Data management for A/B Teams

Jarad Niemi

Iowa State University

December 13, 2018

Outline

- Data horror stories
- Reasonable data management practices
- STRIPS specific information

Duke cancer clinical trials

`https://www.nature.com/articles/nm1107-1276b`

Issues with data management:

1. "The list of genes ... are wrong because of an 'off-by-one' indexing error."
2. "suggesting that most labels are reversed. If the labels are reversed, the model suggests administering the drug only to patients it would not benefit."

`https://bioinformatics.mdanderson.org/Supplements/ReproRsch-Chemo/`

Principles

My (current) guiding principles

- Be consistent
- Have ownership
- Store each piece of data only once
- Don't touch the raw data
- Leave as much of the work as possible to the computer

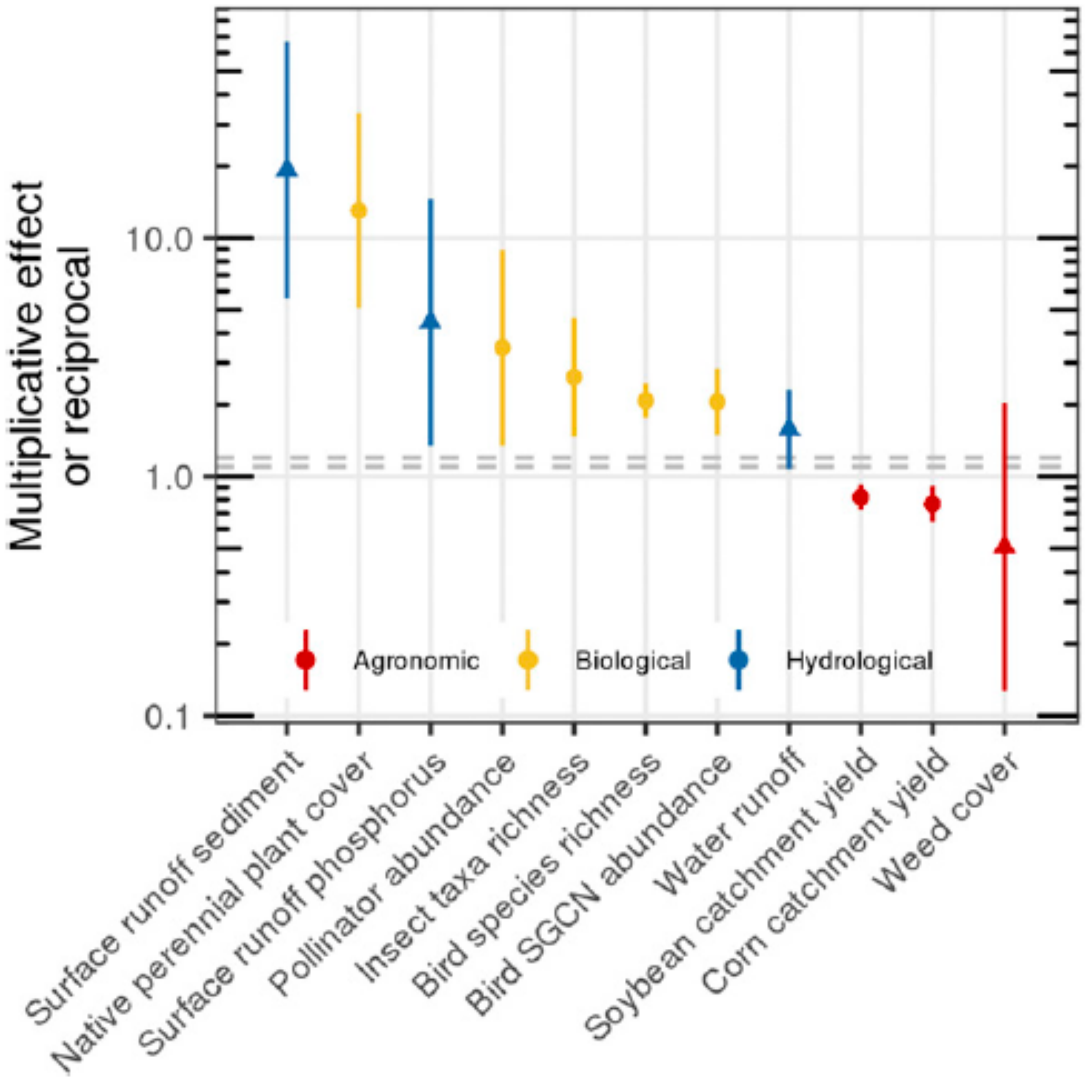
STRIPS Overview

Science-based Trials of Rowcrops Integrated with Prairie Strips (STRIPS) is the idea that we can create enormous benefits if small amount of rowcrop are converted to prairie.

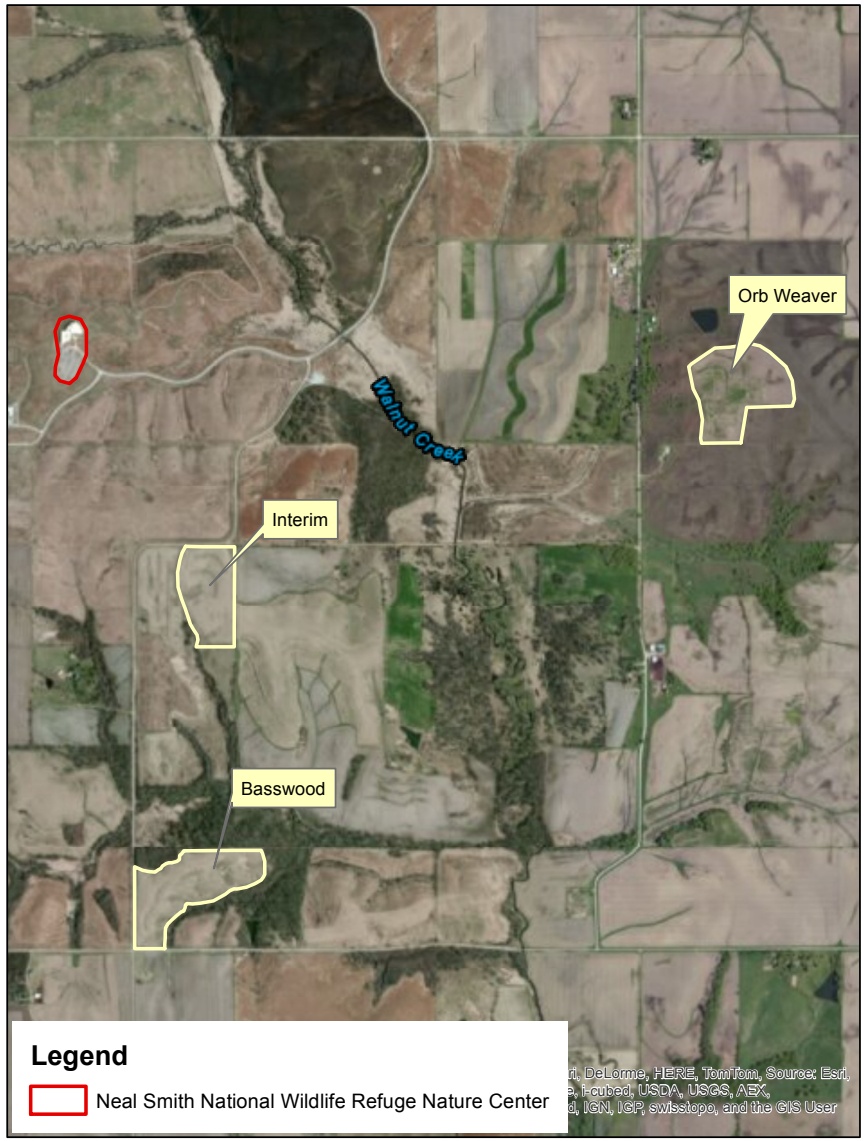
- STRIPS1: 4 treatments at Neal Smith National Wildlife Refuge
- STRIPS2: 2 treatments at commercial farms

`https://www.nrem.iastate.edu/research/STRIPS/`

STRIPS1 outcomes



Iowa State University - STRIPS Research Sites - Jasper County, IA



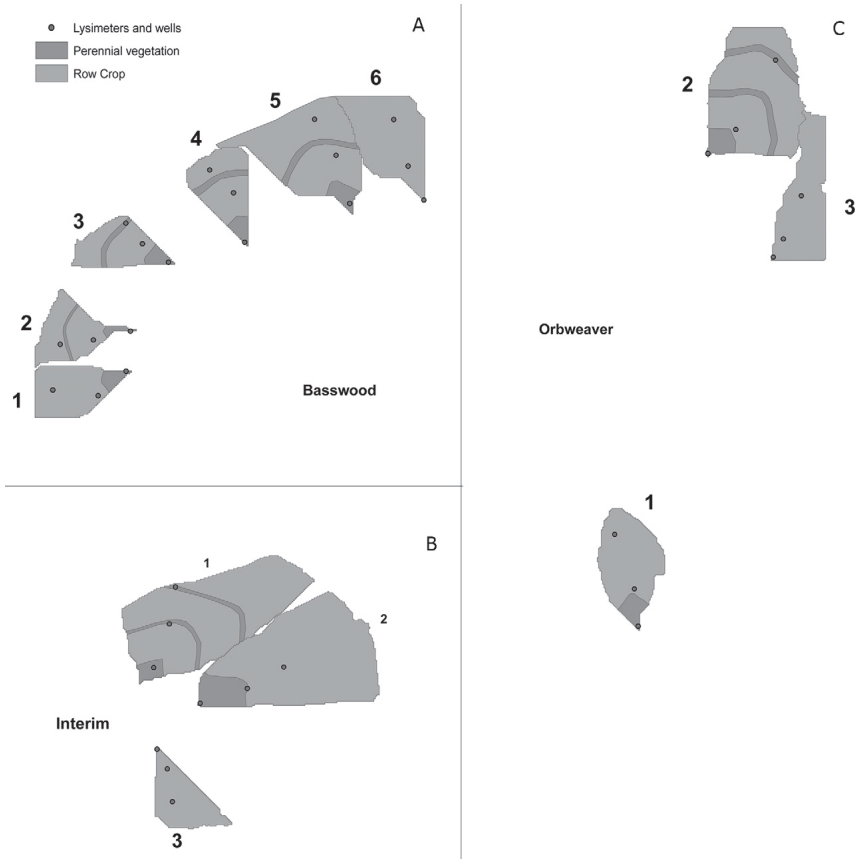


Fig. 2. Experimental design of vegetative filters for the study watersheds at (a) Basswood, (b) Interim, and (c) Orbweaver.

Table 1. Watershed description and experimental design.

Watershed	Size	Slope	Location and percentage of grass filters†
	ha	%	
Basswood-1	0.53	7.5	10% at toeslope
Basswood-2	0.48	6.6	5% at toeslope and 5% at upslope
Basswood-3	0.47	6.4	10% at toeslope and 10% upslope
Basswood-4	0.55	8.2	10% at toeslope and 10% upslope
Basswood-5	1.24	8.9	5% at toeslope and 5% upslope
Basswood-6	0.84	10.5	All rowcrops
Interim-1	3.00	7.7	3.3% at toeslope, 3.3% at sideslope, and 3.3% at upslope
Interim-2	3.19	6.1	10% at toeslope
Interim-3	0.73	9.3	All rowcrops
Orbweaver-1	1.18	10.3	10% at toeslope
Orbweaver-2	2.40	6.7	6.7% at toeslope, 6.7% at sideslope, and 6.7% at upslope
Orbweaver-3	1.24	6.6	All rowcrops

† Percentage of grass filters = area of filters/area of watershed

STRIPS yield

Where are the raw STRIPS1 yield data?

- Who is the owner of the STRIPS1 yield data?
- What are the files stored?
- What is the file organization?

This is my horror story.

STRIPS2 sites

Attempting to maintain farmer anonymity:

- We use 3-letter abbreviations:

```
 [1] "ARM" "BON" "BUE" "DMW" "EIA" "GES" "GOS" "GRE" "GUT"
[10] "HOE" "INH" "ISB" "JUD" "KAL" "MAR" "MCC" "MCN" "MQW"
[19] "MRS" "MUG" "NIR" "NYK" "POW" "RDM" "RHO" "ROD" "SER"
[28] "SLO" "SME" "SMF" "SMI" "SPI" "STN" "STT" "WAT" "WHI"
[37] "WOR"
```

- We keep spatial information on a private server and spatially anonymize reported data.

Archive data in STRIPS Box folder under STRIPS2...?