Forest Monitoring and Improvement Program
Project 4: Baselines, drivers and trends in soil health and stability
DPIE (Science) and University of Sydney, 30 November 2020

Thomas Bishop[1], Sabastine Ugbaje[2]

[1] School of Life & Environmental Science, The University of Sydney
[2] Sydney Informatics Hub, The University of Sydney

Milestone 5: Collation of data cube for forests of NSW and associated R/Python script.

1. Introduction

This report presents the progress made towards achieving Milestone 5, which involves activities for the collation of a space-time data cube and associated scripts for soil carbon modelling in the forest regions of NSW.

The report discusses the following:

- Collated data and gaps
- Exploratory data analysis and initial modelling results

2. Data cube collation

The primary indicator of soil health adopted for this project is soil organic carbon (SOC). We have identified and collated potential space-time predictors of this variable. Table 1 gives an overview of the description of the datasets processed and collated into a data cube for modelling.

Table 1: Overview of data collated for modelling

Data type	Covariate	Source	Resolution	Note
Response	Soil organic carbon (SOC)	SALIS	-	Surface soil
Spatial	DEM, slope	Geoscience Australia	90 m	
	Topographic Wetness Index (TWI), Multi-resolution Valley Bottom Flatness (MrVBF)	ASRIS	90 m	
	Gama-radiometric data • Potassium • Uranium • Thorium • Radiation dose	SLGA	90 m	
	Silica	-	~100 m	
	Clay % (0-5)	SLGA	90 m	
Spatial and temporal	Precipitation	SILO	5 km, monthly	
	Temperature (min and max)	SILO	5 km, daily	

	Solar radiation	SILO	5 km, daily
	NDVI	LANDSAT	30 m, 16-day

Data cube: The data cube consists of SOC measurements, the month and year of profile sampling, as well as the space, and space and time covariates information for the soil profile locations.

In modelling and mapping soil attribute from legacy data such as the one used in this project, certain known sources of variation like differences in depth characteristics of the soil profiles, as well as the use of different analytical methods between laboratories and survey campaigns need to be considered carefully. We dealt with depth variation by extracting SOC measurements for the surface horizon only, since this layer is impacted the most by natural and human disturbances. However, we included the thickness of the surface layer in the data cube to account for the location-to-location variation in horizon thickness. We also accounted for variation in SOC measurements arising from the differences in analytical methods by including a field in the data cube indicating what SOC analytical method was used.

All time-varying covariates (NDVI and climate variables) were aggregated to monthly values. Since the effect of these covariates on soil health dynamics depends on current and past conditions, we applied a weighted aggregation algorithm to sixty months (5 years) of the covariate timeseries prior to when the soil profile was sampled. The algorithm attaches more weight to the most recent observations. Feature extraction by this method instead of taking the mean value over the last 5 years has been shown to create more interpretable and more accurate models (Wimalathunge and Bishop 2019).

Another key feature of the data cube is the incorporation of the proxy for natural and anthropogenic disturbances of SOC dynamics at the time of soil profile sampling compared to conditions at discrete times in the past. The potential effect of these disturbances is represented in the data cube by incorporating NDVI difference features wherein the NDVI of the previous 1, 2, 3,6, and 12 months are subtracted from the NDVI of the month of profile sampling. For example, if there was a fire 2 months ago, we would expect there would be a drop in NDVI between the NDVI today and 3 months ago. This could be useful in situations where we don't have fire or logging spatial data.

While most of the pre-processing and extraction was done in R, a large portion of the NDVI preparation was performed in Google Earth Engine (GEE). This is more efficient and circumvents the need of downloading relatively large amount of Landsat scenes, thus freeing up local storage space. The GEE Java scripts used for the NDVI extraction are available and can be to be used to generate all NDVI covariates.

Data gaps: The potential predictors of SOC assembled in the data cube are by no means exhaustive, thus there will be the need to include additional covariates, particularly to improve model performance. For example, recently, some disturbances layers have been provided and additional data are expected. These datasets will soon be included in the cube, but the general workflow has been established.

3. Exploratory data analysis and initial modelling

This part presents and discusses the spatial distribution of the profile data in the SALIS database, as well as exploratory analysis of the data cube. The section also discusses initial modelling results applying a global model trained from all data points in NSW. It is a prototype model to establish the process which can be improved upon.

3.1 Spatial distribution of soil observations

Figures 1-3 present data for locations where surface organic carbon has been measured and are based on all datasets held by DPI-E. Figures 1 and 2 present their spatial distribution for NSW and the Regional Forest Agreement (RFA) regions for different time slices. Figure 3 presents the same data but as number of observations per RFA region with different time slices.

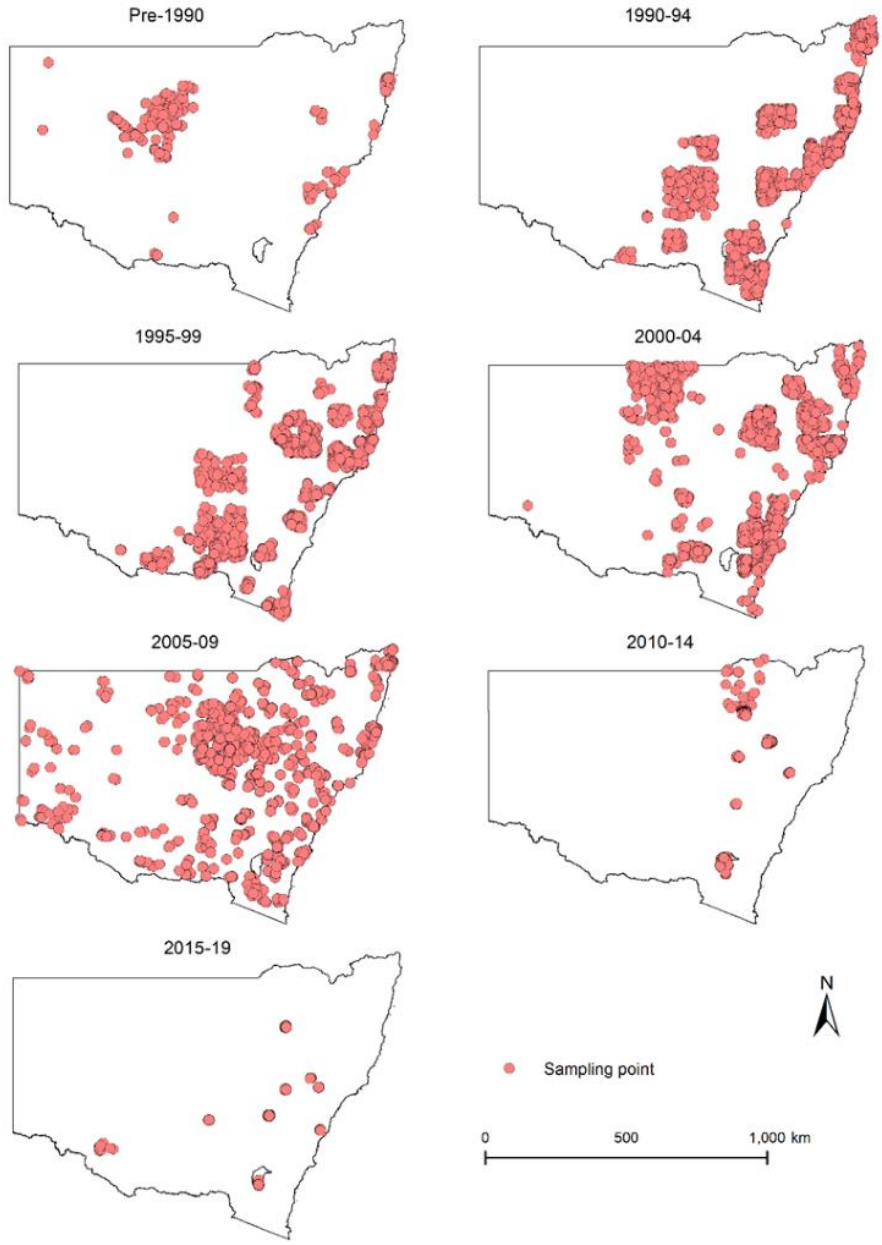

Figure 1. Spatial distribution of surface organic carbon measurements across NSW

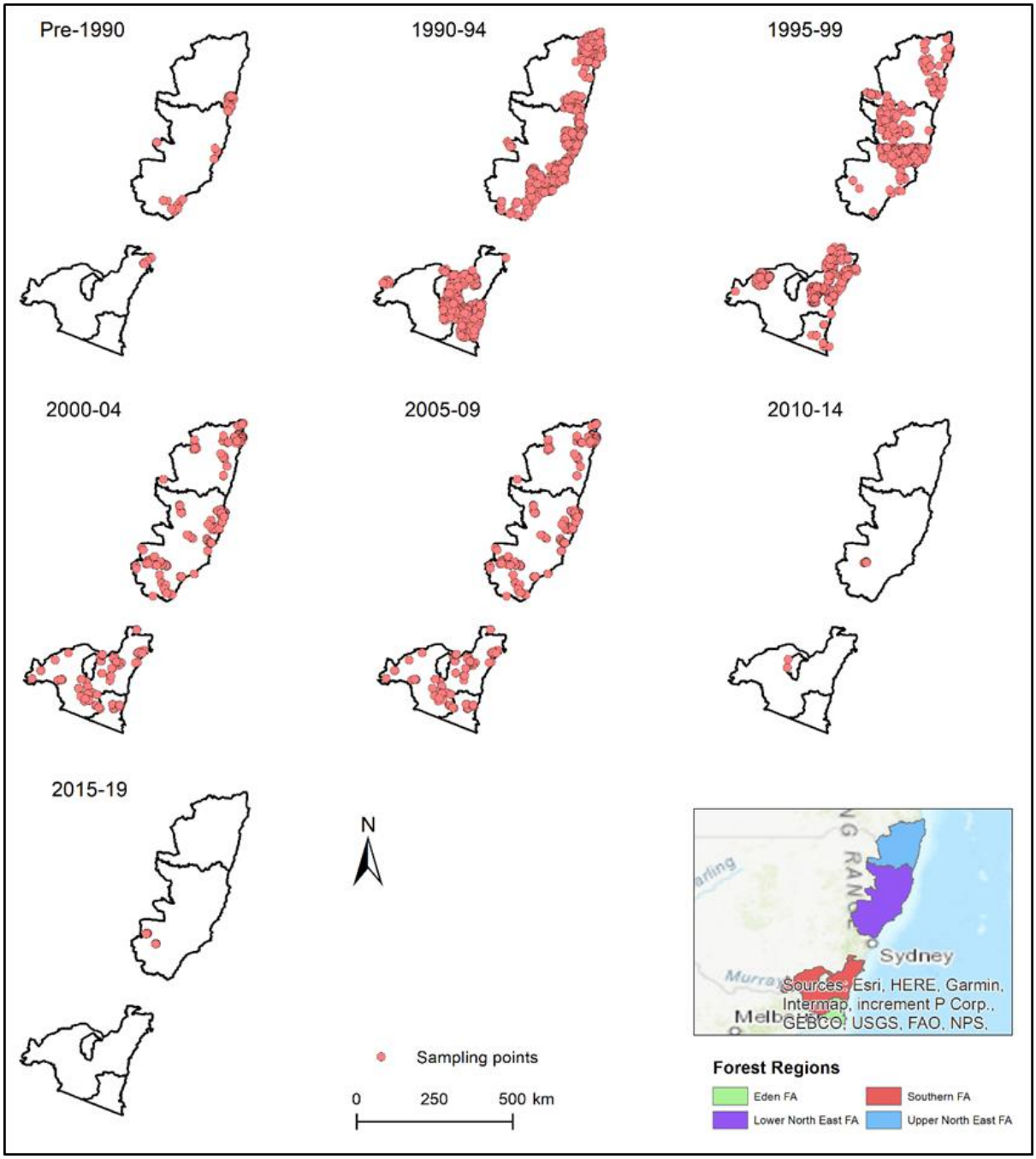

Figure 2. Spatial distribution of surface organic carbon measurements across RFA regions

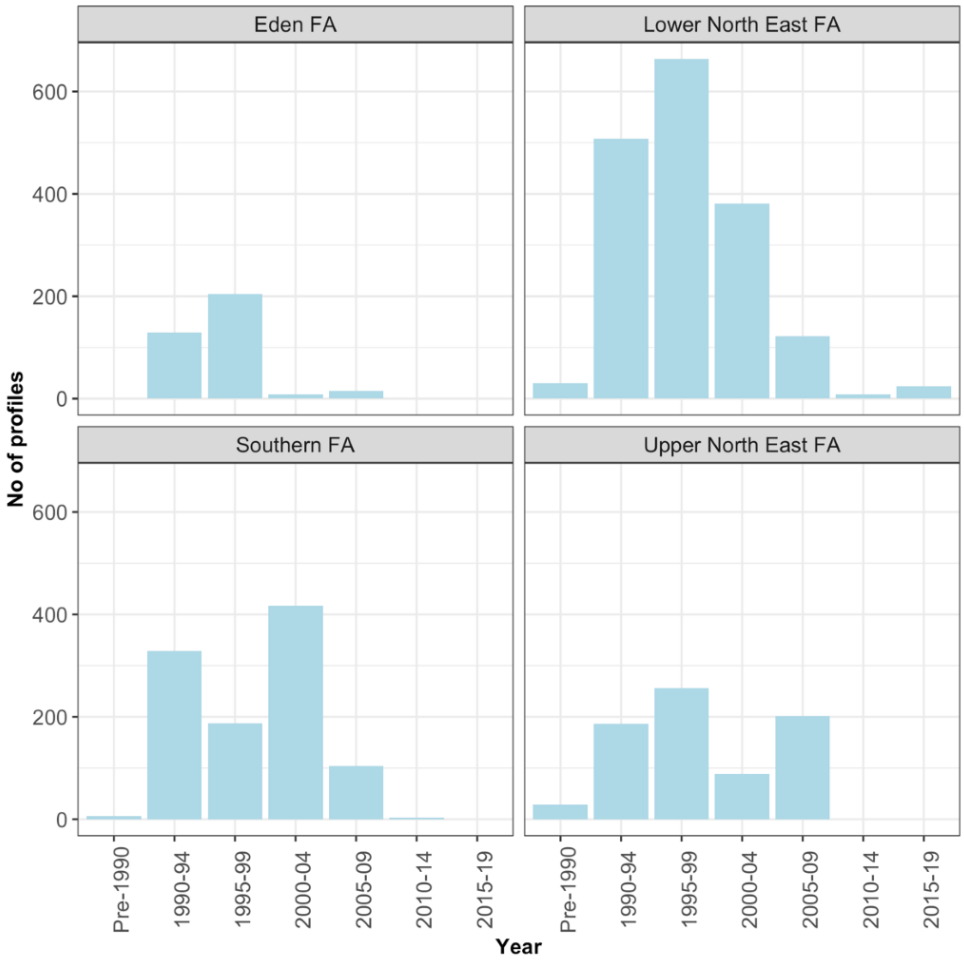

Figure 3. Number of surface organic carbon measurements across RFA regions

In broad terms the number and distribution of soil carbon measurements for each time slice is promising in terms of applying the data cube approach. An obvious issue is the lack of data for 2010 onwards but if the pre-2010 vegetation and weather variation represents the 2010+ period then extrapolation in time may not be such an issue. This will be explored in the next stages of the project.

3.2 Exploratory data analysis of SOC in the data cube

Table 2 presents summary statistics of the SOC observations in the cube. This is for all soil profile locations across NSW that were sampled from 1990 onwards. It can be observed from the results that the data is positively skewed with a significant number of observations below 5%. This can also be observed in the boxplots of the of SOC distribution for each 5-year interval (Figure 4). The presence of the large values (outliers) is likely due to the presence of a small number of profiles from organic soils.

Table 2: Summary statistics of the surface soil organic carbon (%) in the data cube.

Minimum	Maximum	Mean	Median	SD
0.010	56.00	3.01	2.19	3.07

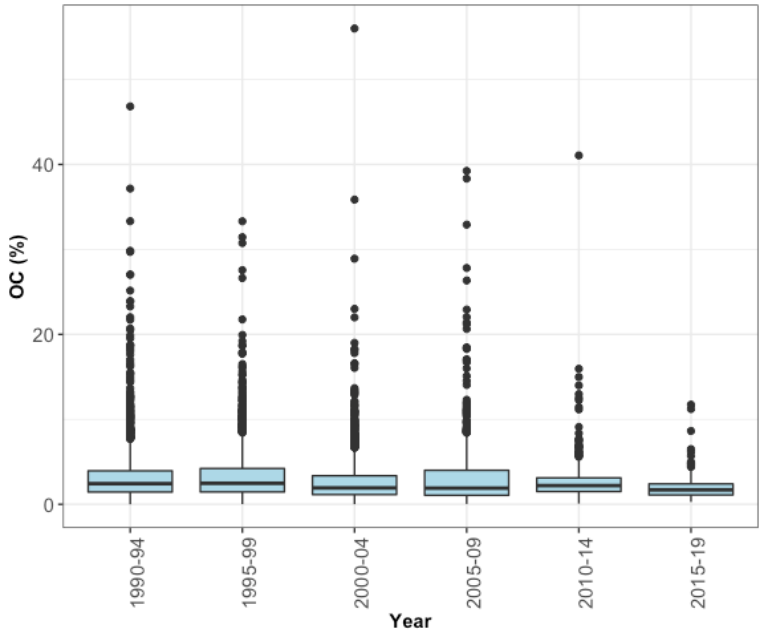

Figure 4: Surface organic carbon distribution at 5-year interval across NSW

3.2 Initial model results and predictions

A model for surface SOC was constructed and evaluated from a random 80:20 split of the data cube for model training and testing respectively. Model estimation was done using Cubist. Figure 5 presents the result of comparing the actual and predicted SOC measurements of the test sample. The root-mean-square error (RMSE) and Lin's concordance correlation coefficient (LCCC) were used in model quality assessment. The LCCC is the fit of the observed and predicted values to the 1:1 line, and is unit-less, making it useful for comparing between models where the magnitude of the predictions may vary. Overall, the approach show somes promise and can be improved.

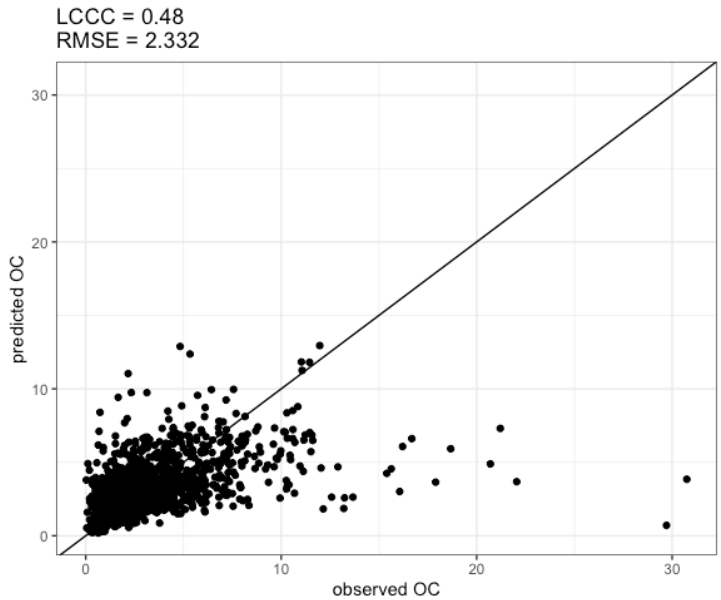

Figure 5: Observed and predicted surface SOC using an 80:20 split of the data cube for model training and testing.

In terms of the contributions of the predictors to the model, the climate variables and NDVI features aggregated using a discount factor of 0.99 are ranked as the most important predictors of SOC in the catchment (Figure 6). This demonstrates the usefulness of space and time variables in modelling dynamic environmental attributes such as SOC.

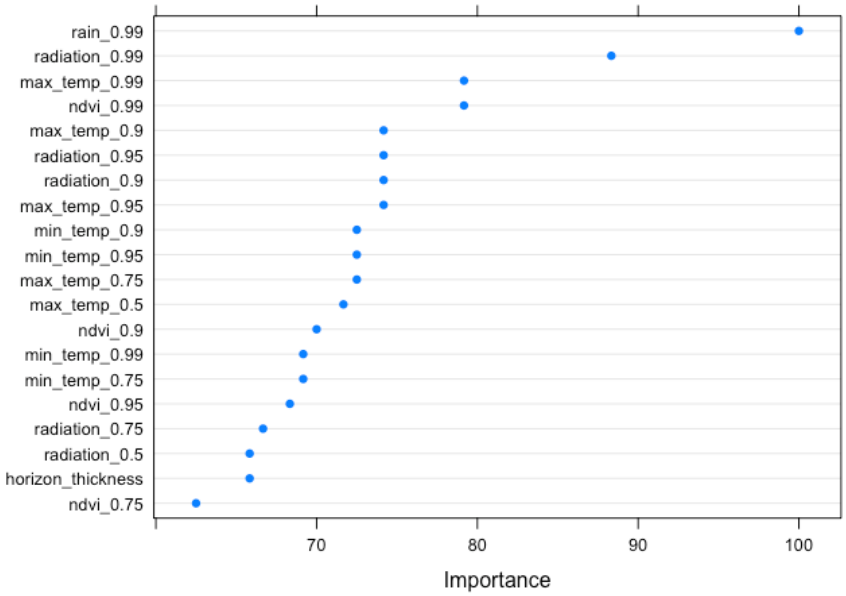

Figure 6: Variable of importance of surface SOC predictors from a cubist model constructed using all location data with complete cases in the data cube.

In order to test the utility of our space-time cube for SOC modelling and prediction, we also mapped SOC in the Muttama catchment in southern NSW where we are in the process of collating independent soil data for model evaluation. For this reason, the entire cube was used to estimate a model of SOC using Cubist and the model was subsequently applied on a stack of covariates to estimate SOC for the month of June 2013 and 2019.

The maps of carbon across Muttama show a large increase in soil carbon in the surface layer which is probably unexpected. We believe this is due to the 2013 surface layers being generally deeper (typically 10-20 cm) than the 2019 survey (typically 5 - 10 cm). This means the 2019 carbon values are greater than 2013 values when mapped. Harmonising the soil carbon values to a common depth interval across the dataset prior to modelling will likely remove this artefact. Despite this the maps show it is possible to predict soil carbon for any spatial location and time using this approach. Further work needs to consider the uncertainty.

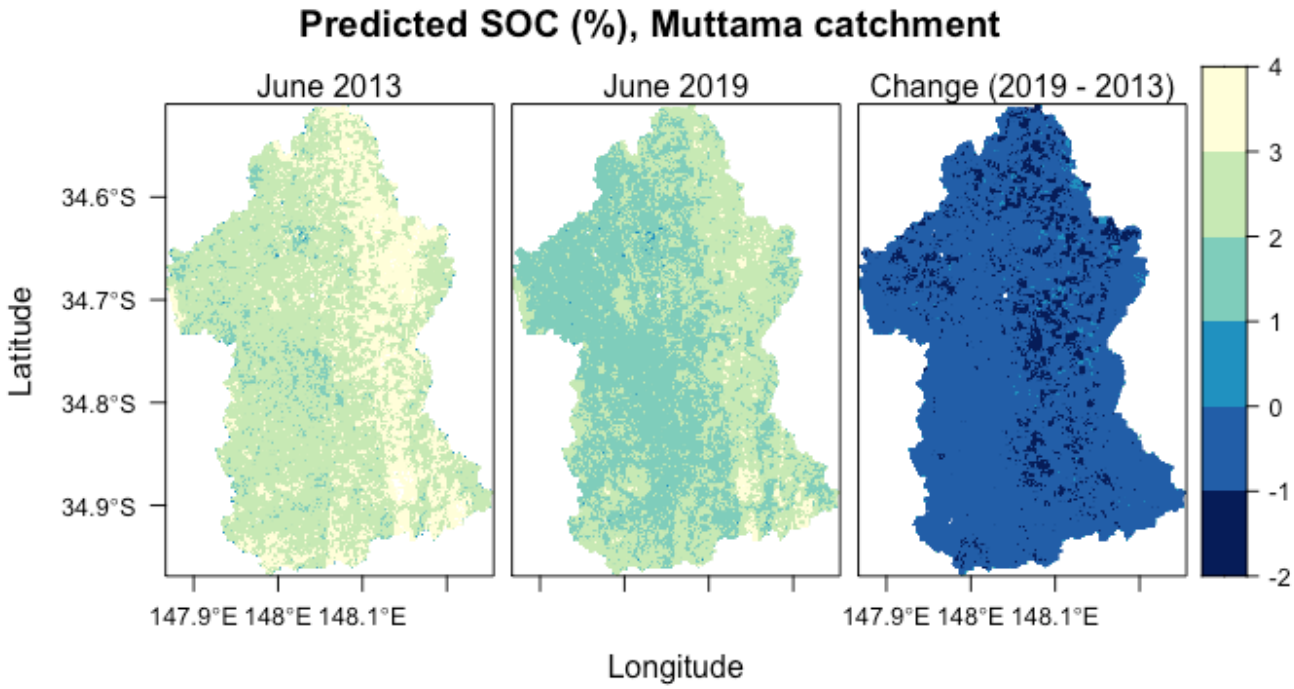

Figure 7: Surface SOC at the Muttama catchment estimated from model constructed on the data cube and applied to a stack of covariates for the catchment.

4. Next steps

We have developed a workflow and implemented a prototype model for all of NSW. An Appendix outlines the script developed. To improve the process, we will consider in the next steps:

- include more covariates
- considering the performance of a model using all of the NSW soil data versus just data situated within the RFA regions
- apply more rigorous machine learning algorithms such as Gaussian Process Regression to improve model performance
- compare the results using the surface layer as performed here versus a depth harmonised estimate of soil carbon for the 0-30 cm layer which is used for greenhouse accounting
- consider regions in the attribute space of the data cube where we have no or few soil carbon measurements and map these to show where any predictions should be interpreted cautiously

References

Wimalathunge, N and Bishop, TFA (2019). A space-time observation system for soil moisture in agricultural landscapes. *Geoderma*, 314, 1-13.

Appendix

This is an overview of scripts generated for Milestone 5 activities for collation of a space-time data cube and associated scripts for initial soil carbon modelling.

Table 1 Overview of data collated for modelling

Activity	Script	Source	Description
Data cube	NDVI_extraction	Java script on Google Earth Engine	Ingests raw Landsat 5, 7 and 8 surface reflectance scenes and computes monthly maximum NDVI composites. Also, extracts and outputs NDVI monthly times series to the soil profile locations. The output is a CSV file.
	extract_covariates and functions	R scripts	The workflow script (extract_covariates) and the function definitions script (functions) process space (e.g., clay and silica rasters) and space and time (climate and NDVI) covariates and extract the values to soil profile locations, creating a space time data cube. It exports a CSV of this data. The processing steps include the aggregation of the timeseries of the space and time variables using a discount algorithm to estimate antecedent conditions and also applying a differencing method on the NDVI time series.
Prediction	prepare_grid	R script	This script creates a regularly spaced grid of all the covariates onto which prediction will be made. Using a baseline grid point derived from one of the NDVI images, values from all the covariates are extracted to the spatial point object. This approach avoids the need to resample covariate rasters prior to stacking them together for mapping. Additionally, the script also aggregates the climate covariates. The output is a data frame.
	differenced_ndvi	Java script on Google Earth Engine	This script creates and exports NDVI differenced images for a given prediction month. The script ingests

			Landsat surface reflectance scenes and compute the difference images. These images are subsequently extracted to a grid with the other covariates using “prepare_grid” R script
	discounted_ndvi	Java script on Google Earth Engine	This script creates and exports NDVI differenced images covariate for a given prediction month. The script also ingests Landsat surface reflectance scenes. The output from this script are subsequently extracted to a grid in the “prepare_grid” R script with the other covariates.
EDA & Modelling	eda_modelling	R Script	This script reads in the data cube performs some basic data exploration and initial modelling