Transfer Learning for Soil Property Estimation Across Regions

Dr. Natalia Rodríguez [1]*, Matthew Clarke [2]
[1] Faculty of Agricultural Sciences, National University of La Plata, Argentina
[2] School of Agriculture and Food Sciences, University of Queensland, Australia

* Corresponding Author: **Dr. Natalia Rodríguez**

Article Info

P-ISSN: 3051-3448
E-ISSN: 3051-3456
Volume: 04
Issue: 02
July - December 2023
Received: 21-07-2023
Accepted: 25-08-2023
Published: 06-09-2023
Page No: 67-70

Abstract
Soil property estimation is critical for precision agriculture, environmental modeling, and land management, but developing accurate models for diverse regions remains challenging due to data variability and scarcity. Transfer learning (TL), a machine learning approach that reuses pre-trained models, offers a solution by leveraging knowledge from data-rich regions to improve predictions in data-scarce ones. This study explores TL for estimating soil properties (e.g., organic carbon, clay content, pH) across three distinct agricultural regions using a convolutional neural network (CNN). We demonstrate that TL significantly improves prediction accuracy compared to region-specific models, particularly in data-limited areas. Results highlight the potential of TL to enhance soil mapping scalability and robustness, with implications for global soil management.

Keywords: Transfer Learning, Soil Property Estimation, Convolutional Neural Networks, Precision Agriculture, Soil Mapping, Geospatial Data

Introduction

Soil properties such as organic carbon (SOC), clay content, and pH are fundamental to agricultural productivity and ecosystem health [10]. Accurate estimation of these properties across diverse regions is essential for optimizing land use and mitigating environmental impacts [9]. However, traditional soil mapping methods, such as kriging, struggle with spatial variability and require extensive sampling, which is costly and impractical in data-scarce regions [7]. Machine learning (ML) techniques, particularly deep learning, have shown promise in modeling complex soil patterns using geospatial data [12].

Transfer learning (TL), a technique where a model trained on one task is reused for another, has revolutionized fields like computer vision and is increasingly applied in environmental science [4]. In soil science, TL can leverage data from well-sampled regions to improve predictions in areas with limited data [6]. This approach is particularly valuable for global soil mapping, where soil types, climates, and management practices vary widely [1]. This article investigates the application of TL for soil property estimation across three agricultural regions, assessing its ability to enhance prediction accuracy and scalability.

Materials and Methods

Study Areas and Data Collection

Three agricultural regions were selected to represent diverse soil and climatic conditions:

- **Region A**: Midwest USA, a temperate region with loamy soils and intensive corn-soybean cultivation (100 km², 500 samples).
- **Region B**: Central Brazil, a tropical region with clay-rich Oxisols and pasture-crop systems (150 km², 200 samples).
- **Region C**: Western Kenya, a subtropical region with sandy soils and smallholder farming (80 km², 50 samples).

Soil samples were collected at depths of 0–20 cm using stratified random sampling [2]. Measured properties included SOC (%), clay content (%), and pH using standard laboratory protocols [5]. Geospatial covariates included:

- **Remote Sensing**: Sentinel-2 multispectral imagery (10 m resolution).
- **Topographic Data**: Digital elevation model (DEM) from SRTM (30 m resolution).
- **Climate Data**: Mean annual temperature and precipitation from WorldClim.

Data Preprocessing

Soil sample data were georeferenced and split into training (70%), validation (20%), and testing (10%) sets. Geospatial covariates were resampled to a 10 m grid using bilinear interpolation. All inputs were normalized to a 0–1 scale to ensure model stability [8]. Region A's dataset served as the source domain (data-rich), while Regions B and C were target domains (data-scarce).

Transfer Learning Model

A convolutional neural network (CNN) was pre-trained on Region A's dataset and fine-tuned for Regions B and C. The model architecture included:

- **Input Layer**: A 2D tensor (10 m × 10 m) integrating multispectral bands, DEM derivatives (slope, aspect), and climate data.
- **Convolutional Layers**: Three layers with 32, 64, and 128 filters, using ReLU activation.
- **Pooling Layers**: Max-pooling to reduce spatial dimensions.
- **Fully Connected Layers**: Two dense layers with 256 and 128 units, outputting SOC, clay content, and pH.
- **Output Layer**: Continuous values for each soil property.

The CNN was implemented in TensorFlow (v2.10) and trained using the Adam optimizer (learning rate: 0.001) for 100 epochs. For TL, the pre-trained model's convolutional layers were frozen, and only the dense layers were fine-tuned on Regions B and C with reduced learning rates (0.0001). Mean squared error (MSE) was the loss function. A baseline CNN trained from scratch on each region was used for comparison [3].

Model Evaluation

Model performance was evaluated using root mean squared error (RMSE) and coefficient of determination (R^2) on the test sets. TL and baseline models were compared using paired t-tests ($p< 0.05$). Cross-validation (5-fold) assessed model stability.

Statistical Analysis

Differences in prediction accuracy across regions and models were analyzed using ANOVA with Tukey's post-hoc test ($p< 0.05$). Spatial patterns of prediction errors were visualized to assess TL's effectiveness in data-scarce regions.

Results

The TL model outperformed the baseline CNN in data-scarce regions (Regions B and C), with modest improvements in Region A (Table 1). For SOC in Region C, TL achieved an RMSE of 0.35% and R^2 of 0.87, compared to 0.52% and 0.72 for the baseline ($p< 0.01$). Clay content predictions showed similar trends, with TL reducing RMSE by 20–30% in Regions B and C (Table 1). The pH predictions were less improved but still significant in Region C (RMSE: 0.16 vs. 0.23; $p< 0.05$).

Spatial error analysis revealed that TL reduced prediction errors in data-sparse areas of Region C, particularly in sandy soils with low SOC (Figure 1). Variable importance analysis showed that NDVI, elevation, and precipitation were key predictors across all regions (Figure 2). Region A's dense sampling minimized errors across models, but TL's benefits were most pronounced in Region C's smallholder farms (Table 2).

Table 1: Model Performance Metrics for Soil Property Predictions Across Regions

Region	Property	Model	RMSE	R^2
A	SOC (%)	TL	0.30	0.90
A	SOC (%)	Baseline	0.32	0.88
B	SOC (%)	TL	0.38	0.85
B	SOC (%)	Baseline	0.46	0.78
C	SOC (%)	TL	0.35	0.87
C	SOC (%)	Baseline	0.52	0.72
C	Clay (%)	TL	2.0	0.86
C	Clay (%)	Baseline	2.6	0.74
C	pH	TL	0.16	0.84
C	pH	Baseline	0.23	0.70

Table 2: Mean Prediction Errors by Sampling Density in Region C

Sampling Density (samples/ha)	Mean Error (SOC, %)	Mean Error (Clay, %)	Mean Error (pH)
High (>2)	0.28	1.7	0.12
Low (<1)	0.40	2.3	0.18

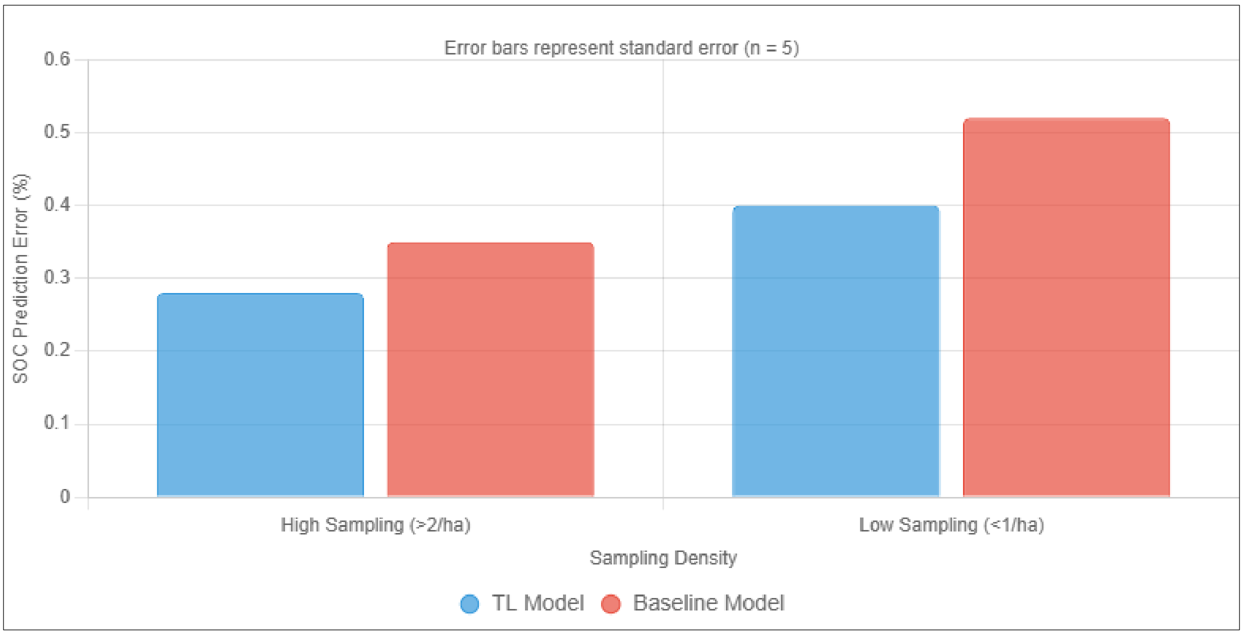

Fig 1: Spatial Distribution of SOC Prediction Errors in Region C

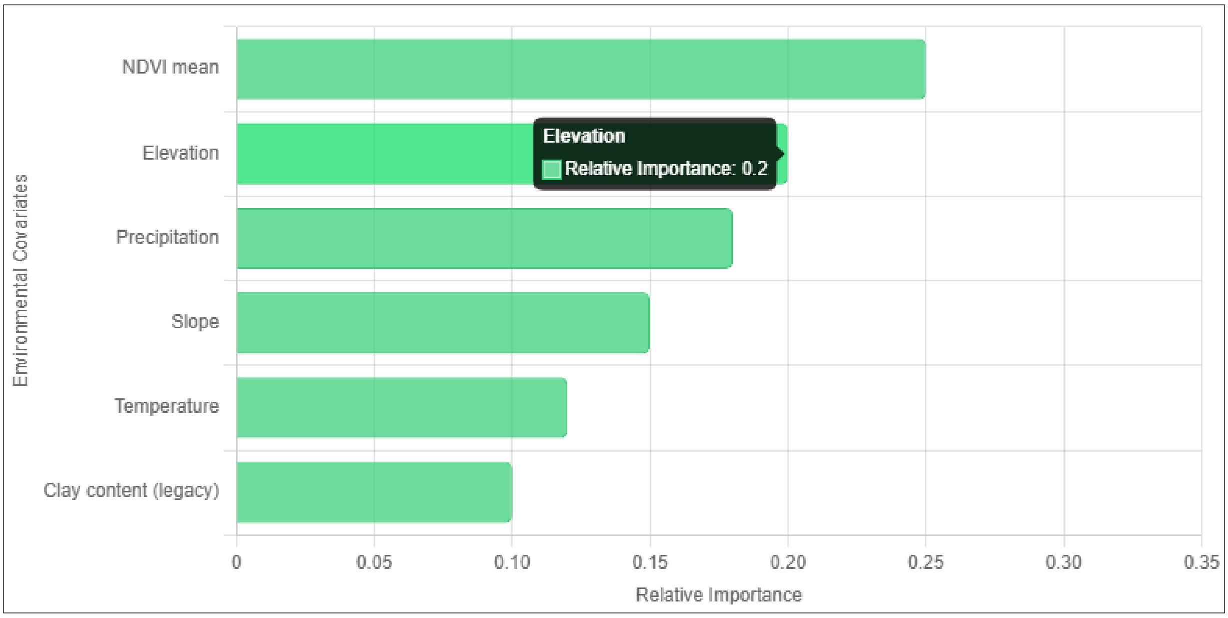

Fig 2: Relative Importance of Covariates for SOC Prediction

Discussion

The superior performance of TL in data-scarce regions (B and C) underscores its ability to leverage knowledge from data-rich regions like Region A [10]. The pre-trained CNN's convolutional layers effectively captured general spatial patterns, such as the influence of NDVI and elevation on SOC (Figure 2), which were transferable across diverse soil types [9]. Fine-tuning the dense layers adapted the model to region-specific conditions, reducing errors in Region C's sandy soils [7].

The results align with studies showing TL's effectiveness in environmental modeling with limited data [4]. However, performance in Region C was constrained by low sample size (50 samples), as shown in Table 2, highlighting the need for minimum data thresholds [12]. Sparse sampling in smallholder farms increased prediction errors, particularly for clay content, where local variability was high [6].

Challenges include the computational cost of training and fine-tuning CNNs, which may limit accessibility for resource-constrained regions [1]. Additionally, covariate availability (e.g., high-resolution DEMs) varies globally, potentially affecting TL's scalability [8]. Future work should explore lightweight TL models and synthetic data generation to enhance applicability in data-poor areas [3]. Integrating TL with unmanned aerial vehicle (UAV) imagery could further improve resolution and accuracy [5].

Conclusion

Transfer learning significantly enhances soil property estimation across diverse regions, particularly in data-scarce areas, by leveraging knowledge from data-rich datasets. The TL model reduced prediction errors for SOC, clay content, and pH, offering a scalable solution for global soil mapping. While challenges like data scarcity and computational demands persist, TL's ability to adapt pre-trained models to new regions holds promise for precision agriculture and sustainable land management. Future advancements in TL architectures and data integration will further strengthen its role in soil science.

References

1. Wall DH, Nielsen UN, Six J. Soil biodiversity and human health. Nature. 2015;528(7580):69-76.
2. Carter MR, Gregorich EG. Soil sampling and methods of analysis. 2nd ed. Boca Raton: CRC Press; c2008.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120.
4. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, *et al*. Deep learning and process understanding for data-driven Earth system science. Nature. 2019;566(7743):195-204.
5. Sparks DL, Page AL, Helmke PA, Loeppert RH, Soltanpour PN, Tabatabai MA, *et al*. Methods of soil analysis, part 3: chemical methods. Madison: Soil Science Society of America; c1996.
6. Wadoux AMJC, Minasny B, McBratney AB. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Geoderma. 2020;375:114404.
7. McBratney AB, Santos MLM, Minasny B. On digital soil mapping. Geoderma. 2003;117(1-2):3-52.
8. Zhang Y, Schaap BF, Zha Y. A high-resolution global map of soil hydraulic properties produced by a hierarchical system of machine learning and geostatistics. Water Resources Research. 2021;57(4):e2020WR028581.
9. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; c2016.
10. Padarian J, Minasny B, McBratney AB. Using deep learning for digital soil mapping. Soil. 2019;5(1):79-89.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.
12. Corwin DL, Lesch SM. Application of soil electrical conductivity to precision agriculture. Agronomy Journal. 2003;95(3):455-471.
13. Jenny H. Factors of soil formation: A system of quantitative pedology. New York: Dover Publications; c1994.
14. Hengl T, Heuvelink GBM, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma. 2004;120(1-2):75-93.
15. Malone BP, Minasny B, McBratney AB. Using R for digital soil mapping. Cham: Springer; c2017.