Datasets: Description and Curation Protocol

Luis Damiano, Jarad Niemi

2018-12-18

Contents

Accessing data

After the package is loaded, two objects `yield` and `yieldExtra` containing point-coordinate level yield data become available. To access the dataset from the PNAS paper [1], please run `pnas_data()`.

name	class	nFactors	min	mean	max	example	units	extra only
site	factor	3	NA	NA	NA	Basswood	NA	Yes
watershed	factor	13	NA	NA	NA	Basswood1	NA	Yes
block	factor	5	NA	NA	NA	BasswoodA	NA	Yes
blockArea	numeric	NA	NA	NA	NA	0.53	Hectare	No
treatment	factor	5	NA	NA	NA	10% prairie bottom	NA	Yes
prairiePercentage	factor	4	NA	NA	NA	10	% (hundreds)	No
prairiePosition	factor	3	NA	NA	NA	bottom	NA	No
slope	numeric	NA	NA	NA	NA	7.5	% (hundreds)	No
year	factor	8	NA	NA	NA	2007	NA	Yes
crop	factor	2	NA	NA	NA	Soybeans	NA	Yes
swath	numeric	NA	2.50	112.31	300.00	300	Unknown	No
record	numeric	NA	0.00	2571.40	9400.00	126	Integer	No
date	Date	NA	13795.00	15679.94	16706.00	2007-10-10	Date	No
timestamp	POSIXct	NA	NA	NA	NA	2007-10-10 07:20:56	POSIX	No
x	numeric	NA	-93.28	-93.26	-93.25	-93.276997	Unknown	Yes
y	numeric	NA	41.54	41.55	41.56	41.537513	Unknown	Yes
elevation	numeric	NA	826.40	881.64	921.90	903.2	Feet	No
speed	numeric	NA	0.02	3.55	6.26	3.3	MPH	No
direction	numeric	NA	NA	NA	NA	NA	Degrees	No
distance	numeric	NA	0.05	57.86	304.00	174	Unknown	No
flow	numeric	NA	0.03	13.24	45.49	9.69	Unknown	No
moisture	numeric	NA	4.40	13.80	25.70	11.9	% (hundreds)	Yes
yield	numeric	NA	4.94	84.49	399.69	57.6314	Unknown	Yes

Terminology

The dataset adheres to the terminology used in [2] to describe the experimental design.

- **Site**: three locations within the Neal Smith National Wildlife Refuge (NSNWR) in central Iowa (namely Basswood, Interim, and Orbweaver).
- **Blocks**: there are four blocks (namely BasswoodA, BasswoodB, Interim, and Orbweaver).
- **Watershed**: twelve experimental units (Basswood 1 to 6, Interim 1 to 3, Orbweaver 1 to 3).
- **Field**: the portion of the whatershed planted with either row crops or perennial vegetation.
- **Treatment**: four watershed-scale treatments having different proportions and topographic positions of PFS (no PFS, 10% PFS at toeslope position, 10% PFS distributed on toe and contour strips, and 20% PFS distributed on toe and contour strips).
- **Coordinate point**: each of the spatial coordinate units with recorded information (number and position of the points vary per year and treatment).

Curation protocol

New datasets

In STRIPSYield v0.2.0, the datasets are distributed in the folders: `data-raw\source\YYYY-site.ext`.

- `legacy`: CSV existing in STRIPSYield v0.1.1 that were produced by a methodology unknown to us.
- `original`: shapefiles as they were transmitted to us. Although we modified the name of these files for clarity, we kept the structure and the content.
- `curated`: new shapefiles originating from the curation protocol described below. Note that this process modifies both the structure and the content of the datasets.

The `original` datasets come from two main sources:

- **2007-2010**: `Research Components\Liebman Yield Data & Analysis\Neal Smith Yield Data & Analysis_Maier\GISdata\CropYield\Original Crop Yield Shapefiles`.
- **2011-2015**: STRIPYield v.0.1.1.

Curation protocol

Because not all the datasets have the same structure and measurement units, we create a curation protocol. We identify two patterns in the data sources, namely Template I (2007-2010 and 2012) and Template II (2013-2015 and 2011). We read the shapefiles from the `original` folder, apply the modifications mentioned below, and store the new shapefiles in the `curated` folder. These editing rules may be broadly classified into four actions:

- **Rename**: we modify the name of the variable but not the content.
- **Reformat**: we modify the name of the variable and the content (e.g. change of measurement unit).
- **Drop**: we discard some content if it is not present in every shapefiles across the years and sites.
- **TBD**: we still need more time until we figure it out.

Although keeping both the original and the curated shapefiles result in significant storage redundancy, this procedure guarantees that no original data is lost in the process.

Shapefiles 2007-2010 and 2012

Since the 2007-2010 and 2012 shapefiles are consistent, we display `2009-basswood` as an example.

name	class	nFactors	min	mean	max	example
ID	integer	NA	0.00	2045.00	4090.00	0
LONGITUDE	numeric	NA	-93.28	-93.27	-93.27	-93.277018
LATITUDE	numeric	NA	41.54	41.54	41.54	41.537171
FLOW	numeric	NA	0.15	11.39	20.79	0.8
TIME	integer	NA	1256130079.00	1256135700.36	1256141544.00	1256130079
CYCLES	integer	NA	1.00	1.80	3.00	1
DISTANCE	integer	NA	2.00	138.66	297.00	24
SWATH	integer	NA	287.00	287.00	287.00	287
MOISTURE	numeric	NA	7.60	15.00	16.70	13
STATUS	integer	NA	33.00	33.00	33.00	33
PASS	integer	NA	14.00	70.82	146.00	14
SERIAL	integer	NA	2007713498.00	2007713498.00	2007713498.00	2007713498
FIELD	factor	1	NA	NA	NA	"F0:BASSWOOD"
LOAD	factor	1	NA	NA	NA	"L0:09/10/21-10:58:13"
CROP	factor	1	NA	NA	NA	"SOYBEANS"
GPS	integer	NA	7.00	7.00	7.00	7
PDOP	integer	NA	0.00	0.00	0.00	0
ALTITUDE	numeric	NA	839.40	875.55	896.30	888.7
DRY_BU_AC	numeric	NA	5.01	50.97	150.00	11.8693
DAY	factor	1	NA	NA	NA	Wednesday
MONTH	factor	1	NA	NA	NA	October
DAYOFMONTH	integer	NA	21.00	21.00	21.00	21
HOUR	integer	NA	8.00	9.08	11.00	8
MINUTE	integer	NA	0.00	29.50	59.00	1
SECOND	integer	NA	0.00	29.39	59.00	19
TIMELAPSE	integer	NA	0.00	2.80	595.00	0
SPEED	numeric	NA	0.09	4.37	6.25	1.36

We apply the following formatting rules:

name	action
ID	Rename
LONGITUDE	Rename
LATITUDE	Rename
FLOW	TBD
TIME	Reformat
CYCLES	TBD
DISTANCE	Rename
SWATH	Drop
MOISTURE	Rename
STATUS	Drop
PASS	TBD
SERIAL	Reformat
FIELD	Rename
LOAD	Reformat
CROP	Rename
GPS	Drop
PDOP	Drop
ALTITUDE	Rename
DRY_BU_AC	Rename
DAY	Drop
MONTH	Drop
DAYOFMONTH	Drop
HOUR	Drop
MINUTE	Drop
SECOND	Drop
TIMELAPSE	Drop
SPEED	TBD

The PROJ4 string defining the CRS of the coordinates recorded in these shapesfiles is `"+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"`.

Shapefiles 2013-2015 and 2011

Since the 2013-2015 and 2011 shapefiles are consistent, we display `2015-basswood` as an example.

name	class	nFactors	min	mean	max	example
Field	factor	1	NA	NA	NA	BASSWOOD
Dataset	factor	1	NA	NA	NA	15/09/28-15:35:37 (2007713498)
Product	factor	1	NA	NA	NA	SOYBEANS
Obj___Id	numeric	NA	1.00	3390.50	6780.00	1
Distance_f	numeric	NA	0.18	5.59	9.12	1.8701
Track_deg_	numeric	NA	0.00	161.18	360.00	116.1
Duration_s	numeric	NA	1.00	1.00	1.00	1
Elevation_	numeric	NA	835.19	873.83	896.28	890.794
Time	Date	NA	16706.00	16706.00	16706.00	2015-09-28
Area_Count	factor	1	NA	NA	NA	On
Swth_Wdth_	numeric	NA	29.00	29.00	29.00	28.9993
Diff_Statu	factor	1	NA	NA	NA	Yes
Crop_Flw_M	numeric	NA	0.21	12.08	21.25	0.8157
Moisture___	numeric	NA	6.00	12.13	18.30	13
Yld_Mass_W	numeric	NA	303.25	3268.34	9925.10	655.1942
Yld_Vol_We	numeric	NA	5.05	54.47	165.42	10.9199
Yld_Mass_D	numeric	NA	303.25	3266.16	9925.10	655.1942
Yld_Vol_Dr	numeric	NA	5.05	54.44	165.42	10.9199
Work_State	factor	1	NA	NA	NA	In
Y_Offset_f	numeric	NA	0.00	0.45	1.00	0.0039
Sky_Cond	factor	1	NA	NA	NA	Unknown
Wind_Speed	numeric	NA	0.00	0.00	0.00	0
Wind_Dir	factor	1	NA	NA	NA	Unknown
Air_Temp___	numeric	NA	32.00	32.00	32.00	32
Humidity___	numeric	NA	255.00	255.00	255.00	255
Soil_Tex	factor	1	NA	NA	NA	Coarse Sand
Soil_Cond	factor	1	NA	NA	NA	Unknown
Soil_Moist	factor	1	NA	NA	NA	Unknown
Crop_Resid	factor	1	NA	NA	NA	Unknown
Nozzle_PN	factor	1	NA	NA	NA	NA
Pass_Num	numeric	NA	2.00	46.78	102.00	2
Speed_mph_	numeric	NA	0.12	3.81	6.22	1.2751
Prod_ac_h_	numeric	NA	0.42	13.40	21.85	4.4819
Crop_Flw_V	numeric	NA	12.37	724.81	1274.79	48.9418
Date	Date	NA	16706.00	16706.00	16706.00	2015-09-28

We apply the following formatting rules:

name	action
Field	Reformat
Dataset	Drop
Product	Reformat
Obj___Id	Rename
Distance_f	Rename
Track_deg_	Rename
Duration_s	Reformat
Elevation_	Rename
Time	Rename
Area_Count	Reformat
Swth_Wdth_	Rename
Diff_Statu	Reformat
Crop_Flw_M	Rename
Moisture___	Rename
Yld_Mass_W	TBD
Yld_Vol_We	TBD
Yld_Mass_D	TBD
Yld_Vol_Dr	TBD
Work_State	Reformat
Y_Offset_f	Reformat
Sky_Cond	Reformat
Wind_Speed	Reformat
Wind_Dir	Reformat
Air_Temp___	Drop
Humidity___	Drop
Soil_Tex	Drop
Soil_Cond	Drop
Soil_Moist	Drop
Crop_Resid	Drop
Nozzle_PN	Drop
Pass_Num	Rename
Speed_mph_	TBD
Prod_ac_h_	TBD
Crop_Flw_V	Reformat
Date	Reformat

The PROJ4 string defining the CRS of the coordinates recorded in these shapesfiles is `"+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"`.

Consolidated shapefile

As the final output does not vary across the years and sites, we display `2015-basswood` as an example.

name	class	units
site	factor	NA
crop	factor	NA
swath	numeric	Unknown
record	numeric	Integer
date	Date	Datetime
timestamp	logical	POSIXct
x	numeric	Unknown
y	numeric	Unknown
elevation	numeric	Unknown (feets?)
speed	numeric	MPH
direction	numeric	Degrees
distance	numeric	Unknown
flow	numeric	Unknown
moisture	numeric	% (hundreds)
yield	numeric	Unknown

To build our consolidated shapefiles, we decided to keep only those variables recorded for every site and year. The only exceptions are `timestamp` (only available for years 2007-2010 and 2012) and direction (only available for years 2013-2015 and 2011), which we kept as partial information may be relevant for our future research.

The PROJ4 string defining the CRS of the coordinates recorded in these shapesfiles is `"+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"` (no projections were needed).

Additional notes

- The shapefiles in the folders `Original Crop Yield Shapefiles\2011 Corn Yield` and `Original Crop Yield Shapefiles\2012 Corn Yield` have the same content as the STRIPYield v0.1.1 shapefiles (full path: `Research Components\Liebman Yield Data & Analysis\Neal Smith Yield Data & Analysis_Maier\GISdata\CropYield\Original Crop Yield Shapefiles\2012 Corn Yield`).
- The shapefiles in the folder `2012 Corn Yield` added by Matthew Helmers on 2018-12-13 have the same content as the STRIPYield v0.1.1 shapefiles (full path: `Research Components\Liebman Yield Data & Analysis\Neal Smith Yield Data & Analysis_Maier\GISdata\2012 Corn Yield`).

Naming convension

File naming convention:

- data-raw/yield_original/YYYY-site.ext
- data-raw/yield_curated/YYYY-site.ext
- Note that we use hyphen to separate words, and site names are lowercase.

Column naming convention:

- Use camelCase (e.g. prairiePosition). Note that the starting letter is lowercase.
- No measurement units in the column names. For measurement units, see this vignette.

Data structure convention:

- All strings as factors.
- All strings start with uppercase. (ex. Soybeans, Orbweaver).
- Dates and timestamps are Date and POSIXct objects respectively.

- Use NA for missing data.

References

[1] Lisa A. Schulte, Jarad B. Niemi, Matthew J. Helmers, Matt Liebman, J. G. Arbuckle, David E. James, Randall K. Kolka, Matthew E. O'Neal, Mark D. Tomer, John C. Tyndall, Heidi Asbjornsen, Pauline Drobney, Jeri Neal, Gary Van Ryswyk, and Chris Witte (2017). "Prairie strips improve biodiversity and the delivery of multiple ecosystem services from corn-soybean croplands" Proceedings of the National Academy of Sciences, 114(42), 11247-11252. (url)

[2] Xiaobo Zhou, Matthew J. Helmers, Heidi J. Asbjornsen, Randy Kolka, and Mark D. Tomer (2010). "Perennial filter strips reduce nitrate levels in soil and shallow groundwater after grassland-to-cropland conversion" Journal of environmental quality, 39(6), 2006-2015.