Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

ELSEVIER

Data Article

Soil Organic Carbon estimates derived from remote sensing data[★]

Ricardo Barros Lourenco[a,*], Camile Sothe[a], Alemu Gonsamo[a], Joyce Arabian[b], James Snider[b]

[a] *School of Earth, Environment & Society, McMaster University, Hamilton, Ontario, Canada*
[b] *World Wildlife Fund Canada, Toronto, Ontario, Canada*

ARTICLE INFO

Keywords:

Soil Organic Carbon
Remote Sensing
Machine Learning

ABSTRACT

This article describes a dataset of Soil Organic Carbon (SOC) estimates for the entire Canadian territory, derived from remote sensing measurements processed by a machine learning model.

This dataset can be used by scientists involved in climate science studies, more specifically in carbon stock modeling, and carbon cycle monitoring.

The data is derived from satellite measurements of temperature, precipitation, elevation, terrain slope, vegetation indexes, and radar polarization. These sources were used to train a machine learning model able to predict SOC concentrations on soil. As target for such modelling process, it were used ground measurements of SOC at different depths.

The main article in which the employed methodology and analysis details are provided in full extent is available at Sothe et al (2021)[1].

[★] Preliminary work and title - as partial requirement for approval at GEOG 512 - Reproducible Research Workflow.

[*] Corresponding author.

e-mail: `barroslr@mcmaster.ca` (Ricardo Barros Lourenco)

http://dx.doi.org/10.1016/j.dib.xxxx.xx.xxx

Specifications Table

Every section of this table is mandatory. Please enter information in the right-hand column and remove all the instructions

Subject	Computers in Earth Sciences
Specific subject area	Machine-learning generated data (precisely Soil Organic Carbon Estimates), using remote sensing data as covariates.
Type of data	Multi-channel Georeferenced Image
How data were acquired	Data generated by a machine learning model (Quantile Regression Random Forests) on a covariate set derived from satellite imagery bands. The target variable are carbon estimates obtained from field surveys across Canada.
Data format	Georeferenced multi-channel raster files stored in WebService.
Parameters for data collection	The remote sensing data was collected for periods between 5-20 years based on availability (for cloud removal), with a temporal upsampling starting on 5 days (of recoverage). The scale of analysis also was set in 250m (with downsampling starting in 1km and upsampling starting on 30m).
Description of data collection	Satellite data is orbital sun-synchronous satellite data, using multispectral sensors. SOC data was acquired on field campaigns over the years.

Data source location	Primary data sources: USGS Landsat 8 Surface Reflectance Tier 1 Red band 0.64–0.67 µm, NIR band 0.85–0.88 µm, SWIR1 band 1.57–1.65 µm, SWIR2 band 2.11–2.29 µm Landsat 8 Collection 1 Tier 1 32-Day NDWI Composite - Normalized Difference Water Index (annual) Landsat 8 Collection 1 Tier 1 8-Day NDSI Composite - Normalized Difference Snow Index (annual) Tasseled cap transformation based on Landsat 8 at-satellite reflectance (annual - brightness/greenness/wetness) USGS Landsat 8 Surface Reflectance Tier 1 Brightness temperature band 10.60–11.19 µm (bimonthly except Nov to Feb) / Brightness temperature band 11.50–12.51 µm (bimonthly except Nov to Feb) MODIS Terra Vegetation Indices 16-Day Global 250 m (NDVI) - Normalized Difference Vegetation Index (bimonthly) MODIS Terra Vegetation Indices 16-Day Global 250 m (EVI) - Enhanced Vegetation Index (annual) MODIS Global Terrestrial Evapotranspiration 8-Day Global 1 km MODIS Terra Land Surface Temperature and Emissivity Daily (Day and Night) Global 1 km MODIS Long-term Land Surface Temperature daytime monthly standard deviation Daymet - Daily surface maximum 2-meter air temperature / Daily surface minimum 2-meter air temperature / Daily incident shortwave radiation flux density / Daily total precipitation, sum of all forms converted to water-equivalent Digital Elevation Model (ALOS) and slope derived from it Global PALSAR-2/PALSAR Yearly Mosaic, converted to decibels (DB) / L-band duo-polarization horizontal transmit/horizontal receive (HH) and horizontal transmit/vertical receive (HV) Calculated topographic position index [2] Calculated terrain ruggedness index [3] Soil types and depths of Soil Landscape of Canada (SLC)
Data accessibility	The dataset will be stored on the Google Earth Engine Platform[5], using either a Google Drive mount, or a Google Cloud Storage. In case of impossibility, it will be stored on premises of McMaster University, and Indexed with Globus [6].
Related research article	This article is directly related with: Camile Sothe, Alemu Gonsamo, Joyce Arabian, James Snider, Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations, Geoderma, Volume 405, 2022, 115402, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2021.115402.

Value of the Data

[Provide 3-6 bullet points explaining why these data are of value to the scientific community. Bullet points 1-3 must specifically answer the questions next to the bullet point, but do not include the question itself in your answer. You may provide up to three additional bullet points to outline the value of these data. Please keep points brief, with ideally no more than 400 characters for each point.]

- This dataset is important, because it is the first attempt to systematically estimate Canadian soil carbon stocks with usage of machine learning and remote sensing.
- Potential beneficiaries are climate change policy makers, and stakeholders involved in carbon cycle studies.
- This data may be used as input in other models, given that the uncertainties need to be propagated across models (since it is a model run output).
- Broader impacts of this dataset may lie on the ability for policymakers and the general public to assess current carbon stock location, and due to that evaluate proposed policies, and risks/benefits derived from actions of the government to mitigate climate change.

Data Description

[*This is still preliminary work. Once I reprocess the data, I intend to add up to this section.*]

Experimental Design, Materials and Methods

[*This is still preliminary work. Once I reprocess the data, I intend to add up to this section.*]

Ethics Statement

This work does not involve either human subjects, or animal experiments.

Acknowledgments

This work was supported by World Wildlife Fund Canada (WWF-Canada).

A.G. acknowledges funding from Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2020-05708) and the Canada Research Chairs Program.

R. B. L. acknowledges the careful review provided by professors and colleagues at GEOG 512 - Reproducible Data Workflow.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Camile Sothe, Alemu Gonsamo, Joyce Arabian, James Snider, Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations, Geoderma, Volume 405, 2022, 115402, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2021.115402.

[2] Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis - Principles and Applications. 512p., ISBN: 978-0-471-32188-0.

[3] Riley, Shawn J., Stephen D. DeGloria, and Robert Elliot. ”Index that quantifies topographic heterogeneity.” intermountain Journal of sciences 5.1-4 (1999): 23-27.

[4] National Soil Database. Soil Landscape of Canada version 3.2. (2011) http://sis.agr.gc.ca/cansis/nsdb/slc/index.html

[5] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202, 18-27.

[6] Rachana Ananthakrishnan, Ben Blaiszik, Kyle Chard, Ryan Chard, Brendan McCollam, Jim Pruyne, Stephen Rosen, Steven Tuecke, and Ian Foster. 2018. Globus Platform Services for Data Publication. In Proceedings of the Practice and Experience on Advanced Research Computing (PEARC ’18). Association for Computing Machinery, New York, NY, USA, Article 14, 1–7. DOI:https://doi.org/10.1145/3219104.3219127