ITSEW 2019
International Total Survey Error Workshop

Applications of R Shiny to Explore, Evaluate and Improve Total Survey Quality

Location	(-93.6842, 41.9883)
Year	2017
Category	Soybeans
Value	5
Color	

Xiaodan Lyu
Center for Survey Statistics & Methodology

Joint work with Heike Hofmann, Emily Berg, Jie Li

IOWA STATE UNIVERSITY

Introduction

- Focus on non-sampling errors
  - Sources: data collection, data processing, modeling/estimation
  - Solutions: iterative review and editing, …
- 9 dimensions of total survey quality (Biemer, 2010)
  - accuracy, credibility, comparability, usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness, and coherence

Introduction

- R Shiny (Chang et al., 2018)
  - An R package for developing reactive dashboards
  - Direct and immediate interaction with data in a web-browser
  - Shiny user showcases *https://shiny.rstudio.com/gallery/*
  - Low cost and simple to start with
  - Password-protected Shiny Apps hosted on internal servers
  - Application to survey: a social-network based survey (Joblin and Mauerer, 2016)

National Resources Inventory

- A longitudinal survey on non-federal US land
  - conducted by USDA-NRCS and ISU-CSSM
  - PSU = .5 mi x .5 mi segment, SSU = 3 point locations per PSU
- Estimation of change over time
  - surface area by land cover/use
  - average water and wind erosion on cropland and pastureland
- Record level data set (*pointgen*)
  - location with a single weight and complete data

National Resources Inventory

- Conservation Effects Assessment Project (CEAP)
  - On-site study subsampled from NRI cropland or pastureland
  - Farmer interview (crop management, conservation practice, …)
  - Agricultural Policy Environmental eXtender (APEX) model
  - Output: measurements of soil erosion and chemical runoff
- Small Area Estimation (SAE, Rao and Molina, 2015)
  - Direct estimates for small domains are unreliable
  - Model-based SAE uses population-level auxiliary information

iNtr: an interactive NRI table review tool

Table 2 - Land Cover/use of non-Federal rural land, by State and year
In thousands of acres, with margins of error

State	Year	Cropland	CRP land	Pastureland	Rangeland	Forest land	Other rural land	Total rural land
Alabama	**1982**	4,464.7 ±176.9	--	3,793.9 ±186.4	53.7 ±47.3	20,876.8 ±184.4	523.5 ±73.2	29,712.6 ±112.8
	1987	3,944.7 ±187.9	207.5 --	3,643.8 ±157.9	52.8 ±45.6	21,160.9 ±181.1	491.2 ±74.0	29,500.9 ±116.9
	1992	3,126.2 ±192.4	535.2 --	3,753.3 ±147.5	52.7 ±45.6	21,250.7 ±189.7	611.7 ±84.8	29,329.8 ±124.5
	1997	2,915.5 ±209.9	522.2 --	3,558.0 ±134.4	53.8 ±46.7	21,325.9 ±200.6	590.9 ±78.9	28,966.3 ±139.8
	2002	2,508.5 ±183.0	504.6 --	3,452.2 ±194.7	50.7 ±108.9	21,550.4 ±250.2	505.8 ±83.3	28,572.2 ±151.7
	2007	2,200.2 ±180.6	459.8 --	3,434.9 ±177.3	50.7 ±108.9	21,668.5 ±262.4	554.6 ±90.2	28,368.7 ±164.9
	2012	2,217.0 ±189.0	329.2 --	3,302.9 ±175.9	50.7 ±108.9	21,787.8 ±264.1	591.6 ±85.8	28,279.2 ±168.1
	2015	2,274.9 ±194.2	225.6 --	3,220.2 ±187.2	50.7 ±108.9	21,887.4 ±263.4	591.6 ±85.2	28,250.4 ±168.1
Arizona	**1982**	1,253.0 ±146.6	--	83.6 ±50.4	33,366.4 ±1,026.6	4,572.6 ±862.1	1,711.8 ±577.2	40,987.4 ±270.6
	1987	1,234.6 ±145.5	0.0 --	76.6 ±40.2	33,395.7 ±1,046.3	4,553.8 ±860.8	1,784.4 ±594.7	41,045.1 ±286.7
	1992	1,199.9 ±148.8	0.0 --	83.3 ±36.0	33,796.0 ±1,069.5	4,434.9 ±881.6	1,801.6 ±569.8	41,315.7 ±295.4

Table 14 - Estimated average annual sheet and rill erosion on non-Federal rural land, by State and year
Tons per acre per year with margins of error

State	Year	Cropland			CRP land	Pastureland
		Cultivated	Non-Cultivated	Total		
Alabama	1982	5.02 ±0.29	0.41 ±0.25	4.73 ±0.28	--	0.65 ±0.08
	1987	4.39 ±0.29	0.32 ±0.07	4.07 ±0.27	2.32 ±1.39	0.51 ±0.07
	1992	4.86 ±0.26	0.37 ±0.17	4.35 ±0.23	0.62 ±0.27	0.49 ±0.06
	1997	4.72 ±0.25	0.35 ±0.16	4.18 ±0.22	0.76 ±0.38	0.52 ±0.06
	2002	4.47 ±0.30	0.41 ±0.11	3.62 ±0.29	0.68 ±0.42	0.53 ±0.07
	2007	4.34 ±0.47	0.36 ±0.09	3.34 ±0.41	0.51 ±0.24	0.44 ±0.04
	2012	3.87 ±0.34	0.39 ±0.07	3.03 ±0.33	0.62 ±0.37	0.42 ±0.06
	2015	4.22 ±0.43	0.39 ±0.07	3.31 ±0.42	0.51 ±0.41	0.44 ±0.08
Arizona	1982	0.59 ±0.06	0.45 ±0.04	0.57 ±0.06	--	0.15 ±0.08
	1987	0.65 ±0.06	0.44 ±0.05	0.62 ±0.04	0.00 –	0.11 ±0.03
	1992	0.67 ±0.05	0.27 ±0.03	0.60 ±0.05	0.00 –	0.15 ±0.04

2015 NRI Table Review

- Reasons
  - Multiple estimation runs before final publication
- Differences
  - The 2015 NRI versus the final 2012 NRI
  - A new 2015 estimation versus an earlier 2015 estimation
- Results
  - Expected differences: updated algorithms, data edits, …
  - Surprising differences: problematic data input, …

annielyu.com/#shiny

NRI Data Review Visualization

Info +

Options

Please select V1 with new values: Final_2015

Please select V2 with old values: Final_2012

Table number 2

Table cell (•) level ( ) se

Color scale (•) absolute relative difference ( ) absolute difference

Table 2 (level): Land Cover/use of non-Federal rural land, by State and year, in thousands of acres.
Filter: diff > 0.1 of new val && new val >= 1

Difference Table - US +

Difference Map - US

[ ] Apply filter

Please click a cell in the 1st panel (Difference Table - US) to see the corresponding difference map.

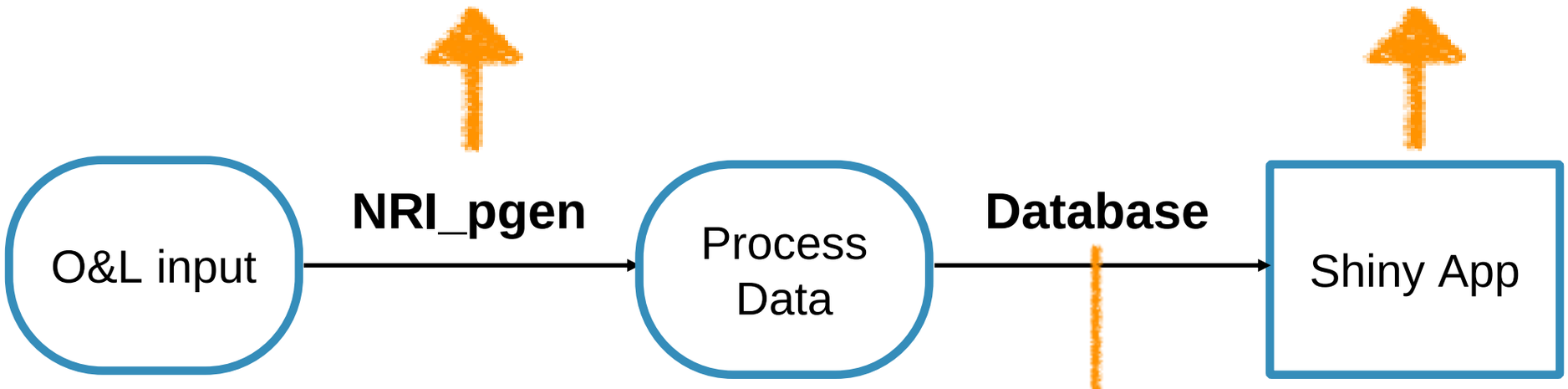
Difference Table - State

[x] Apply filter [ ] Enable hover

Please click a state polygon in the 2nd panel (Difference Table - State) to see the corresponding difference table

- NRI_Data
  | - V1
    | - AL_pgen.txt
    | - …
    | - WY_pgen.txt
  | - V2
  | - …

- app.r
- template.r
- help.r
- NRItables_by_version_state_year.csv
- table_structure.csv
- us_nri_mapdf.rds



Key-value pairs

table	row	column	level	se	state	setname
2	1982	Cropland	4474.3	91.68367	AL	Final_2012
2	1982	CRP land	*NA*	*NA*	AL	Final_2012
2	1982	Forest land	20855.8	93.87755	AL	Final_2012
2	1982	Other rural land	524.4	39.33673	AL	Final_2012
2	1982	Pastureland	3789.4	101.63265	AL	Final_2012
2	1982	Rangeland	70.3	27.24490	AL	Final_2012

viscover: visualize soil and crop data and their overlay

Motivation

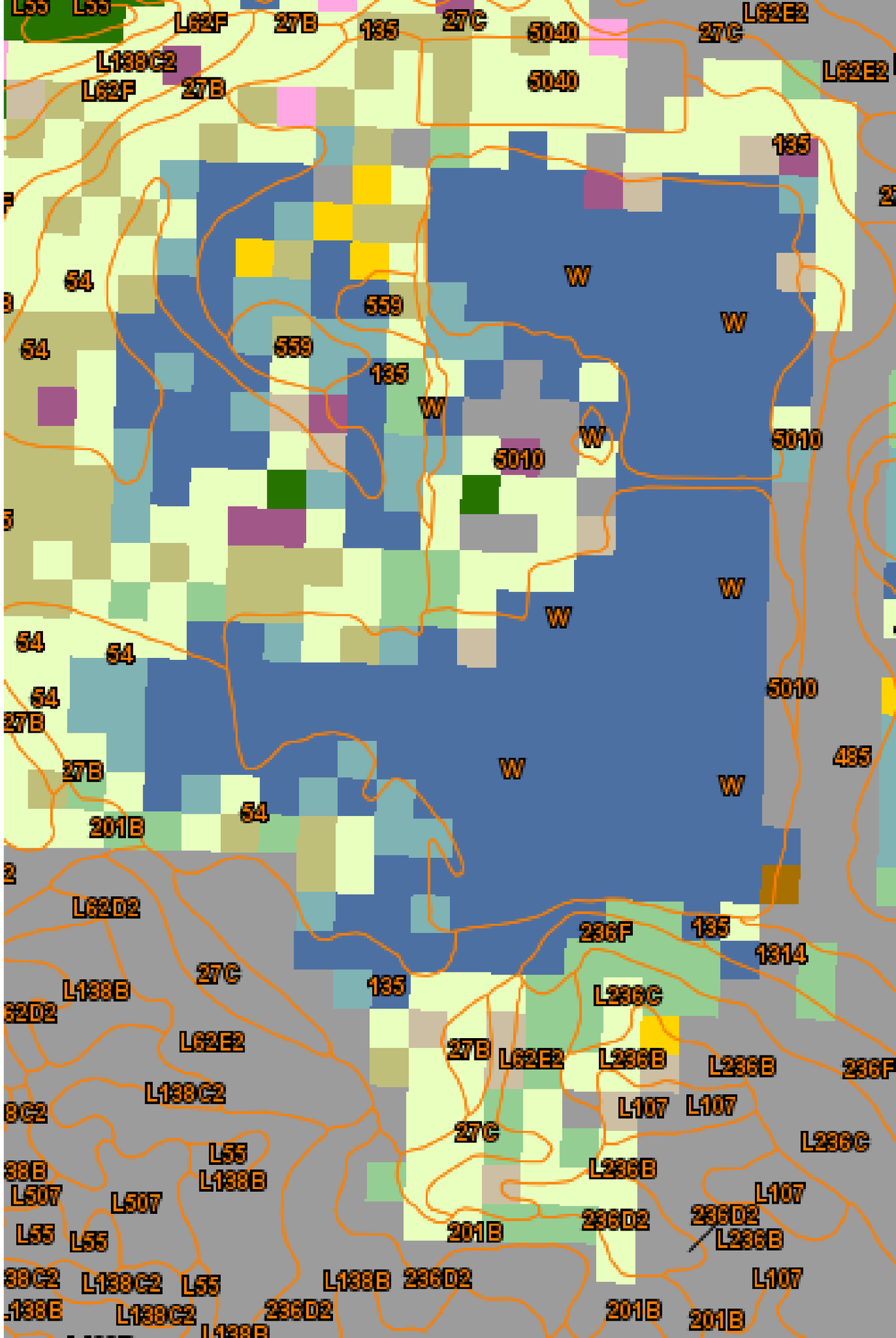
- CEAP Sample: unit-level RUSLE2
- Parameter of interest: county-level RUSLE2
- SAE population-level covariates (soil and crop)
  - data quality of auxiliary variables
  - integrity of overlay operation
- Fitted SAE Model (Lyu, Berg and Hofmann, submitted)

$$\log(Y_{pos}) = b_0 + 2.08 * \mathbf{logR} + 0.48 * \mathbf{logK} + 0.48 * \mathbf{logS} + (1 \,|\, \mathbf{county})$$

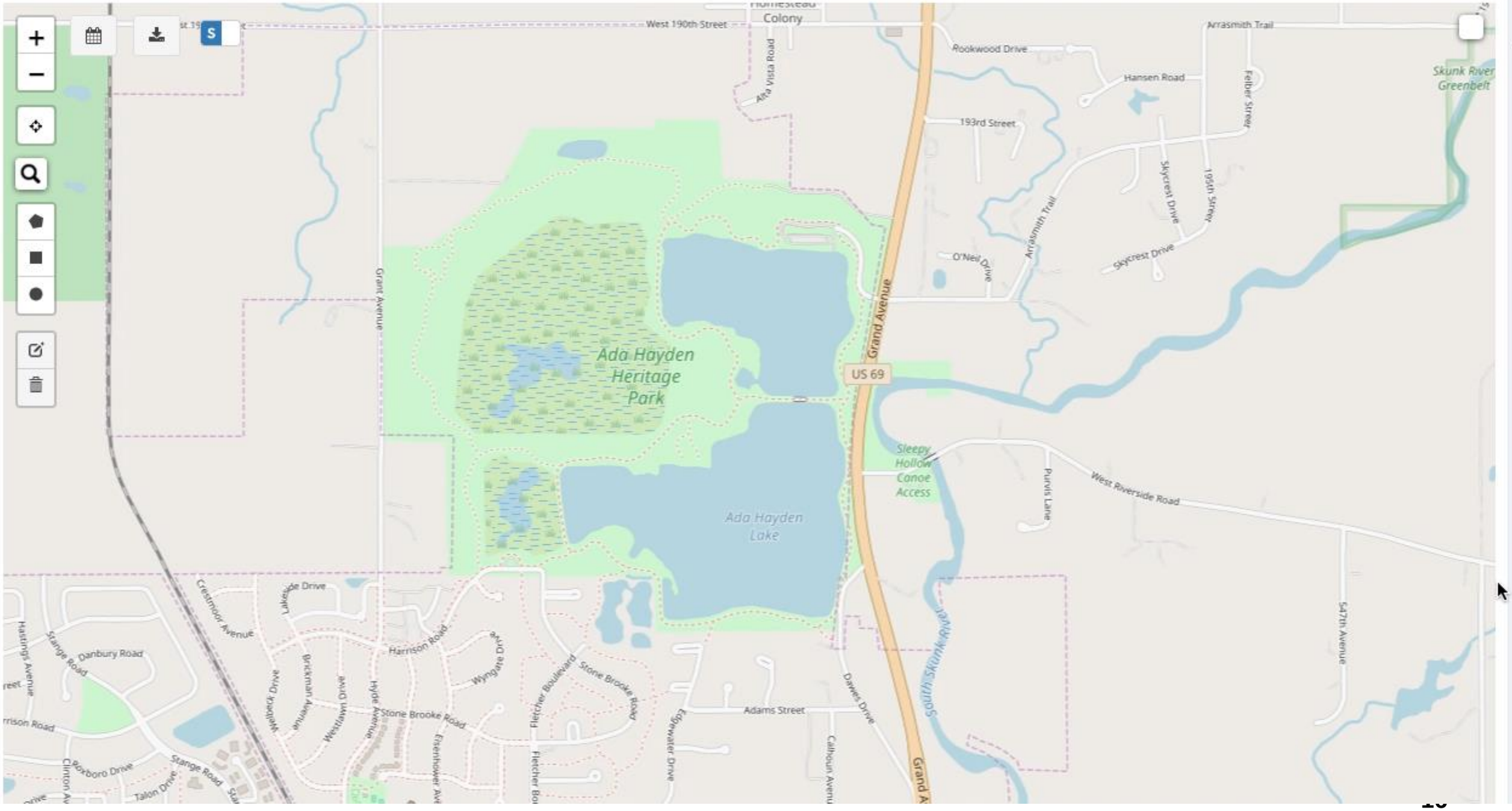
$$\mathbf{logit}(P(Y_{obs} = 1)) = a_0 + 5.04 * \mathbf{logR} + 0.38 * \mathbf{logS} + 0.7 * \mathbf{is.soybean} + 0.95 * \mathbf{is.sprwht} + (1 \,|\, \mathbf{county})$$

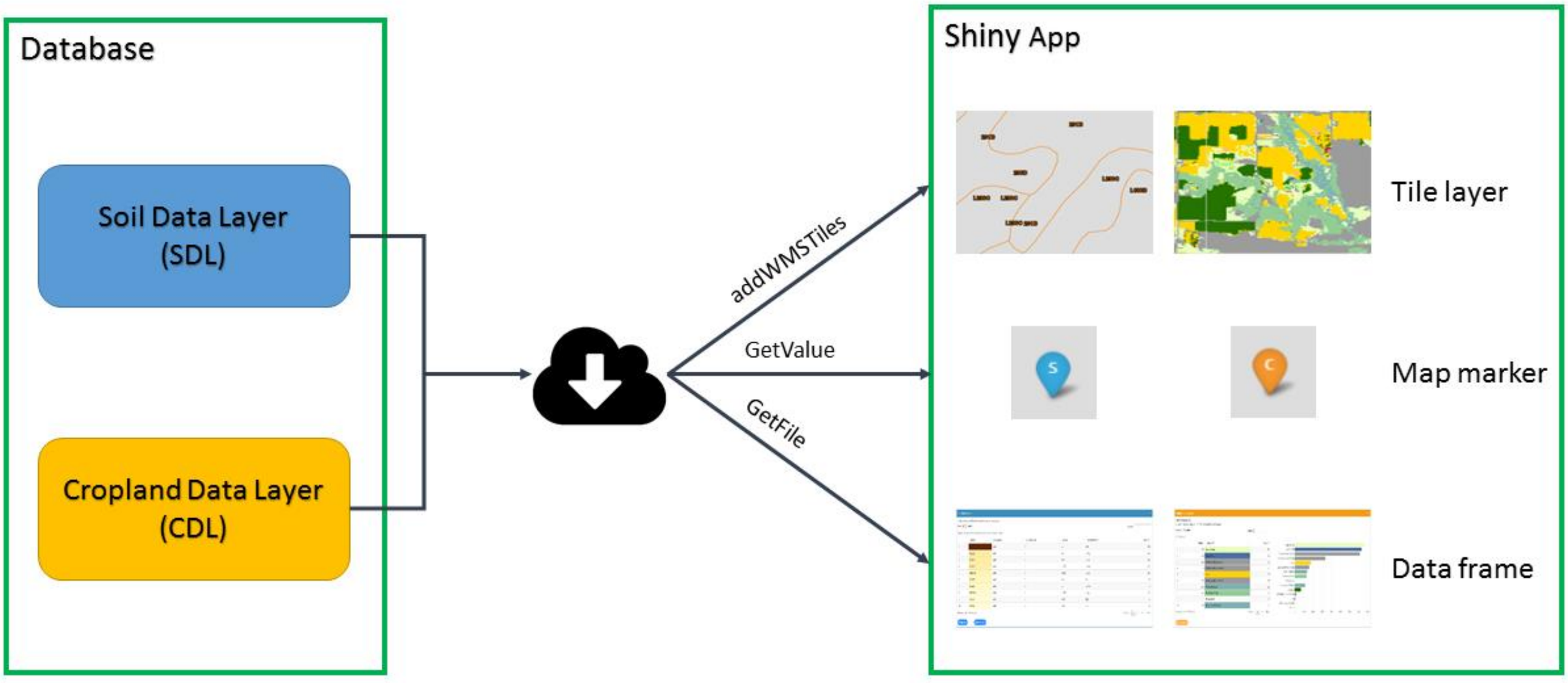
Cropland/Soil Data Layer

- Cropland data layer (CDL)
  - Annual data product for the contiguous United States
  - Geo-referenced crop-specific land cover data layer
- Soil data layer (SDL)
  - Soil Survey Geographic Data (SSURGO)
  - Soil component data on topology and erodibility
  - Available for the United States and the Territories



annielyu.com/#shiny





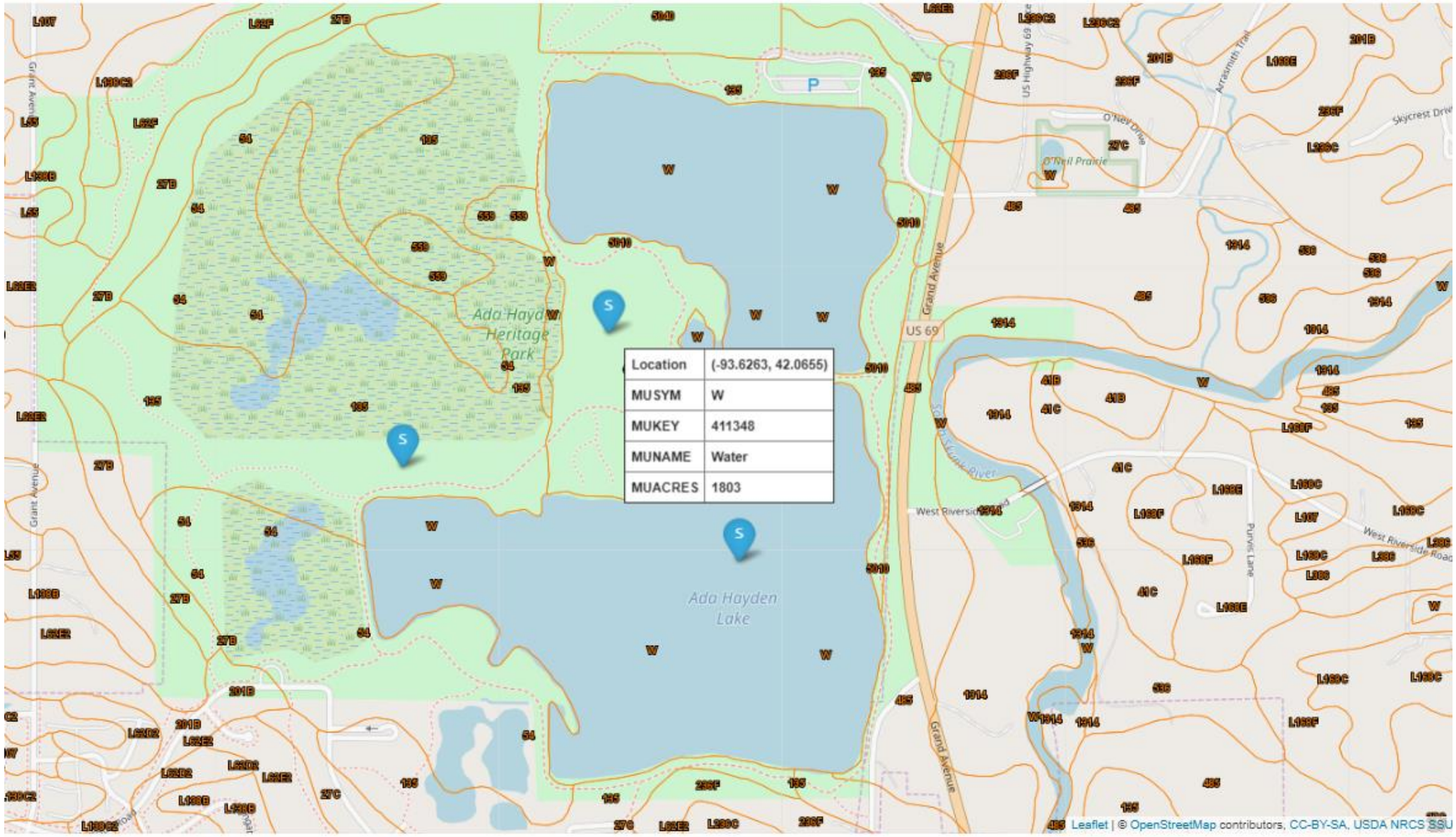
Flowchart of *viscover*.

viscover: an R package

- Installation
  - devtools::install_github(“XiaodanLyu/viscover”)
- Functions
  - run the interactive tool: runTool()
  - fetch data: GetCDLFile, GetCDLValue, GetSDLValue
  - CDL color mapping: cdlpal
- Data
  - CDL category codes: cdl.dbf

Conclusion

- iNtr
  - *Accuracy* - locate issues in NRI data collection and computer programs
  - *Timeliness* - more efficient table review, on schedule for release
  - *Comparability* - geographically hierarchical comparison
- viscover
  - A*ccuracy* - explore the data quality of covariates for small area models
  - *Comparability* - visualize and integrate complex geospatial datasets
  - *Usability* - open source, freely available
  - *Accessibility* - mouse events, customized graphic and tabular output



“A picture is worth a thousand words.”

References

1. P. P. Biemer. Total survey error: Design, implementation, and evaluation. Public Opinion Quarterly, 74(5):817–848, 2010.
2. Rao J, Molina I. Small Area Estimation. John Wiley & Sons, 2015.
3. W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. shiny: Web Application Framework for R, 2018. URL https://CRAN.R-project.org/package=shiny.
4. M. Joblin, and W. Mauerer. "An Interactive Survey Application for Validating Social Network Analysis Techniques." *R Journa*l 8.1 (2016).
5. U.S. Department of Agriculture. 2018. Summary Report: 2015 National Resources Inventory, Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa.
6. X. Lyu, E. J. Berg, and H. Hofmann. Empirical bayes small area prediction of sheet and rill erosion under a zero-inflated lognormal model. 2019+. Manuscript submitted for publication.

Discussion

1. Can our data tools be applicable or generally useful to your project?

2. How could such data tools be applied to reducing sampling errors?

3. What are appropriate outlets where we can publish such kind of applied work?

- annielyu.com
- http://bit.ly/itsew19