

Research Article

Pedometric Mapping of Soil Classes: A Case Study of San Mateo de Otao, Peru

Carlos J. Mestanza  and **Julio C. Nazario** 

Soil Department, National Agrarian University La Molina, Lima 15024, Peru

Correspondence should be addressed to Carlos J. Mestanza; cmestanza@lamolina.edu.pe

Received 12 February 2022; Revised 7 April 2022; Accepted 18 April 2022; Published 30 April 2022

Academic Editor: Balram Ambade

Copyright © 2022 Carlos J. Mestanza and Julio C. Nazario. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Conventional soil maps are designed based on expert criteria, a characteristic that reduces their reproducibility and generates subjective uncertainty. Pedometric mapping uses mathematical and statistical principles, which makes it the opposite of conventional mapping. It was proposed to apply the pedometric mapping in San Mateo de Otao and find out its characteristics against the conventional one. Satellite and field data were used to extract covariables (soil-forming factors) and soil classes. The data were modeled with Naïve Bayes, global uncertainty was calculated by resubstitution, cross-validation and retention, and local uncertainty with the confusion and Shannon indices. A low uncertainty map was obtained with six identified soil classes, relief, and parent material having the most important covariates. We conclude that pedometric mapping has considerable advantages over conventional mapping and its application is possible under the context of soil survey in Peru.

1. Introduction

Conventional soil maps present the location and boundaries of soil groups (mapping units); they are consolidated as a basic input for the creation of new maps that diagram the properties of the soils used for various purposes. Its main use is for the development of land use plans, its evaluation, and prediction of the effects of these uses [1]. The main limitations they have are the lack of an uncertainty value, soil classes cannot be mapped and the relationship of the soil with its formation factors is understood subjectively. Even if they are well designed with geographic information systems, Brevik et al. [2] emphasized that this method preserves the limitations of paper maps, since the use of human criteria for the synthesis and interpretation of the model of soil formation factors leads to an accumulation of error that is impossible to measure and far from randomness. Conventional soil maps are considered obsolete and are opposed to pedometric soil maps [3] that apply mathematical, statistical, and computational methods to model, interpret, and map soils. Several countries use this method to produce soil classes maps on national scales, Germany [4], India [5], Iran

[6], France [7], and the United States [8] are some examples. The common characteristic of all studies is a large number of predictors and samples used to generate and validate the models. The rugged topography of the Peruvian Andes limits the sampling process to a low density of soil samples, which limits the use of computationally sophisticated mapping methods. The recent increase in the number and types of remote sensing platforms is an opportunity to improve the number of predictors, staying as a challenge using machine learning models to increase the pedological understanding [9]. Few soil studies in Peruvian territory use statistical [10] or computational [11, 12] methods to map soils, commonly conventional mapping is chosen [13]. Bayesian methods can be used for small datasets; Steinbuch [14] showed that in small datasets, an appropriate prior selection improves overall accuracy compared to non-Bayesian methods. A common and simple machine learning method that incorporates the Bayesian approach is Naïve Bayes. Incorporating pedometric mapping into Peruvian soil surveys can provide new knowledge of the relationship of soil and their formation factors and the uncertainty value of the maps. It was proposed applying pedometric mapping in the district of San

Mateo de Otao to map soil classes using Naïve Bayes model and understand the differences from conventional mapping.

2. Materials and Methods

2.1. Location. The study area is a small catchment (77.48 km²) located within the limits of San Mateo de Otao district of the Huarochiri Province, Lima, Peru. The main Holdridge life zones (Figure 1(a)) were tropical montane moist forest (TM-mf), tropical premontane perarid desert (TP-pd), tropical montane steppe (TM-s), tropical lower montane thorn steppe (TLM-ts), tropical lower montane desert scrub (TLM-ds), and tropical premontane desert scrub (TPM-ds). All life zones, except the TM-mf, are arid ecosystems (evapotranspiration > precipitation). The elevation varies from 1200 (TP-pd) to 4350 (TM-mf) meters above sea level; as a consequence, the air temperature is very variable. TP-ds and TPM-ds are the warmest areas (22–24°C), TLM-ds and TLM-ts are cool zones (15–17°C), and TM-s and TM-mf are the coldest (10–12°C) zones [15]. The relative area occupied by each life zone was 21.60 (TM-mf), 2.12 (TP-pd), 21.59 (TM-s), 33.69 (TLM-ts), 9.79 (TLM-ds), and 11.22 (TPM-ds) percent. Geology (Figure 1(b)) consists of alluvial deposits (Qp-al), porphyritic andesite (P-r), rhyolitic tuff (Nm-h), andesites (N-and, J-ar), tonalities, diorites, and granodiorites (Ki-bc/p-di, tn; Ki-bc/sr-tn, gd; PN-tn, gd) [16, 17].

2.2. Data Source. All the Landsat 8 images collection 1 and level 1 (courtesy of the United States Geological Service) of 2018 were downloaded. The dark object subtraction atmospheric correction was applied using the semiautomatic classification plugin [18] of the QGIS 3.10; clouds and shadows were removed using the quality assessment (QA) band. The images were averaged to later perform the index calculation. ASTER L1T satellite image from 07/12/2015, Jaxa/Meti digital elevation model, WorldClim 2.0 climate grids [19], and the integrated geology map [20] were used as base inputs for the generation of representative variables of the soil formation factors.

2.3. Soil Sampling and Analysis. The sampling design was by stratification of study area and free mapping (locations were selected by convenience within the strata). The limits of the strata were equivalent to the limits of the homogeneous land units. The geological units were grouped into extrusive, intrusive igneous rocks, and alluvial sediments. Life zones were grouped by climatic affinity in four keys: 1 (TP-pd), 2 (TLM-ds and TPM-ds), 4 (TLM-ts), 5 (TM-s), and 10 (TM-mf). The landscape was classified using Zinck methodology [21] in valleys (Va), extrusive mountain slopes (Le), and intrusive mountain slopes (Li). The homogeneous land units were obtained by the intersection of the landscape and grouped life zones (Figure 1(c)). A total of 64 samples (soil profiles) were selected by free mapping and were taken within each homogeneous land unit. Each sample (soil profile) was evaluated by describing the pedon's internal and external characteristics. Then, the subsamples (horizons)

were evaluated in the soil laboratories of the National Agrarian University La Molina. The physical-chemical analyses were carried out based on the Soil Survey Staff protocols ([22, 23]). Particle size distribution was carried with hydrometer method, pH, and electrical conductivity in suspension soil:water 1:1, total equivalent carbonates by acid digestion, soil organic carbon by Walkley and Black method, cation exchange capacity with ammonium acetate at pH 7.0, and exchangeable cations by atomic absorption spectrophotometry with a Perkin Elmer model AAnalyst 200 equipment. The soil classes were obtained with the soil taxonomy system [24] down to the subgroup level.

2.4. Predictor Variables. The digital elevation model was processed with the basic terrain analysis module of the SAGAGIS 6.4.0 to extract ten representative variables of the relief factor. The climate factor was extracted from WorldClim v 2.0. and the ET was calculated using the climate and weather-evapotranspiration module of SAGAGIS 6.4.0, Hargreaves method. The ASTER images were used to calculate the quartz, carbonate, and mafic index [25] and Landsat 8 for the Normalized Difference Vegetation Index and soil biological crust [26]. The complete description of the variables obtained is found in Table 1.

2.5. Sampling Quality. The objective of the sampling is to obtain a fraction of representative individuals of the population whose distribution function is close to the population distribution. This comparison was made graphically by observing the histograms of the population and the sample with the Scott method for the number of bins and using the Kolmogorov–Smirnov hypothesis test at a significance level of five percent ($\alpha = 0.05$).

2.6. Modeling and Uncertainty. The Naïve Bayes method was used including the soil classes as the response variable (qualitative) and the predictor variables. The method requires conditional independence between the predictors [27], but all of them were used for our model. Before modeling, a Yeo-Johnson standardization and transformation were performed. The database presented unbalanced classes, a problem that was solved using an oversampling balancing in the response variable. We use a Laplace smoother of 100 and the kernel method to estimate the density function without adjusting the bandwidth (value of one).

General uncertainty measures were obtained with three procedures: (1) by resubstitution, generating a confusion matrix with the same data that the model was generated; (2) by retention, dividing the total data set into 70 percent to estimate the model and 30 percent for validation; and (3) 10-fold cross-validation with 100 repetitions. Local uncertainty was measured using the confusion and Shannon indices. The first was proposed by Burrough et al. [28] and measures the degree of confusion between the two soil taxa with the highest probability of occurrence. A class close to zero means that two soil taxa are similarly likely to occur and

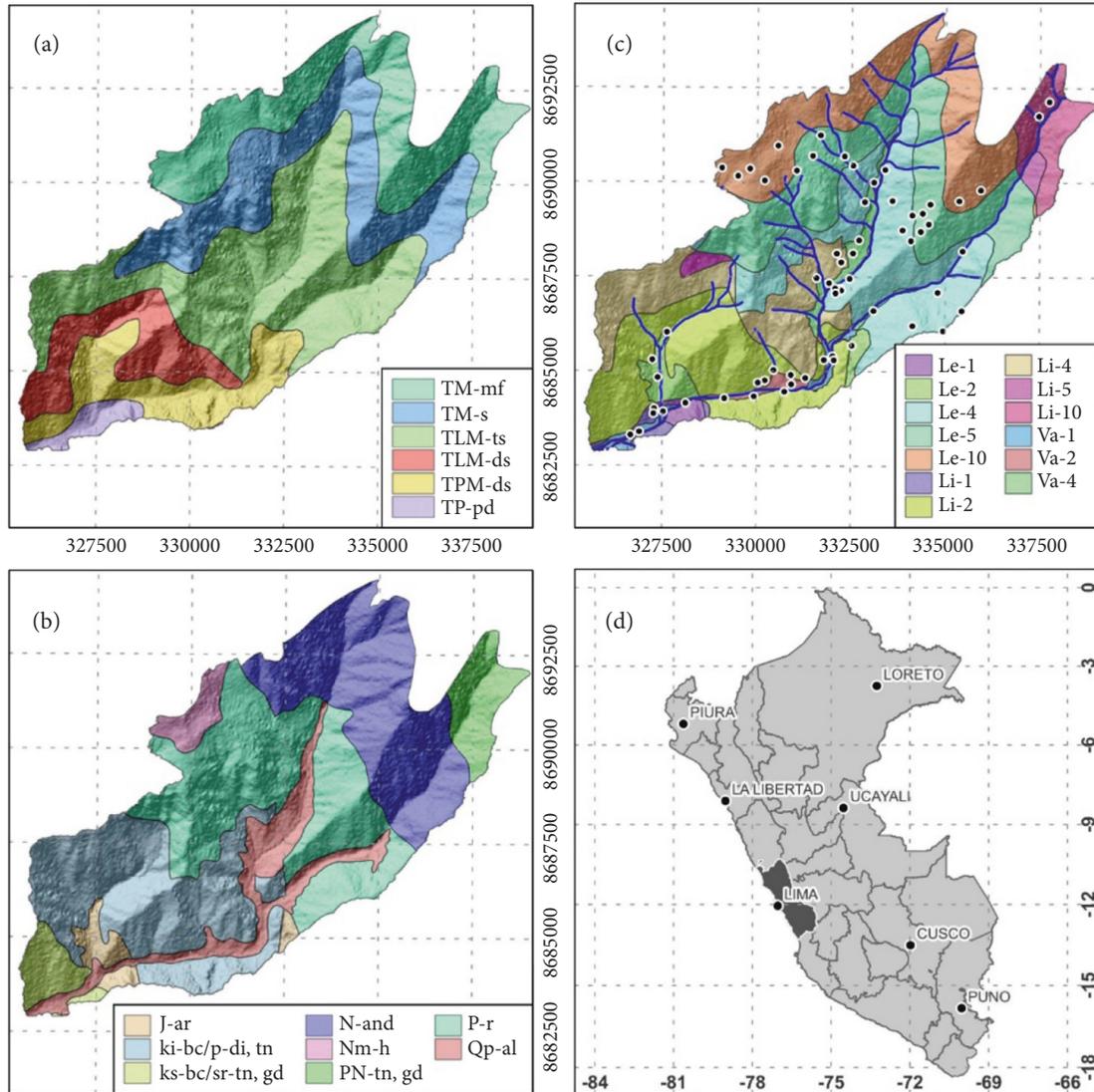


FIGURE 1: Holdridge life zones (a), geological map units (b), homogeneous land units, channel network and sample locations (c), and the location of the study area (d).

values close to one that only one soil taxa dominate. The Shannon Index is very common in ecology and is used to measure the diversity of species; we use the original formula proposed by Shannon [29] with base e for logarithm and the result units were bits. Values close to zero indicate the predominance of few soil taxa and high values the existence of several taxa at the same pixel. The feature importance based on the ROC curve was calculated using caret package, the result was scaled between 0 and 100, and radar plots were used to summarize the importance of the variables in the prediction of each soil class. All procedures previously described were carried out in R language [30] and caret package [31].

3. Results and Discussion

3.1. Soil Description. Six soil classes were determined at the subgroup level: twenty-three Lithic Torriorthents (LT), thirteen Typic Torriorthents (TT), eleven Typic

Haplocambids (TH), five Sodic Haplocambids (SH), six Fluventic Haplocambids (FH), and six Lithic Haplocambids (LH). They are distributed in colluvial-alluvial parent materials (4.3% of study area), residual intrusive materials (35.4%), and residual extrusive materials (60.3%). The intrusive material was the product of the weathering of granodiorites and diorites while extrusive material was of andesite and volcanic tuff. The sediments are coarse-textured with alkali pH, a product of the interaction with the arid climate.

3.2. Soil Sampling Quality. The results of the Kolmogorov–Smirnov test (Table 2) indicate that the ICO, RSP, PPT, and RS differ from the population distribution. In the graphic analysis (Figure 2), no marked differences are observed, which indicates that the samples are representative of the population and will not have negative repercussions on the estimation of the proposed model. This way of evaluating

TABLE 1: Predictor variables description by soil-forming factor. Mean (μ) plus standard deviation (σ) and range (minimum, maximum) are reported.

Name	Units	$\mu \pm \sigma$	Range
Relief			
Elevation (MDE)	m.a.s.l.	2739 \pm 689.53	[1202, 4349]
Slope (PD)	Percentage	70.97 \pm 29.82	[0, 387.49]
Aspect (OR)	Grades	194.99 \pm 78.29	[2.62, 360]
Convergence Index (ICO)	No units	-0.04 \pm 4.39	[-56.72, 93.72]
Plan curvature (CPP)	No units	0 \pm 0.02	[-0.17, 0.17]
Profile curvature (CPL)	No units	0 \pm 0.02	[-0.19, 0.16]
LS factor (LS)	No units	12.16 \pm 4.94	[0, 53.57]
Relative slope position (RSP)	No units	0.38 \pm 0.35	[0, 1]
Topographic Wetness Index (TWI)	m ²	5.52 \pm 1.48	[1.86, 22.52]
Valley depth (VD)	m	200.11 \pm 163.82	[0, 663.53]
Climate			
Annual cumulative evapotranspiration (ET)	mm	1258.52 \pm 98.99	[1066.92, 1414.09]
Annual cumulative precipitation (PPT)	mm	348.53 \pm 101.40	[135.95, 609.54]
Annual mean temperature (TM)	°C	11.25 \pm 3.03	[5.39, 16.59]
Water vapour pressure (WVP)	kPa	1.06 \pm 0.20	[0.71, 1.48]
Wind speed (WS)	m.s ⁻¹	2.97 \pm 0.28	[2.49, 3.54]
Total solar radiation (RS)	kJ.m ⁻² .día ⁻¹	17241 \pm 445	[16357, 18295]
Parent material			
Quartz Index (QI)	No units	1.0 \pm 0.0	0.99-1.01
Carbonates Index (CI)	No units	1.01 \pm 0.0	1.00-1.02
Mafic Index (MI)	No units	0.99 \pm 0.0	0.98-1.02
Organisms			
Normalized Difference Vegetation Index (NDVI)	No units	0.42 \pm 0.16	0-0.96
Soil biological crust (CB)	No units	0.84 \pm 0.16	0.07-1.97

TABLE 2: Population and sample comparison with Kolmogorov-Smirnov test. The null hypothesis is that sample and population have the same distribution. Bold values are those that the hypothesis was rejected.

Variable	Statistical	p value
MDE	0.075	0.864
PD	0.072	0.897
OR	0.069	0.925
ICO	0.229	0.002
CPP	0.103	0.508
CPL	0.088	0.708
LS	0.082	0.778
RSP	0.362	< 0.01
TWI	0.145	0.137
VD	0.156	0.088
ET	0.127	0.577
PPT	0.284	0.005
TM	0.081	0.967
WVP	0.103	0.820
WS	0.114	0.720
RS	0.243	0.023
QI	0.085	0.764
CI	0.087	0.719
MI	0.100	0.540
NDVI	0.154	0.098
CB	0.102	0.520

the quality of the sampling based on the predictors was successfully tested by Carvalho [32].

3.3. *Uncertainty Analysis.* All validation results show that the model is good (Table 3); global accuracy values are

greater than the noninformative ratio which confirms the statement. The result for resubstitution is the highest; unfortunately, the numerical value is not very useful because when using the same data in modeling and validation, it tends to be overestimated [33]. The retention values were 0.67 and 0.60 for global accuracy and Kappa Index, respectively, and are the lowest of the three methodologies. These have the disadvantage of sacrificing samples for validation in a context where samples less than 100 units are not convenient. Accuracy with cross-validation is lower than retention, but the confidence intervals overlap, showing that both results are not statistically different. Cross-validation is one of the most widely used methodologies in soil classes pedometric mapping [34-36] which reinforces that cross-validation is the most appropriate method to estimate global uncertainty in the conditions of our study.

The results are very promising compared to similar studies. Silva et al. [37] obtained values of 0.49 and 0.33 for the uncertainty coefficients using the random forest in 74 samples. Taghizadeh et al. [38] modeled the subgroups of 194 samples with ensemble models and obtained 0.43 of global accuracy and Mirakzehi et al. [39] obtained 0.44 with 108 samples and random forest. Jeune et al. [40] obtained a Kappa Index of 0.45 with multiple linear regression and 0.42 with random forest. All cases used k-fold cross-validation to obtain the parameters, but the values are much lower than ours.

The local uncertainty (Figure 3) shows that in the lower and upper part of the microbasin (lower left and upper right), there is little diversity of soils (Shannon Index), but there is a cooccurrence of the two most probable taxa (Confusion

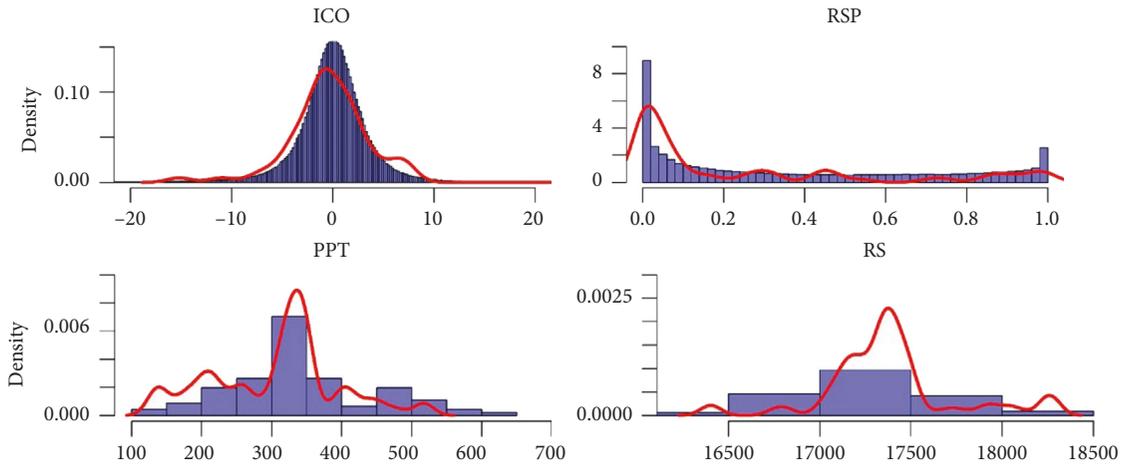


FIGURE 2: Histograms of population density (blue bars) and sample density (red lines).

TABLE 3: Global uncertainty parameters. Confidence interval (CI) and no information ratio (NIR) are reported for global accuracy analysis.

	Global accuracy		Kappa Index	NIR
	Mean	CI	Mean	
Resubstitution	0.8610	0.7920–0.9140	0.8330	0.1680
Cross validation	0.7433	0.6260–0.8606	0.6906	
Retention	0.6290	0.4490–0.7850	0.5520	

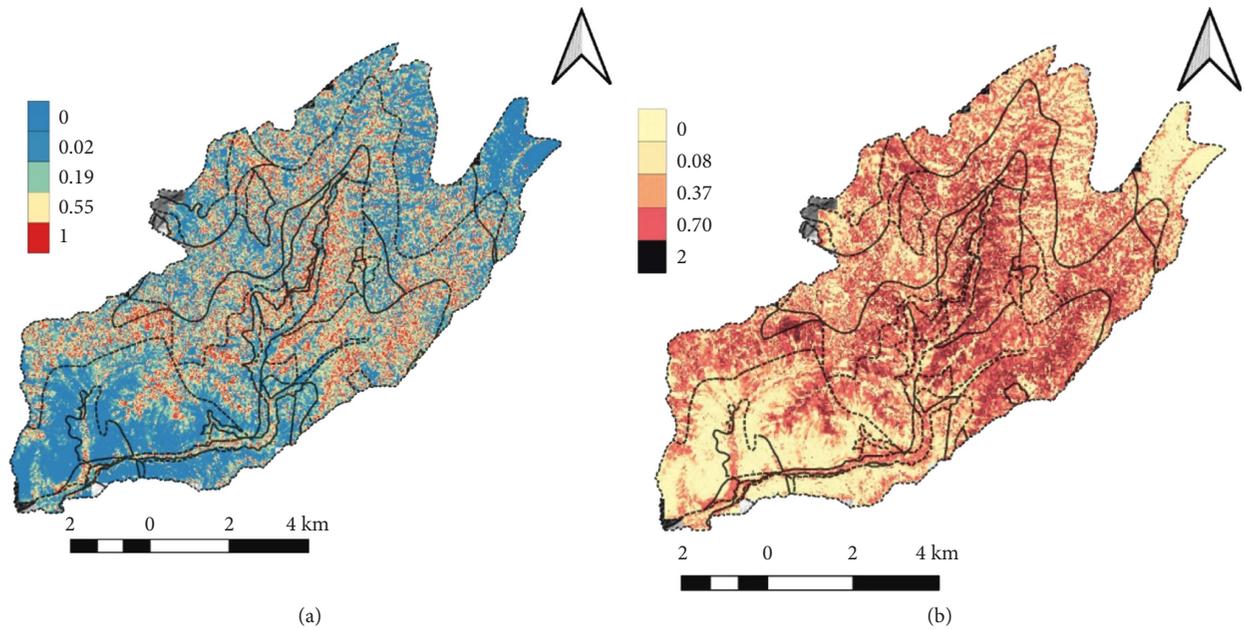


FIGURE 3: Spatial distribution of local uncertainty. The Confusion Index (a) shows the cooccurrence of soil classes and the Shannon Index (b) is an indicator of the presence of more than two soil classes.

Index). These concepts can be used to delimit cartographic consociations and associations. The mean values of indexes are 0.31 and 0.43; in general, we can say that the study area has low pedodiversity. Lamichhane et al. [41] obtained similar results for little soil diversity and a slight predominance of a single taxon, with the exception of mountainous areas.

However, their study was carried out on a national scale (1: 1000000) and as a result of a disaggregation process to estimate the distribution of taxa based on the cartographic units of the soil. Odgers et al. [42] mentioned values of 0.1 and 0.9 to decide if there is confusion or not, these are very conservative, but using them as a reference, it is possible to affirm

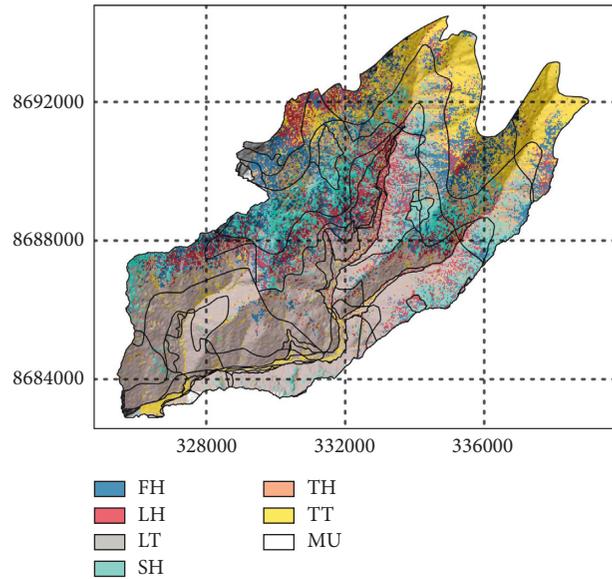


FIGURE 4: Spatial distribution of soil classes at the subgroup level in soil taxonomy. The black lines represent the boundaries of homogeneous land units (MU).



FIGURE 5: The relative importance of predictors on soil classes.

that the blue areas have greater uncertainty. This coincides with the areas where the density of samples was lower due to the harsh accessibility conditions.

3.4. Soil Classes Prediction. The cross-validation model was used to predict the probabilities of each soil class. A

threshold of 0.5 was used to decide the soil class to be represented per pixel (Figure 4). Fluventic Haplocambids cover 13.36 percent, Lithic and Typic Haplocambids occupy 11.28 percent each; 15.45 is occupied by the Typic Torriorthents, 10.79 by the Sodic Haplocambids, and 37.84 by the Lithics Torriorthents. Arid conditions, steep slopes, and

absence of vegetation favor the regressive evolution of soils and result in the predominance of Torriorthents, especially Lithic Torriorthents. The map shows that Haplocambids have random distribution and Torriorthents are spatially aggregated. Despite the favorable climatic conditions for the development of Aridisols, the other factors locally conditioned their distribution.

Studies that seek to predict soil classes use an average of 20 covariates [43] and the most widely used are those related to the digital elevation model, climate grids, and vegetation [44]. Our study used 21 covariates where Topographic Wetness Index (TWI), derived from the digital elevation model, stands out in importance for the prediction of Lithic Haplocambids, Sodic Haplocambids, Typic Torriorthents, and Lithic Torriorthents (Figure 5). Fluventic and Typic Haplocambids are more influenced by Total Solar Radiation (RS), quartz (QI), and carbonate (CI) indices. The irregularity of the importance of predictors does not allow us to indicate which soil-forming factor is the most influential for the study area; however, in broad strokes, we can highlight the relief and parent material.

Conventional soil maps represent the spatial distribution of mapping units and abstractions of the soil boundaries determined by photointerpretation [1]. In contrast, pedometric maps show the distribution of taxonomic units or soil classes. Consequently, conventional maps significantly underestimate the area occupied by each soil class [45] having a direct effect on the interpretation of the stakeholders. Pedometric mapping also allows quantifying the uncertainty of the final maps [46], data that is difficult to obtain with conventional methods. Reproducibility is another factor outperformed by pedometric mapping [47], using the same data and computer codes, different users obtain the same result. Some strong disadvantages are that pedologists require additional skills such as computational and statistical knowledge [48]. In places where the training of pedologists does not include these skills, it leads to rejection in the adoption of these methods. The development of applied research on pedometric mapping facilitates the migration from conventional methods and shows the advantages of using the methodology in local conditions.

4. Conclusion

Pedometric mapping applying the Naïve Bayes model produced the soil map of San Mateo de Otao with low uncertainty (accuracy=0.77) and quantified the area occupied by each soil class. The Lithic Torriorthents and Sodic Haplocambids were the largest and smallest subgroups in the district. An aggregate (Torriorthents) and random (Haplocambids) distribution was observed present in the maps, and relief and parent material were identified as the most important factors in predicting soil classes. Everything suggests a considerable advantage of pedometric mapping compared to conventional mapping. The application on San Mateo de Otao and under the conditions of the soil survey of Peru suggest that it can be used in the national territory. It is recommended to continue investigating the application of

pedometric mapping at different cartographic scales and with the most modern machine learning models.

Data Availability

The soil data, covariates, and code used to support the findings of this study have been deposited in the FigShare repository (<https://doi.org/10.6084/m9.figshare.c.5839364.v1>).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the laboratory for analysis of soils, plants, water, and fertilizers (LASPAP) of the National Agrarian University La Molina (UNALM). The authors thank Vanessa Llamoya, Rosel Aguilar, Michael Sandoval, Jose Tarrillo, Sandro Martinez, and the San Mateo community for their support in the field and laboratory evaluations.

References

- [1] Soil Survey Division Staff, C. Ditzler, K. Scheffe, and H. Monger, *Soil Survey Manual (Handbook 18)*, Government Printing Office, Washington, D.C, 2017.
- [2] E. C. Brevik, C. Calzolari, B. A. Miller et al., "Soil mapping, classification, and pedologic modeling: history and future directions," *Geoderma*, vol. 264, pp. 256–274, 2016.
- [3] A. A. Nikiforova, M. E. Fleis, M. V. Nyrtsov et al., "Problems of modern soil mapping and ways to solve them," *Catena*, vol. 195, Article ID 104885, 2020.
- [4] T. Behrens and T. Scholten, "Digital soil mapping in Germany—a review," *Journal of Plant Nutrition and Soil Science*, vol. 169, no. 3, pp. 434–443, 2006.
- [5] S. Dharumarajan, R. Hegde, N. Janani, and S. K. Singh, "The need for digital soil mapping in India," *Geoderma Regional*, vol. 16, Article ID e00204, 2019.
- [6] M. Zeraatpisheh, A. Jafari, M. Bagheri Bodaghabadi et al., "Conventional and digital soil mapping in Iran: past, present, and future," *Catena*, vol. 188, Article ID 104424, 2020.
- [7] D. Arrouays, A. C. Richer-de-Forges, F. Héliès et al., "Impacts of national scale digital soil mapping programs in France," *Geoderma Regional*, vol. 23, Article ID e00337, 2020.
- [8] J. A. Thompson, S. Kienast-Brown, T. D'avello, J. Philippe, and C. Brungard, "Soils 2026 and digital soil mapping - a foundation for the future of soils information in the United States," *Geoderma Regional*, vol. 22, Article ID e00294, 2020.
- [9] R. Searle, A. McBratney, M. Grundy et al., "Digital soil mapping and assessment for Australia and beyond: a propitious future," *Geoderma Regional*, vol. 24, Article ID e00359, 2021.
- [10] M. G. Velásquez Ramírez, J. A. G. Barrantes, E. Thomas et al., "Heavy metals in alluvial gold mine spoils in the peruvian amazon," *Catena*, vol. 189, Article ID 104454, 2020.
- [11] H. I. Mejia Cabrera, D. Vilchez, V. A. Tuesta, M. G. Forero, A. G. Tescher, and T. Ebrahimi, "Soil salinity estimation of sparse vegetation based on multispectral image processing

- and machine learning,” *Presented at the Applications of Digital Image Processing*, vol. XLIII, 2020.
- [12] O. Wigmore, B. Mark, J. McKenzie, M. Baraer, and L. Lautz, “Sub-metre mapping of surface soil moisture in proglacial valleys of the tropical Andes using a multispectral unmanned aerial vehicle,” *Remote Sensing of Environment*, vol. 222, pp. 104–118, 2019.
- [13] D. Nikitina, L. Remizova, and R. L. Losco, “A preliminary investigation of the soils and geomorphology of a portion of the madre de Dios region, Peru,” *Soil Horizons*, vol. 52, no. 2, pp. 40–2, 2011.
- [14] L. Steinbuch, D. J. Brus, and G. B. M. Heuvelink, “Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors,” *Geoderma*, vol. 316, pp. 56–69, 2018.
- [15] Instituto Nacional de Recursos Naturales, *Mapa Ecológico del Perú: Guía Explicativa*, 195 pages, Instituto Nacional de Recursos Naturales, Lima, 1995.
- [16] H. Salazar-Díaz, *Geología de los cuadrángulos de Matucana y Huarochirí*, 74 pages, Instituto Geológico, Minero y Metalúrgico del Perú, Lima, 1983.
- [17] O. Palacios-Moncayo, J. Caldas-Vidal, and C. Vela-Velásquez, *Geología de los cuadrángulos de Lima, Lurín, Chancay y Chosica*, 162 pages, Instituto Geológico, Minero y Metalúrgico del Perú, Lima, 1992.
- [18] L. Congedo, “Semi-Automatic Classification Plugin: a Python tool for the download and processing of remote sensing images in QGIS,” *Journal of Open Source Software*, vol. 6, no. 64, 64 pages, Article ID 3172, 2021.
- [19] S. E. Fick and R. J. Hijmans, “World clim 2: new 1-km spatial resolution climate surfaces for global land areas,” *International Journal of Climatology*, vol. 37, no. 12, pp. 4302–4315, 2017.
- [20] Instituto geológico Minero y Metalúrgico, *GEOCATMIN: Sistema de Información Geológico y Catastral Minero* Instituto geológico Minero y Metalúrgico, Lima, 2012, <https://geocatmin.ingemmet.gob.pe/>.
- [21] J. A. Zinck, G. Metternicht, G. Bocco, and H. F. Del Valle, *Geopedology*, Springer, Berlin, Germany, 2016.
- [22] Soil Survey Staff, *Soil Survey Field and Laboratory Methods Manual: Soil Survey Investigations Report No 51*, Natural Resources Conservation Service, United States, 2nd edition, 2014.
- [23] Soil Survey Staff, *Kellogg Soil Survey Laboratory Methods Manual: Soil Survey Investigations Report No 42*, Natural Resources Conservation Service, United States, 5th edition, 2014.
- [24] Soil Survey Staff, *Keys to Soil Taxonomy*, Natural Resources Conservation Service, Washington D.C, 12th edition, 2014.
- [25] X. P. Maldague, Y. Ninomiya, and A. E. Rozlosnik, “Mapping quartz, carbonate minerals, and mafic-ultramafic rocks using remotely sensed multispectral thermal infrared ASTER data,” *Thermosense*, vol. XXIV, 2002.
- [26] A. Karnieli, “Development and implementation of spectral crust index over dune sands,” *International Journal of Remote Sensing*, vol. 18, no. 6, pp. 1207–1220, 2010.
- [27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning (Springer Series in Statistics)*, Springer, Berlin, Germany, 2009.
- [28] P. A. Burrough, P. F. M. Van Gaans, and R. Hootsmans, “Continuous classification in soil survey: spatial correlation, confusion and boundaries,” *Geoderma*, vol. 77, no. 2-4, pp. 115–135, 1997.
- [29] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [30] R Foundation, *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria, 2021, <https://www.R-project.org/>.
- [31] M. Kuhn, “Building predictive models in R Using the caret Package,” *Journal of Statistical Software*, vol. 28, no. 5, 2008.
- [32] W. D. Carvalho Júnior, C. D. S. Chagas, A. Muselli, H. S. K. Pinheiro, N. R. Pereira, and S. B. Bhering, “Método do hiper cubo latino condicionado para a amostragem de solos na presença de covariáveis ambientais visando o mapeamento digital de solos,” *Revista Brasileira de Ciência do Solo*, vol. 38, no. 2, pp. 386–396, 2014.
- [33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning (Springer Texts in Statistics)*, Springer, Berlin, Germany, 2021.
- [34] L. Zhang, L. Yang, T. Ma, F. Shen, Y. Cai, and C. Zhou, “A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data,” *Geoderma*, vol. 384, Article ID 114809, 2021.
- [35] I. Esfandiarpour-Boroujeni, M. S. Shamsabadi, H. Shirani, Z. Mosleh, M. Bagheri Bodaghabadi, and M. H. Salehi, “Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations,” *Catena*, vol. 191, Article ID 104581, 2020.
- [36] S. J. Forghani, M. R. Pahlavan-Rad, M. Esfandiari, and A. M. Torkashvand, “Spatial prediction of WRB soil classes in an arid floodplain using multinomial logistic regression and random forest models, south-east of Iran,” *Arabian Journal of Geosciences*, vol. 13, no. 13, 2020.
- [37] B. P. C. Silva, M. L. N. Silva, F. A. P. Avalos, M. D. De Menezes, and N. Curi, “Digital soil mapping including additional point sampling in Posses ecosystem services pilot watershed, southeastern Brazil,” *Scientific Reports*, vol. 9, no. 1, 2019.
- [38] R. Taghizadeh-Mehrjardi, B. Minasny, N. Toomanian, M. Zeraatpisheh, A. Amirian-Chakan, and J. Triantafyllis, “Digital mapping of soil classes using ensemble of models in is Fahan Region, Iran,” *Soil Systems*, vol. 3, no. 2, 37 pages, 2019.
- [39] K. Mirakzehi, M. R. Pahlavan-Rad, A. Shahriari, and A. Bameri, “Digital soil mapping of deltaic soils: a case of study from Hirmand (Helmand) river delta,” *Geoderma*, vol. 313, pp. 233–240, 2018.
- [40] W. Jeune, M. R. Francelino, E. d. Souza, E. I. Fernandes Filho, and G. C. Rocha, “Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti,” *Revista Brasileira de Ciência do Solo*, vol. 42, no. 0, 2018.
- [41] S. Lamichhane, L. Kumar, and K. Adhikari, “Updating the national soil map of Nepal through digital soil mapping,” *Geoderma*, vol. 394, Article ID 115041, 2021.
- [42] N. P. Odgers, W. Sun, A. B. McBratney, B. Minasny, and D. Clifford, “Disaggregating and harmonising soil map units through resampled classification trees,” *Geoderma*, vol. 214–215, pp. 91–100, 2014.
- [43] A. M. J.-C. Wadoux, B. Minasny, and A. B. McBratney, “Machine learning for digital soil mapping: applications, challenges and suggested solutions,” *Earth-Science Reviews*, vol. 210, Article ID 103359, 2020.
- [44] D. Kidd, R. Searle, M. Grundy et al., “Operationalising digital soil mapping-lessons from Australia,” *Geoderma Regional*, vol. 23, Article ID e00335, 2020.

- [45] A. P. Zhidkin, M. A. Smirnova, A. N. Gennadiev, S. V. Lukin, Y. A. Zazdravnykh, and N. I. Lozbenev, "Digital mapping of soil associations and eroded soils (prokhorovskii district, belgorod oblast)," *Eurasian Soil Science*, vol. 54, no. 1, pp. 13–24, 2021.
- [46] S. Chen, D. Arrouays, V. Leatitia Mulder et al., "Digital mapping of GlobalSoilMap soil properties at a broad scale: a review," *Geoderma*, vol. 409, Article ID 115567, 2022.
- [47] I. Y. Savin, A. V. Zhogolev, and E. Y. Prudnikova, "Modern trends and problems of soil mapping," *Eurasian Soil Science*, vol. 52, no. 5, pp. 471–480, 2019.
- [48] D. Arrouays, A. McBratney, J. Bouma et al., "Impressions of digital soil maps: the good, the not so good, and making them ever better," *Geoderma Regional*, vol. 20, Article ID e00255, 2020.