Geoderma 436 (2023) 116568

Contents lists available at ScienceDirect

Geoderma

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

Downscaling legacy soil information for hydrological soil mapping using multinomial logistic regression

I.E. Smit [a,*], G.M. Van Zijl [b], E.S. Riddell [c,d], J.J. Van Tol [a]

[a] Department of Soil-, Crop- and Climate Sciences, University of the Free State, Bloemfontein 9300, South Africa
[b] Unit for Environmental Sciences and Management, North-West University, Potchefstroom 2520, South Africa
[c] South African National Parks, Kruger National Park, Skukuza, South Africa
[d] Centre for Water Resources Research, University of KwaZulu-Natal, Pietermaritzburg, South Africa

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Hydropedology
K-means clustering
Digital soil mapping
SMOTE

ABSTRACT

In South Africa, there is a growing demand for large scale detailed hydrological soil maps for modelling and management purposes. However, imbalanced legacy soil information often impedes the accurate creation of such maps by not being representative of the environmental complexity of large-scale catchments and containing imbalanced soil class distributions, often resulting in the loss of minority soil classes, which are often of great hydrological importance (e.g., wetland and riparian soils). In this study, we proposed a new downscaling approach to handle spatially localised legacy soil data within a larger low resolution legacy soil dataset to create an accurate hydrological soil map of the macro-scale (5790 km^2) Sabie-Sand catchment using multinomial logistic regression (MNLR). The spatially localised legacy data was downscaled using k-means clustering and added to the broader legacy dataset. Five levels of legacy soil data were analysed in their representation of environmental covariates using QQ-plots and a Welsh's *t*-test and their mapping accuracy using confusion matrix's and Kappa coefficient statistics. However, MNLR also requires balanced soil classes. The value of the best performing legacy soil dataset was also compared to using all available soil information after both had their soil class distributions fully balanced using Synthetic Minority Oversampling Technique (SMOTE). The 500 ha/observation-SMOTE dataset resulted in the most accurate hydrological soil map with a validation point accuracy of 73% and a Kappa coefficient of 0.60, substantially outperforming the other downscaled soil maps as well as the SMOTE balanced dataset using all available soil information. This was due to the decreased variation between observations and catchment means, where the 500 ha/observation dataset yielded the least variation between soil observation and catchment datasets and well as reducing the class imbalance within the legacy soil data. Downscaling spatially localised legacy soil data for environmental representation is an effective tool to improve digital soil mapping accuracy using MNLR.

1. Introduction

Digital Soil Mapping (DSM) involves mathematical models for predicting soil continuous and soil class properties using environmental covariates as predictors. The modelling procedure can be implemented using the framework of DSM (McBratney et al., 2003) to relate the environmental covariates with the target soil variable or class. Advances in DSM coincided with the need for detailed and accurate spatial soil information, which has widely been realized to provide appropriate solutions for conserving and managing agricultural and environmental resources. This is achieved by soil information's ability to improve modelling, policy making and scenario analysis at different spatial extents from catena to catchment and from national to global scales (Häring et al., 2012; Arrouays et al., 2014; Lamichhane et al., 2021).

DSM has therefore developed into a cost-effective tool for creating large scale detailed soil maps. This is due to the expansion of highly detailed remotely sensed covariate data, the drastic increase in desktop computing power and its availability to users as well as the availability of large swaths of legacy soil data (Häring et al., 2012). In South Africa, DSM has also been used for a wide range of agricultural and environmental applications including precision agriculture, land degradation studies, land capability studies, determining irrigation potential as well as in Environmental Impact Assessments (EIAs) and town planning projects (Van Zijl, 2019). DSM has also been increasingly used in the

* Corresponding author.
E-mail address: smitie@ufs.ac.za (I.E. Smit).

https://doi.org/10.1016/j.geoderma.2023.116568
Received 19 March 2023; Received in revised form 2 June 2023; Accepted 8 June 2023
Available online 15 June 2023

field of hydropedology, where micro-, and *meso*-scale hydrological soil maps have been created for modelling purposes and have been shown to improve hydrological modelling accuracy (Van Tol et al., 2015; Van Tol et al., 2020; Harrison et al., 2022; Smit and Van Tol, 2022). The value of hydrological soil information was tested by Van Tol et al. (2015), where the use of more detailed soil descriptions using hydropedological insight within the Agriculture Catchments Research Unit (ACRU) model led to an increase in Nash–Sutcliffe efficiency (NSE) of between 9% and 52% (Van Tol et al., 2015). Van Tol et al. (2020) also found that hydropedological soil information improved modelling accuracy at three different catchment sizes compared to readily available soil information, and that therefore hydrological soil information can improve hydrological modelling accuracy at various scales. The value of hydrological soil information may also extend beyond the ability to accurately model long-term streamflow predictions. The argument is that hydrological soil information may serve as an effective "soft data" tool, to better represent internal hydrological processes within a catchment. Smit and Van Tol (2022) illustrated these benefits where hydrological soil information improved the representation of internal catchment processes in a 230 km^2 catchment in South Africa.

In this light, the Water Research Commission of South Africa has seen the potential of large-scale hydrological soil maps for hydrological modelling and water resource management purposes and has also authorised a project which builds towards a hydrological soil map of the country as underpinned by Van Tol and Van Zijl (2022) which provided rough steps to create a national hydrological soil map for South Africa. However, the creation of such large-scale hydrological soil maps remains reliant on using and interpreting large amounts of legacy soil data, from various different soil surveys, which often use different classification systems at different spatial resolutions.

DSM provides unique challenges in balancing legacy soil datasets because soils are never evenly distributed throughout a landscape, being a product of variations within complex soil forming factors. Brungard et al. (2015) reported that although machine learning models, such as Multinomial logistic regression (MNLR), were generally more accurate than simple models, the accuracy of these models are highly dependent upon the number of soil classes and the frequency of their distribution. MNLR has been widely used for DSM purposes, specifically for large scale mapping endeavours (Campling et al., 2002; Bailey et al., 2003; Hengl et al., 2007; Kempen et al., 2009; Debella-Gilo and Etzelmüller 2009; Van Zijl, 2019). For more information regarding the MNLR algorithm see Venables and Ripley (2002). However, the performance of standard classification algorithms, such as the MNLR algorithm, is often poor when learning from imbalanced data, which is well documented in the field of categorical data modelling (López et al., 2013; Haixiang et al., 2017; Li et al., 2022).

Legacy soil datasets are often not representative of the environmental complexity of large-scale catchments, due to a limited number of soil observations, also containing imbalanced soil class distributions, leading to poorly performing predictive models. This is due to most legacy soil datasets containing spatially localised sampling locations which were purposively selected according to the aim and purpose of the soil survey (Ma et al., 2019). For example, most agricultural soil surveys would refrain from sampling within specific low potential or prohibited positions in the landscape. Creating hydrological soil maps from legacy soil data the minority soil class are often of great importance (e.g., undersampled wetlands and riparian soils) and predictive models based off these spatially imbalanced legacy soil datasets, would fail to map these ecologically important soils.

When addressing the challenge of imbalanced datasets, three main approaches have been recognised. Firstly, there is the data level approach where different resampling methods are used to create a balanced dataset. These methods include, undersampling, where the majority soil class is reduced to balance observations. Oversampling, which creates replications of the minority soil class and synthetic data generation, where new artificial data of the minority observations are created to balance class distributions within datasets. The second approach is at the algorithm level, where algorithms are selected which are capable of handling imbalanced datasets, these include algorithms which incorporate cost sensitive learning and active learning (Chawla et al., 2002). The third approach is to apply a combination of both data-level and algorithm-level approaches to produce the most accurate results (He and Garcia, 2008; García and Herrera, 2009).

Within this paper we will be focussing on the data level approach to create a balanced soil dataset. Fairly limited data balancing research specific to the field of soil science has been conducted (Heung et al., 2016; Sharififar et al., 2019a; Taghizadeh-Mehrjardi et al., 2020). Recently, Sharififar et al. (2019b) analysed the use of random oversampling (ROS) and random undersampling (RUS) methods to balance soil datasets using various machine learning models, where majority soil observations were down-sampled and minority soil observations up-sampled, preserving proportionality, to deal with the issue of imbalanced soil data with 452 profiles observations in an on an area of about 12,000 ha. They concluded that balancing soil datasets using a combined approach of both RUS and ROS significantly improved MNLR accuracy and decreased mapping uncertainty. However, these simplistic random resampling techniques, such as RUS and ROS, potentially increases the likelihood of overfitting by discarding potentially useful observations, especially when large difference occur between the number of majority and minority soil classes (Zhu et al., 2017; Piri et al., 2018).

Taghizadeh-Mehrjardi et al. (2020) analysed eight resampling strategies on five of the most used machine learning algorithms on a national scale for Iran (1,648,195 km^2) using 7,664 soil observations. The researchers concluded the that highest increase in prediction accuracy was achieved using the synthetic minority oversampling technique (SMOTE) which as the name suggests generates synthetic examples of the minority soil classes. SMOTE uses the existing minority samples and interpolates between samples and their covariate attributes to generate new samples of the specific class (Chawla et al., 2002). Taghizadeh-Mehrjardi et al. (2020) also established that oversampling approaches of the minority soil class outperformed undersampling techniques. This is due to the fact that useful information, which was obtained by costly, time consuming and labour-intensive soil sampling, in the majority soil classes are ignored, leading to the degradation of classifier performance. ROS was also shown to provide poor performance results when legacy datasets contained large differences between majority and minority soil classes in accordance with Peri et al. (2018). ROS creates exact copies of the minority soil classes, which leads to a small decision region for the minority soil classes compared to the majority soil class and therefore the likelihood of overfitting increases substantially (Chawla et al., 2002; Zarinabad et al., 2017). Sharififar et al. (2019b) also cited the need for more research regarding the effects of different balancing techniques on different classifiers within the field of DSM.

The main aim of this study was to create an accurate hydrological soil map of a strategic water source macro-scale catchment in South Africa using MNLR and legacy soil information in accordance with the objectives set out by the Water Research Commission of South Africa. This aim was achieved by providing an approach to handle large spatially localised legacy soil datasets for DSM purposes using MNLR. Our first objective was to address the use of a highly spatially localised and imbalanced legacy soil dataset within the legacy soil data available. This was achieved by first downscaling localised data to improve environmental covariate representativeness of the broader legacy soil dataset prior to balancing soil classes. Our second objective was to analyse the value of the downscaling legacy soil data by firstly, analysing mapping results without balancing soil classes, and secondly, comparing mapping results when soil classes are balanced between our representativeness legacy soil dataset and a legacy soil dataset containing all available soil observations, using synthetically generated soil observations.

2. Material and methods

2.1. Study area

The 5790 km^2 Sabie-Sand catchment is in the Mpumalanga province of South Africa (Fig. 1) and forms part of the larger transboundary Incomati river basin. The catchment stretches from the Drakensberg Escarpment in the west at an altitude of 2200 m above sea level (m.a.s.l.) and gradually flattens towards the east with an altitude of 150 m.a.s.l. before the Sabie River flows into Mozambique.

With a semi-arid warm and hot climate in the east of the catchment and a temperate warm climate in the west of the catchment, a strong rainfall gradient, ranging from 1600 mm in the west to 450 mm in the east exists within the catchment. Rainfall occurs mainly in the austral summer (November through to March) and normally results from convective thunderstorms, although periodic high-intensity rainfall events do occur from cyclones that form over the Indian Ocean and track inland, where the orographic effect of the Drakensberg Escarpment creates severe localised flooding (Kruger et al., 2002).

The main bio-regions of the catchment consists of savanna at lower altitudes and montane grasslands and montane forests in the mountainous regions, which have been heavily altered by commercial forestry applications (Mucina and Rutherford, 2006). The area comprises various ranges of bedrock lithologies including, quartzites, granites, basalts, conglomerates, andesites, gneiss and shales (Council for Geosciences, 2007). The Sabie-Sand system is one of the most biodiverse river systems in South Africa and thus a flagship catchment for the provision of environmental flows to the Kruger National Park, and transboundary flows to the Republic of Mozambique.

2.2. Soil data

The legacy soil data of the Sabie-Sand catchment in total amounts to 12,875 soil observations from various soil surveys. The Agricultural Research Council (ARC) made 380 observations within the study area as part of the national soil survey (Land Type Survey Staff, 1976–2006), 108 soil observations by an infield hydropedological survey and 118 observations from various research and consultancy projects from various private and state-owned enterprises. However, the majority of soil observations within the catchment originate from a single forestry soil survey done by the South African Forestry Company Limited (SAFCOL) where 12,269 soil observations we made on 38,000 ha. The total soil observation density of the SAFCOL legacy data is 3.4 ha per soil observation compared to the 960 ha per soil observation for the remaining legacy soil information within the catchment.

The spatial representation (Fig. 2) of legacy soil observations within the catchment illustrates a clear bias in the mountainous (western) section of the catchment, visually illustrating the spatially localised nature of the SAFCOL soil dataset.

Due to the large size of the catchment area and large number of soil observations, a wide variety of soils occur within the catchment, driven by the differences in soil forming factors (parent material, climate, topography, organisms, and time) between the afromontane regions and the semi-arid lowveld regions of the catchment. This spatial imbalance also results in an imbalanced number of the different soil types and therefore mapping classes within the catchment.

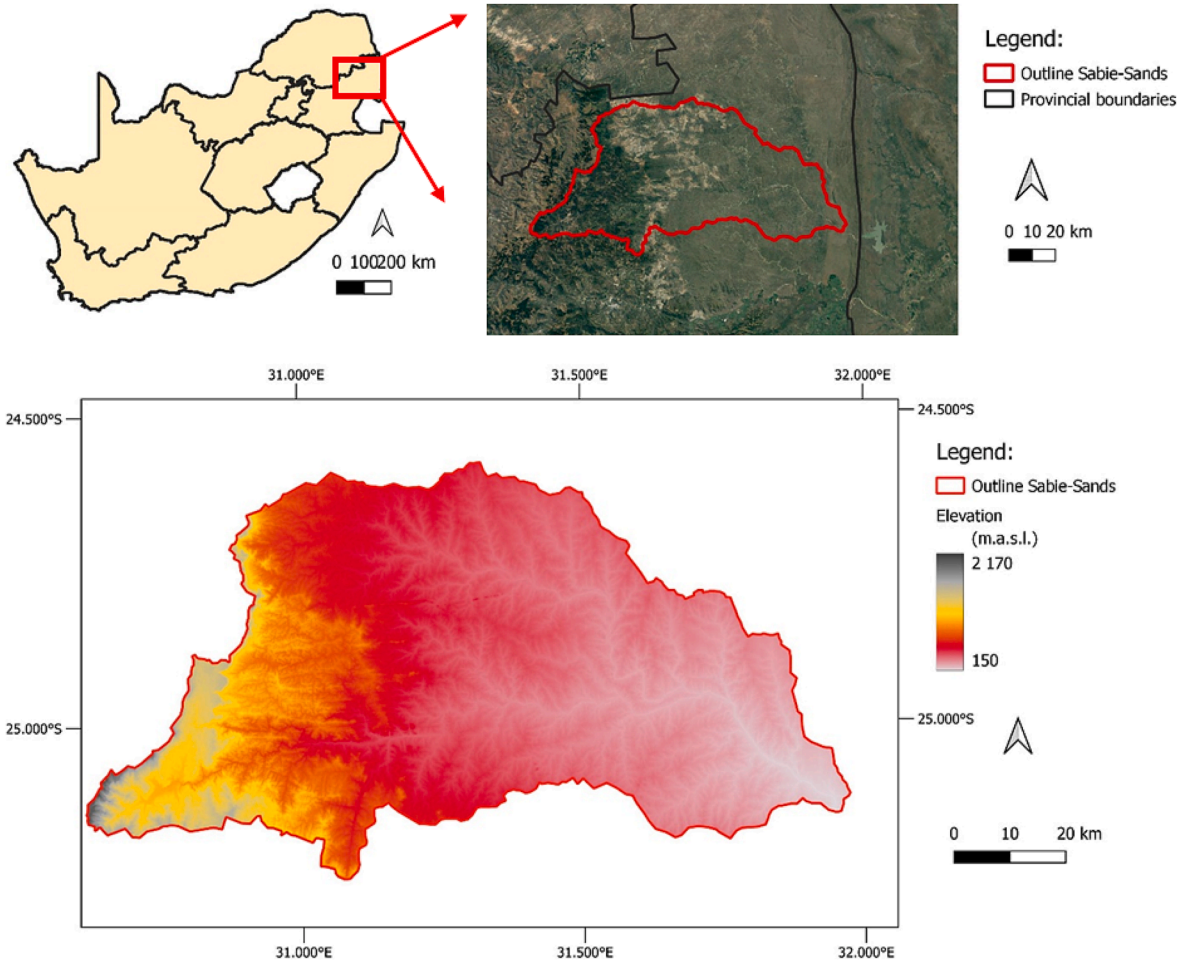

Fig. 1. The location of the study area.

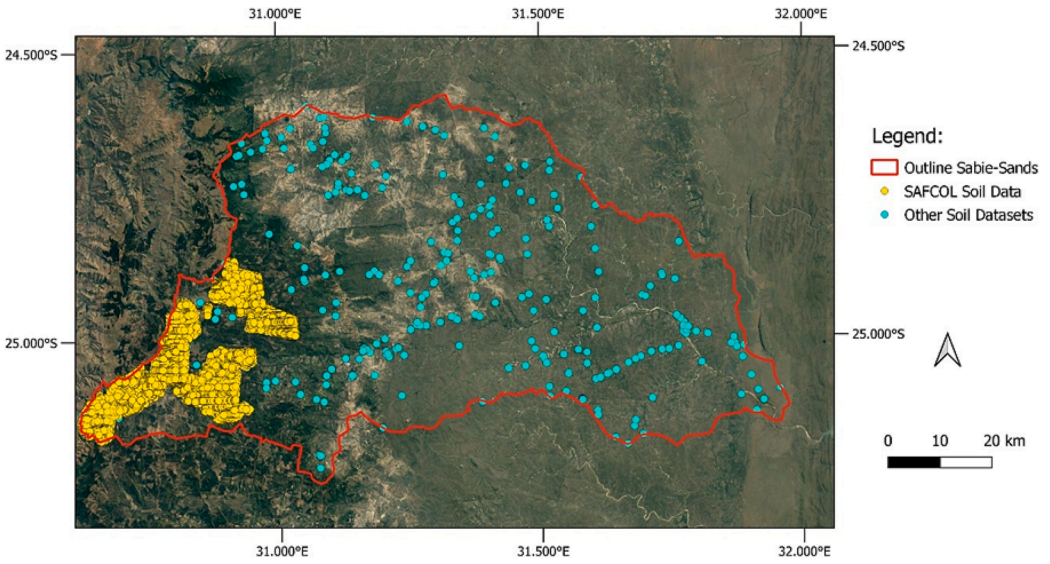

Fig. 2. The spatial distribution of legacy soil datasets within the Sabie-Sand.

Using the hydropedological groupings of South African soils (Van Tol and Le Roux, 2019) the soils of the Sabie-Sand were divided into six hydrological soil types (Table 1). Recharge deep soils comprise approximately 90% of the total soil samples, recharge shallow soils 3%, responsive saturated soils comprise approximately 0.5%, responsive shallow soils 5%, A/B interflow soils comprise approximately 0.6% and soil/bedrock interflow soils 1%.

The mountainous soils of the Sabie-Sand are characterized by well weathered soils being either deep apedal soils (Acrisols and Nitisols) or shallow apedal soils (Leptosols) on midslope and hillcrest positions depending on the parent material. Alluvial (Fluvisols) and saturated high clay (Gleysols) soils are also prominent of footslope and valley bottom terrain positions depending on upslope environmental covariates.

The lowland soils are comparatively far shallower than the mountainous soils, gravellier, being less well weathered and primarily controlled by differences in parent material. These soils show a more distinct toposequence with apedal soils (Acrisols and Nitisols) on hillcrest, albic soils (Arenosols) on midslope and footslope terrain positions. A small number of duplex soils (Solonetz) are also present on footslope and valley bottom terrain positions. Apedal soils (Nitisols and Fluvisols) are also present on valley bottom terrain positions as floodplains form on the major river networks (IUSS Working Group WRB, 2015).

The hydropedological grouping of different soil types into six conceptual classes decreases the number of soil mapping classes but also creates larger variation between the number of soil observations per mapping class and thus further adds to the imbalance within the legacy soil dataset. This is especially true when a majority of the soil observations are spatially localised as in the Sabie-Sand catchment, where the majority soil class (e.g., recharge deep: 90%) massively overshadows the smallest minority soil class (e.g., responsive saturated: 0.5%). Creating a national hydrological soil map would be highly reliant on these spatially localised legacy soil datasets, especially in mountainous regions where the majority of soil observations consist of commercial forestry surveys. These detailed surveys would add a substantial amount of soil observations to the available soil mapping resources in South Africa. Therefore, developing a protocol for handling these localised datasets to improve DSM accuracy is imperative moving forward.

2.3. Multinomial logistic regression

MNLR forms a part of the broader family of generalized linear models and is applied when the target variable contains more than two categorical variables. This helps predict the probability of the occurrence of each unique soil mapping unit. As explained by Kempen et al. (2009), if a given variable Y_i represents the observed soil mapping unit at a given observation location, with $i = 1,\ldots, n$ and n is the number of soil mapping units within the study area. In case n equals 2 and Y has outcomes Y_1 and Y_2. Both the counts of Y_1 and Y_2 therefore follow a binomial distribution. The probability of occurrence of Y_1 is π_1 and that of Y_2 is π_2. Logistic regression relates probability π_1 to a set of predictors, in our case environmental covariates, using the logit link function:

$$logit(\pi_1) = ln\frac{\pi_1}{\pi_2} = \ln\left(\frac{\pi_1}{1-\pi_1}\right) = x'\beta \quad (1)$$

where x is a vector of environmental covariates, and β is a vector of model coefficients that are typically estimated by maximum likelihood. Therefore, Eq. (1) can also be rewritten as:

$$\frac{\pi_1}{1-\pi_1} = \exp(x'\beta) = \exp(n) \quad (2)$$

The quotient in Eq. (2) is referred to as the *odds*. Eq. (2) can then be reinterpreted as follows:

$$\pi_1 = \frac{\exp(n)}{1+\exp(n)} \quad (3)$$

The binomial logistic regression model is then be generalized to the multinomial case. Where, there are n soil mapping units and also n variables $Y_1,\ldots, Y_n$ with corresponding probabilities of occurrence π_1, $\ldots, \pi_n$. Analogous to binomial logistic regression the odds $\pi_1/\pi_n, \ldots, \pi_{n-1}/\pi_n$ are modelled by means of $\exp(n_1),\ldots, \exp(n_{n-1})$. From $\sum_{i=1}^{n}\pi_1 = 1$ it then follows that:

$$\pi_i = \frac{\exp(n_i)}{\exp(n_1)+\exp(n_1)+\cdots+\exp(n_n)} \quad (4)$$

where $n_n = 0$. This model ensures that all probabilities are in the interval [0,1] and that the probabilities sum to 1.

2.4. Covariate data and statistical analysis

A comprehensive environmental covariate dataset to describe the soil forming factors within the *scorpan* model (McBratney et al., 2003) was required for the MNLR algorithm to predictively map the different hydrological soil types of the Sabie-Sand catchment. These covariates were all resampled to a resolution of 30 m × 30 m, regardless of their

Table 1
The defining characteristics of the hydrological mapping units of the Sabie-Sand catchment.

Hydrological mapping unit	Soil form	WRB Reference Groups	Number of Obs.	Defining hydrological characteristic
Recharge Deep	Hutton, Longtom, Kranskop	Acrisols, Nitisols, Fluvisols	11,582	Deep soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying bedrock is the dominant flow direction.
Recharge Shallow	Glenrosa, Nomanci	Leptosols	401	Shallow soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying fractured bedrock is the dominant flow direction.
Responsive Saturated	Katspruit, Champagne	Gleysols	68	Soils with morphological evidence of long periods of saturation promoting the generation of overland flow due to saturation excess.
Responsive Shallow	Mispah, Graskop	Leptosols	649	Shallow soils overlying relatively impermeable bedrock. Limited storage capacity results in the generation of overland flow after rainfall events.
A/B interflow	Estcourt, Sterkspruit	Solonetz	85	Duplex soils where the textural discontinuity facilitates build-up of water in the topsoil, with discharge in a predominantly lateral direction.
Soil/bedrock Interflow	Fernwood, Cartref	Arenosols	173	Soils overlying relatively impermeable bedrock. Hydromorphic properties signify temporal build of water on the soil/bedrock interface and slow discharge in a predominantly lateral direction.

original resolution.

Elevation was obtained from a 30 m × 30 m Shuttle Radar Topography Mission Digital Elevation Model (USGS, 2022). The covariates used to train the models included elevation, a 1: 250 000 geology map (Council for Geosciences, 2007), a broad landtype map (Land Type Survey Staff, 1976–2006), planform curvature, profile curvature, vertical distance to channel network, topographical wetness index, climate covariates such as mean annual minimum temperature, mean annual maximum temperature, and mean annual precipitation (Schulze, 2007). Slope, relative slope position, multiresolution valley bottom flatness, multiresolution ridge top flatness, LS-factor, longitudinal profile, flow accumulation index, cross-sectional curvature, convergence index and channel network distance were also incorporated. All topographic covariates were developed using the System for Automated Geoscientific Analysis (Conrad et al., 2015), from the digital elevation model (DEM) raster.

Additionally spectral covariates were also developed for the Sabie-Sand catchment from Sentinel 2A satellite imagery to further differentiate between different soil types. These spectral covariates included brightness index, coloration index, redness index, saturation index and NDVI values for both the wet and dry season (Table 2) as described by in (Bannari et al., 1995; Ray et al., 2004; Flynn et al., 2019).

2.5. Calibration and validation datasets

The SAFCOL data was used to create five levels of legacy soil information of which three were created using a downscaling approach. The downscaled legacy soil datasets were created using the base R software (R Core Team, 2022) in conjunction with the prospectr package (Stevens and Ramirez-Lopez, 2022) for k-means clustering, trained on a comprehensive environmental covariate dataset, and the nnet package (Venables and Ripley, 2002) for running the MNLR algorithm.

The k-means clustering is a simple unsupervised non-linear clustering algorithm, where the algorithm seeks to partition the observations into a pre-specified number of k clusters. These clusters try to maximize the difference between clusters whilst also minimizing the difference within clusters using the Euclidean distance between soil observation covariate data. K-means clustering is therefore an effective tool to downscale legacy soil data to a specific user defined number of soil observations (number of clusters), while also retaining the most representative soil observations within the legacy soil dataset. When downscaling the legacy soil data, we therefore defined the number of clusters (k) within R to be equal to the number of soil observations defined by the predetermined observation density. For more information regarding the development of the k-means clustering algorithm refer to Hartigan and Wong (1979) and its use within the R software refer to Stevens and Ramirez-Lopez (2022).

The least detailed level of soil observation was using only the legacy soil information (Legacy) which excluded all of the SAFCOL soil data. The most detailed level of soil information included using all available soil observations (All observations) within the study area at 12,875 soil observations, regardless of the balance or representativeness of soil observations. Observation density was then used to create the downscaled SAFCOL soil datasets at a further three levels. The third level of soil information included the legacy soil information as well as the SAFCOL data downscaled to the same observation density (960 ha/observation). Therefore, our spatially localised dataset was downscaled to the same observation density as the soil observation density in rest of the catchment. The fourth and fifth level of soil information downscaled the SAFCOL data to 500 ha per observation (500 ha/observation) and 100 ha per observation (100 ha/observation) in addition with the

Table 2
Spectral bands, spectral covariates, and their development.

Bands	**Band origin (μm)**	**Symbol**
Blue	0.490	B
Green	0.560	G
Red	0.665	R
Near infrared (NIR)	0.842	NIR
Covariates	**Equation**	**Property**
Brightness index	$(R^2 + G^2 + B^2)/3^{0.5}$	Reflectance
Coloration index	$(R - G)/(R + G)$	Colour
Redness index	$R^2/(B * G^3)$	Hematite
Saturation index	$(R - B)/(R + B)$	Spectral slope
NDVI	$(NIR - R)/(NIR + R)$	Chlorophyll

remaining legacy soil data. These observation densities were selected by the researchers to establish if observation density affected the hydrological soil mapping accuracy and to potentially determine how many soil observations are required to create an accurate hydrological soil map of a macro-scale catchment.

All five levels of soil information was still split into a calibration dataset (75%) and validation dataset (25%). However, a completely independent validation dataset (152 soil observations) was initially created consisting only of legacy soil information which excluded the SAFCOL soil data, which was removed from the different calibration datasets prior to creation. This allowed for the differences in the downscaling procedures to not affect the number of observations within the validation dataset or its internal hydrological soil class distribution.

The MNLR algorithm in conjunction with the above-mentioned environmental covariates were then applied. The resulting hydrological soil maps were analysed using a confusion matrix as well as a Kappa coefficient statistic for both the calibration and validation datasets. How the different downscaled legacy soil datasets affected the representation of the larger covariate soil dataset was tested using a Welsh's *t*-test analysing the mean annual precipitation, slope, topographic wetness index and NDVI (dry season), with a p-value of 0.05 value being used as a threshold for significant difference between the datasets.

These four select covariates each represent different soil forming factors, giving an indication of the representativeness of legacy soil datasets. The Welch's *t*-test compares the two means between datasets, the basic null hypothesis is that the means are equal. The QQ-plot was constructed by determining the (k/n + 1)-th quantiles of the large dataset, (where k = 1, …n and n is the number of observations of the legacy soil dataset) and plotting those quantiles against the values of the observation's covariates of the legacy soil datasets, sorted from small to large. The closer the values of the QQ-plots to the identity line (x = y), the more representative the legacy soil dataset is of environmental covariates.

However, the MNLR algorithm requires a balanced soil mapping class dataset rather than one balanced by the respective environmental covariates. The most representative legacy soil dataset therefore still needed to be balanced by hydrological soil type and assessed.

As undersampling has shown to decrease classifier performance by losing potentially useful information of the majority soil class and because the majority soil class has already been downscaled, only an oversampling approach was used to balance soil classes and compare soil mapping accuracy. However, because ROS causes overfitting when dealing with large differences between majority and minority soil classes, which still exists after downscaling, this approach would be nonsensical (Peri et al., 2018; Taghizadeh-Mehrjardi et al., 2020). Based on the available literature, the synthetic minority oversampling technique (SMOTE) to generate synthetic data of the minority soil classes was selected to balance hydropedological soil classes. This technique was applied for both the most representative and best performing legacy soil dataset and all available soil observations using the smotefamily package (He et al., 2008). This allows us to compare if the downscaling of spatially localised soil observations for covariate representatives adds value to the DSM process, or if all legacy soil observations should simply be balanced using SMOTE for the best results.

The two SMOTE-balanced legacy soil datasets were then also applied within the MNLR algorithm to create a hydrological soil map of the Sabe-Sand catchment. The resulting hydrological soil maps were also analysed using the same validation dataset and statistical measurements (confusion matrix and Kappa coefficient).

3. Results and discussion

3.1. Downscaling legacy soil data

Fig. 3 illustrates the QQ-plots of select environmental covariates of the different legacy soil datasets and the corresponding covariates of the entire catchment. The QQ-plot of the mean annual precipitation (MAP) covariate illustrates the observation bias of the All Observations and 100 ha/observation datasets for values ranging from 1200 to 1600 mm per annum, resulting from the localised SAFCOL data within the higher rainfall regions of the catchment.

The 500 ha/observation, 960 ha/observation and Legacy datasets all provide more representative datasets, with the 500 ha/observation and 960 ha/observation datasets yielding the most representative results for MAP values.

The same trend can be observed when analysing the slope values between datasets, where the 500 ha/observation and 960 ha/observation datasets yielded the most representative results, meaning these datasets spatially provided the most accurate representation of slope values across the catchment. Due to the bias of the legacy datasets towards the mountainous regions of the catchment where large slope values are present, the All Observations and 100 ha/observation datasets are biased to higher slope values, whereas the Legacy soil dataset and 960 ha/observation are slightly biased to low slope values common in the east of the catchment.

For topographical wetness index values all datasets excluding the All Observations dataset yielded accurate representations of the specific environmental covariate for the landscape. The All Observations dataset remains biased to low topographic wetness index values, which is due to the fact that these observations were focussed on upslope terrain positions because proportionately limited SAFCOL observations were made in valley bottom positions.

The 500 ha/observation yielded the most accurate representation of catchment covariate values within the NDVI QQ-plot, which once again followed the same trend where both the All Observations and 100 ha/observation datasets were biased to high NDVI values, indicative of the evergreen forestry activities in the mountainous areas in the catchment. The Legacy and 960 ha/observation were slightly biased toward low NDVI values, indicative of the lack of vegetation in the savanna dry season.

Table 3 illustrates the results of a Welsh's *t*-test between the mean annual precipitation (MAP), slope, topographic wetness index (TWI) and NDVI values (dry) of the different soil datasets and the corresponding catchment environmental covariate.

With all p-values below 0.05 the means of the Legacy, All Observations and the 100 ha/observation soil datasets were significantly different than catchment environmental covariates. This is to be expected for catchments the size of the Sabie-Sand (5790 km^2) where large ranges of environmental covariates exist, and legacy soil datasets are relatively small in comparison. However, the means of the 960 ha/observation were not significantly different for mean annual precipitation and topographical wetness index with p-values of 0.258 and 0.094, respectively. The means of the 500 ha/observation were not significantly different for all four catchment covariates with p-values of 0.369, 0.034, 0.141 and 0.300, respectively.

The t-value, which measures the size of the difference of the observation data relative to the variation in our catchment data, where t-values closest to 0 indicates the lowest variation between the catchment covariates and soil observations covariates, improves as the soil data is downscaled due to the improved representation of catchment environmental covariates.

The QQ-plots and Welsh's *t*-test result of mean annual precipitation, slope, topographic wetness index and NDVI values between the datasets illustrates the improved catchment covariate representation which can be achieved using a downscaling approach on localised legacy soil information. The 500 ha/observation and 960 ha/observation datasets provided a substantially improved representation of catchment environmental covariates compared to the Legacy, All Observations, and 100 ha/observation datasets, and in particular the 500 ha/observation dataset provided the most accurate representation of environmental covariates within the Sabie-Sand catchment.

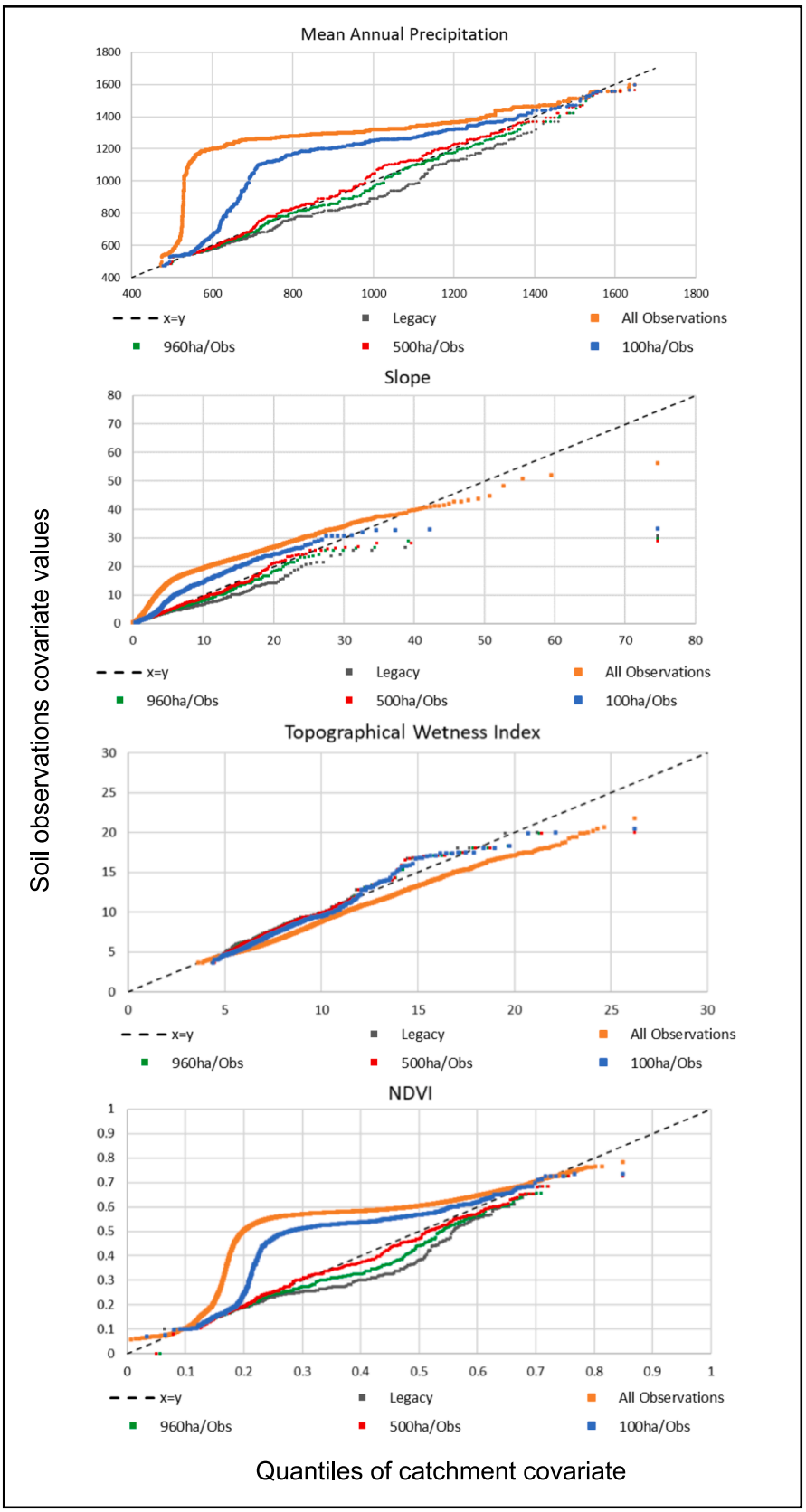

Fig. 3. The QQ-plots of the different legacy soil datasets and select environmental covariates.

Table 3
The Welsh's *t*-test of select environmental covariates.

Covariate	Legacy dataset	Welsh's *t*-test			
		p-value	t-value	Mean soil dataset	Mean covariate dataset
MAP	Legacy	<0.002	−4.09	758.52	800.43
	All Observations	<0.002	315.42	1233.34	
	960 ha/ Observation	0.258	−1.11	788.36	
	500 ha/ Observation	0.369	0.90	810.25	
	100 ha/ Observation	<0.002	14.82	951.59	
Slope	Legacy	<0.002	−8.06	4.15	5.51
	All Observations	<0.002	101.54	11.84	
	960 ha/ Observation	<0.002	−3.94	4.75	
	500 ha/ Observation	0.034	−2.07	5.09	
	100 ha/ Observation	<0.002	8.34	7.35	
TWI	Legacy	0.008	2.63	8.07	7.82
	All Observations	<0.002	−61.48	6.75	
	960 ha/ Observation	0.094	1.68	7.98	
	500 ha/ Observation	0.141	1.48	7.96	
	100 ha/ Observation	<0.002	−3.25	7.57	
NDVI (dry)	Legacy	<0.002	−5.05	0.24	0.26
	All Observations	<0.002	176.75	0.48	
	960 ha/ Observation	0.003	−2.96	0.25	
	500 ha/ Observation	0.300	−1.03	0.26	
	100 ha/ Observation	<0.002	12.25	0.33	

MAP = Mean Annual Precipitation.
TWI = Topographic Wetness Index.
NDVI = Normalize Difference Vegetation Index.

3.2. Digital soil mapping

Table 4 illustrates the five different legacy soil datasets mapping accuracy in relation to their calibration and validation datasets. Focussing on the calibration datasets the best performing soil dataset was the Legacy soil dataset with a confusion matrix accuracy of 72% and Kappa coefficient of 0.52, whereas using All Observations resulted in the highest confusion matrix accuracy (88%) but the lowest Kappa coefficient (0.21). The three downscaled approaches provided modest calibration accuracy results with 960 ha/observation and 500 ha/observation yielding confusion matrix values and Kappa coefficient values of 50% and 0.48 and 64% and 0.47, respectively. However, the 100 ha/observation dataset yielded the east accurate calibration dataset within the different downscaling approaches. Although calibration accuracy should not be considered when assessing mapping accuracy, insight can be gained on how the models learned from the calibration legacy soil data.

Table 4
The statistical accuracy of the legacy soil datasets.

Legacy dataset	Dataset used	Point accuracy (%)	Kappa coefficient
Legacy	Calibration	72	0.52
	Validation	50	0.34
All Observations	Calibration	88	0.21
	Validation	46	0.28
960 ha/observation	Calibration	50	0.48
	Validation	48	0.42
500 ha/observation	Calibration	64	0.47
	Validation	62	0.46
100 ha/observation	Calibration	63	0.27
	Validation	54	0.33

When analysing the validation results of the five levels of soil information it is the 500 ha/observation dataset which yielded the most accurate hydrological soil map with a confusion matrix value of 62% and Kappa coefficient values 0.46, compared to the two control datasets (Legacy and All Observations) with confusion matrix values and Kappa coefficient values of 50% and 0.34 and 46% and 0.28, respectively. Both the 960 ha/observation and 500 ha/observation datasets represent a moderate strength of agreement with reality, outperforming both control datasets which represent a fair agreement with reality. The 100 ha/observation only slightly outperformed the All Observations dataset also indicating a fair agreement with reality (Landis and Koch, 1977).

The 500 ha/observation legacy soil dataset yielded the most accurate mapping results (Table 4) as well as being the most representative legacy soil dataset for the selected environmental covariates (Fig. 3 and Table 3). This dataset was therefore selected for further balancing and comparison using the SMOTE technique to balance soil mapping classes.

Fig. 4 illustrates the mapping difference between the best downscaled soil dataset, after SMOTE balancing (500 ha/observations-SMOTE) and using the same approach using all available observations (All Observations-SMOTE) and their accompanying validation confusion matrix and Kappa coefficient accuracy.

The All Observations-SMOTE dataset yielded a validation point accuracy of 53% and a Kappa coefficient of 0.47, which is similar to results achieved by the 500 ha/observation dataset prior to SMOTE balancing and represents a moderate strength of agreement with reality. Although the All Observations-SMOTE hydrological soil map provided relatively accurate distributions of dominant hydrological soil types within the SAFCOL dataset such as recharge deep, recharge shallow, responsive shallow and responsive saturated, the ability to map the hydrological soils outside of these areas was poor. Especially, the prediction of A/B interflow soils, where the MNLR algorithm vastly overestimated its presence within the catchment area to the detriment of recharge shallow and soil/bedrock interflow soils (Fig. 4a and Table 5) in the lowveld areas of the catchment.

This overprediction is most likely due to the specific nature of A/B interflow soils relative to the other hydrological soil types, where a limited range of soil forming factors result in these very specific soils compared to a far wider range resulting in soil/bedrock interflow and recharge shallow soils. Therefore, when SMOTE generated a substantial amount of synthetic data from the existing A/B interflow data, the resulting data was comparative to random oversampling with a small region of specific examples being created, which lead to overfitting.

The 500 ha/observation-SMOTE map (Fig. 4b) yielded a validation point accuracy of 72% and a Kappa coefficient of 0.60, which was the most accurate hydrological soil map of the Sabie-Sand catchment and resulted in a substantial agreement with reality (Landis and Koch, 1977). These results are comparable with the bulk of the other hydrological soil maps created in South Africa, such as Van Zijl (2019) with a point accuracy of 69% and Kappa coefficient value of 0.59; Van Zijl et al. (2012), with a point accuracy of 69% and Kappa coefficient value of 0.59; Van Zijl et al. (2014), with a point accuracy of 88% and Kappa coefficient value of 0.82 as well as Smit and Van Tol (2022) with a point accuracy of 74% and Kappa coefficient of 0.68. These results are also comparable to other digital soil mapping projects globally, such as MacMillan et al. (2010) with 69% point accuracy and Zhu et al. (2008) with a point accuracy of 76%.

The confusion matrix for the All Observations-SMOTE hydrological soil map (Table 5) indicated that not all the soil classes were sufficiently mapped, with user's accuracy below 50% for recharge shallow and responsive shallow and soil/bedrock interflow classes. In general, producer's mapping accuracy performed slightly better with only A/B

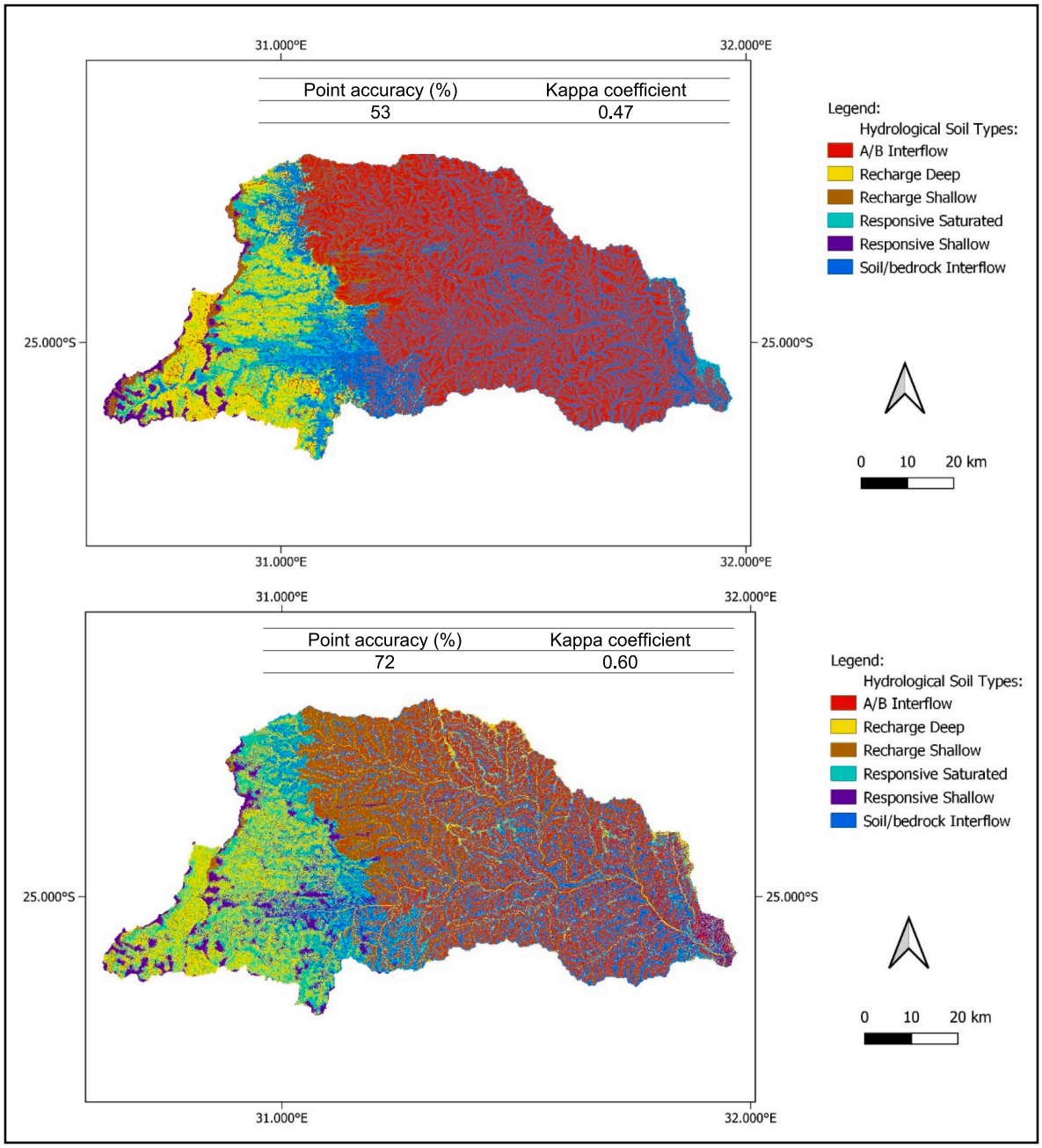

Fig. 4. The hydrological soil types of a) All Observations-SMOTE and b) 500 ha/observation-SMOTE maps and their accompanying validation accuracy.

interflow soils yielding an accuracy value below 50%.

The confusion matrix for the 500 ha/observations-SMOTE hydrological soil map (Table 6) indicated that all the soil classes were sufficiently mapped, with all the user's and producer's accuracies above 50%. In general, mapping accuracy decreased with decreased validation observations as seen with the 57% user's accuracy of responsive saturated soils where only seven observations were present compared to the 84% user's accuracy of recharge deep soils where 73 observations were present.

The 500 ha/observation-SMOTE hydrological soil map (Fig. 4b) contains a distinct toposequence in the mountainous as well as lowland areas within the catchment. In the mountainous regions in the catchment responsive shallow soils dominate the steepest of slopes with recharge shallow soils also occur on the hillcrest terrain positions on the highest peaks where overland flow and relatively quick recharge conditions are the dominant hydrological processes, potentially based on differences in parent material between soil classes. The less steep midslope and footslope terrain positions being dominated by recharge deep soils where the vertical drainage of water through the soil profile is the dominant hydrological response, this vertical drainage is most likely followed up by the lateral movement of water within the shallow groundwater aquifer.

These deep soils resulted from the processes of illuviation and colluviation, where the clayey material has been predominantly removed from midslope positions, with colluvial deposits forming at footslope terrain positions. The valley bottom terrain positions are primarily dominated by responsive saturated soils typical of riparian and wetland areas, where saturation excess leads to the overland flow of water on top of the soil surface. These soils most likely originate due to the process of eluviation which results in the addition of clays from upslope terrain positions to these soils, expressed as the gleyed subsoils clay rich of these terrain positions.

Table 5
Confusion matrix of the All Observations-SMOTE hydrological soil map.

		User Accuracy								
		A/B interflow	Recharge deep	Recharge shallow	Responsive saturated	Responsive shallow	Soil/bedrock interflow	Total	Correct	%
Producers Accuracy	A/B interflow	23	14	6	1	4	10	58	23	40
	Recharge deep	2	50	4	1	1	3	61	50	82
	Recharge shallow	0	2	7	0	1	0	10	7	70
	Responsive saturated	0	1	2	5	0	0	8	5	62
	Responsive shallow	0	1	0	0	4	0	5	4	80
	Soil/bedrock interflow	0	5	0	0	0	5	10	5	50
	Total	25	73	19	7	10	18	152	62	
	Correct	23	50	7	5	4	5	62		
	%	92	69	37	71	40	28			

Table 6
Confusion matrix of the 500 ha/observations-SMOTE hydrological soil map.

		User Accuracy								
		A/B interflow	Recharge deep	Recharge shallow	Responsive saturated	Responsive shallow	Soil/bedrock interflow	Total	Correct	%
Producers Accuracy	A/B interflow	15	3	1	0	0	1	20	15	75
	Recharge deep	5	61	4	2	3	5	80	61	76
	Recharge shallow	3	2	14	1	1	2	23	14	61
	Responsive saturated	0	1	0	4	0	0	5	4	80
	Responsive shallow	2	1	0	0	6	0	9	6	67
	Soil/bedrock interflow	0	5	0	0	0	10	15	10	67
	Total	25	73	19	7	10	18	152	72	
	Correct	15	61	14	4	6	10	72		
	%	60	84	74	57	60	5			

The lowland hillcrest and midslope terrain positions are dominated by soil/bedrock interflow and recharge shallow soils where the lateral movement of water at the soil/bedrock interface, more prevalent in the east and southeast, and vertical drainage to the shallow aquifer, more prevalent in the northwest and west, is the dominant hydrological responses. A/B interflow soils occur on footslope positions where the dominant hydrological process is the lateral movement of water at the topsoil/subsoil interface where the textural discontinuity between soil horizons lead to the build-up and lateral movement of water through the soil profile. These soils were created by the eluviation of clays from the recharge shallow and soil/bedrock interflow upslope positions. Valley bottom terrain positions are dominated by a mixture of recharge deep, due to sandy floodplains on the major river networks forming during periodic flooding caused by cyclone events, and responsive saturated soils on the valley bottom positions which are not exposed to the same periodic flooding.

3.3. . The value of downscaling and balancing legacy soil data

In South Africa a large volume of available legacy soil data exists, which remains largely untapped within commercial and semi-commercial sources that have not been freely available in the past (Patterson et al., 2015). However, these sources have become more frequently available in recent years by the improved cooperation between various public and private sector stakeholders as seen with the acquisition of the SAFCOL legacy soil dataset within this paper. The opportunity to add substantial amounts of additional spatially localised legacy soil data for use in DSM across South Africa should also coincide with research regarding how best to apply these datasets for DSM purposes.

The 500 ha/observation downscaled dataset statistically improved the existing legacy soil dataset and provided the best representation of environmental covariates within the catchment, resulting in improved mapping accuracy prior to further balancing. Therefore, the downscaling of spatially localised legacy soil information to improve environmental covariate representation is an effective tool to improve the representation of legacy soil datasets. However, these results only consider environmental covariate representation and not necessarily the representativeness of the minority soil classes, which was why SMOTE was still required to improve the representation of the minority soil classes.

However, simply applying SMOTE balancing of mapping units using all available soil information still provided comparative results when evaluating point accuracy and Kappa coefficient values. When balancing soil classes using SMOTE was applied to our best representative legacy soil dataset, the resulting hydrological soil mapping accuracy was significantly improved compared to using only SMOTE balancing on all available soil information. These improved results using SMOTE are in accordance with results from Taghizadeh-Mehrjardi et al. (2020) as well as other results outside of the field of soil science (Chawla et al., 2002; Tantithamthavorn et al., 2020). The value of downscaling spatially localised legacy soil information therefore must coincide with additional class balancing when using MNLR. These results also reaffirm the importance of balancing legacy soil information for mapping units across the entire catchment area because imbalanced data affects the predictive ability of the MNLR algorithm. Our results also reaffirm the ability of the SMOTE resampling technique to handle major class imbalances.

However, simply finding and adding additional highly imbalanced legacy soil data and balancing the minority soil classes using SMOTE did not improve mapping accuracy using MNLR compared to the best performing environmental covariate balanced map. Therefore, there is a

limit to the capability of the SMOTE for resampling and care should be taken when applying synthetic data generation techniques to balance legacy soil data. Emphasis should remain on using a representative legacy soil dataset for DSM purposes.

Sharififar et al. (2019b) did not encounter this problem because the researchers collected soil samples from a grid spacing of 500 m across their entire study area of 12,000 ha, meaning that environmental covariate representation of soil samples was guaranteed. However, as the size of the study area increases, grid sampling becomes too costly and time consuming, where legacy soil datasets are not necessarily representative of environmental covariates. Taghizadeh-Mehrjardi et al. (2020) used the national soil database of Iran consisting of 7,664 samples, which was created using stratified random sampling (~87%), grid sampling (~8%) and the conditioned latin hypercube sampling approach (~5%), with the assumption being that these samples are representative of the soils and environmental covariates of Iran. However, the readily available soil observations in the national soil database of South Africa (Land Type Survey Staff, 1976–2006) amounts to 2500 modal profile observations, less than half of that of Iran. This study therefore provides a relevant protocol to use highly spatially localised legacy soil datasets to improve DSM accuracy by downscaling and adding additional soil observations that improve overall representation and balance of environmental covariates data within the legacy soil dataset.

Therefore, downscaling highly spatially localised legacy soil data using k-means clustering for improving environmental covariate representation is an effective method to improve DSM accuracy. Our best performing hydrological soil map therefore more accurately represents the dominant hydrological processes throughout the Sabie-Sand catchment, then the readily available soil information in South Africa. An improved representation of these internal catchment processes could hold the key to improved climate- and landuse change scenario analyses, and improved water resource management practices at catchment scale due to the fact that the major hydrological processes are better understood both spatially and temporally. This approach could potentially aid in the mapping of the hydrological soils for macro-scale catchments, by optimally using large spatially localised imbalanced legacy soil datasets, such as soil surveys within the forestry, mining, and agricultural sectors. The approach may be particularly applicable in South Africa where large amounts of spatially localised legacy soil information exists within large scale mapping projects, such as creating a hydrological soil map of South Africa.

Balancing spatially localised legacy soil datasets for use in large scale DSM projects extends beyond merely downscaling and upscaling soil observations of the majority and minority soil classes. The representativeness of soil observations within the catchment environmental covariates are measurable and is indicative of how accurate the resulting soil maps should be. Downscaling highly spatially localised legacy soil observations should strive to improve catchment representation and is also easily repeatable across soil datasets of various sizes and various resolutions.

A problem in dealing with spatially localised legacy soil observations is that comparatively small validation dataset is used to validate mapping results as can be seen from this study where only 152 observations were used which creates further uncertainty regarding the accuracy of minority soil classes within the different hydrological soil maps as limited observations are available to validate the minority soil classes (Sharififar et al., 2019b). Future research should focus on different balancing approaches. At the data level, different upscaling, and downscaling approaches across different catchment sizes with different imbalanced legacy soil datasets preferably with larger validation datasets should be applied. Research should also be conducted at the algorithm level, where algorithms specifically developed to handle class imbalanced dataset, such as boosting algorithms, should also be considered as viable alternatives to well established DSM algorithms. Recently, Sharififar and Sarmadian (2022) looked at dealing with imbalance soil data using the algorithm level approach and found that a one-class support vector machine combined with multi-class classification yielded the most accurate soil map and adequately represents the minority soil class. Research should also be conducted at the algorithm level, where algorithms specifically developed to handle class imbalanced dataset, such as boosting algorithms, should also be considered as viable alternatives to well established DSM algorithms. Lastly, a combination of these approaches should also be researched once the best approaches of each level have been firmly established in the field of soil science.

4. Conclusion

An accurate hydrological soil map of the macro-scale Sabie-Sand catchment was created using machine learning based digital soil mapping and legacy soil information. Th downscaling of spatially localised legacy soil datasets to improve the representation of environmental covariates was applied using k-means clustering, where the 500 ha/observation dataset resulted in the best improved representation of catchment environmental covariates. However, the improved catchment representation does not necessarily result in improved mapping accuracy, especially when dealing with imbalanced soil mapping classes. Further balancing the imbalanced soil classes of the 500 ha/observation dataset using SMOTE, significantly improved mapping accuracy compared to using SMOTE on all available soil information.

Therefore, downscaling spatially localised legacy soil is an effective tool to improve legacy soil data covariate balance and representation which leads to improve DSM accuracy. This approach is of value where large spatially localised datasets exists. Our main recommendation would be to focus future research on further testing the use of downscaling spatially localised legacy soil information to improve digital soil mapping accuracy across different catchment sizes with different legacy soil datasets.

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.

Data availability

Data will be made available on request.

Acknowledgements

This work could not have been possible without legacy soil datasets provided by the Agricultural Research Council, Digital Soils Africa and the South African Forestry Company Limited. We would also like to acknowledge the South Africa National Parks for logistical support and the Water Research Commission project: C2020 2021-00455 for partially funding this research.

References

Arrouays, D., McKenzie, N., de Forges, A.R., Hempel, J., McBratney, A.B., 2014. GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press/Balkema, Leiden.

Bailey, N., Clements, T., Lee, J.T., Thompson, S., 2003. Modelling soil series data to facilitate targeted habitat restoration: a polytomous logistic regression approach. J. Environ. Manage. 67 (4), 395–407.

Bannari, A., Morin, D., Bonn, F., Huete, A.R., 1995. A review of vegetation indices. Remote Sens. Rev. 13, 95–120. https://doi.org/10.1080/02757259509532298.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68–83.

Campling, P., Gobin, A., Feyen, J., 2002. Logistic modeling to spatially predict the probability of soil drainage classes. Soil Sci. Soc. Am. J. 66 (4), 1390–1401.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analysis (SAGA). In: Geoscientific Model Development, https://doi.org/10.5194/gmd-8-1991-2015.
Council for Geoscience, 2007. Geological Data 1:250 000. Council for Geoscience, Pretoria, South Africa.
Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: examples from Vestfold County, Norway. Catena 77 (1), 8–18.
Flynn, T., Van Zijl, G.M., Van Tol, J.J., Botha, C., Rozanov, A., Warr, B., Clarke, C., 2019. Comparing algorithms to disaggregate complex soil polygons in contrasting environments. Geoderma 352, 171–180. https://doi.org/10.1016/j.geoderma.2019.06.013.
García, S., Herrera, F., 2009. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evol. Comput. 17, 275–306. https://doi.org/10.1162/evco.2009.17.3.275.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239.
Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schr¨oder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. Geoderma 185–186, 37–47.
Harrison, R.L., Van Tol, J.J., Toucher, M.L., 2022. Using hydropedological characteristics to improve modelling accuracy in Afromontane catchments. J. Hydrol.: Reg. Stud. 39, 100986.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a K-means clustering algorithm. Appl. Stat. Royal Stat. Soc. 28 (1), 100.
He, H., Bai, Y., Garcia, E., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference, 1322–1328.
He, H., Garcia, E.A., 2008. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284.
Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma 140 (4), 417–427.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.
IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106. FAO, Rome.
Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 151 (3-4), 311–326.
Kruger, A.C., Makamo, L.B., Shongwe, S., 2002. An analysis of Skukuza climate data. Koedoe 45, 87–92. https://doi.org/10.4102/koedoe.v45i1.16.
Lamichhane, S., Kumar, L., Adhikari, K., 2021. Updating the national soil map of Nepal through digital soil mapping. Geoderma 394, 115041.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.
Staff, L.T.S., 1972–2006. Land types of South Africa: Digital map (1: 250 000 scale) and soil inventory datasets. ARC-Institute for Soil, Climate and Water, Pretoria.
Li, Y., Adams, N., Bellotti, T., 2022. A relabeling approach to handling the class imbalance problem for logistic regression. J. Comput. Graph. Stat. 31 (1), 241–253. https://doi.org/10.1080/10618600.2021.1978470.
López, V., Fernandez, A., García, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. (Ny) 250, 113–141.
Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70 (2), 216–235.
MacMillan, R.A., Moon, D.E., Coupé, R.A., Phillips, N., 2010. In: Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Netherlands, Dordrecht, pp. 337–356.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1-2), 3–52.
Mucina, L., Rutherford, M.C., 2006. The vegetation of South Africa, Lesotho and Swaziland, Strelitzia 19. South African National Biodiversity Institute, Pretoria.
R Core Team, 2022. R: A language and environment for statistical computing. https://www.R-project.org.
Piri, S., Delen, D., Liu, T., 2018. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. Decis. Support Syst. 106, 15–29.
Ray, S.S., Singh, J.P., Das, G., Panigrahy, S., 2004. Use of high resolution remote sensing data for generating site-specific soil mangement plan. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.
Schulze, R.E., 2007. South African Atlas of Climatology and Agrohydrology. Water research Commission, Pretoria. WRC report 1489/1/06.
Sharififar, A., Sarmadian, F., Malone, B.P., Minasny, B., 2019. Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma 350, 84–92. https://doi.org/10.1016/j.geoderma.2019.05.016.
Sharififar, A., Sarmadian, F., 2022. Coping with imbalanced data problem in digital mapping of soil classes. Eur. J. Soil Sci. 74 (3).
Sharififar, A., Sarmadian, F., Minasny, B., 2019. Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling technique. Comput. Electron. Agric. 159, 110–118.
Smit, I.E., Van Tol, J.J., 2022. Impacts of soil information on process-based hydrological modelling in the upper Goukou catchment, South Africa. Water 14 (3), 407. https://doi.org/10.3390/w14030407.
Stevens, A., & Ramirez-Lopez, L. (2022). An introduction to the prospectr package. R package Vignette. R package version 0.2.6.
Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N., Scholten, T., 2020. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. Eur. J. Soil Sci. 71, 352–368.
Tantithamthavorn, C., Hassan, A.E., Matsumoto, K., 2020. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. IEEE Trans. Softw. Eng. 46 (11), 1200–1219.
USGS, 2022. Landsat images. accessed 17 April 2022. http://landsat.usgs.gov.
Van Tol, J.J., Le Roux, P.A.L., 2019. Hydropedological grouping of South African soil forms. S. Afr. J. Plant Soil 36, 233–235. https://doi.org/10.1080/02571862.2018.1537012.
Van Tol, J.J., Van Zijl, G.M., 2022. South Africa needs a hydrological soil map: a case study from the upper uMngeni catchment. Water SA 48 (4), 335–347. https://doi.org/10.17159/wsa/2022.v48.i4.3977.
Van Tol, J.J., Van Zijl, G.M., Riddell, E.S., Fundisi, D., 2015. Application of hydropedological insights in hydrological modelling of the Stevenson-Hamilton Research Supersite, Kruger National Park, South Africa. Water SA 41 (4), 525.
Van Tol, J., Van Zijl, G., Julich, S., 2020. Importance of detailed soil information for hydrological modelling in an urbanized environment. Hydrology 7, 34. https://doi.org/10.3390/hydrology7020034.
Van Zijl, G.M., 2019. Digital soil mapping approaches to address real world problems in southern Africa. Geoderma 337, 130–1308.
Van Zijl, G.M., Le Roux, P.A.L., Smith, H.J.C., 2012. Rapid soil mapping under restrictive conditions in Tete, Mozambique. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond. CRC Press, Balkema, pp. 335–339.
Van Zijl, G.M., Bouwer, D., Van Tol, J.J., Le Roux, P.A.L., 2014. Functional digital soil mapping: a case study from Namarroi, Mozambique. Geoderma 219–220, 155–161.
Venables, B., Ripley, B.D., 2002. Modern applied statistics with S, 4th ed. Springer.
Zarinabad, N., Wilson, M., Gill, S.K., Manias, K.A., Davies, N.P., Peet, A.C., 2017. Multiclass imbalance learning: Improving classification of pediatric brain tumors from magnetic resonance spectroscopy. Magn. Reson. Med. 77 (6), 2114–2124.
Zhu, B., Baesens, B., Vanden Broucke, S.K.L.M., 2017. An empirical comparison of techniques for the class imbalance problem in churn prediction. Inf. Sci. (Ny) 408, 84–99.
Zhu, A.-X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E., Zhou, C., 2008. Purposive sampling for digital soil mapping for areas with limited data. In: Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M.D.L. (Eds.), Digital Soil Mapping With Limited Data. Springer, Dordrecht.