



Data Science Support: Rwanda

download the presentation at: <https://osf.io/5jnnp/>



This document by the [Africa Soil Information Service](#) is licensed under a [Creative Commons Attribution 4.0 Unported License](#).

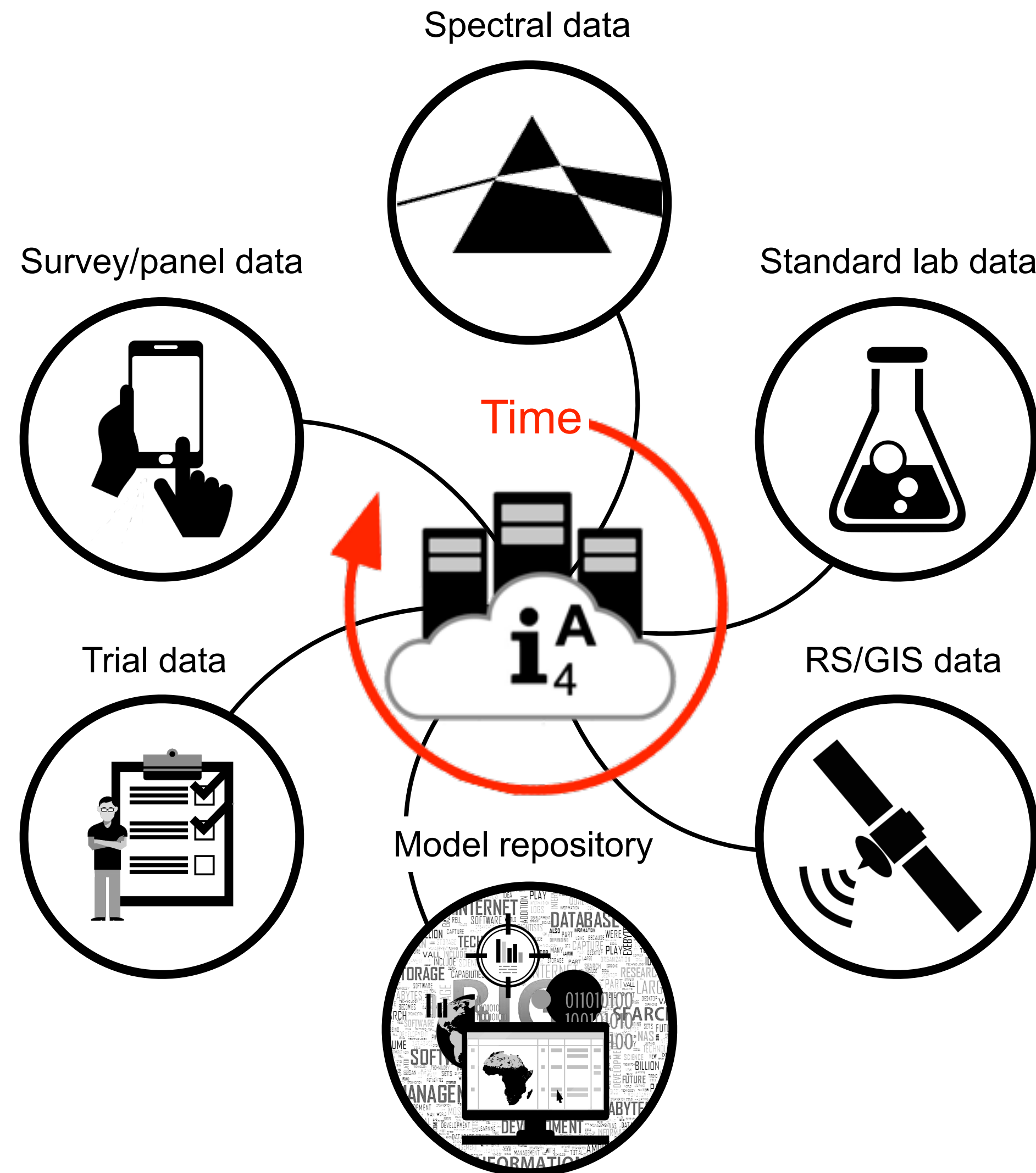
Outline

- ▶ Introduction
 - Applied predictive modeling
 - Typical workflow
- ▶ Examples of FAIR data & code
 - Databases
 - Reproducible workflows
- ▶ Optional: Live 5 minute walk-through FAIR data and code via Google search: [site:osf.io "Rwanda Soil Information Service"](#)

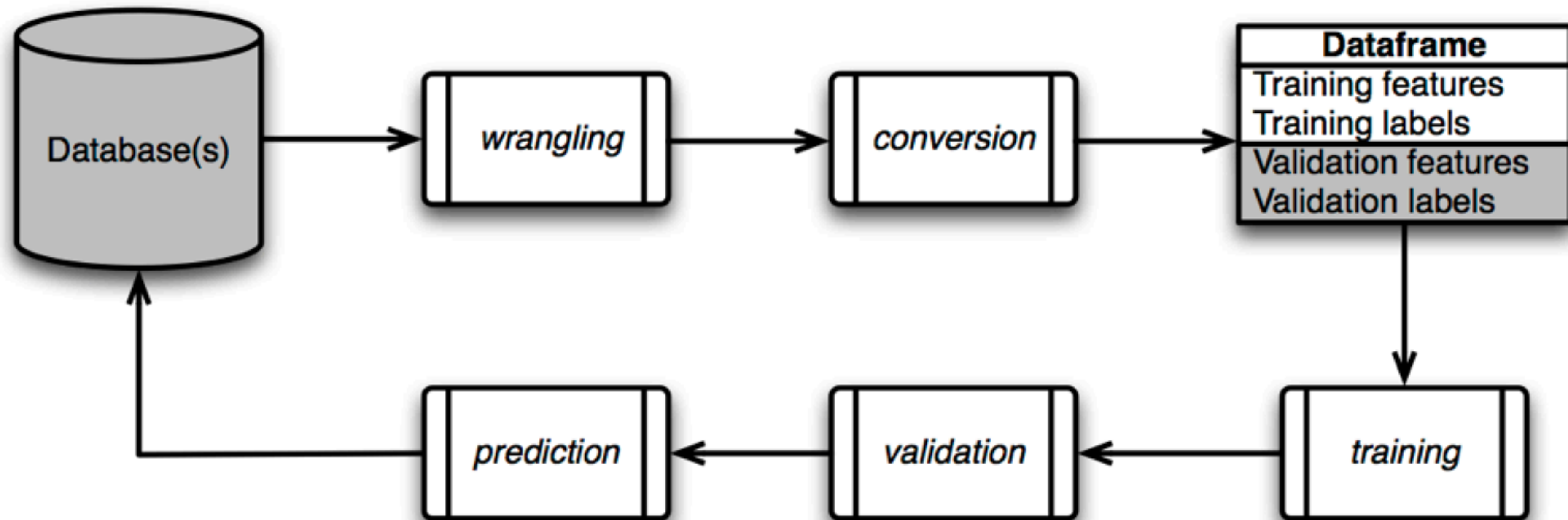
Applied predictive modeling

- ▶ The agricultural sciences still largely use stochastic data models under the assumption that models with high interpretability have inherently high predictive value. **This is neither true nor useful ... most of the time!**
- ▶ The main aims of predictive models are to use readily gathered information to predict responses and diagnostics from (e.g., wet-lab, spectral, remote sensing, spatial, panel, survey, observational, photographic, omic, species distribution, text and sentiment data, etc).
- ▶ Data and predictive modeling approaches are not mutually exclusive ... depending on the data you may be able to do both.

Data “Washing Machine”



Typical predictive modeling workflow



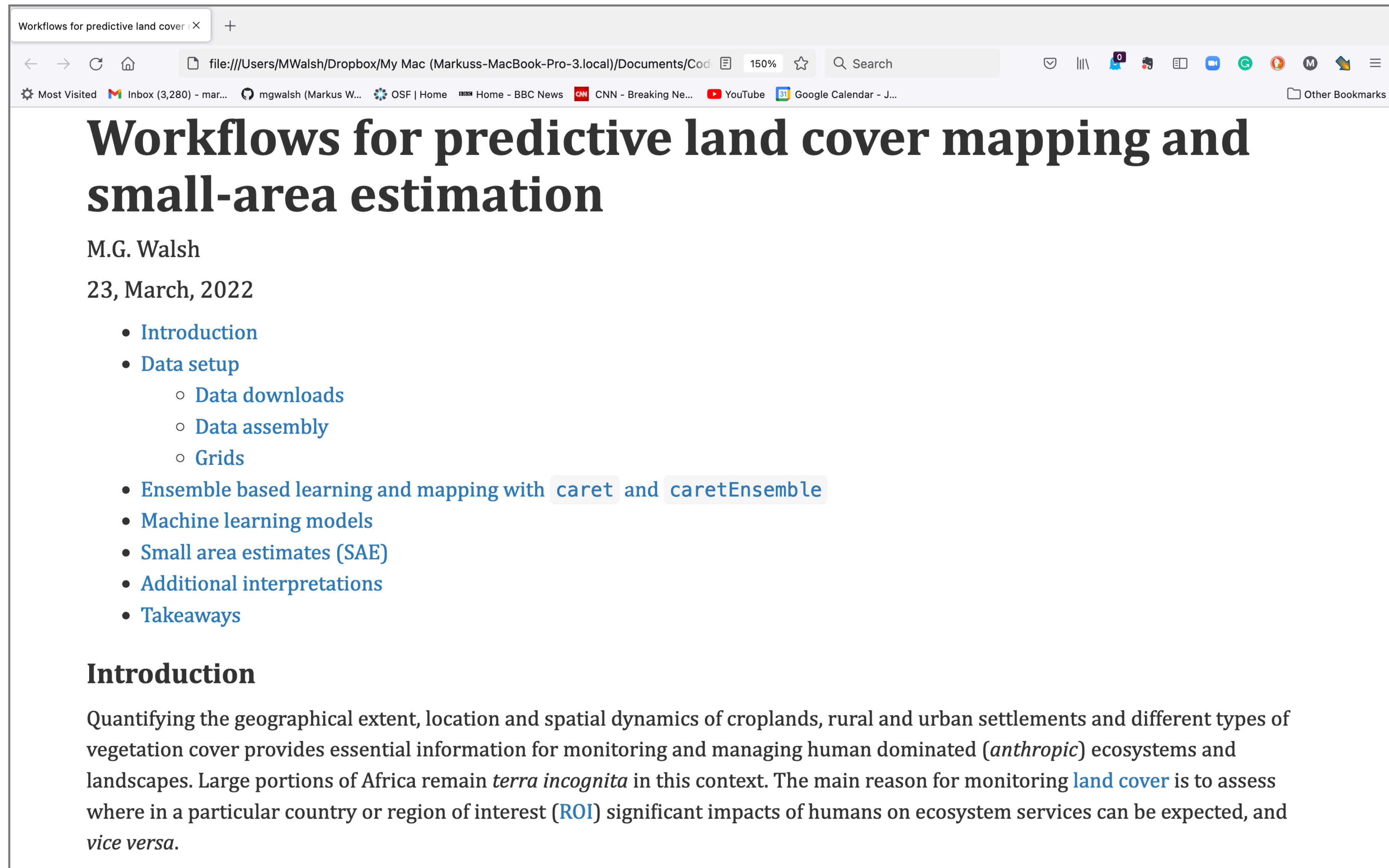
... resulting in many possible predictive models of the data

Example FAIR Data & Code

- ▶ Small-area land cover prediction (at: <https://osf.io/shkxp/>)
- ▶ Priority crop distribution predictions (at: <https://osf.io/ub6ar/>)
- ▶ Predictive soil mapping (at: <https://osf.io/3a5z6/>)
- ▶ RwaSIS cropland sampling frame (at: <https://osf.io/nrb5e/>)
- ▶ Staple food crop association rules (at: <https://osf.io/7rt6c/>)
- ▶ Spectral prediction of lime requirements (at: <https://osf.io/2v46w/>)
- ▶ Meta-analysis of liming trials (at: <https://osf.io/cngwx/>)
- ▶ Landscape soil aggregate stability ratings (at: <https://osf.io/q6ste/>)
- ▶ ... links to more.

Small-area land cover prediction

download the notebook at: <https://osf.io/shkxp/>



Workflows for predictive land cover | × +

file:///Users/MWalsh/Dropbox/My Mac (Markuss-MacBook-Pro-3.local)/Documents/Cod 150% ☆ Search

Most Visited Inbox (3,280) - mar... mgwalsh (Markus W... OSF | Home Home - BBC News CNN - Breaking Ne... YouTube Google Calendar - J... Other Bookmarks

Workflows for predictive land cover mapping and small-area estimation

M.G. Walsh

23, March, 2022

- [Introduction](#)
- [Data setup](#)
 - [Data downloads](#)
 - [Data assembly](#)
 - [Grids](#)
- [Ensemble based learning and mapping with `caret` and `caretEnsemble`](#)
- [Machine learning models](#)
- [Small area estimates \(SAE\)](#)
- [Additional interpretations](#)
- [Takeaways](#)

Introduction

Quantifying the geographical extent, location and spatial dynamics of croplands, rural and urban settlements and different types of vegetation cover provides essential information for monitoring and managing human dominated (*anthropic*) ecosystems and landscapes. Large portions of Africa remain *terra incognita* in this context. The main reason for monitoring **land cover** is to assess where in a particular country or region of interest (**ROI**) significant impacts of humans on ecosystem services can be expected, and *vice versa*.

Rwanda GeoSurvey land cover labels (on +23k quadrats)

https://geosurvey.qed.ai/admin/survey/resultchange/1216483/

geosurvey
qed.ai

Menu

Buildings present?
 Yes No Don't know

Tag every building

Cropland present?
 Yes No Don't know

Woody cover >60%?
 Yes No Don't know

Update

View sample

Survey Discuss

Longitude: 27.006964
Latitude: -16.793931
Zoom Level: 18
Country: —
Side length of box: 250m

https://geosurvey.qed.ai/admin/survey/resultchange/1216389/

geosurvey
qed.ai

Menu

Buildings present?
 Yes No Don't know

Cropland present?
 Yes No Don't know

Tag cropland on 4x4 grid

Conservation structures present?
 Yes No Don't know

Woody cover >60%?
 Yes No Don't know

Update

View sample

Survey Discuss

Longitude: 33.203169
Latitude: -11.816827
Zoom Level: 18
Country: —
Side length of box: 250m

https://geosurvey.qed.ai/admin/survey/resultchange/1216459/

geosurvey
qed.ai

Menu

Buildings present?
 Yes No Don't know

Cropland present?
 Yes No Don't know

Woody cover >60%?
 Yes No Don't know

Update

View sample

Grid
Color: Blue

Show markers

Markers in a row: 4

Survey Discuss

Longitude: 24.658301
Latitude: -15.574436
Zoom Level: 18
Country: —
Side length of box: 250m

https://geosurvey.qed.ai/admin/survey/resultchange/1216477/

geosurvey
qed.ai

Menu

Buildings present?
 Yes No Don't know

Cropland present?
 Yes No Don't know

Woody cover >60%?
 Yes No Don't know

Update

View sample

Grid
Color: Blue

Show markers

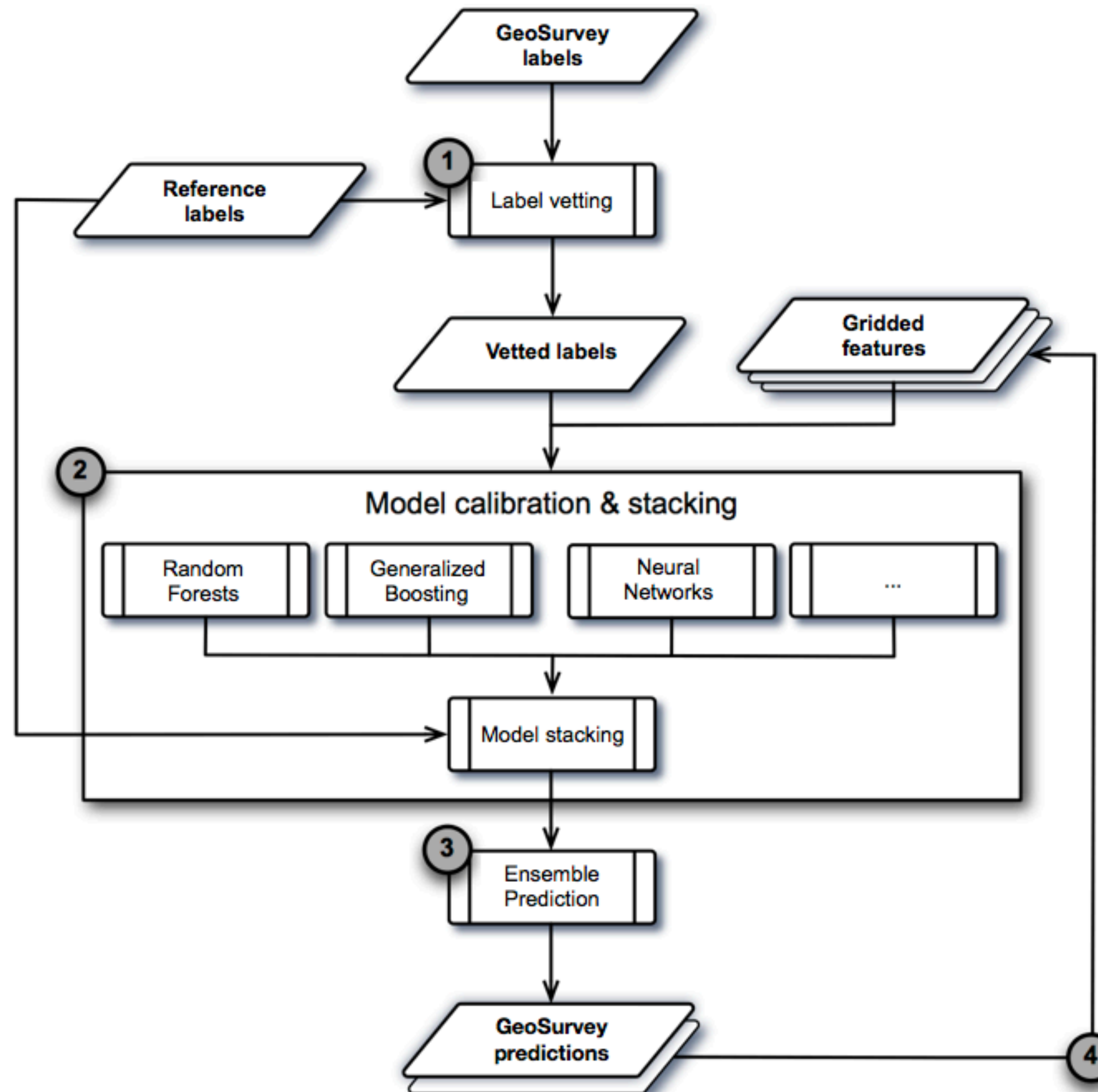
Markers in a row: 4

Survey Discuss

Longitude: 25.941938
Latitude: -14.975861
Zoom Level: 18
Country: —
Side length of box: 250m

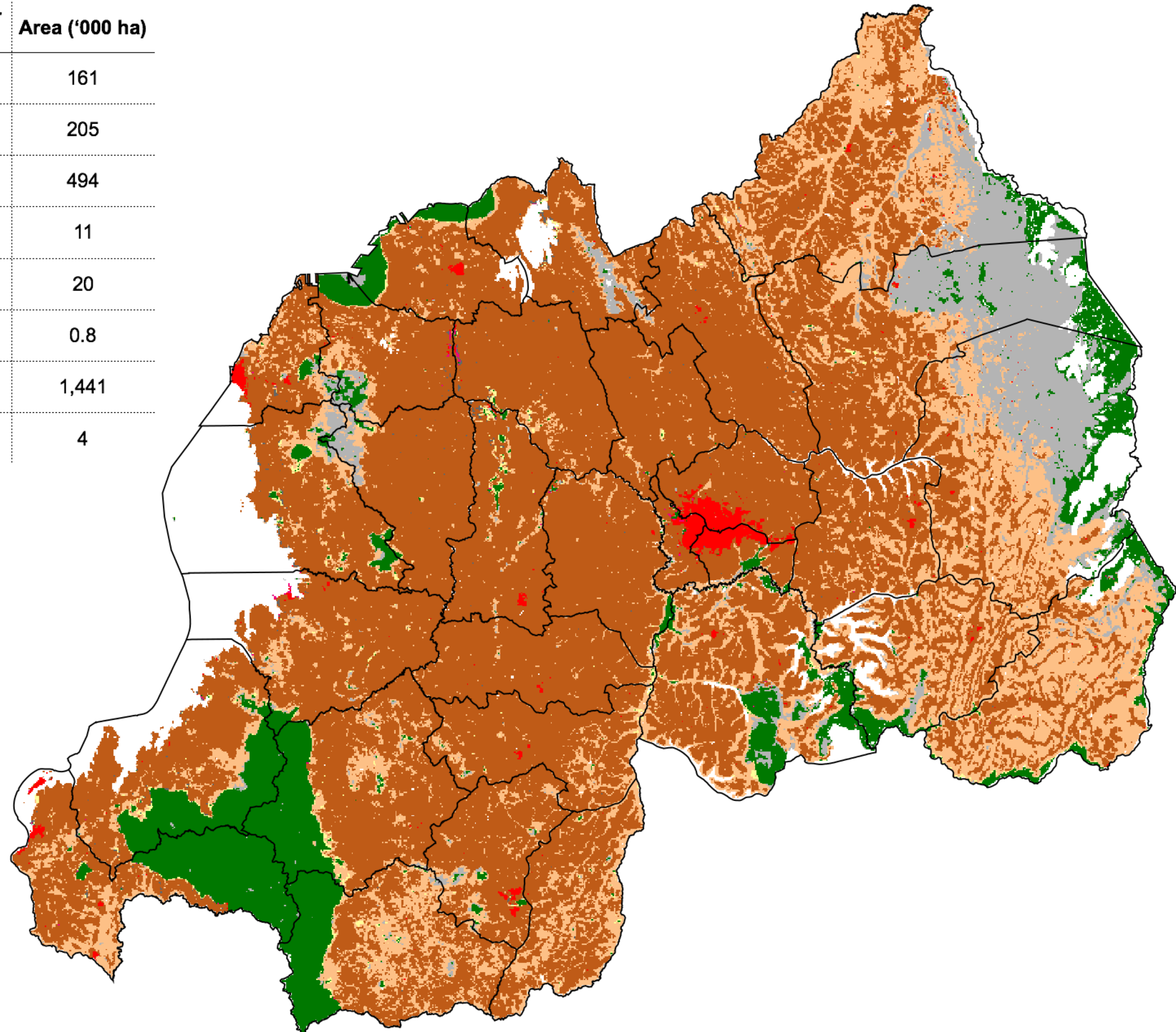
GeoSurvey landcover prediction workflow

see the notebook at: <https://osf.io/shkxp/>



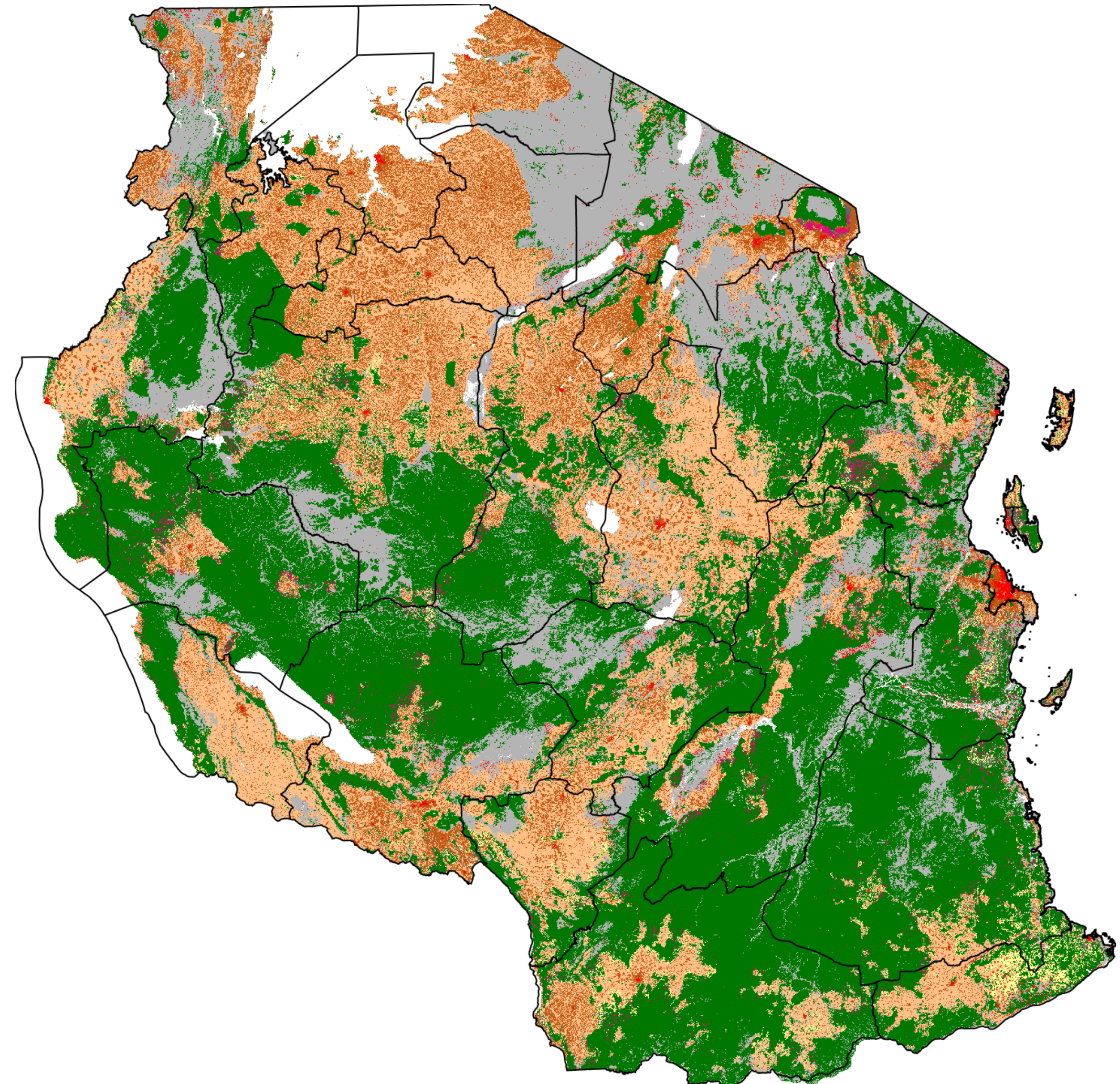
Rwanda GeoSurvey landcover classification (2020)

	Buildings present?	Cropland present?	Woody cover > 60%?	Area ('000 ha)
Light Gray	No	No	No	161
Dark Green	No	No	Yes	205
Light Orange	No	Yes	No	494
Yellow	No	Yes	Yes	11
Red	Yes	No	No	20
Magenta	Yes	No	Yes	0.8
Brown	Yes	Yes	No	1,441
Dark Gray	Yes	Yes	Yes	4



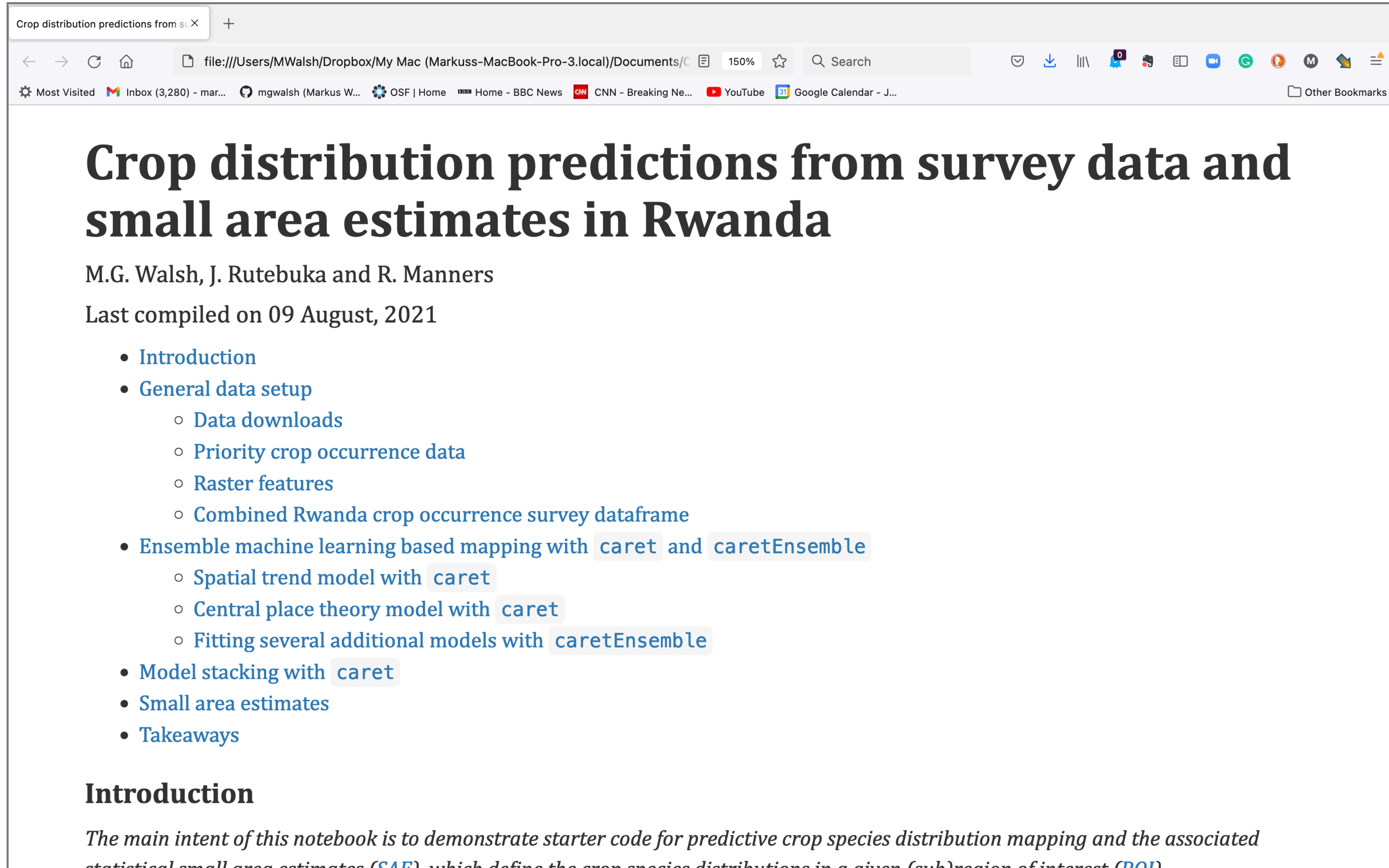
Tanzania GeoSurvey landcover classification (2020)

	Buildings present?	Cropland present?	Woody cover > 60%?	Area ('000 ha)
	No	No	No	17,599
	No	No	Yes	38,882
	No	Yes	No	18,348
	No	Yes	Yes	1,484
	Yes	No	No	662
	Yes	No	Yes	811
	Yes	Yes	No	9,266
	Yes	Yes	Yes	732



Priority crop distribution predictions

download the notebook at: <https://osf.io/ub6ar/>



Crop distribution predictions from survey data and small area estimates in Rwanda

M.G. Walsh, J. Rutebuka and R. Manners

Last compiled on 09 August, 2021

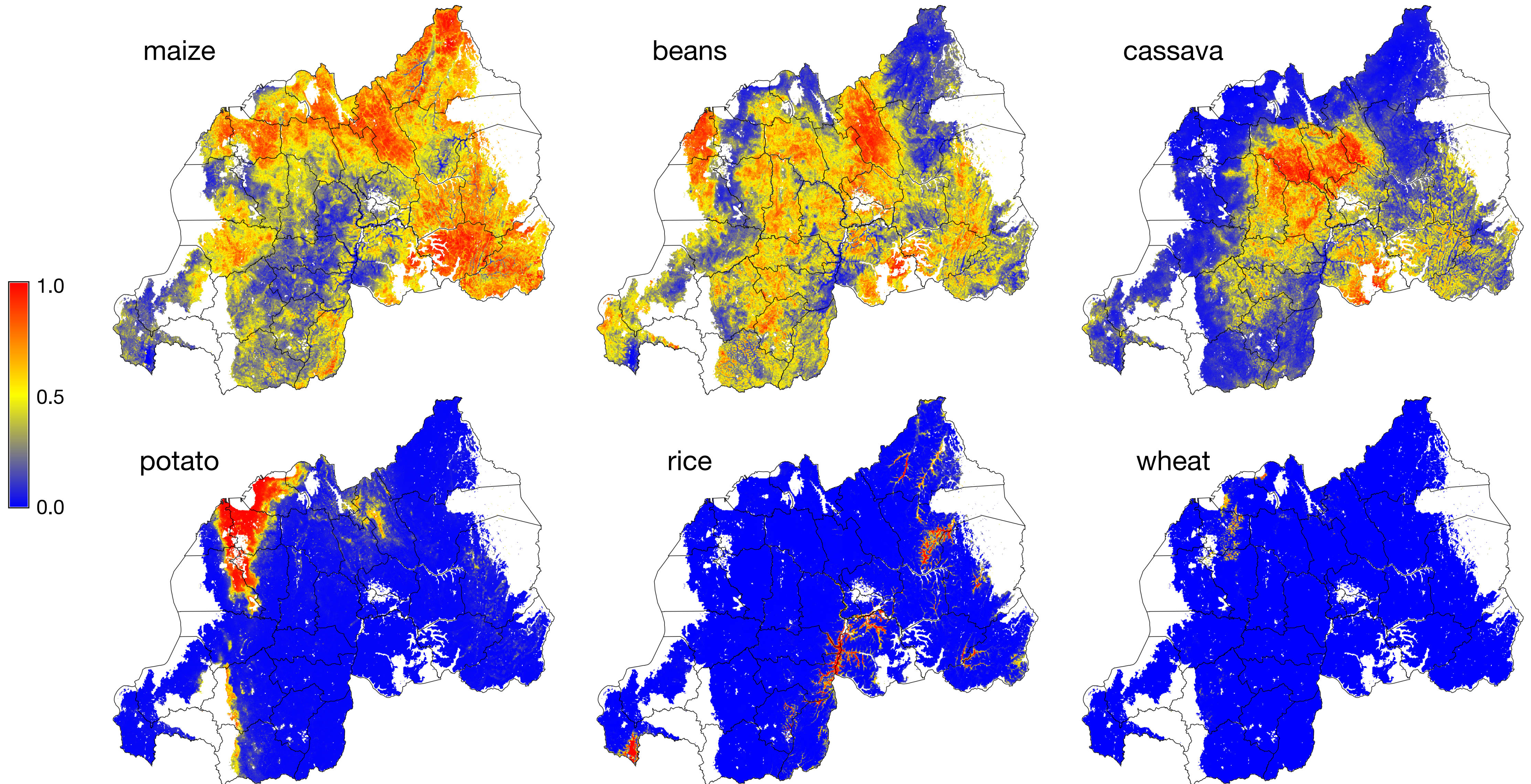
- [Introduction](#)
- [General data setup](#)
 - [Data downloads](#)
 - [Priority crop occurrence data](#)
 - [Raster features](#)
 - [Combined Rwanda crop occurrence survey dataframe](#)
- [Ensemble machine learning based mapping with `caret` and `caretEnsemble`](#)
 - [Spatial trend model with `caret`](#)
 - [Central place theory model with `caret`](#)
 - [Fitting several additional models with `caretEnsemble`](#)
- [Model stacking with `caret`](#)
- [Small area estimates](#)
- [Takeaways](#)

Introduction

The main intent of this notebook is to demonstrate starter code for predictive crop species distribution mapping and the associated statistical small area estimates (SAE) which define the crop species distributions in a given (sub)region of interest (ROI)

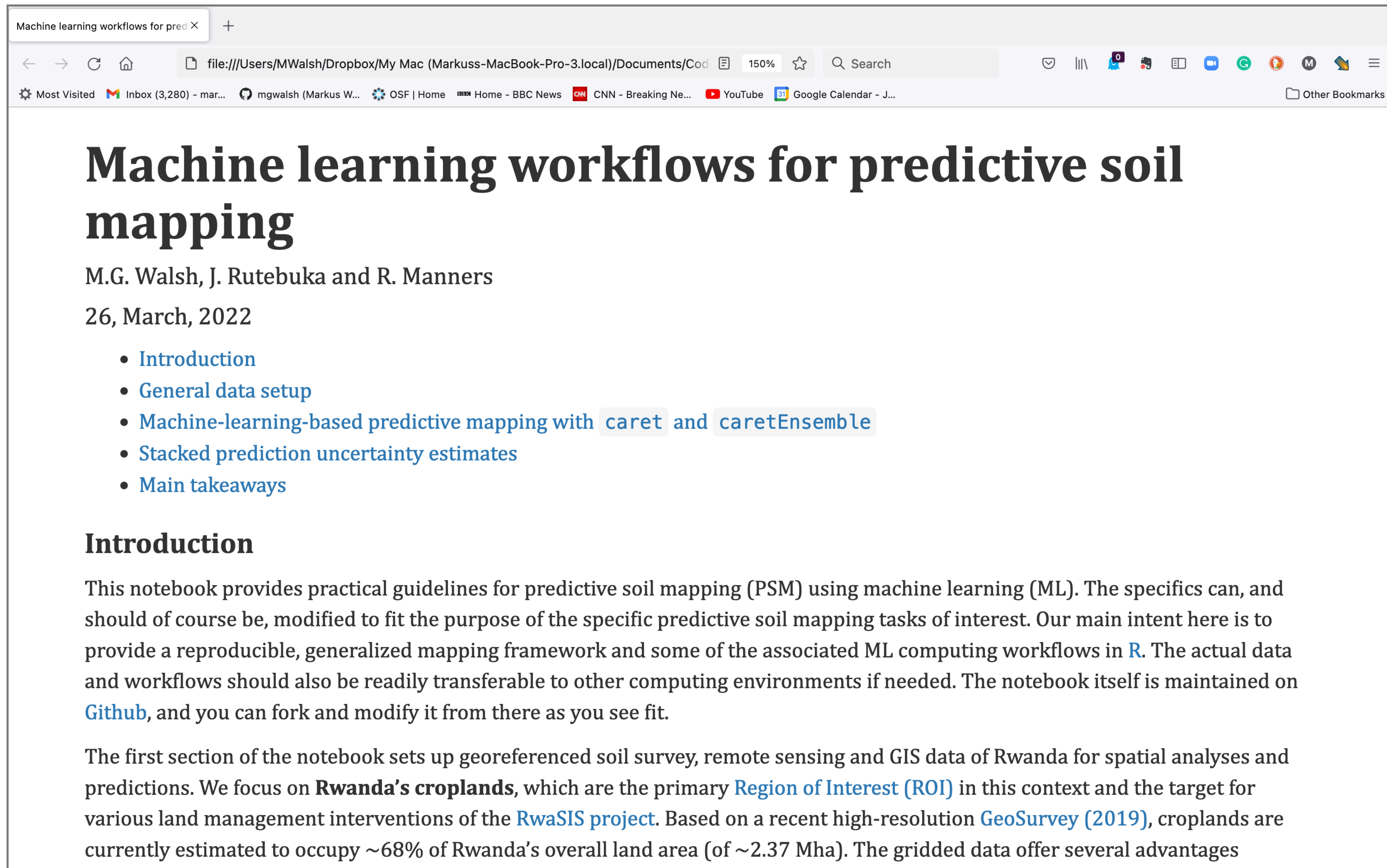
RAB priority crop distribution predictions

download the notebook at: <https://osf.io/ub6ar/>



Predictive soil mapping

download the notebook at: <https://osf.io/hpebu/>



The image is a screenshot of a web browser window. The browser's address bar shows a file path: `file:///Users/MWalsh/Dropbox/My Mac (Markuss-MacBook-Pro-3.local)/Documents/Cod`. The browser's bookmark bar contains several items, including 'Most Visited', 'Inbox (3,280) - mar...', 'mgwalsh (Markus W...', 'OSF | Home', 'Home - BBC News', 'CNN - Breaking Ne...', 'YouTube', and 'Google Calendar - J...'. The main content of the page is a document titled 'Machine learning workflows for predictive soil mapping' by M.G. Walsh, J. Rutebuka, and R. Manners, dated 26, March, 2022. The document includes a table of contents with five items: 'Introduction', 'General data setup', 'Machine-learning-based predictive mapping with caret and caretEnsemble', 'Stacked prediction uncertainty estimates', and 'Main takeaways'. The 'Introduction' section is highlighted in bold. The text of the introduction states that the notebook provides practical guidelines for predictive soil mapping (PSM) using machine learning (ML), and that the data and workflows are transferable to other computing environments. It also mentions that the notebook is maintained on Github and can be forked and modified. The first section of the notebook sets up georeferenced soil survey, remote sensing, and GIS data of Rwanda for spatial analyses and predictions, focusing on Rwanda's croplands as the primary Region of Interest (ROI) for the RwaSIS project. It notes that croplands are currently estimated to occupy ~68% of Rwanda's overall land area (of ~2.37 Mha) and that the gridded data offer several advantages.

Machine learning workflows for predictive soil mapping

M.G. Walsh, J. Rutebuka and R. Manners

26, March, 2022

- [Introduction](#)
- [General data setup](#)
- [Machine-learning-based predictive mapping with `caret` and `caretEnsemble`](#)
- [Stacked prediction uncertainty estimates](#)
- [Main takeaways](#)

Introduction

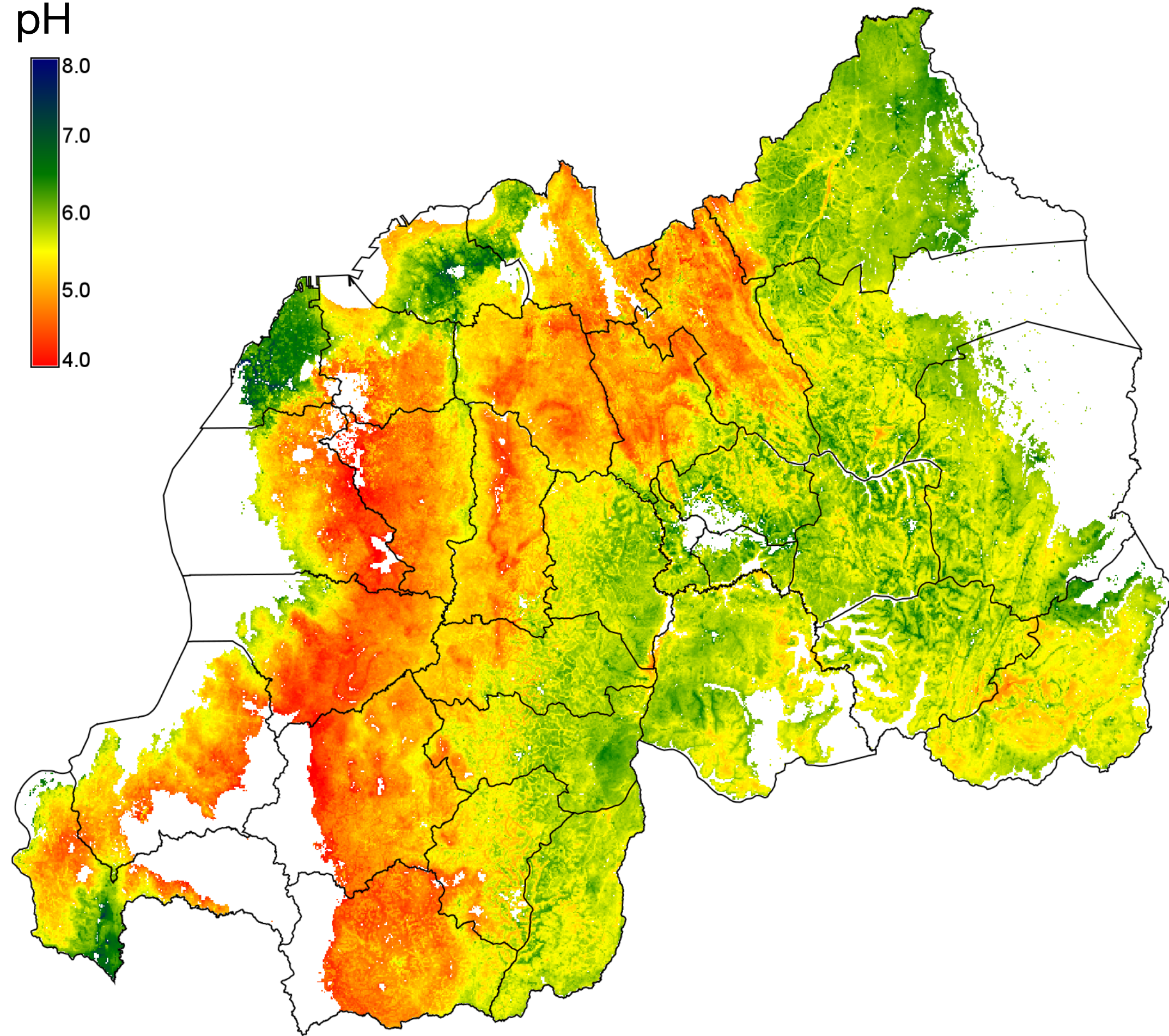
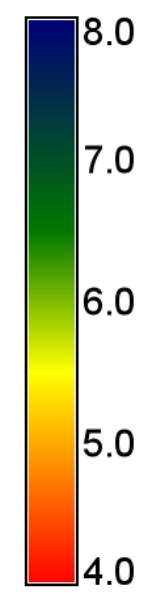
This notebook provides practical guidelines for predictive soil mapping (PSM) using machine learning (ML). The specifics can, and should of course be, modified to fit the purpose of the specific predictive soil mapping tasks of interest. Our main intent here is to provide a reproducible, generalized mapping framework and some of the associated ML computing workflows in `R`. The actual data and workflows should also be readily transferable to other computing environments if needed. The notebook itself is maintained on [Github](#), and you can fork and modify it from there as you see fit.

The first section of the notebook sets up georeferenced soil survey, remote sensing and GIS data of Rwanda for spatial analyses and predictions. We focus on **Rwanda's croplands**, which are the primary [Region of Interest \(ROI\)](#) in this context and the target for various land management interventions of the [RwaSIS project](#). Based on a recent high-resolution [GeoSurvey \(2019\)](#), croplands are currently estimated to occupy ~68% of Rwanda's overall land area (of ~2.37 Mha). The gridded data offer several advantages

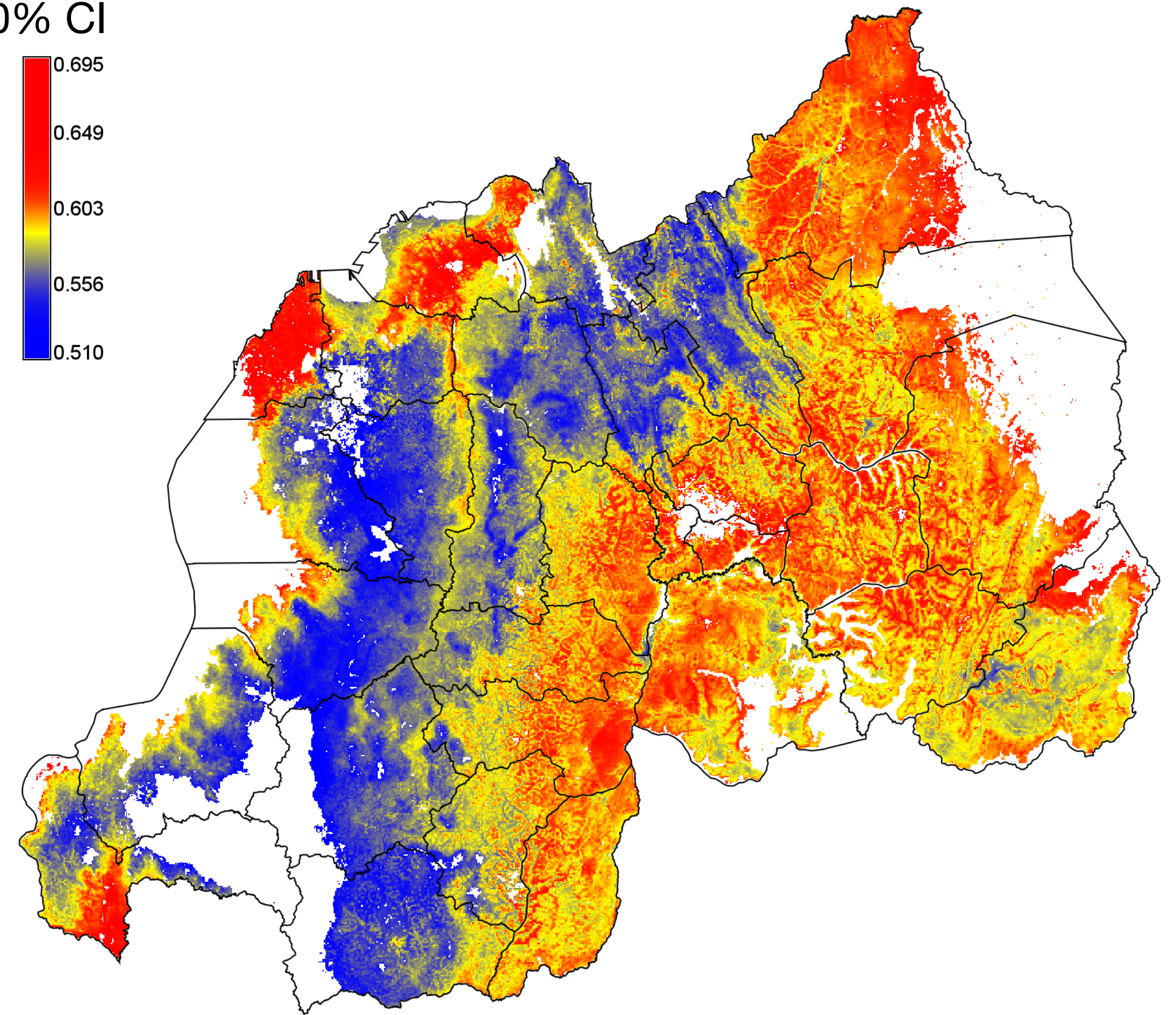
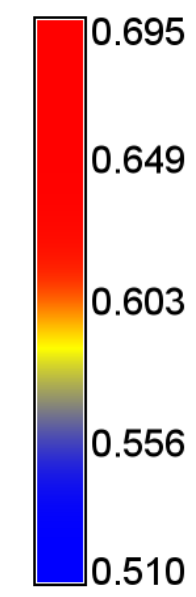
Predictive soil mapping

see the notebook at: <https://osf.io/hpebu/>

pH

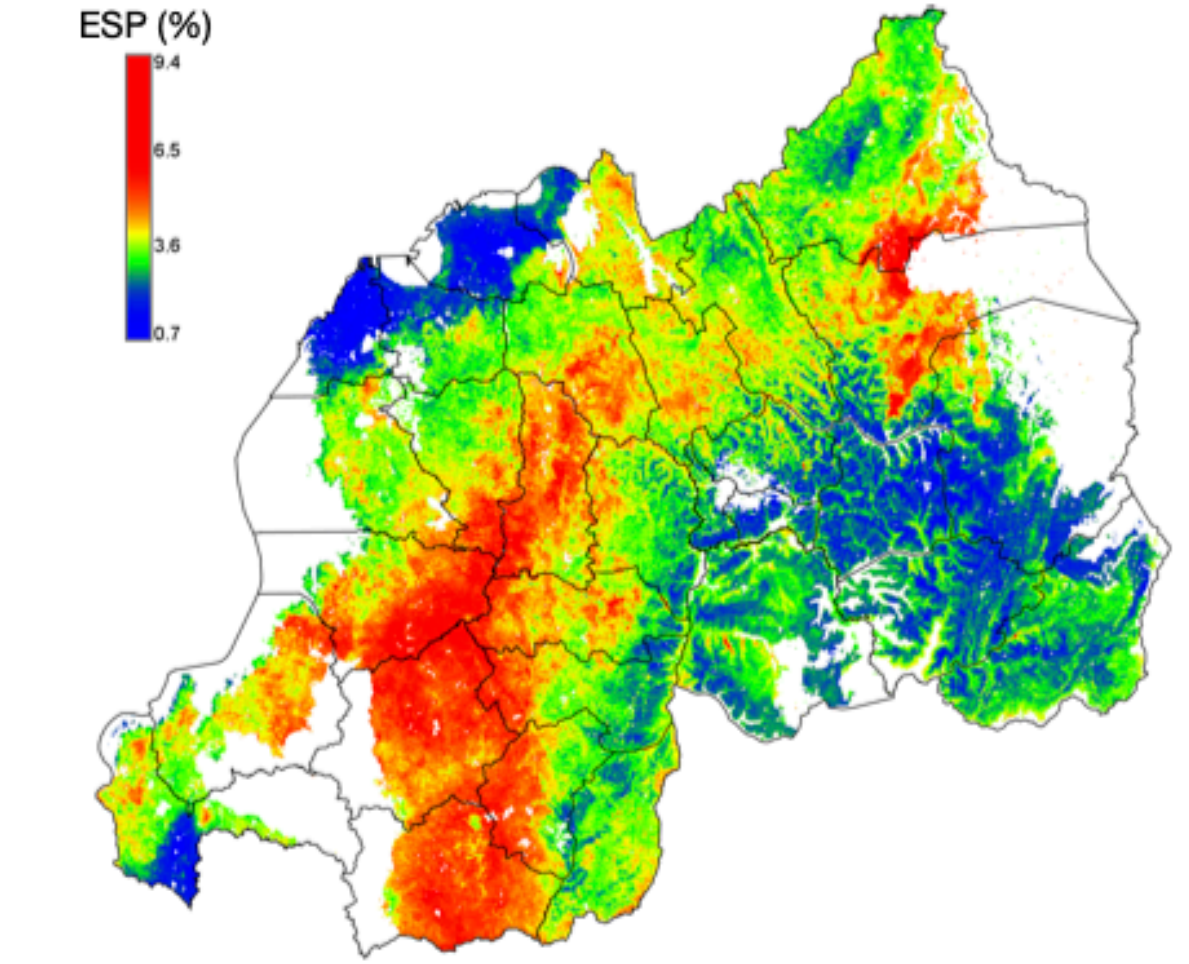
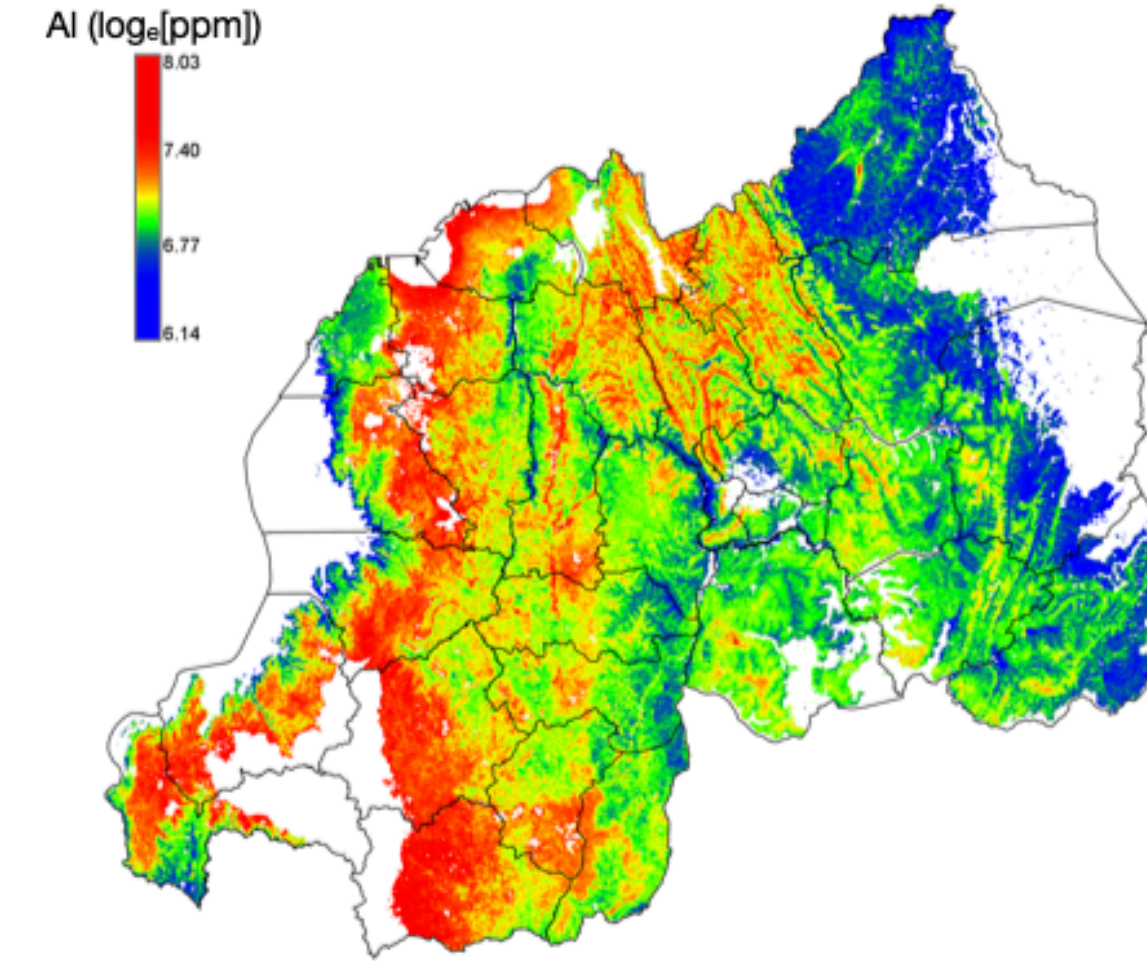
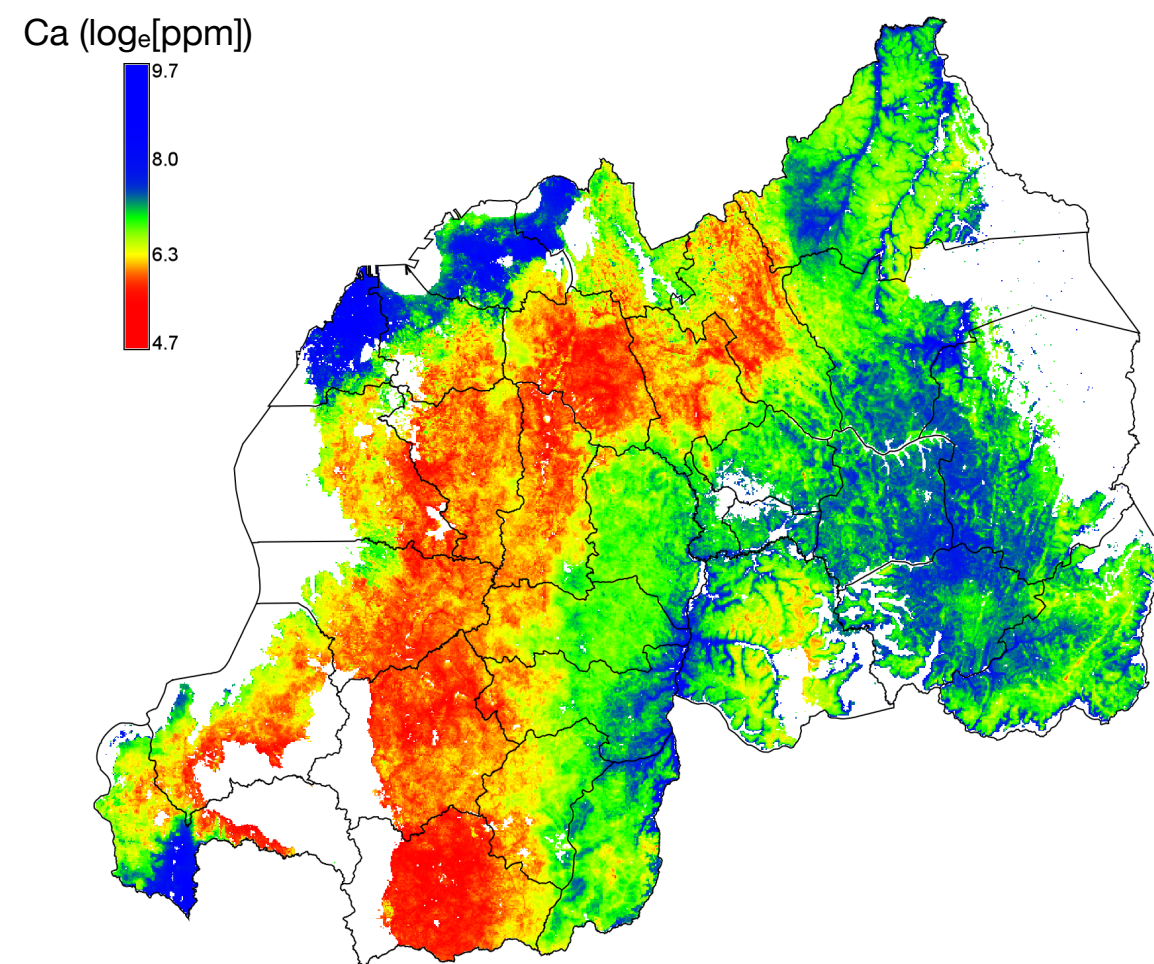
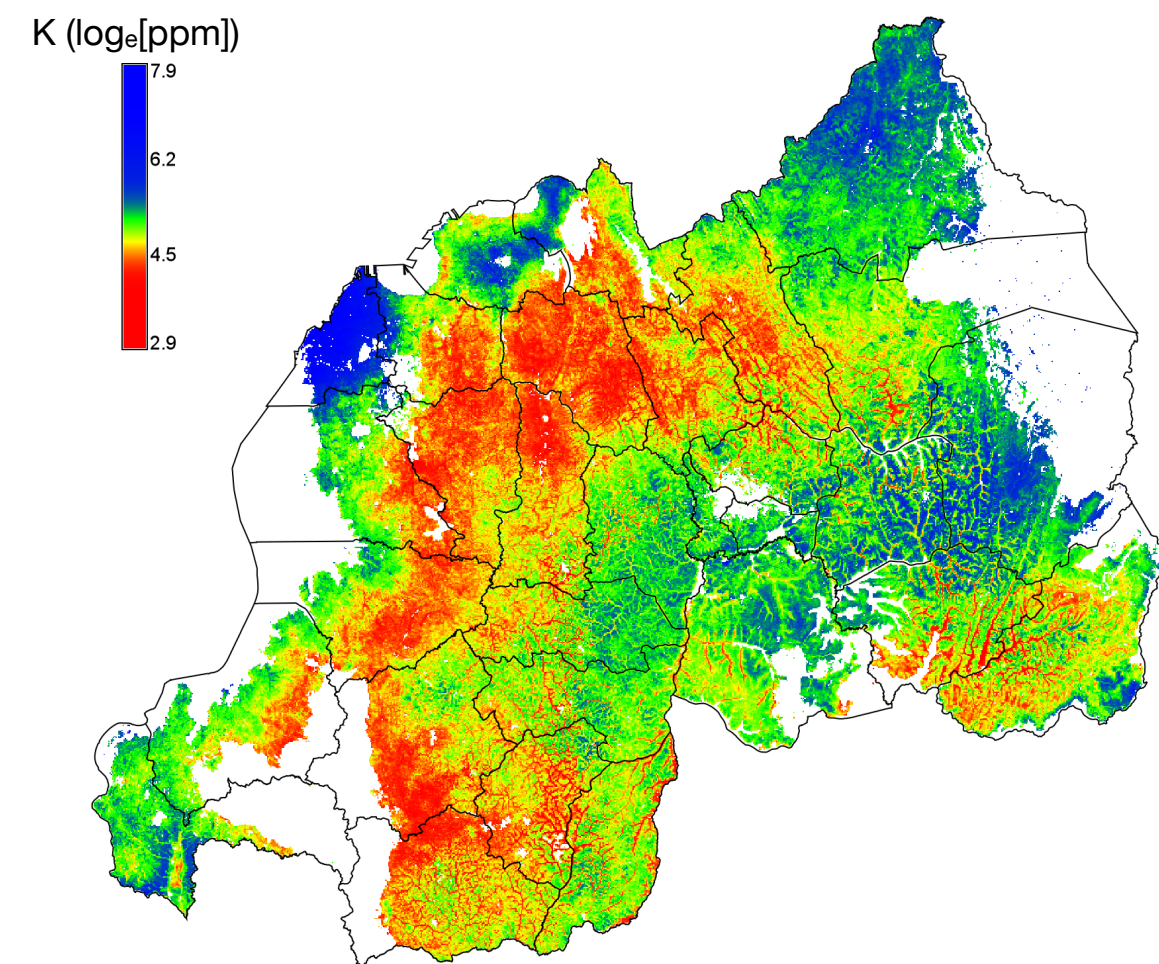
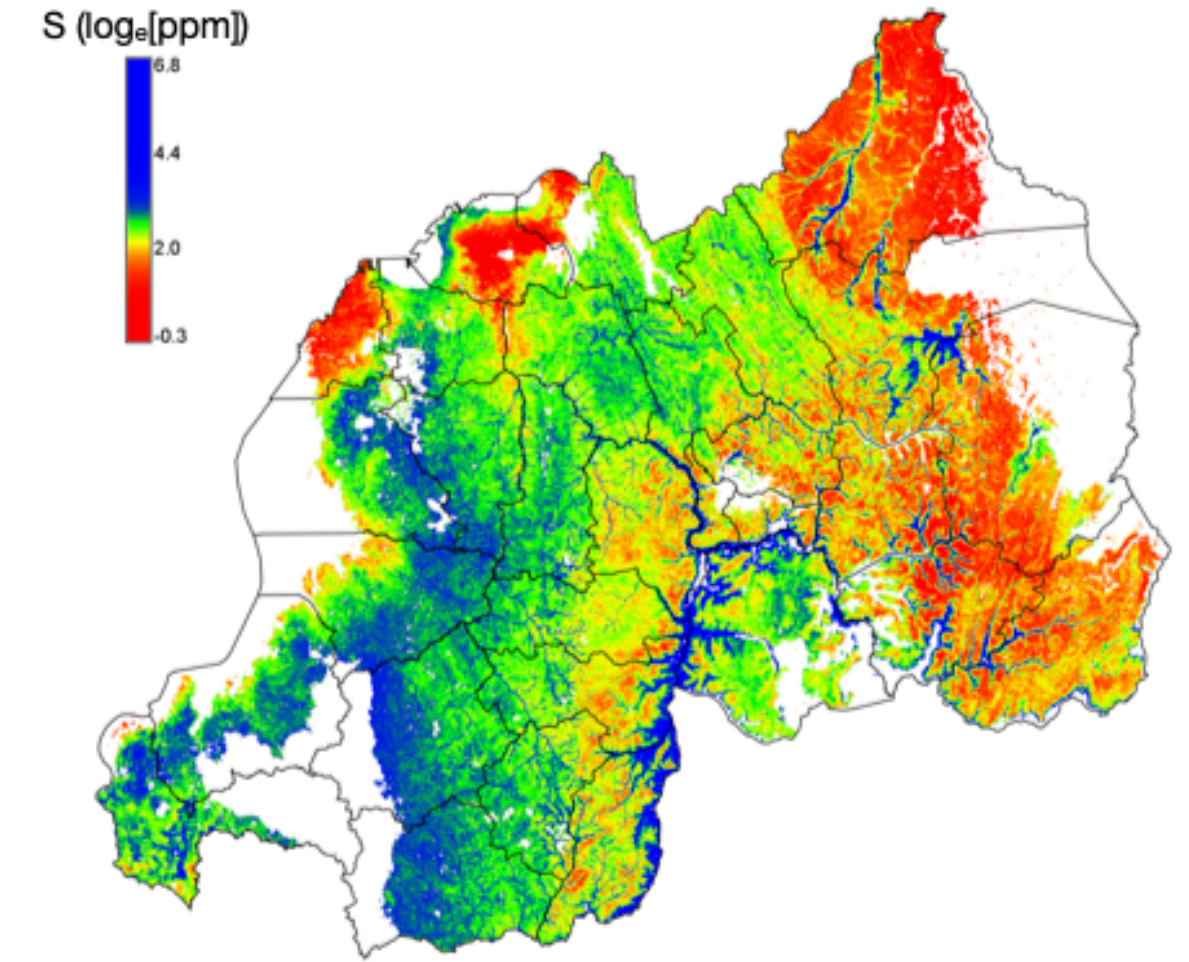
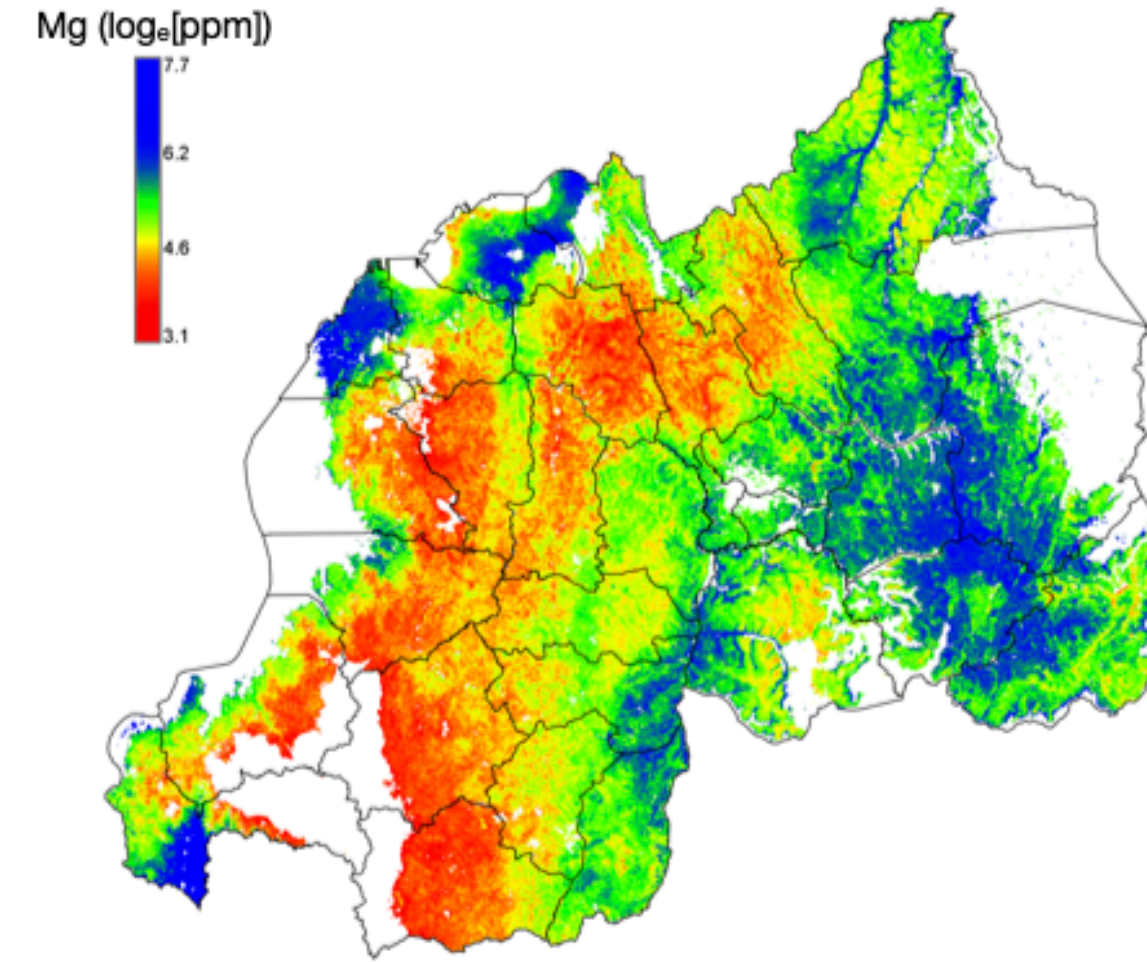
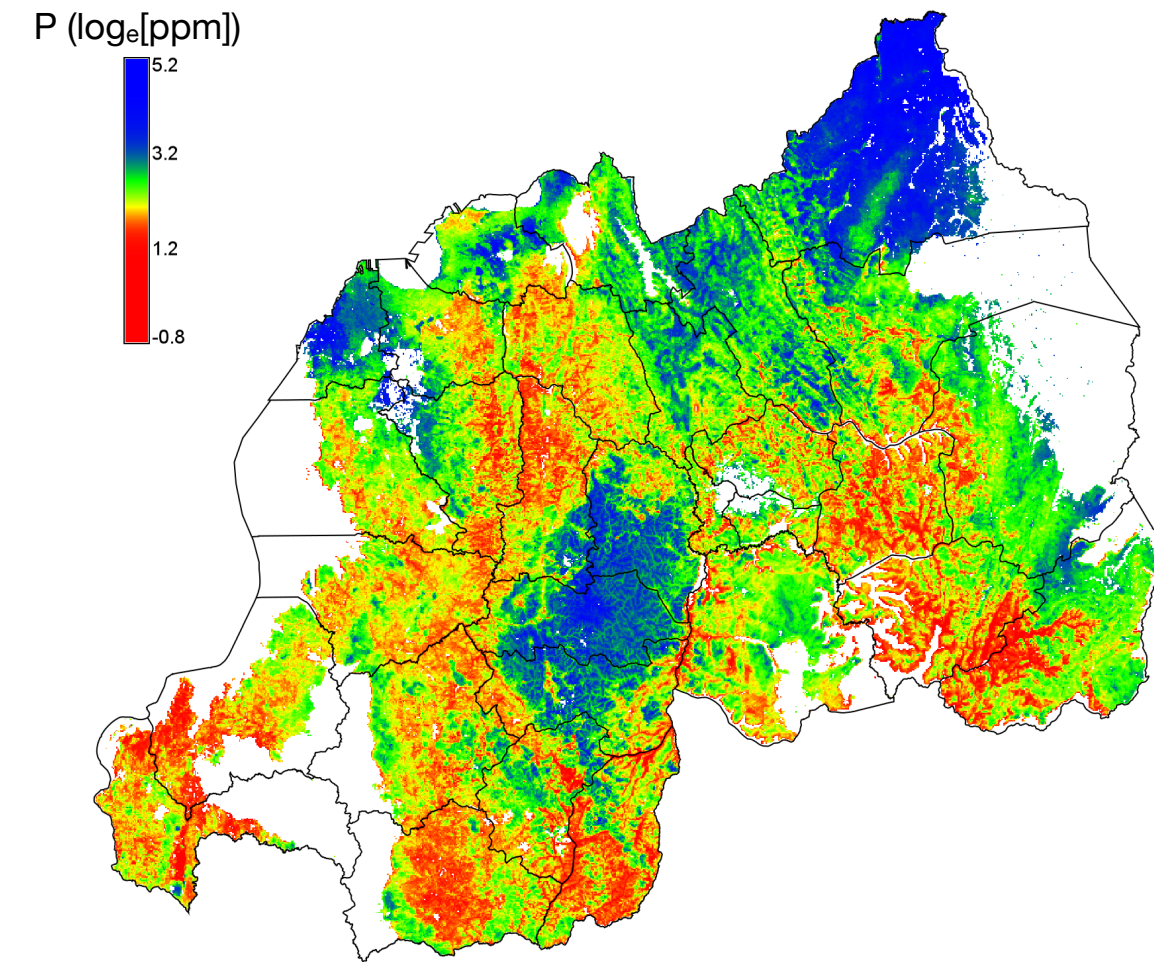
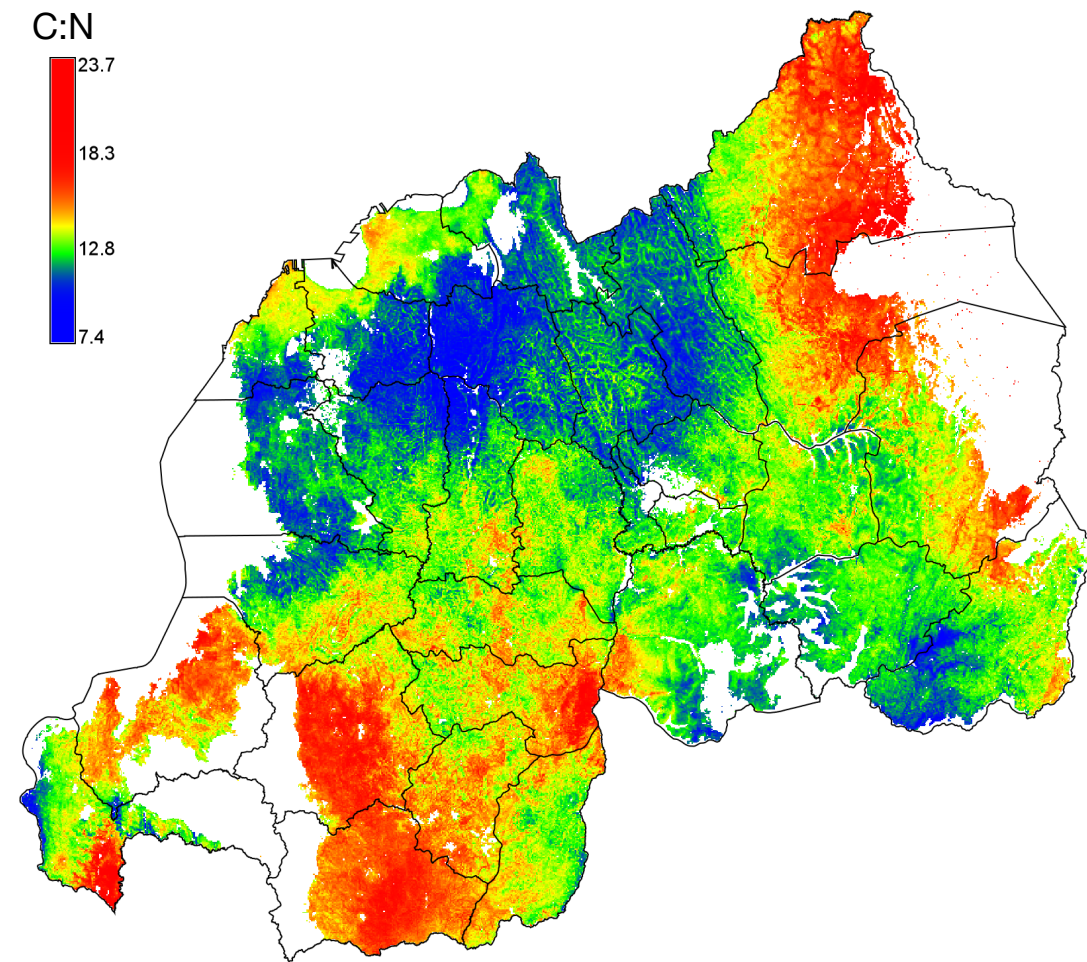


90% CI



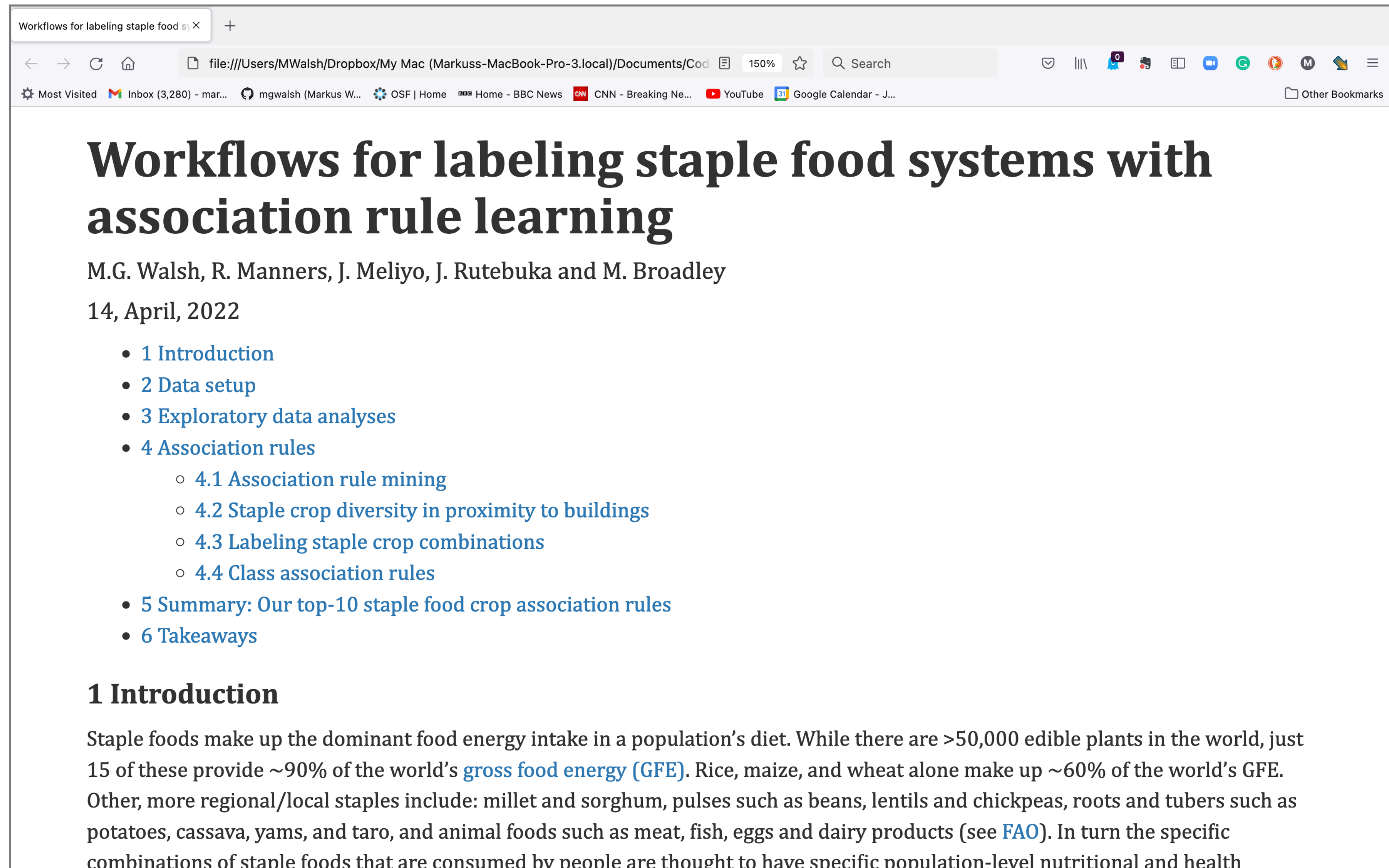
Predictive soil mapping

see the notebook at: <https://osf.io/hpebu/>



Staple food crop association rules

download the notebook at: <https://osf.io/hpebu/>



Workflows for labeling staple food systems with association rule learning

M.G. Walsh, R. Manners, J. Meliyo, J. Rutebuka and M. Broadley

14, April, 2022

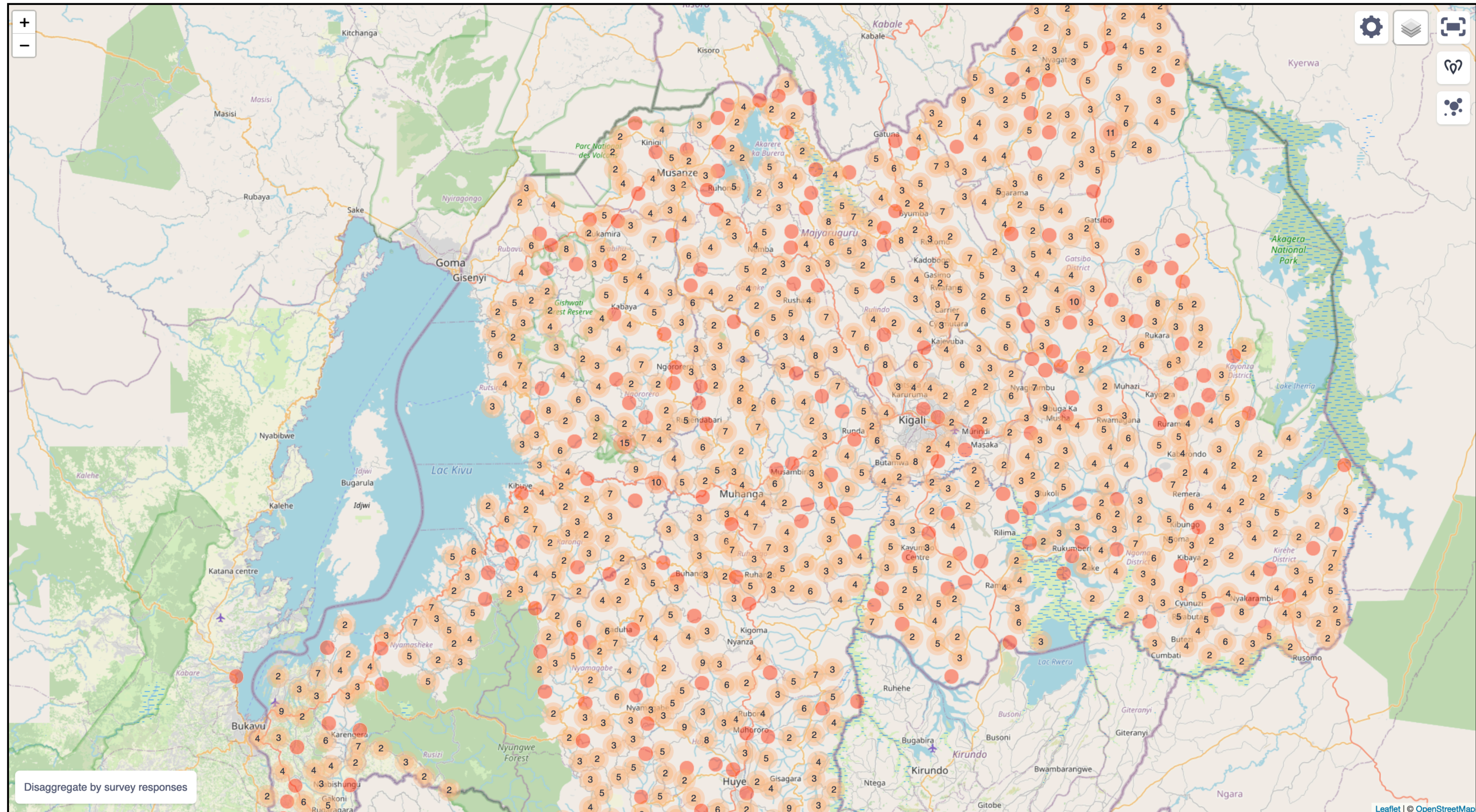
- [1 Introduction](#)
- [2 Data setup](#)
- [3 Exploratory data analyses](#)
- [4 Association rules](#)
 - [4.1 Association rule mining](#)
 - [4.2 Staple crop diversity in proximity to buildings](#)
 - [4.3 Labeling staple crop combinations](#)
 - [4.4 Class association rules](#)
- [5 Summary: Our top-10 staple food crop association rules](#)
- [6 Takeaways](#)

1 Introduction

Staple foods make up the dominant food energy intake in a population's diet. While there are >50,000 edible plants in the world, just 15 of these provide ~90% of the world's [gross food energy \(GFE\)](#). Rice, maize, and wheat alone make up ~60% of the world's GFE. Other, more regional/local staples include: millet and sorghum, pulses such as beans, lentils and chickpeas, roots and tubers such as potatoes, cassava, yams, and taro, and animal foods such as meat, fish, eggs and dairy products (see [FAO](#)). In turn the specific combinations of staple foods that are consumed by people are thought to have specific population-level nutritional and health

Rwanda collocated soil and crop survey locations (Apr - Aug, 2021)

raw data available at: <https://kobo.humanitarianresponse.info/#/forms/a3jwzkBTtDvdaZJgR6YjAe>



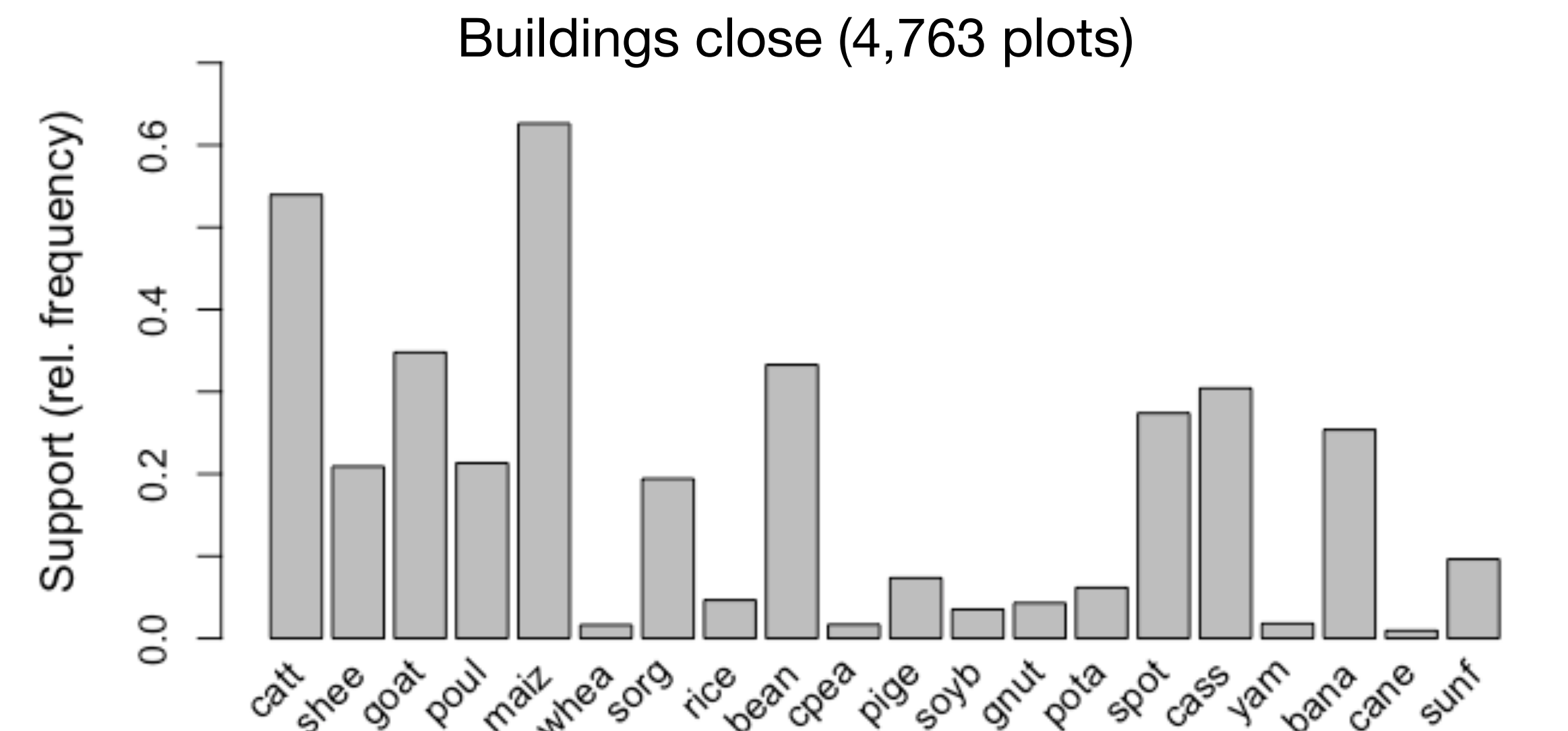
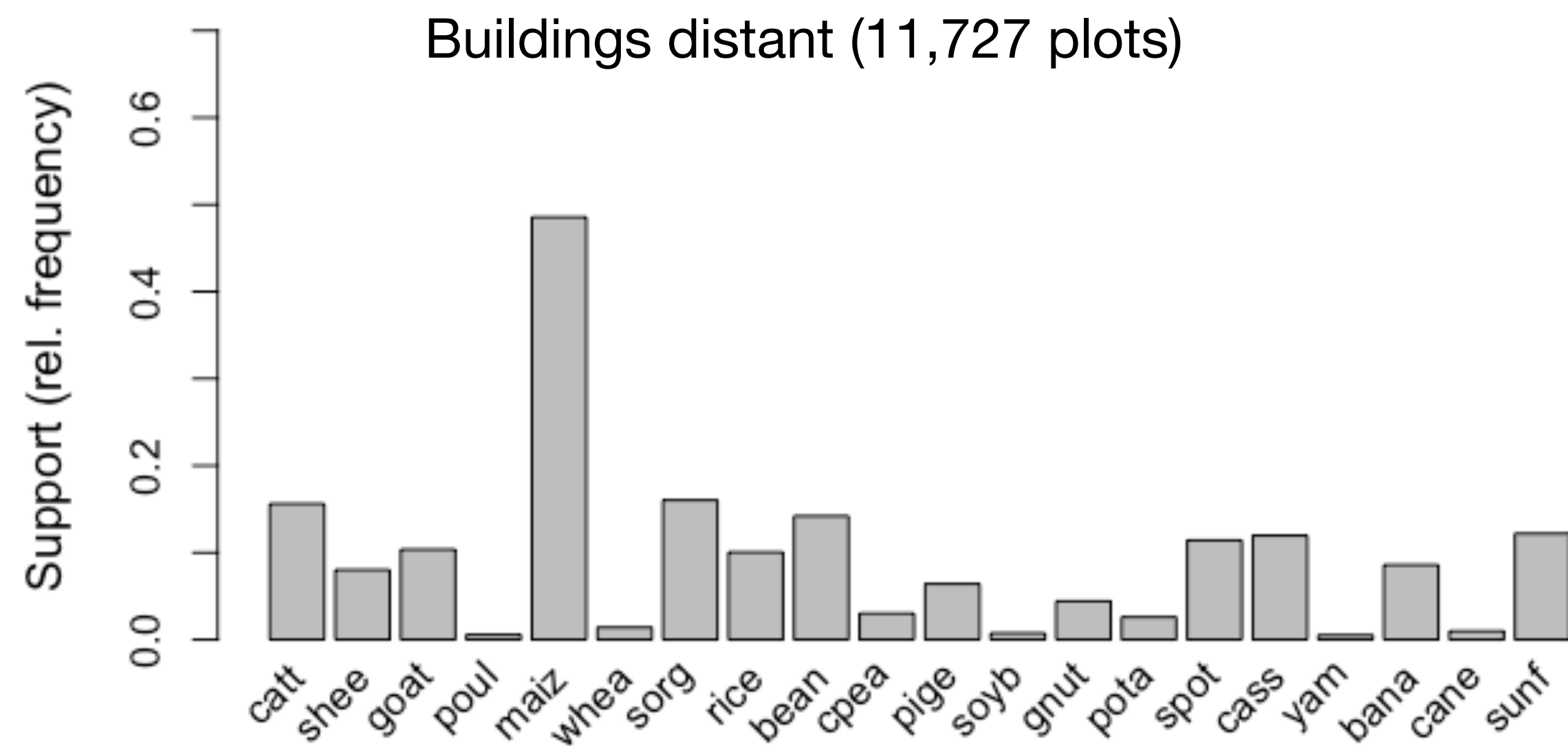
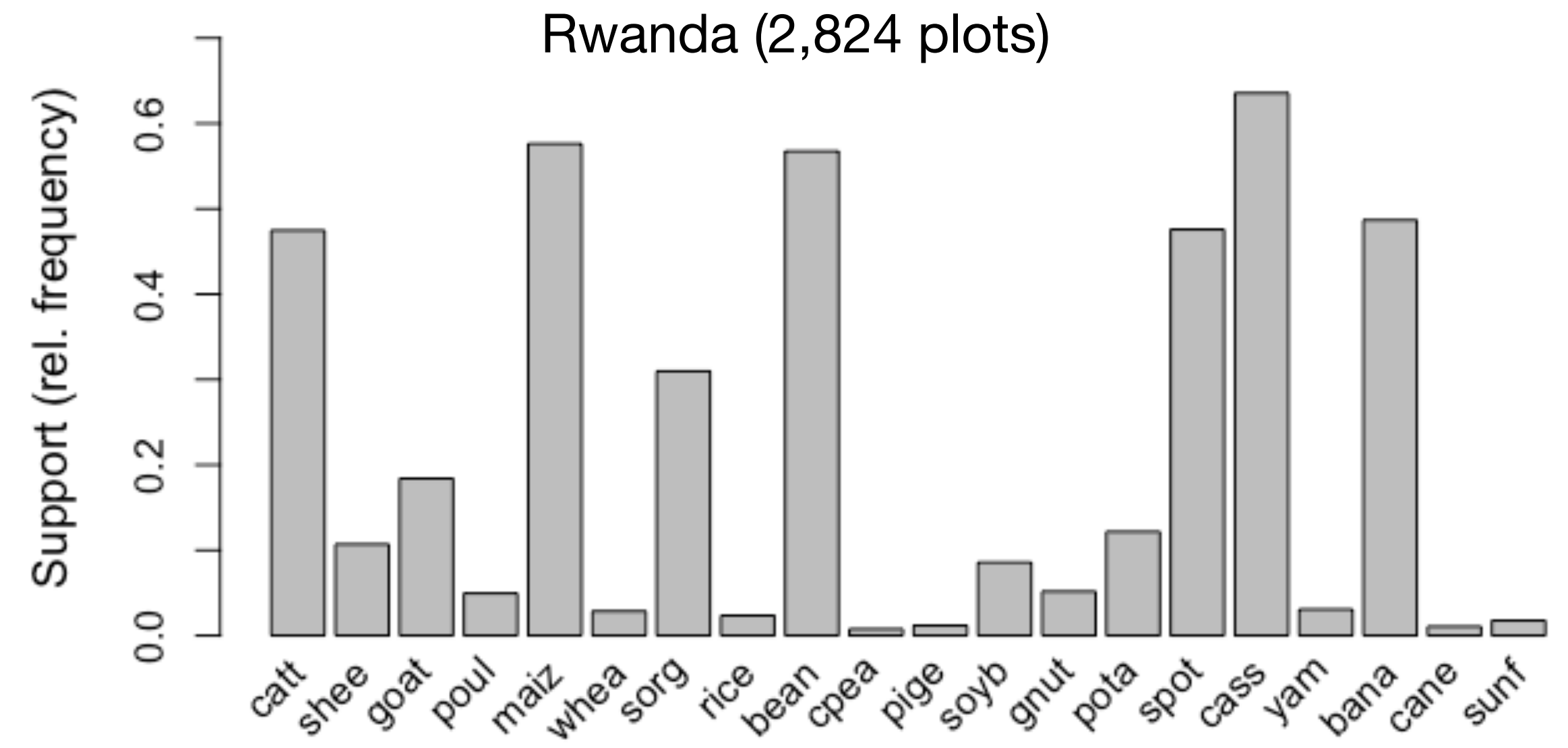
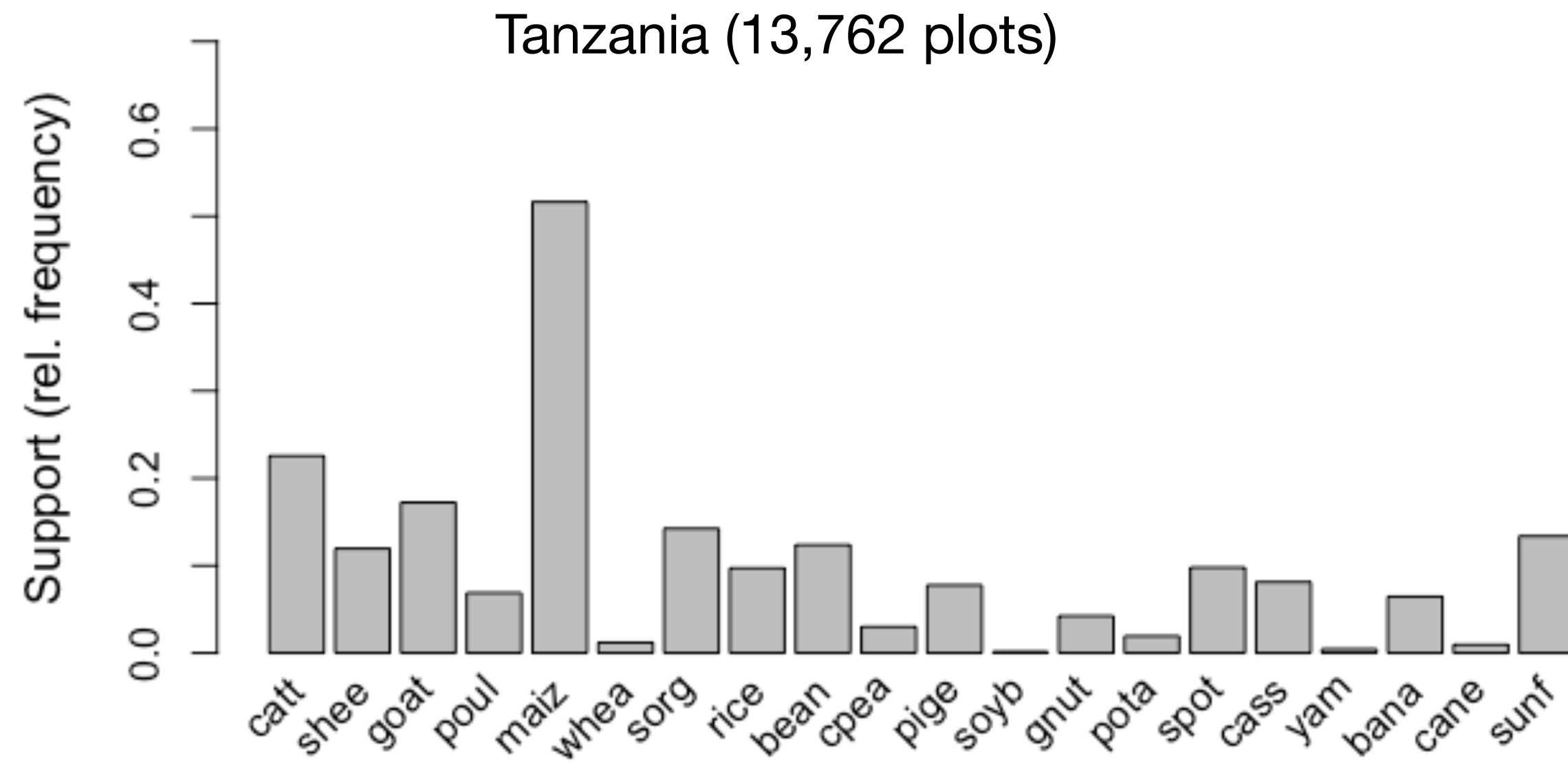
Staple food crop diversity in Rwanda (2,824 plots currently)

georeferenced, field tagged data & photos available at: <https://osf.io/hpebu/>



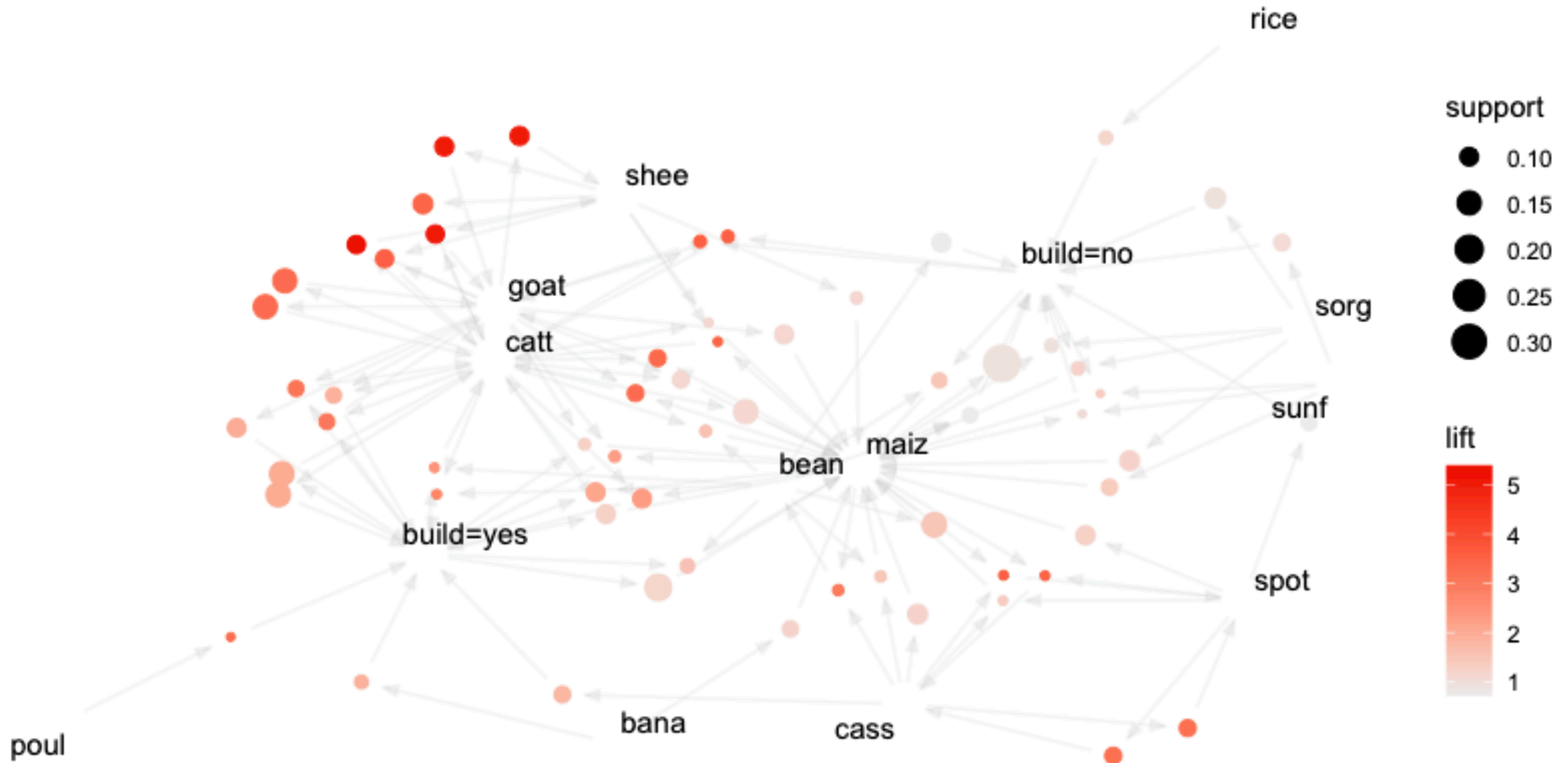
Staple food crop “market basket” analyses

notebook at: <https://osf.io/hpebu/>



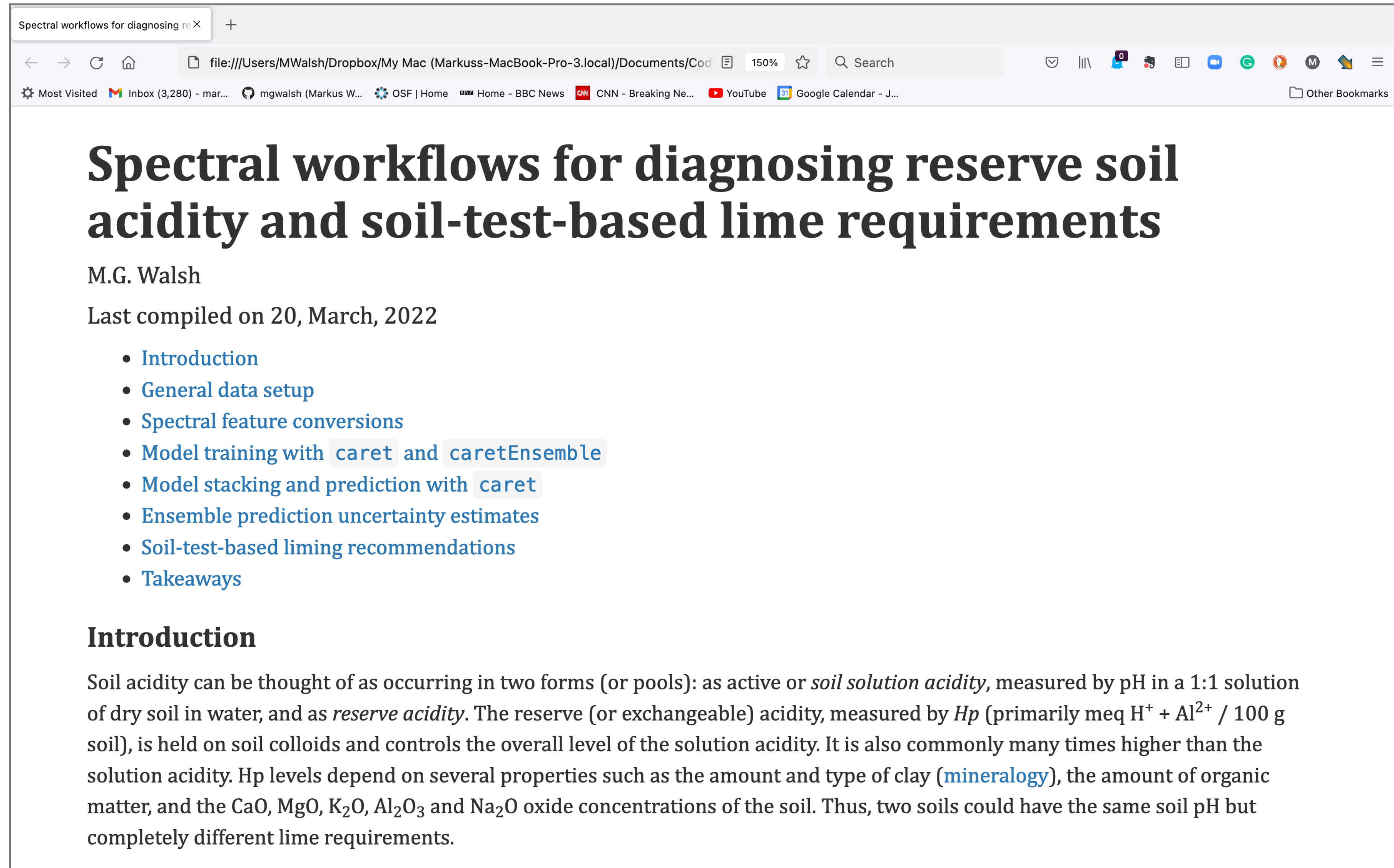
Association rule mapping of main staple food cropping systems

notebook at: <https://osf.io/hpebu/>



Spectral prediction of lime requirements

download the notebook at: <https://osf.io/2v46w/>



Spectral workflows for diagnosing re X

file:///Users/MWalsh/Dropbox/My Mac (Markuss-MacBook-Pro-3.local)/Documents/Cod 150% Search

Most Visited Inbox (3,280) - mar... mgwalsh (Markus W... OSF | Home Home - BBC News CNN - Breaking Ne... YouTube Google Calendar - J... Other Bookmarks

Spectral workflows for diagnosing reserve soil acidity and soil-test-based lime requirements

M.G. Walsh

Last compiled on 20, March, 2022

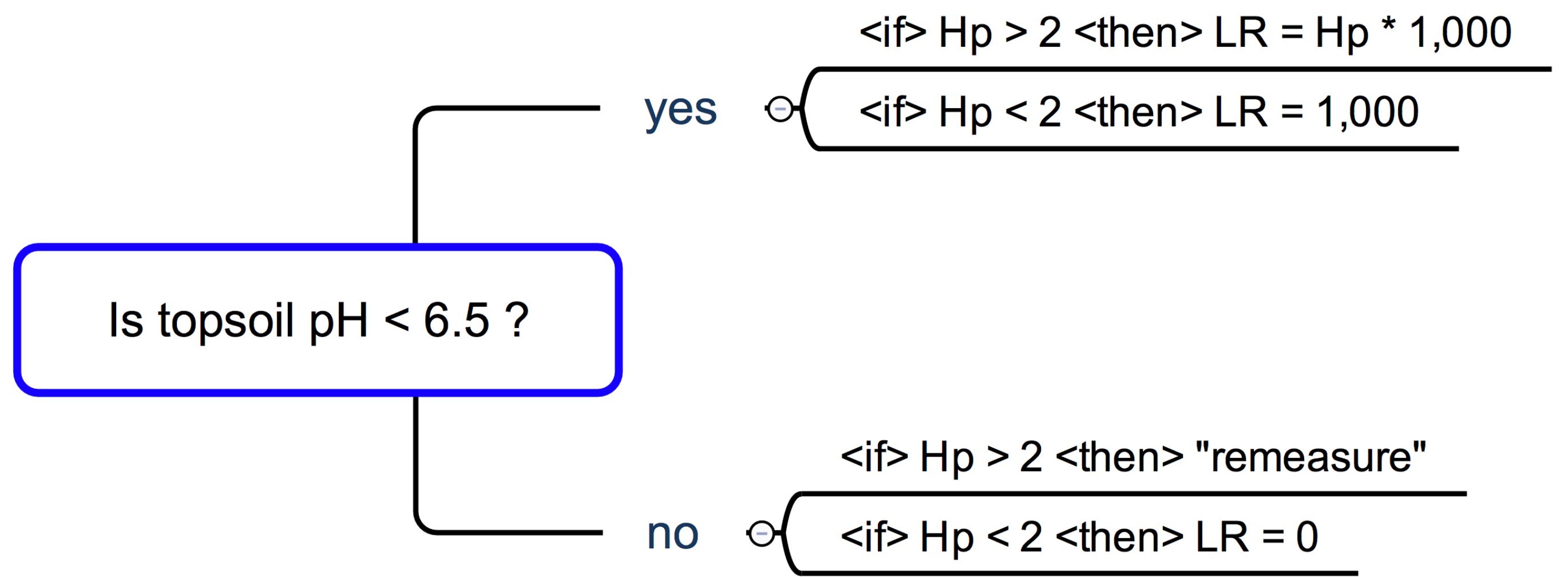
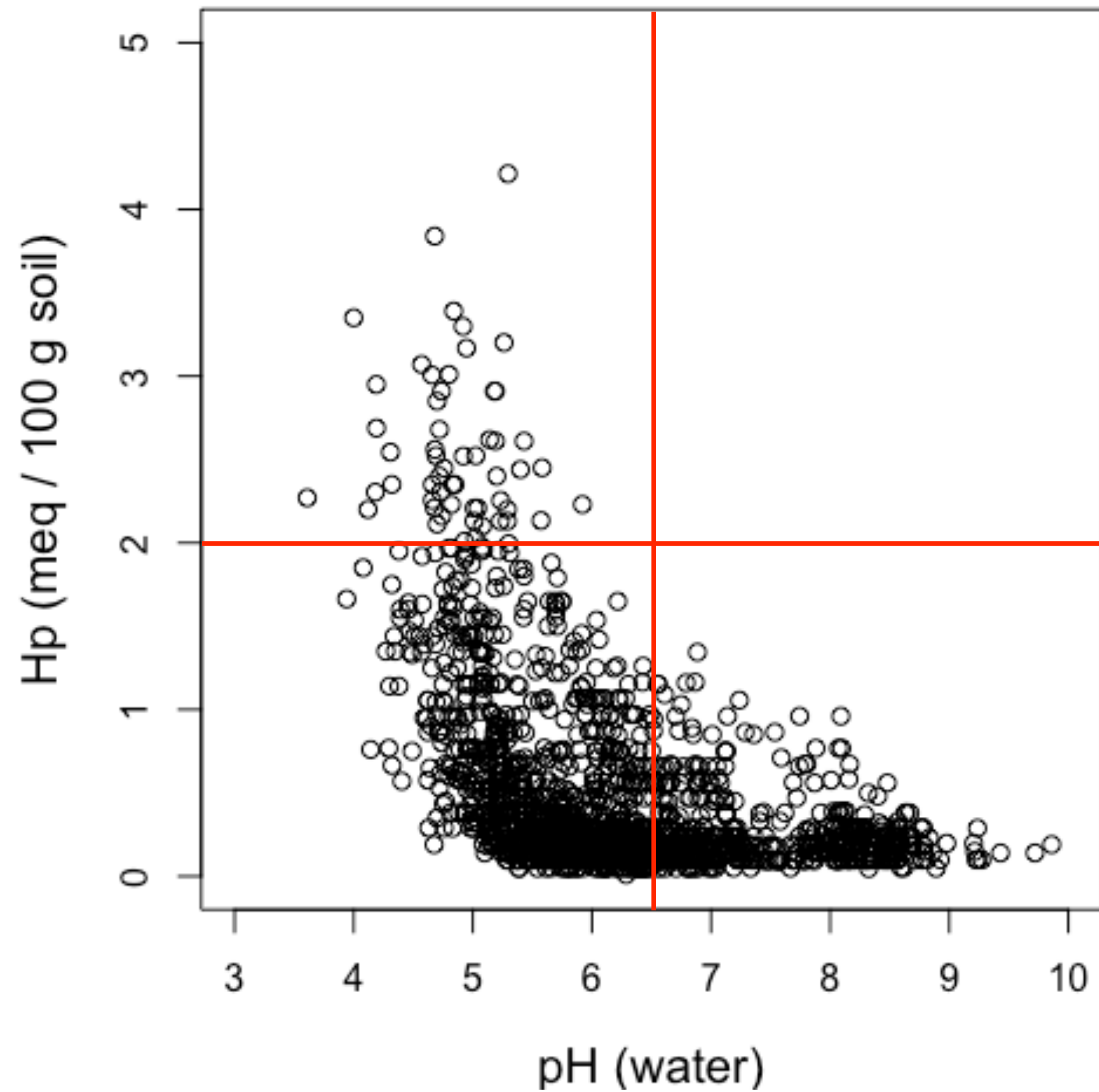
- [Introduction](#)
- [General data setup](#)
- [Spectral feature conversions](#)
- [Model training with `caret` and `caretEnsemble`](#)
- [Model stacking and prediction with `caret`](#)
- [Ensemble prediction uncertainty estimates](#)
- [Soil-test-based liming recommendations](#)
- [Takeaways](#)

Introduction

Soil acidity can be thought of as occurring in two forms (or pools): as active or *soil solution acidity*, measured by pH in a 1:1 solution of dry soil in water, and as *reserve acidity*. The reserve (or exchangeable) acidity, measured by H_p (primarily $\text{meq H}^+ + \text{Al}^{2+} / 100 \text{ g soil}$), is held on soil colloids and controls the overall level of the solution acidity. It is also commonly many times higher than the solution acidity. H_p levels depend on several properties such as the amount and type of clay ([mineralogy](#)), the amount of organic matter, and the CaO , MgO , K_2O , Al_2O_3 and Na_2O oxide concentrations of the soil. Thus, two soils could have the same soil pH but completely different lime requirements.

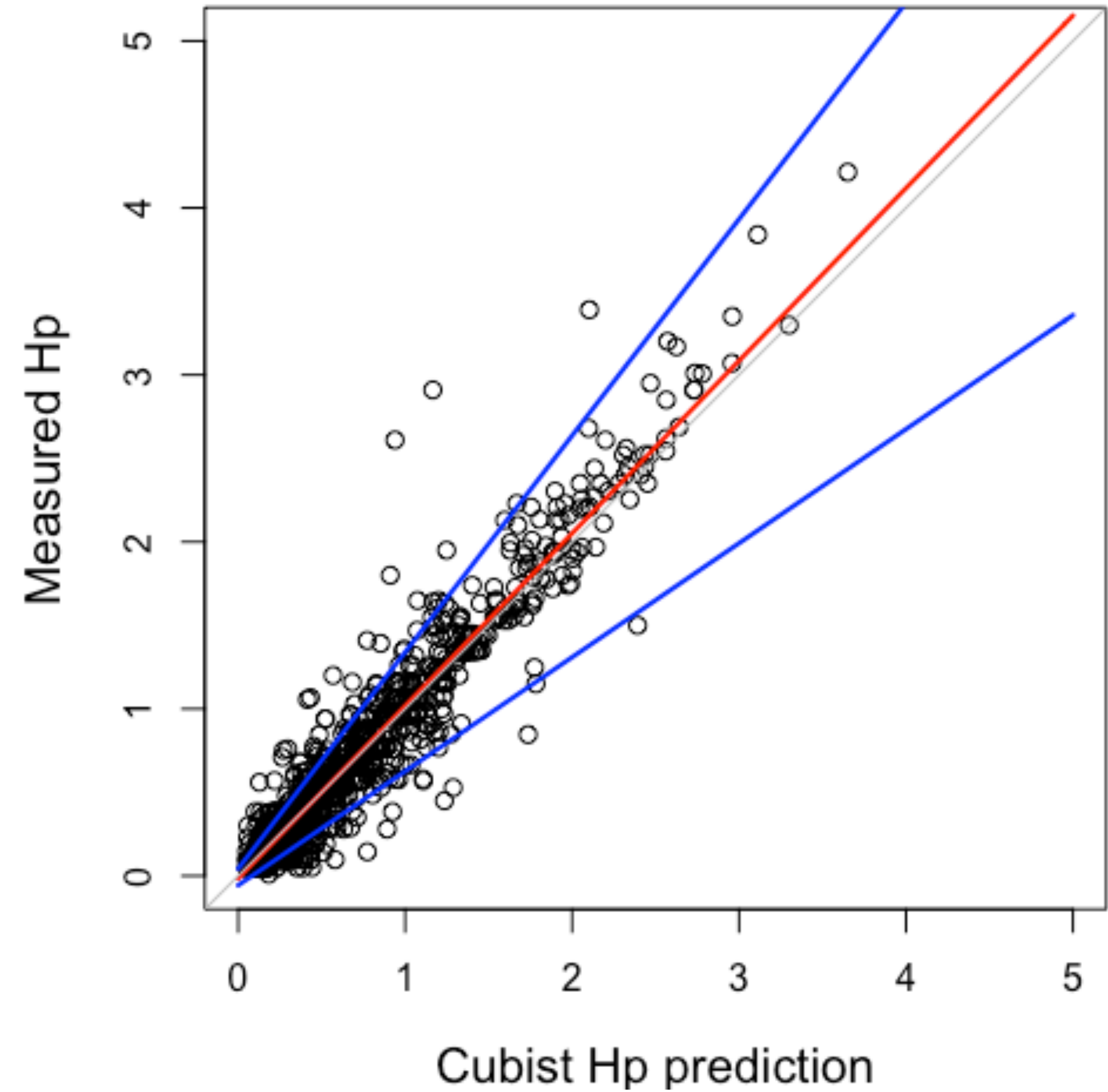
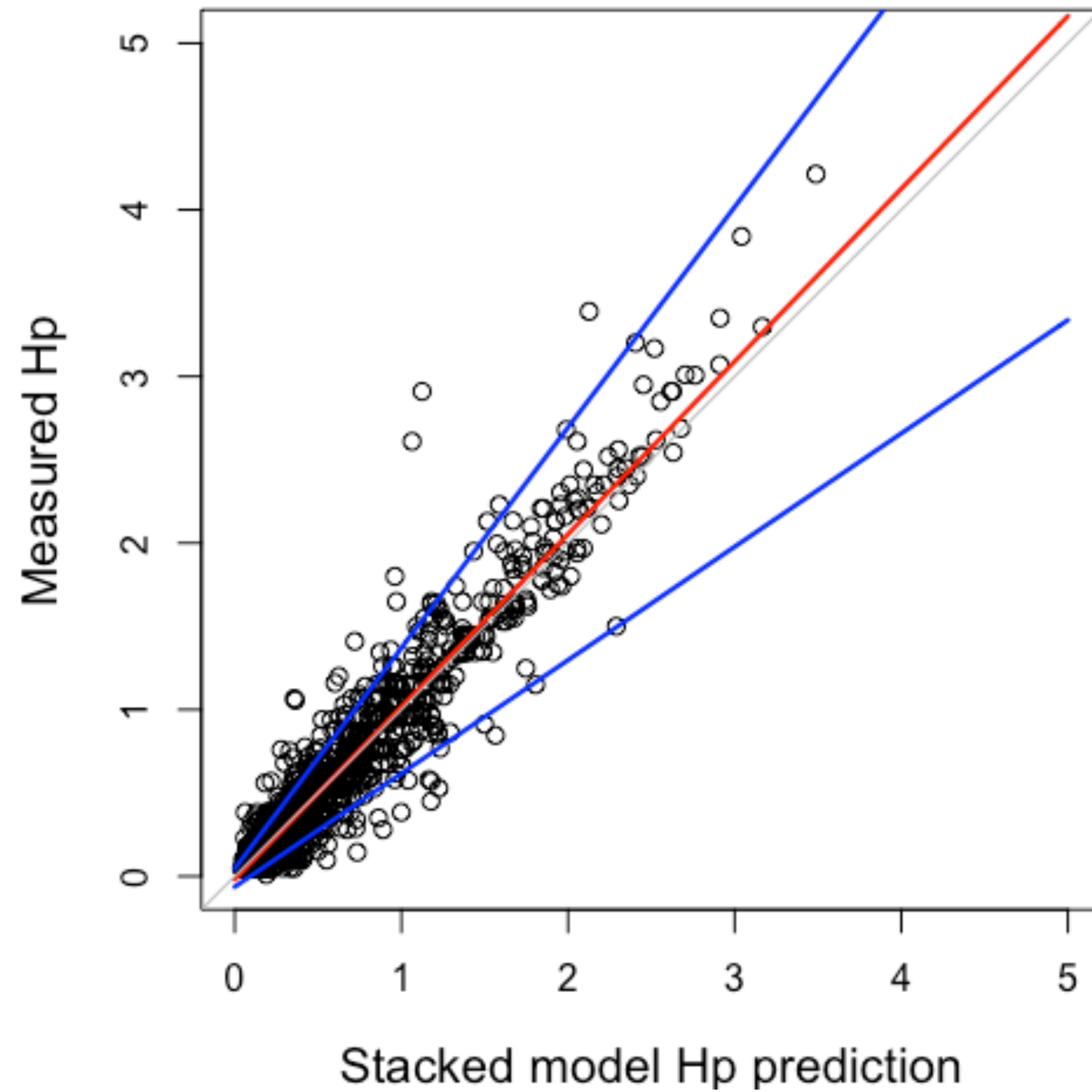
Soil-test based lime requirement heuristics

notebook at: <https://osf.io/2v46w/>



Soil reserve acidity predictions from spectral measurements

notebook at: <https://osf.io/2v46w/>



Landscape soil aggregate stability ratings

download the notebook at: <https://osf.io/q6ste/>

Rating landscape soil aggregate stability

file:///Users/MWalsh/Dropbox/My Mac (Markuss-MacBook-Pro-3.local)/Documents/Cod 150% Search

Most Visited Inbox (3,280) - mar... mgwalsh (Markus W... OSF | Home Home - BBC News CNN - Breaking Ne... YouTube Google Calendar - J... Other Bookmarks

Rating landscape soil aggregate stability from laser diffraction particle size data

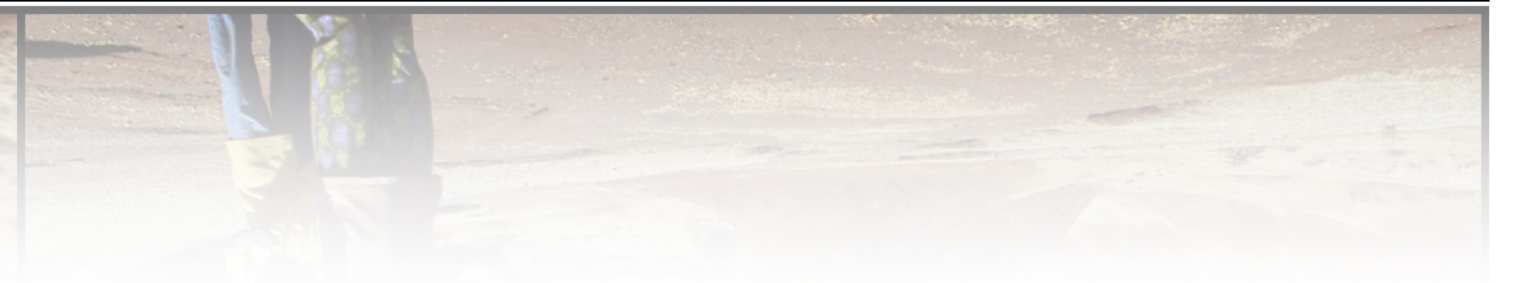
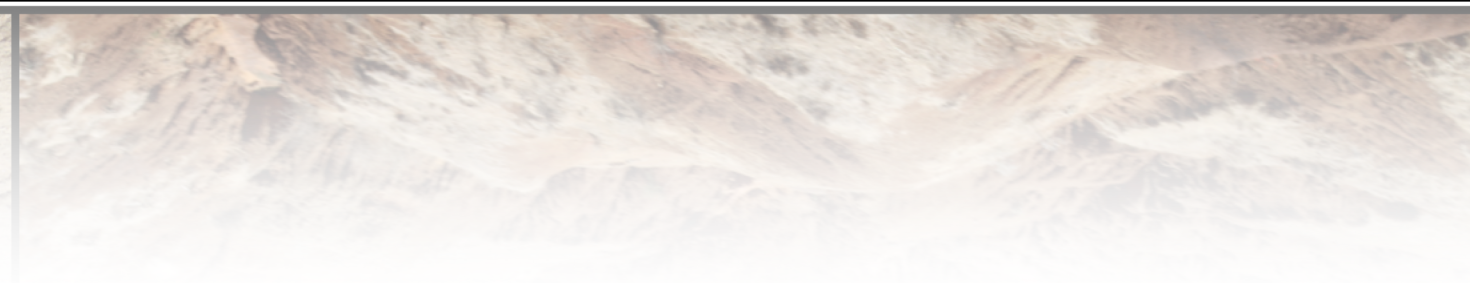
M.G. Walsh and A. Sila

Last compiled: 09, April, 2022

- [1 Introduction](#)
- [2 Initial data setup](#)
 - [2.1 Data dictionary](#)
 - [2.2 Spatial distribution of soil samples](#)
 - [2.3 Soil particle size fractions by treatment combinations](#)
 - [2.4 Encoding treatment effects with cumulative link indices](#)
- [3 Landscape-level soil aggregate stabilities](#)
 - [3.1 Fixed-effects](#)
 - [3.2 Mixed-effects](#)
- [4 Takeaways and next steps](#)

1 Introduction

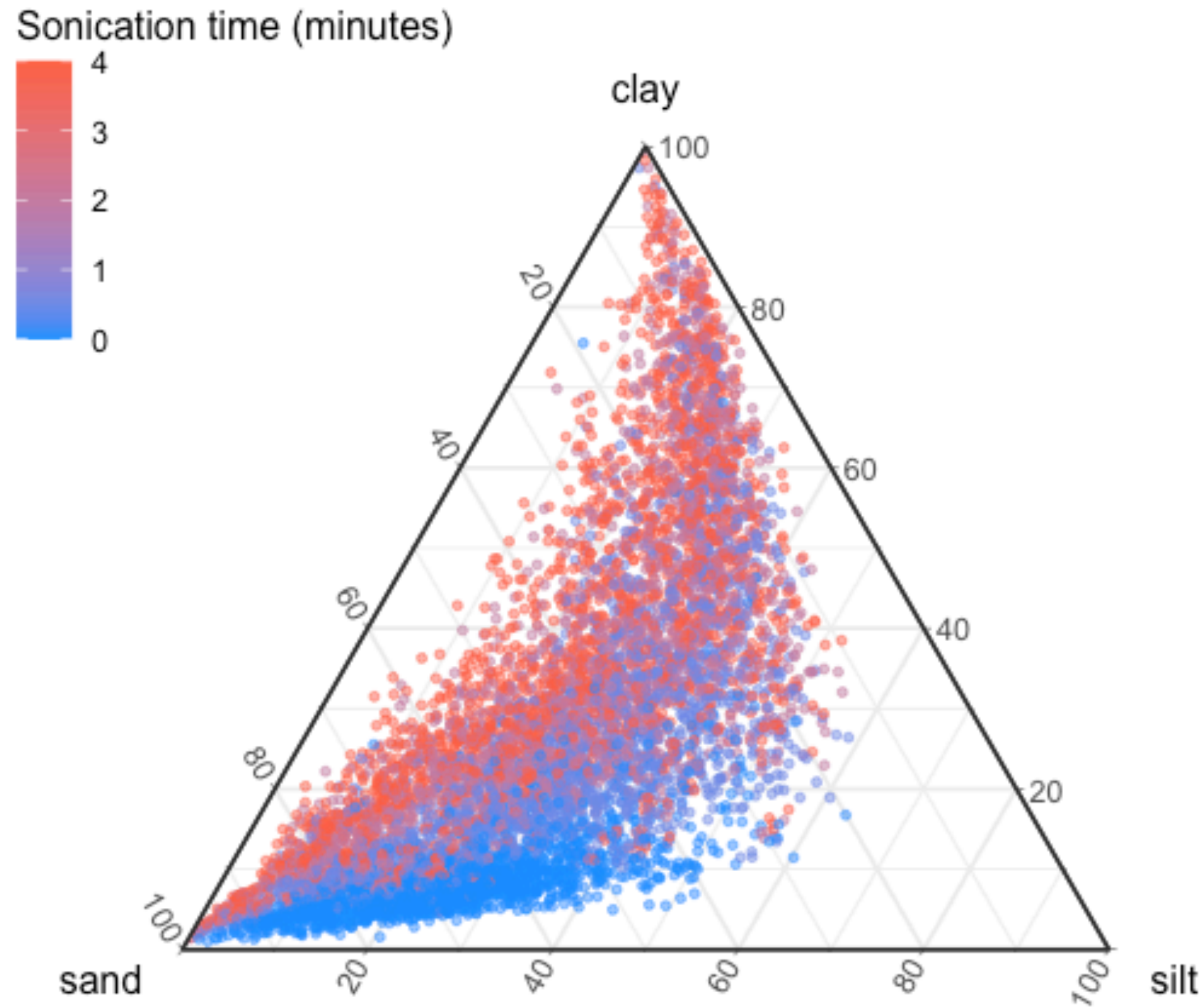
Soil aggregate stability refers to the ability of soil aggregates to resist disintegration when disruptive forces associated with e.g., tillage operations and erosion occur. Aggregate stability indicates how well soils can resist compaction, wind abrasion, rainfall detachment and atmospheric and/or overland transport. It is a dynamic soil physical chemistry property, which is important for water infiltration, retention and drainage, soil aeration, microbial activity, organic matter storage and stabilization and plant root growth, among others. When soil aggregates disintegrate e.g. during tillage operations or rainstorms, dispersed particles fill soil pore



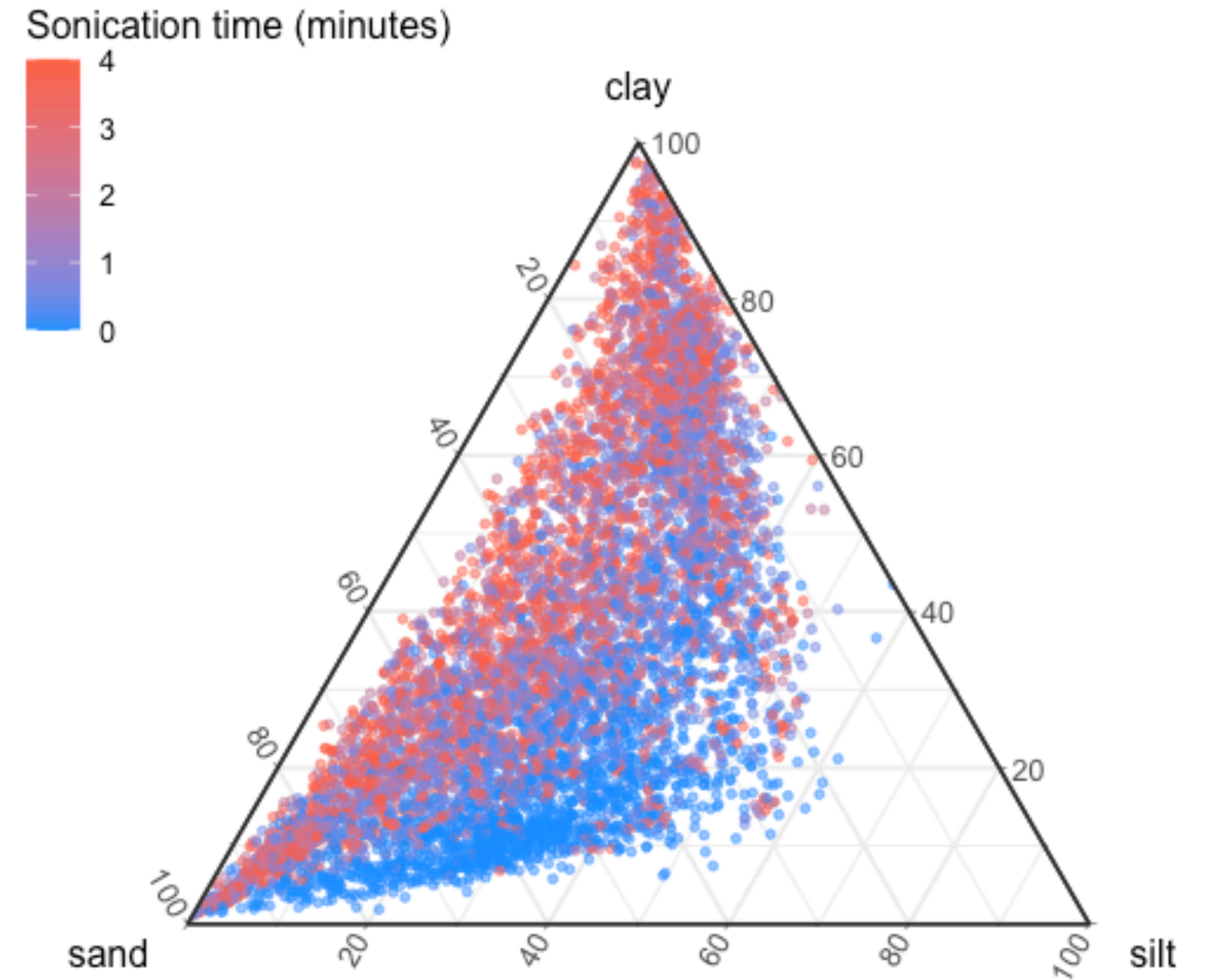
Africa-wide Laser Diffraction Particle Size Analysis data

notebook at: <https://osf.io/q6ste/>

Dispersed in calgon

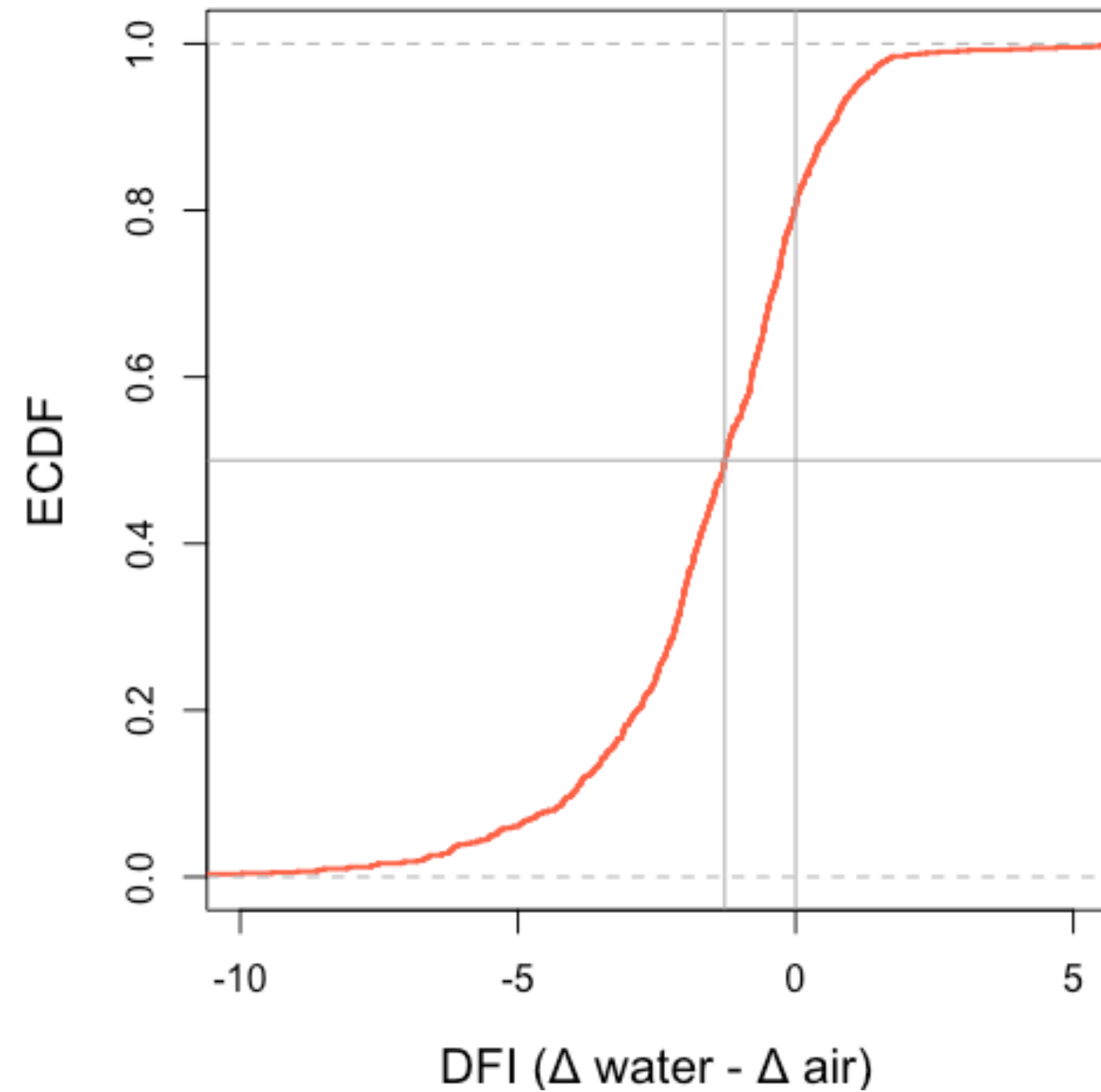
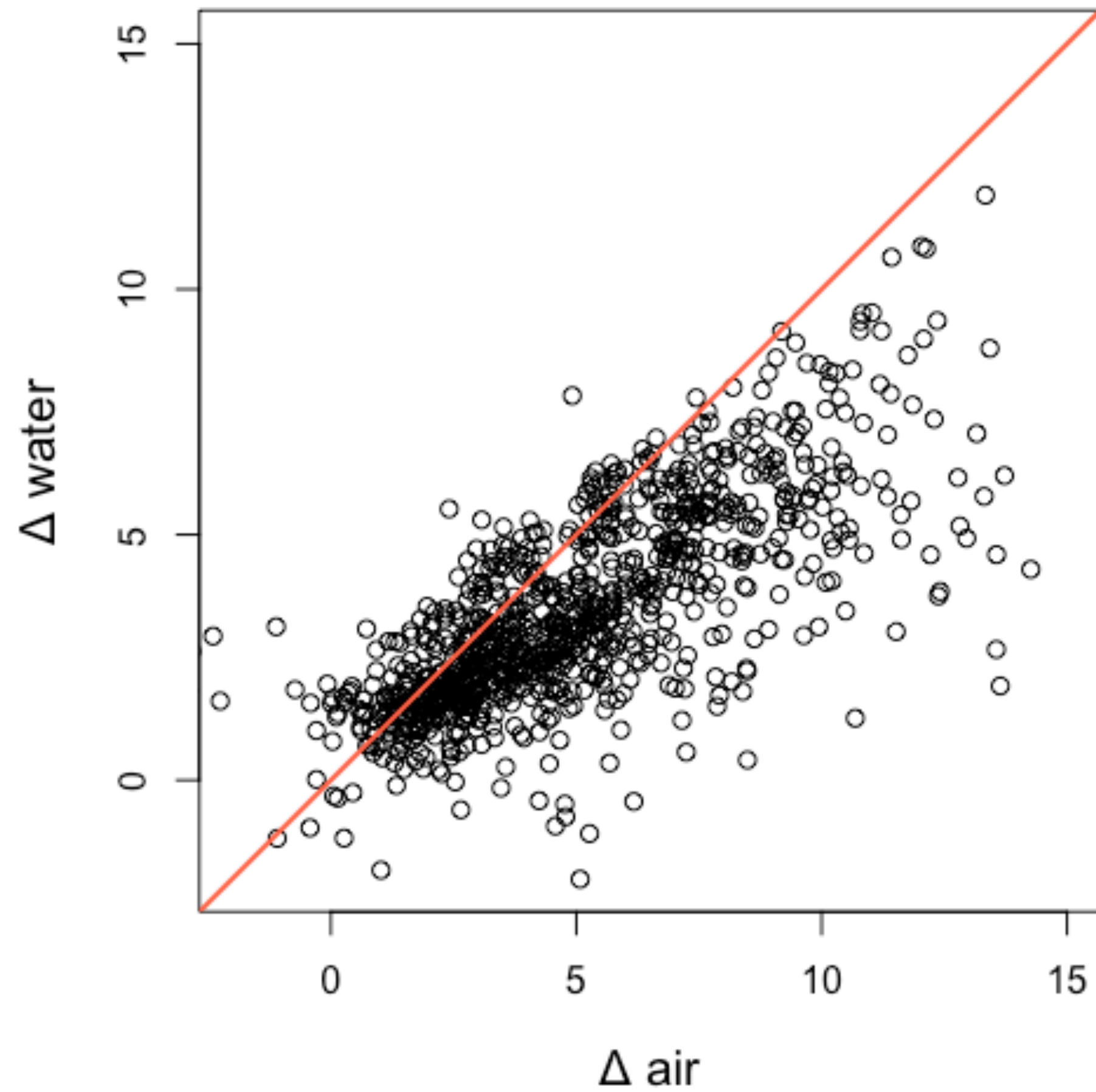


Dispersed in water



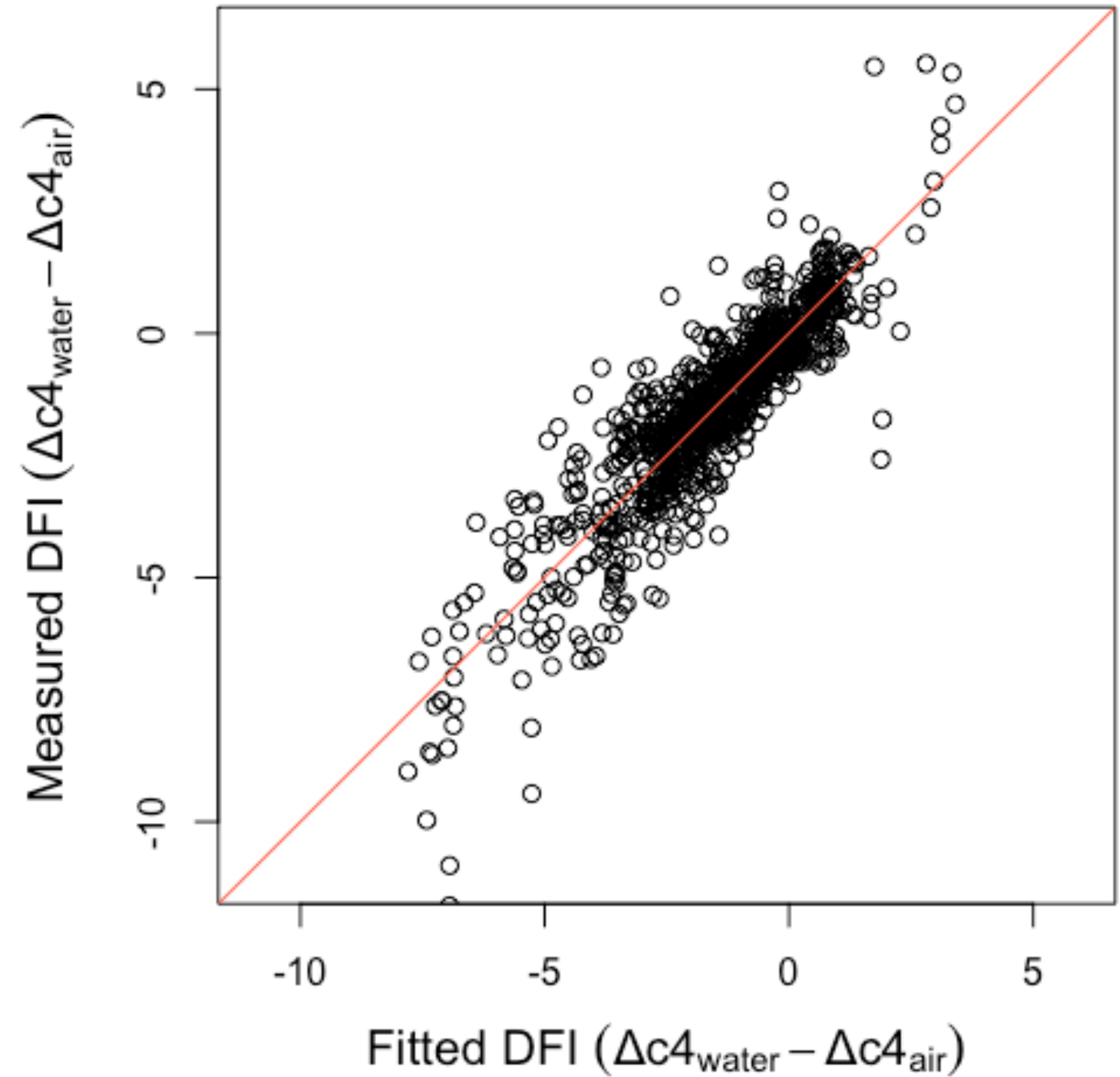
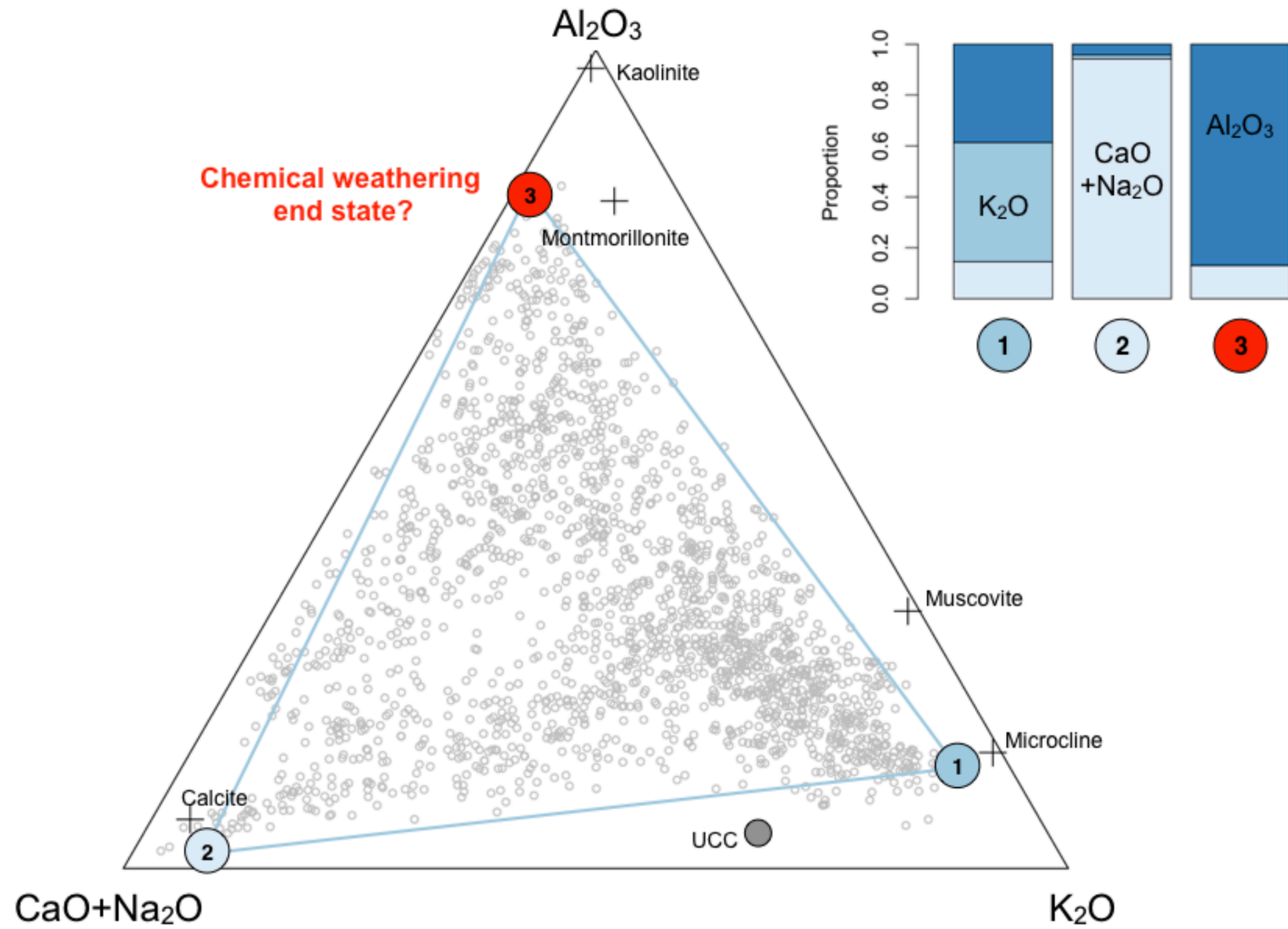
Dispersion / Flocculation Index (DFI, relative to calgon treatment)

notebook at: <https://osf.io/q6ste/>



Relationship between soil weathering (metal oxide) indicators and DFI

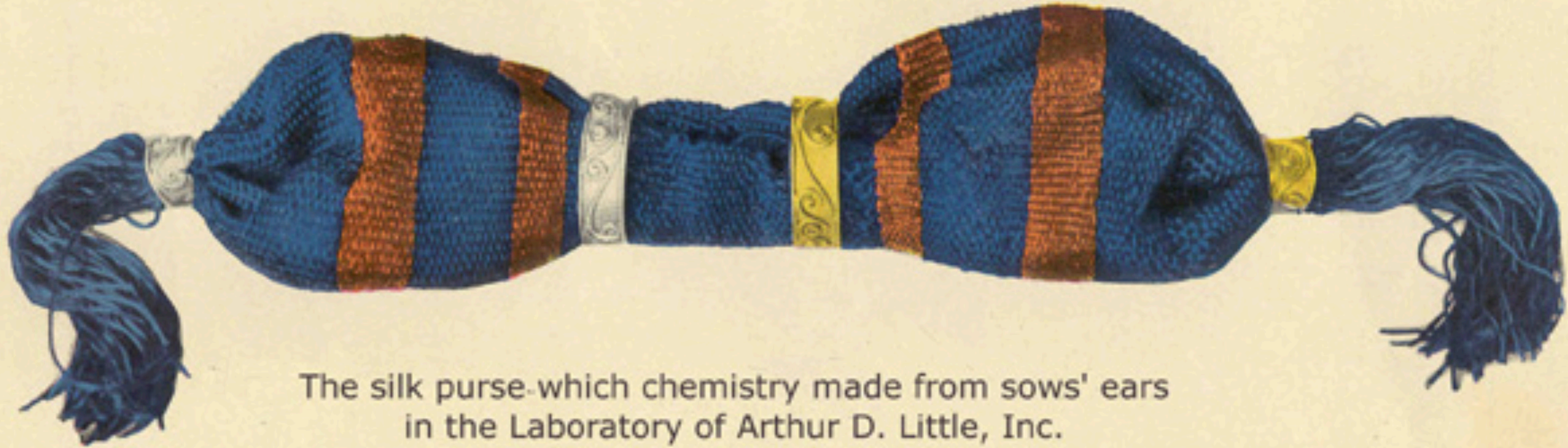
notebook at: <https://osf.io/q6ste/>



Takeaways

- ▶ Complex projects such as RwaSIS and GAIA should maintain all their data, code, bibliographies, documents, reports, publications, etc. available in one easily accessible place (e.g. at: <https://osf.io/>).
- ▶ Use Rmarkdown (and/or Jupyter) notebooks to document and version all of your data work including import, wrangling, prediction and reproducible reporting/publication in an interoperable project lifecycle framework.
- ▶ Predictive models typically provide far better accuracies for classification and regression tasks than data models. They also occasionally yield better information about any underlying data generating mechanisms.

About legacy data ...



The silk purse-which chemistry made from sows' ears
in the Laboratory of Arthur D. Little, Inc.