsheet not available: Adobe Acrobat OCR (ZAMZAR)

  "With optical character recognition (OCR) in Adobe Acrobat, you can extract text and convert scanned documents into editable, searchable PDF files instantly. "

- **Main message: there is not an standard way to get this done!**

Converting PDF and spreadsheet files to standard format

Liana Nice

- A marathon of coding ($\sim 10k$):
- R base functions: `paste`, `list.files`, `grep`, `c()`, `strsplit`, `name`, `length`, `gsub` `if(){}else{}`, `for(i in x)`, `lapply`, `sapply`
- Made difficult by non-standard formats across years. Manual?
- Spreadsheet not available: Adobe Acrobat OCR (ZAMZAR)

  "With optical character recognition (OCR) in Adobe Acrobat, you can extract text and convert scanned documents into editable, searchable PDF files instantly. "

- **Main message: there is not an standard way to get this done!**

Converting PDF and spreadsheet files to standard format

Liana Nice

- A marathon of coding ($\sim 10k$):
- R base functions: `paste`, `list.files`, `grep`, `c()`, `strsplit`, `name`, `length`, `gsub` `if(){}else{}`, `for(i in x)`, `lapply`, `sapply`
- Made difficult by non-standard formats across years. Manual?
- Spreadsheet not available: Adobe Acrobat OCR (ZAMZAR)

  "With optical character recognition (OCR) in Adobe Acrobat, you can extract text and convert scanned documents into editable, searchable PDF files instantly. "

- **Main message: there is not an standard way to get this done!**

Converting PDF and spreadsheet files to standard format

Liana Nice

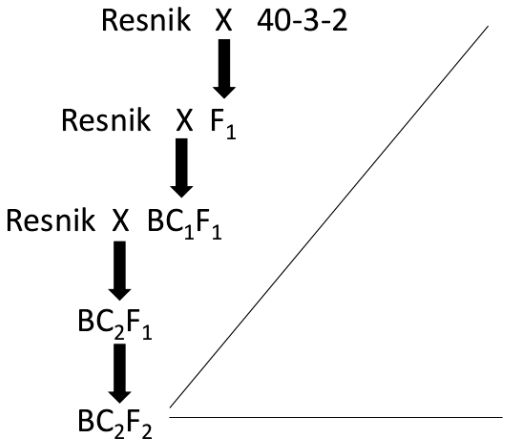
- A marathon of coding ($\sim 10k$):
- R base functions: `paste`, `list.files`, `grep`, `c()`, `strsplit`, `name`, `length`, `gsub` `if(){}else{}`, `for(i in x)`, `lapply`, `sapply`
- Made difficult by non-standard formats across years. Manual?
- Spreadsheet not available: Adobe Acrobat OCR (ZAMZAR)

  "With optical character recognition (OCR) in Adobe Acrobat, you can extract text and convert scanned documents into editable, searchable PDF files instantly. "

- **Main message: there is not an standard way to get this done!**

Curation - Typos, genotype and location names

Liana Nice, Ben Campbell, Ana Poets, Cleiton Wartha, and Matheus Krause



- (Resnik*2/40-3-2)
- Resnik BC2F2
- Resnick BC2F2
- Resnik2
- Resnik(RR)
- Resnik RR
- ResnikRR
- RR

From Ben Campbell (PAG 2019) - Example of "Breeders' creativity"

Curation - Similarity between two strings

```
library(RecordLinkage) # It is based on Levenshtein distance
names <- c("(Resnik*2/40-3-2)", "Resnik BC2F2", "Resnick BC2F2",
           "Resnik2","Resnik(RR)", "Resnik RR", "ResnikRR", "RR")
levenshteinSim("Resnik(RR)", names)
[1] 0.41 0.50 0.46 0.60 1.00 0.80 0.80 0.20
```

When crosschecking names between phenotypic and genotypic data:

M06-338016 | M06_338016 = 0.90
M09-242072 | M09_242072 = 0.90
OAC_13-64C-ChCdN | OAC_13_64C_ChCdn = 0.81
LD 02-4485 | LD02_4485 = 0.80

Curation - Similarity between two strings

```
library(RecordLinkage) # It is based on Levenshtein distance
names <- c("(Resnik*2/40-3-2)", "Resnik BC2F2", "Resnick BC2F2",
           "Resnik2","Resnik(RR)", "Resnik RR", "ResnikRR", "RR")
levenshteinSim("Resnik(RR)", names)
[1] 0.41 0.50 0.46 0.60 1.00 0.80 0.80 0.20
```

When crosschecking names between phenotypic and genotypic data:

```
M06-338016 | M06_338016 = 0.90
M09-242072 | M09_242072 = 0.90
OAC_13-64C-ChCdN | OAC_13_64C_ChCdn = 0.81
LD 02-4485 | LD02_4485 = 0.80
```

Curation - Typos, genotype and location names

Important lesson: standardize EVERYTHING

- Remove white/empty spaces: `gsub(" ", "", names)`
- Lower/upper cases: `gsub("-", "\_", tolower(names))`
- Remove special characters:

```
"a1~!@#$%^&*(){}_+:\"<>?,./;'[]-="
```

Similar procedure with location names

```
gsub("burkeyfarms_ia", "boone_ia", Location)
gsub("boonecounty_ia", "boone_ia", Location)
gsub("finchfarms_ia", "ames_ia", Location)
```

Curation - Typos, genotype and location names

Important lesson: standardize EVERYTHING

- Remove white/empty spaces: `gsub(" ", "", names)`
- Lower/upper cases: `gsub("-", "\_", tolower(names))`
- Remove special characters:

```
"a1~!@#$%^&*(){}_+:\"<>?,./;'[]-="
```

Similar procedure with location names

```
gsub("burkeyfarms_ia", "boone_ia", Location)
gsub("boonecounty_ia", "boone_ia", Location)
gsub("finchfarms_ia", "ames_ia", Location)
```

Curation and cleaning - Is it an endless process?

Source: Saraswathi Analytics

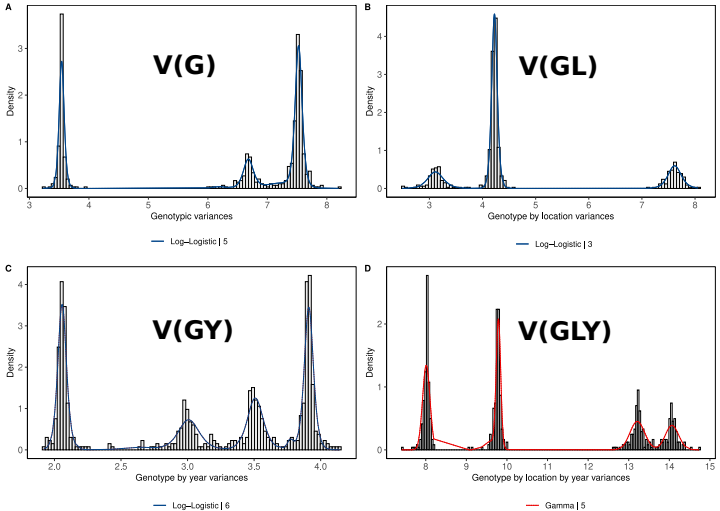
bioRxiv - under review in Field Crops Research

Using soybean historical field trial data to study genotype by environment variation and identify mega-environments with the integration of genetic and non-genetic factors

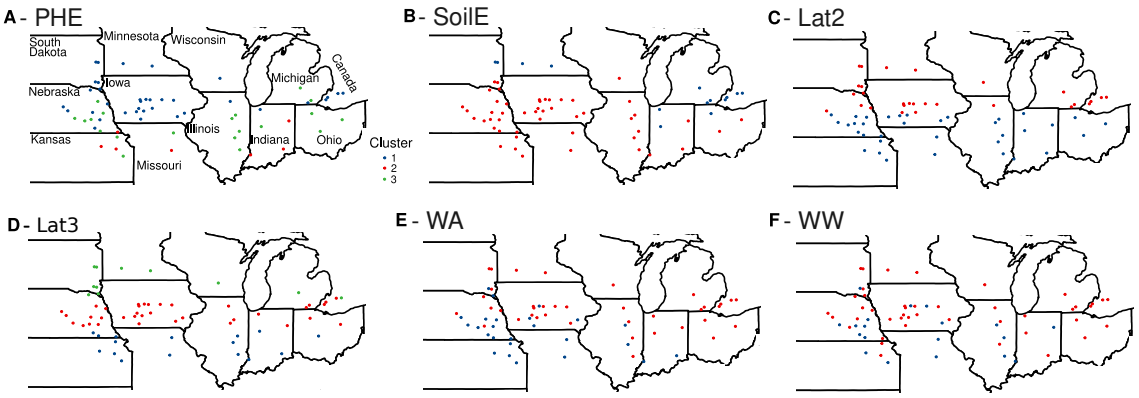
Matheus Dalsente Krause, Kaio Olimpio das Gracas Dias, Asheesh K Singh, William D Beavis

doi: https://doi.org/10.1101/2022.04.11.487885

Empirical density functions for variance components (simulation studies)



Mega-environments using phenotypes, soil, and weather data



SoyURT package - 1,008 downloads from CRAN! (01/14/2023)

SoyURT: USDA Northern Region Uniform Soybean Tests Dataset

Data sets used by 'Krause et al. (2022)' <doi:10.1101/2022.04.11.487885>. It comprises phenotypic records obtained from the USDA Northern Region Uniform Soybean Tests from 1989 to 2019 for maturity groups II and III. In addition, soil and weather variables are provided for the 591 observed environments (combination of locations and years).

Version:	1.0.0
Depends:	R (≥ 3.5.0)
Suggests:	spelling
Published:	2022-06-13
Author:	Matheus Dalsente Krause [aut, cre], William Dale Beavis [aut]
Maintainer:	Matheus Dalsente Krause <krause.d.matheus at gmail.com>
License:	CC BY 4.0
URL:	https://github.com/mdkrause/soyurt
NeedsCompilation:	no
Language:	en-US
Citation:	SoyURT citation info
Materials:	README
CRAN checks:	SoyURT results

Next steps

- Recover the data from 1941 to 1988. Is it useful?
- The R package `SoyURT` is available
- The raw data can be downloaded from Soybase
- Biological insights: our approach can be used in any crop with multiple-environment trials
- Multiple traits are available

Acknowledgements

- William D. Beavis
- Asheesh K. Singh
- Kaio O. G. Olímpio
- Aaron Lorenz
- Ben Campbell
- Liana Nice
- Cleiton Wartha
- Rex Nelson
- ISU Research IT

IOWA STATE UNIVERSITY
Department of Agronomy

NCSRP
NORTH CENTRAL SOYBEAN RESEARCH PROGRAM

NSF National Science Foundation WHERE DISCOVERIES BEGIN

SoyBase
Integrating Genetics and Genomics to Advance Soybean Research

Let's chat - I am open to new opportunities

Poster PE0310

Estimating the Realized Rate of Genetic Gain in Soybean Breeding Programs Using Routine Field Trials

about.me