Geoderma 350 (2019) 84–92

Contents lists available at ScienceDirect

Geoderma

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

Addressing the issue of digital mapping of soil classes with imbalanced class observations

Amin Sharififar[a,*], Fereydoon Sarmadian[a], Brendan P. Malone[b], Budiman Minasny[c]

[a] Department of Soil Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
[b] CSIRO, Agriculture and Food, Canberra, ACT, Australia
[c] Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, NSW, Australia

ARTICLE INFO

Handling Editor: Morgan Crisitne L.S.

Keywords:
Imbalanced classification
Digital soil mapping
Uncertainty assessment
Data resampling
Categorical soil mapping
Machine learning

ABSTRACT

Considering the nature of soils distribution, an important modeling issue in soil class mapping is imbalanced class observations. Imbalanced number of data in observed soil classes in an area can result in the underestimation or loss of minority classes and an overestimation of the majority classes in predictive modeling. The effect of this phenomenon is that an area of land with comparatively fewer soil profile observations could be unmapped in the digital maps. To address this problem, this paper investigated the usefulness of data pretreatment techniques called over- and under-sampling of data applied on three predictive models including decision trees (DT), random forest (RF), and multinomial logistic regression (MNLR). The study area is situated in the northwest of Iran with 452 profiles observations on a regular grid covering about 12,000 ha. This area has 8 USDA soil great groups with an imbalanced frequency distribution. Results showed that modeling using imbalanced distribution of class observation caused uncertain maps with minority classes being lost and relatively poor accuracies. After data treatment, with over- and under-sampling, all models showed significant improvement in maintaining the minority classes, in both calibration and validation evaluations. Balancing the classes led to a notable decrease in uncertainty of all 3 models by decreasing the confusion index and raising the probability of occurrence for the soil classes in the final maps. Comparing the 3 models, decision trees showed the largest calibration and validation accuracies with and without data treatment. RF has an issue of overestimation of some of the majority classes. Data resampling technique can be a useful solution for dealing with imbalanced class observations to produce more certain digital soil maps.

1. Introduction

Soil type maps are an essential tool for targeted land use planning and soil management. For example, maps of soil types can help in assigning suitable land management practices and plans that are based on soil type-specific conditions and capabilities. Digital soil mapping (DSM) (McBratney et al., 2003) has been an excellent tool for contemporary soil mapping efforts as it leverages and exploits the availability of geospatial datasets and model-based approaches. Statistical and geostatistical models are routinely applied throughout the world for mapping of soil classes in different spatial scales and extents (Brungard et al., 2015; Heung et al., 2016; Ma et al., 2019).

Various modeling techniques such as multinomial logistic regression, random forest, and decision trees models have been applied in predicting and mapping soil classes (Adhikari et al., 2014; Grunwald, 2009). A study by Brungard et al. (2015) found that complex machine learning models were generally more accurate than simple models; however, the accuracy of the model depends upon the number of classes and the frequency distribution of soil observations. The number of soil classes and frequency distribution of the classes in an area are largely a function of the environmental complexity and the nature of the soil taxonomic classification system in that area. Hence, an important issue that affects the accuracy of a digital soil model is the imbalanced number of observations among classes. This phenomenon affects the predictive models, in such a way that usually some of the minor soil classes get omitted in the resulting maps (Ma et al., 2019). When the minor class is important (e.g., a rare or endemic soil class) (Baker et al., 2016), such models do not preserve pedo-diversity (Costantini and L'Abate, 2016) and lead to an unreliable soil map. Furthermore, imbalanced class observations are difficult to deal with when setting aside portions of data for training and testing the applied models. For example, it is hard to make sure that all classes are included

* Corresponding author.
E-mail address: Sharififar1988@gmail.com (A. Sharififar).

https://doi.org/10.1016/j.geoderma.2019.05.016
Received 17 October 2018; Received in revised form 22 April 2019; Accepted 11 May 2019

in both calibration and validation datasets without omission of the minority class or classes. While imbalanced classification is a recognized problem in the machine learning discipline for categorical data modeling (Haixiang et al., 2017; Nayal et al., 2017) this issue has not been well addressed in soil mapping.

In DSM, a lot of effort has been made to compare different machine learning models to seek out the most accurate or optimal model configuration (Brungard et al., 2015; Heung et al., 2016; Taghizadeh-Mehrjardi et al., 2015). Here, we compare three important and well-known models including random forest, multinomial logistic regression, and decision trees by investigating them for soil class mapping in the case of imbalanced datasets to see how they produce a soil type map. These models are widely used for digital soil mapping despite that they are known to be inefficient in handling imbalanced class data input in other fields of science. It would be helpful for DSM studies to compare to what extent the problem exists with these models and how they would respond to a treatment.

To deal with the imbalanced classification issue, several suggestions have been proposed for improving the model training performance and result of classification such as data-level solutions, algorithm-level solutions and ensemble solutions (Zhu et al., 2017). Data-level solutions are resampling techniques such as random over- and under-sampling for manipulating the observed data to become as close to a balanced distribution as possible (Chawla et al., 2002; Abdi and Hashemi, 2016). In an algorithm-level solution, modifications on models functions are applied to raise the ability of models to maintain the minority class, especially through cost-sensitive learning (e.g., Siers and Islam, 2018). In an ensemble solution, classification is often improved by combining several classifiers to obtain a new and better classifier (Galar et al., 2012). The data-level solution seems the easiest for the purpose of this study, and as these types of techniques are simpler to apply compared to others mentioned above, we chose them for the present study. However, each of the proposed solutions has advantages and disadvantages and discipline-specific challenges, on which some issues have been discussed in López et al. (2013).

To address the issue of imbalanced data in soil class mapping, this paper investigated naturally imbalanced soil type classes for use in 3 models mentioned above for prediction of USDA soil great groups. The data were treated with both over and under-sampling techniques to improve classification results.

2. Methods and materials

2.1. Study area

The study area is located in the northwest of Iran, in a semi-arid region according to de Martonne climate classification (de Martonne, 1926). The average annual rainfall is 271 mm and the mean annual temperature is 15 °C. Mean altitude is 255 m above sea level, and main physiographic units include plateau and hills with piedmont plains to a lesser extent. The main soil orders according to the USDA soil taxonomy (USDA, 2010) are Aridisols and Entisols. Land cover types of the area include rangelands and agriculture, which are of high importance because of typical production and income for the rural population. Soil samples were collected from 452 profiles to the depth of 1.5 m on a regular grid, covering an area of approximately 12,000 ha. The grid spacing was 500 m, but in some situations, the site had to be relocated to a nearby site, because of access issues in the intended location. Fig. 1 shows the study area location and sampling points. Morphological description and physicochemical analysis of a range of properties were conducted to classify the soils according to the USDA soil taxonomy key for classification (USDA, 2010). Soils were allocated to 8 great groups: (A) Calcigypsids, (B) Argigypsids, (C) Natrigypsids, (D) Haplogypsids, (E) Haplocalcids, (F) Haplocambids, (G) Torrifluvents, and (H) Torriorthents (Table 1). The two last soil classes were considered as minority classes, as they have a much lower number of observations (7% and 2% of the total observed data) compared to the other classes. Calcigypsids followed by Haplocambids and Haplogypsids were considered the majority classes with 35%, 18.5% and 17% of the whole study area observations. These classes had higher frequencies compared to other soil classes (Table 1; Fig. 2).

2.2. Digital soil mapping

The procedure of DSM (McBratney et al., 2003) relies on relating soils to proxies of soil forming factors (available environmental data), which captures inherent soil spatial variation. By adopting a similar approach, in this study soil types were estimated spatially across a given mapping extent via an empirical model-based approach. The environmental covariates used in this study include a digital elevation model (DEM), which was obtained from the freely available ASTER satellite image (ASTER GDEM, METI, and NASA; http://earthexplorer.usgs.gov), and several environmental variables that were derived from this DEM. Based on scientific and local expert knowledge of the study area, among a number of covariates, six covariate maps were selected for DSM. For instance, vegetation cover created from Landsat satellite image was not effective in explaining the soil variations, as it showed a very low variation, thus it was not used in the study. The 6 covariates include digital elevation model, terrain ruggedness index, relative slope position, channel network base level, landforms, surface texture, and valley depth. All these covariates have a 32 m resolution. Fig. 3 shows the maps of the covariates used in this study.

Three models including decision trees, multinomial logistic regression and random forest were used for producing a digital map of soil great groups. For the decision trees model, we used the C5.0 function in the C50 package in R (Kuhn et al., 2018). This function is based on Quinlan's C5.0 algorithm (Quinlan, 1993). For the MNLR model, *multinom* function in the *nnet* package in R (Ripley and Venables, 2015) was used. Likewise, for RF, we used *randomForest* function in the *randomForest* package (Breiman, 2006). Using all of these models, we fit soil class observations with the raster stack of the six considered environmental covariates. The soil classes (our target variable) are then predicted as a function of the environmental covariates.

The models of DT and RF were executed with 100 bootstrap iterations to resample the training dataset for 100 times. For the MNLR model, training was performed with 100 iterations, and each of the 8 classes probabilities were computed through the *predict* function for the map sites. The resulting multiple maps were used to calculate the most probable class map and the average probabilities for each of the classes according to the procedure of Odgers et al. (2014) as a way of quantifying model uncertainties.

2.3. Data treatment using oversampling and under-sampling methods

Before applying any treatment on the data, 30% of the dataset was randomly selected for a validation set and 70% for calibration of the models (This ratio was variable for different classes due to imbalanced number of observations) (Table 1). This was done by manually separating each of the classes and then setting aside a portion of each class observations randomly using a computerized function, so that we make sure that every class exists in both validation and calibration datasets.

First, the models were executed on the untreated data, to see how class imbalance affects the models' performance and resultant map accuracy and uncertainty. Afterward, two data treatment functions namely random oversampling on the minority soil classes and random under-sampling on the majority soil classes were executed. These were performed using the *ubOver()* and *ubUnder()* functions, respectively, from the "unbalanced" package (Dal Pozzolo et al., 2015) in the R software (R Development Core Team, 2011). Under-sampling was performed for the majority classes of A, D and F and oversampling for the minority classes of G and H (Table 1 and Fig. 2). The majority classes were under-sampled to half and the minority class G and H were

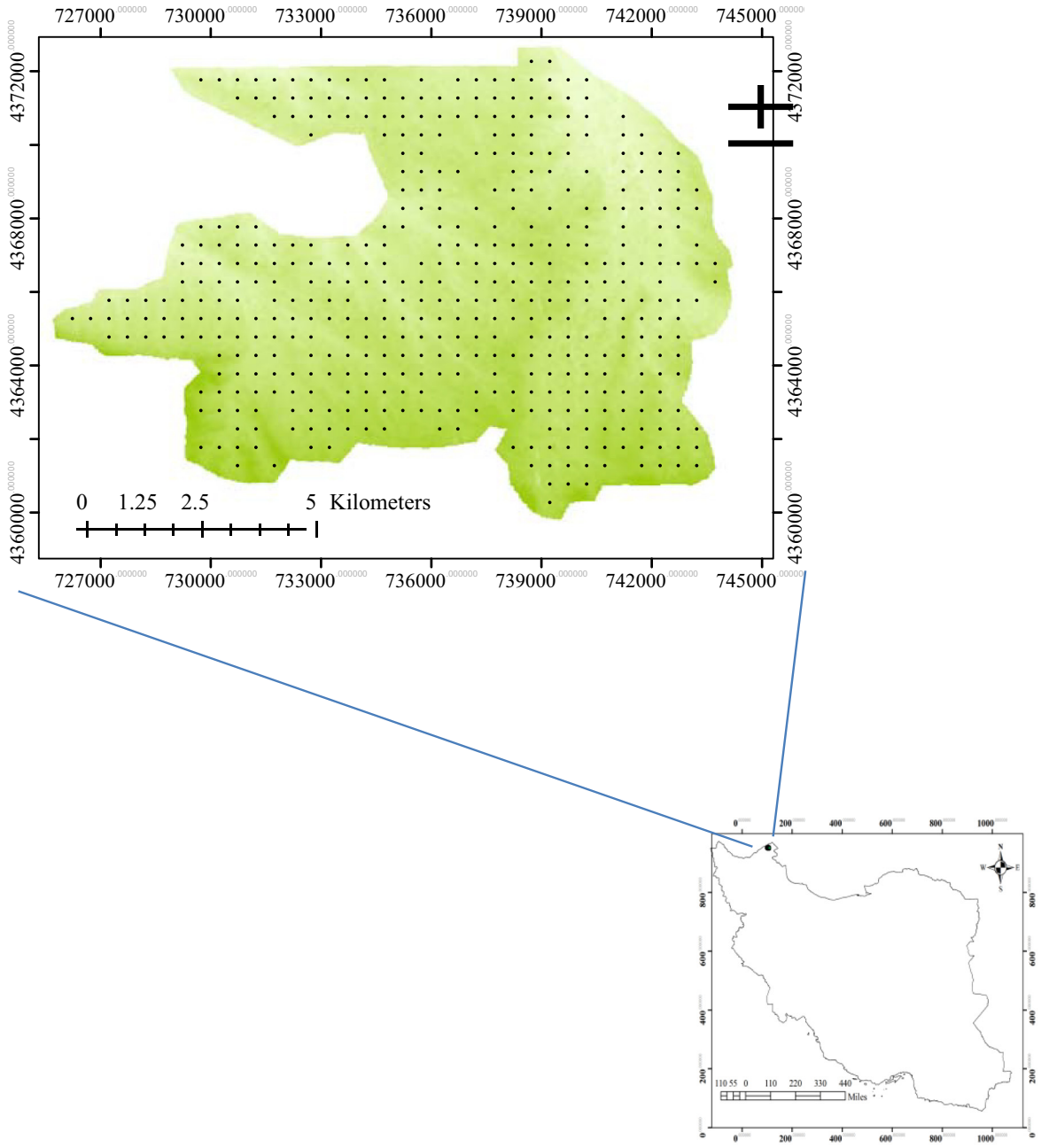

Fig. 1. Study area location and sampling points.

Table 1
Soil great groups and number of observations.

Soil class code	Taxonomic class (great group level)	No of observations (and percentage of observation)	Calibration (number and percentage)	Validation (number and percentage)
A[a]	Calcigypsids	160 (35.39%)	112 (70%)	48 (30%)
B	Argigypsids	37 (8.18%)	26 (70%)	11 (30%)
C	Natrigypsids	23 (5.08%)	17 (70%)	7 (30%)
D	Haplogypsids	79 (17.47%)	56 (70%)	23 (30%)
E	Haplocalcids	60 (13.27%)	40 (67%)	20 (33%)
F	Haplocambids	84 (18.58)	59 (70%)	25 (30%)
G	Torrifluvents	7 (1.5%)	4 (60%)	3 (40%)
H	Torriorthents	2 (0.44%)	1 (50%)	1 (50%)

[a] Codes used throughout the paper to represent the soil classes for ease.

oversampled by 5 folds and 15 folds, respectively, to make the distribution as close to a normal one as possible without significantly scrambling the original proportion of different classes.

The under-sampling procedure randomly decreases the number of observations in the majority classes. While in the oversampling treatment, observations in minority soil classes are repeated, which means that the added observations already exist in the area and no irrelevant data is added. The objective is to approximately balanced the distribution of the soil classes data (Abdi and Hashemi, 2016; Dal Pozzolo et al., 2015). The frequency bar chart of the soil classes before and after data resampling treatments is shown in Fig. 2.

2.4. Models evaluation

2.4.1. Accuracy assessment and validation

To assess the accuracy of the models, we used four measures

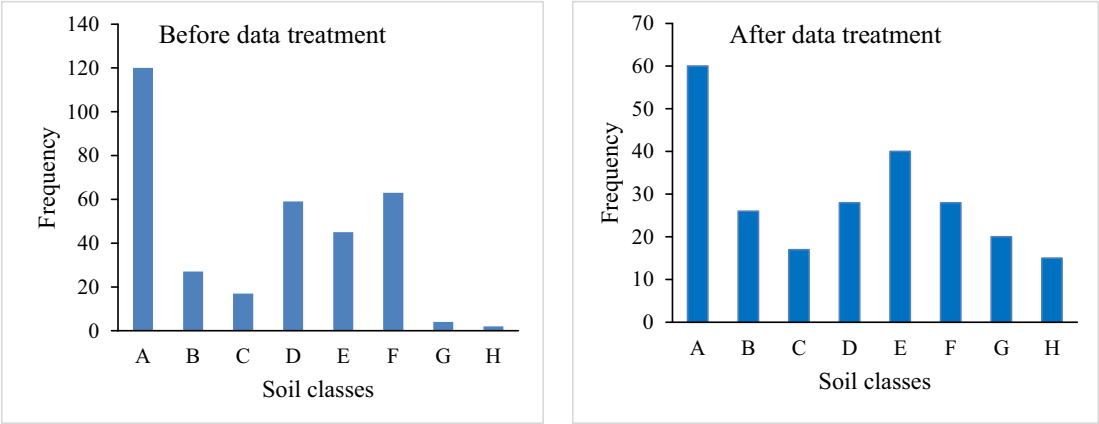

Fig. 2. Frequency of the different soil type classes before (left) and after (right) data resampling. (The letters A–H are described in Table 1).

including overall accuracy, user's accuracy, producer's accuracy and Kappa coefficient of agreement (Congalton, 1991). Overall accuracy is obtained by dividing the total correctly predicted number of classes by the total number of observations. Producer's accuracy is the correctness of predictions for a certain class, obtained by dividing the total number of correct predictions of a class to the total number of observations of that class. User's accuracy is also used for an individual class accuracy assessment, which is calculated as the total number of correct predictions of a class divided by the total number of predictions that were predicted in that class. Finally, the Kappa coefficient is a measure that shows the difference between observed agreement and expected agreement by chance, obtained by the following:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{3}$$

where, p_0 is the overall or observed accuracy, and p_e is the expected accuracy, where:

$$p_e = \sum_{i=1}^{n} \left(\frac{colSum_i}{TO}\right) \times \left(\frac{rowSum_i}{TO}\right) \tag{4}$$

Here, $colSum_i$ and $rowSum_i$ are the summations of the columns and rows of classes in the confusion matrix. *TO* is the total number of observations and *n* is the number of classes. These measures were computed by applying the *goofcat* function within *ithir* package in the R software (Malone et al., 2017).

2.4.2. Uncertainty assessment

To evaluate the uncertainty of the predictive maps obtained by the models mentioned above, we calculated the confusion index, which is the difference between the most and second most probable class (Burrough et al., 1997). This index is calculated as follows:

$$CI = 1 - (p_{max} - p_{max-1})$$

where p_{max} is the probability of the most probable soil class and p_{max-1} is the probability of the second most probable class (Burrough et al., 1997; Odgers et al., 2014). The lower the difference between the most probable and second-most-probable soil class, the more the predictive models are uncertain. Also, the spatial average of the probabilities of each class in the most probable outcome of the models is compared before and after the resampling treatments among the three different models. The probability of occurrence of a class and the confusion index indicate what is the correct class to be determined in a given location and how confused we are about that prediction.

3. Results

3.1. Results of predictive mapping using imbalanced soil classes

3.1.1. Accuracy assessment

The kappa coefficient of agreement test showed that DT is more accurate compared to RF and MNLR in both the calibration (K = 0.95) and validation (K = 0.14) datasets. The kappa coefficient for the RF model was poor: −0.04 and −0.01 for the calibration and validation datasets, respectively (Table 2). This means that agreement between observations and predictions was less than expected by chance for this model, indicating a systematic disagreement.

The user's and producer's accuracy tests on the calibration dataset (Table 3) showed a much better performance for DT model compared to MNLR and RF models. However, the Torriorthents minority class (class H) and Natrigypsids (class C) were omitted in the DT and MNLR, respectively, but, the minority classes were predicted by RF, albeit with zero correct number of predictions. Also, using the validation dataset, producer's and user's accuracies showed more number of classes with relatively higher accuracy results for DT, compared to other models (Table 4). Using the validation dataset, two classes in DT and four classes in MNLR were omitted when imbalanced classes were used, as indicated by user's accuracy test. These results confirmed that an imbalanced number of class observations affects machine learning results negatively. RF model seems to have over-fitted the majority class A and hence, have probably overestimated this class, as can be seen from the most probable map that shows a very high ratio of this class when the imbalanced classes were used (Fig. 4).

3.1.2. Uncertainty assessment

The average probability of occurrence for each of the 8 soil classes was calculated within the most probable map of the whole study area (Table 5). Results showed that using the imbalanced data, the RF model had a very low (0.01 to 0.001) average probability of occurrence for the eight soil great groups in the most probable map. The DT model showed classes with higher probability of occurrence compared to the other two models. Yet, Torriorthents (class H) and Natrigypsids (class C) had no probability of occurrence in DT and MNLR models, respectively.

The average confusion index of the whole study area pixels was calculated for comparison of the models. The confusion index showed a lower value for the DT model compared to the other two models. The average confusion was 0.73 for DT and 0.99 for both MNLR and RF models (Table 6). This index shows that the DT model produces more certain results. Also, the percentage of area with high confusion was the smallest for DT (Table 6). The confusion index did not show a significant variation within the study area; hence, the mean statistic was

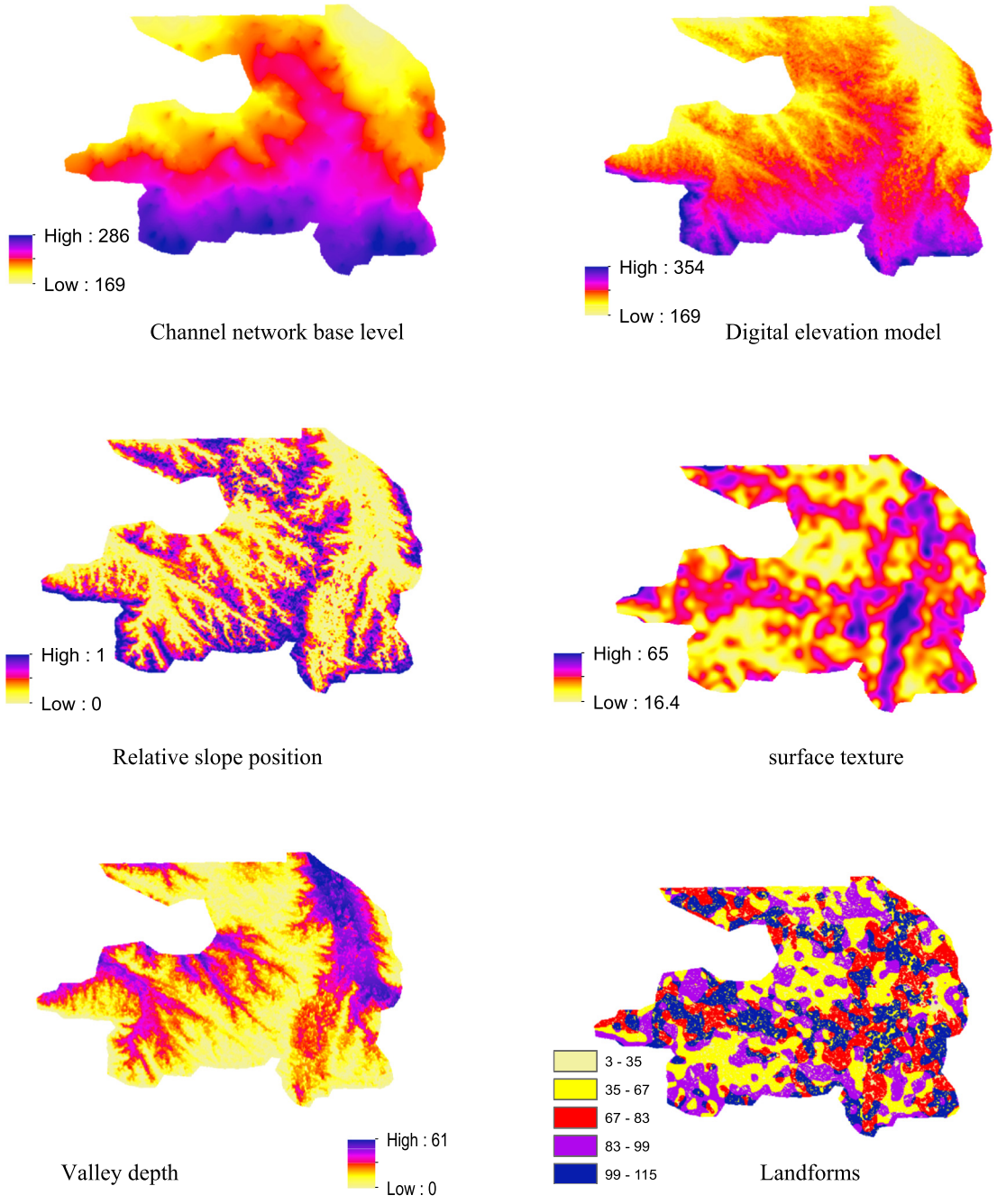

Fig. 3. Maps of the covariates used in the study.

Table 2
General accuracy results of the predictive models for balanced and imbalanced datasets.

		Imbalanced dataset		Balanced dataset (treated data)	
Model	Dataset	Overall accuracy	Kappa coefficient	Overall accuracy	Kappa coefficient
Decision tree	Calibration	96	0.95	82	0.78
	Validation	37	0.14	29	0.14
Random forest	Calibration	27	−0.04	55	0.48
	Validation	28	−0.01	14	0.06
MNLR	Calibration	46	0.23	47	0.34
	Validation	39	0.12	33	0.11

sufficient to represent the confusion index map situation.

3.2. Results of predictive mapping using balanced soil classes

3.2.1. Accuracy assessment

After applying random oversampling and under-sampling on the data, the above-mentioned tests were performed again. Results showed a remarkable improvement for the RF model with regards to the kappa coefficient for calibration and validation datasets compared to when the imbalanced class observations were used (Table 2). Also, the calibration performance for the MNLR model improved to some extent. However, the DT model did not show any improvement in terms of Kappa and overall accuracy; nevertheless, it remains the best performing model compared to RF and MNLR models in terms of Kappa coefficient. The most probable maps achieved by DT, RF and MNLR models using

Table 3
Producer's and user's accuracy results for the 3 models using the calibration dataset.

		Soil classes[a]							
Model	Accuracy test	A	B	C	D	E	F	G	H
DT	Producer's acc.[b] with imbalanced data	100	97	100	95	89	94	100	0
	Producer's with balanced data	84	77	95	78	54	93	100	100
	User's acc. with imbalanced data	93	100	100	100	100	94	100	NaN[c]
	User's acc. with balanced data	77	75	82	86	93	77	89	100
RF	Producer's acc. with imbalanced data	70	4	0	2	5	2	0	0
	Producer's with balanced data	2	66	100	73	49	45	100	100
	User's acc. with imbalanced data	33	34	0	20	16	6	0	0
	User's acc. with balanced data	100	28	46	65	36	100	100	100
MNLR	Producer's acc. with imbalanced data	84	15	0	4	3	70	40	100
	Producer's with balanced data	42	43	0	33	14	77	80	100
	User's acc. with imbalanced data	46	50	NaN	29	20	48	50	100
	User's acc. with balanced data	39	36	0	31	32	52	80	100

[a] Soil classes codes are defined in Table 1.
[b] Accuracy.
[c] NaN: not a number; means that no prediction was made for this class.

resampled data are shown in Fig. 4.

Producer's and user's accuracy tests results, which show models performance for each class individually, showed a notable improvement of calibration for 7 out of the 8 soil classes for the RF model, compared to when imbalanced classes were used (Table 3). In the DT results, class H (Torriorthents), which was omitted when trained using the imbalanced dataset, was well predicted with a user's accuracy of 100% using the treated data. Also, this minor class showed a producer's accuracy of 100%, compared to zero accuracy when trained using imbalanced data. MNLR model improved over calibration for 5 out of the 8 classes (producer's test) and also got succeeded in keeping the Natrigypsids (class C), in spite of incorrect predictions for this class (user's test).

User's accuracy test using the validation dataset revealed that DT and MNLR models succeeded to maintain one and four omitted classes, respectively, compared to when imbalanced classes were used (Table 4). For the RF model, producer's accuracy showed an improvement in prediction for 5 out of 8 soil classes in validation test. Half of the classes were also improved for this model in the user's accuracy test. However, three of the classes were not maintained in prediction using the validation dataset.

3.2.2. Uncertainty assessment

Computing the spatial average for probability of occurrence of each class in the most probable map using the balanced data input for the models showed a notable improvement for all soil classes in DT and RF and some classes for MNLR (Table 5). For MNLR model results, 5 of the classes showed improvement in their probability of occurrence. One of those classes had obtained no probability of occurrence when the imbalanced classes were used. In general, the DT model showed the highest probability of occurrence for all the classes in comparison with the other two models with a probability of 0.99 (rounded up value) for every one of the classes.

The average confusion index of the whole study area map for the 3 models showed a high improvement for the DT and RF models results compared to when imbalanced data were used. The average confusion index decreased from 0.73 to 0.02 for DT, and from 0.99 to 0.10 for RF model (Table 6). That means there is comparatively lower confusion and less uncertainty over the map produced by DT model after data resampling, which shows higher consistency of modeling. This index did not show noticeable improvement for MNLR model. Also, the area percentage with high confusion in the confusion index map decreased noticeably for DT and RF models.

4. Discussion

Researchers have reported improvements in classification problems using resampling techniques in other fields of research. For example, significant improvement was reported for different class types after oversampling in a research on various classification types (Sáez et al.,

Table 4
Producer's and user's accuracy results for the 3 models using the validation dataset.

		Soil classes[a]							
Model	Accuracy test	A	B	C	D	E	F	G	H
DT	Producer's acc.[b] with imbalanced data	58	30	0	10	7	62	0	0
	Producer's with balanced data	31	60	50	25	0	55	0	0
	User's acc. with imbalanced data	39	75	0	11	25	50	NaN[c]	NaN
	User's acc. with balanced data	54	60	40	7	0	24	0	NaN
RF	Producer's acc. with imbalanced data	73	10	0	0	14	0	0	0
	Producer's with balanced data	0	30	85	25	29	5	0	0
	User's acc. with imbalanced data	34	17	0	0	40	0	0	0
	User's acc. with balanced data	NaN	25	13	8	20	20	NaN	NaN
MNLR	Producer's acc. with imbalanced data	73	10	0	0	0	67	0	0
	Producer's with balanced data	33	10	0	30	0	72	0	0
	User's acc. with imbalanced data	38	50	NaN	0	NaN	44	NaN	NaN
	User's acc. with balanced data	63	5	0	15	0	32	0	0

[a] Soil classes codes are defined in Table 1.
[b] Accuracy.
[c] NaN: not a number; means that no prediction was made for this class.

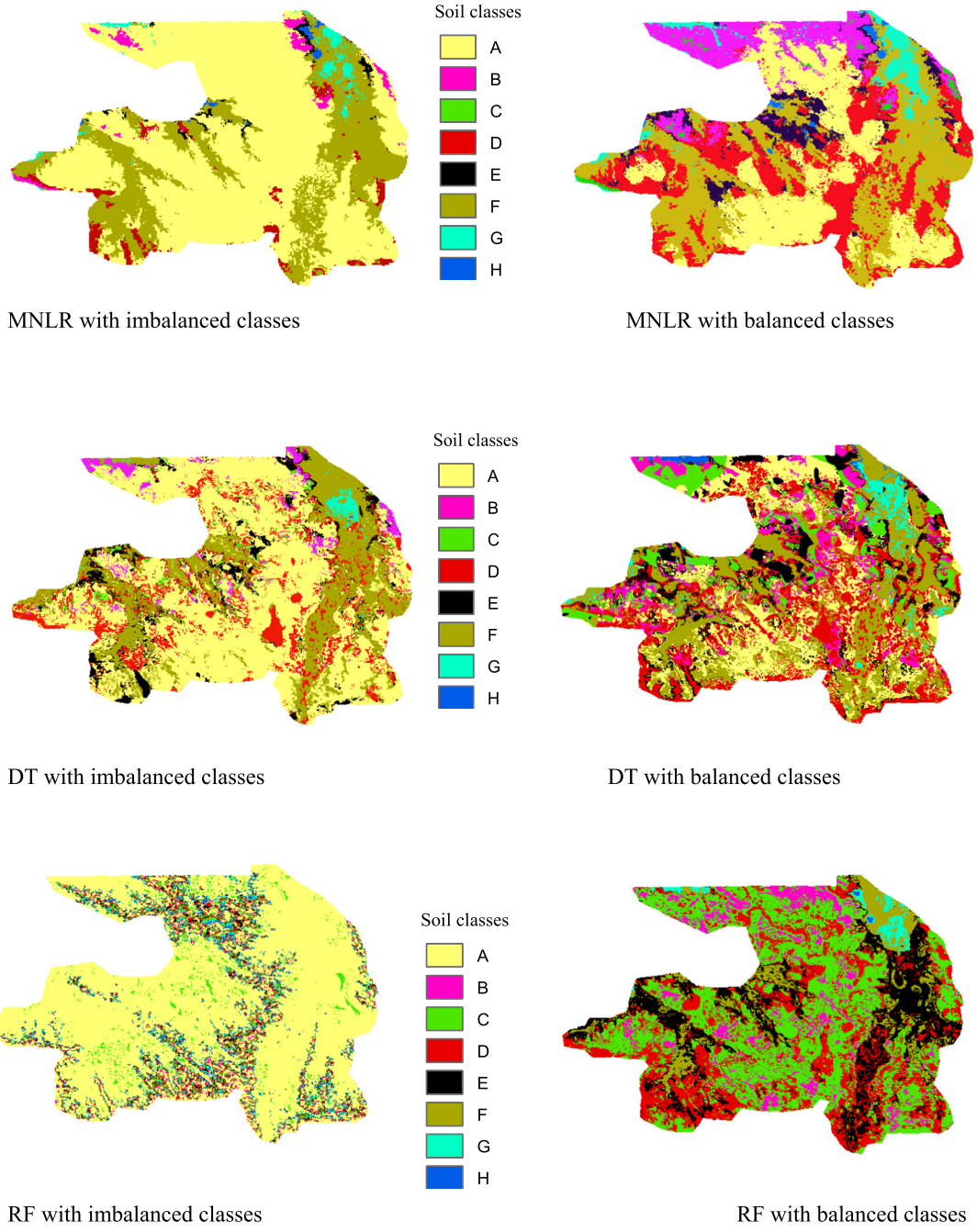

Fig. 4. Comparison of the most probable maps produced by the three models with and without balancing the classes.

Table 5
Average probability of occurrence for each class, obtained by the 3 predictive models.

		Soil classes[a]							
Model	Average probability	A	B	C	D	E	F	G	H
DT	Probability with balanced dataset	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Probability with imbalanced dataset	0.53	0.41	0.35	0.44	0.40	0.55	0.53	NaN[b]
RF	Probability with balanced dataset	0.51	0.93	0.95	0.94	0.96	0.96	0.97	0.98
	Probability with imbalanced dataset	0.001	0.01	0.01	0.01	0.01	0.01	0.01	0.01
MNLR	Probability with balanced dataset	0.39	0.33	0.38	0.34	0.29	0.72	0.72	0.82
	Probability with imbalanced dataset	0.46	0.34	NaN	0.33	0.27	0.48	0.65	0.91

[a] Soil classes codes are defined in Table 1.
[b] NaN: not a number; means that no prediction was made for this class.

Table 6
Confusion index statistics for the models before and after data treatment.

Statistic	DT with imbalanced classes	DT with balanced classes	RF with imbalanced classes	RF with balanced classes	MNLR with imbalanced classes	MNLR with balanced classes
Average CI[a]	0.73	0.02	0.99	0.10	0.99	0.98
Area% of map with CI ≥ 0.99	0.04	0.001	71	0.001	99	99

DT: Decision trees; **RF**: random forest; **MNLR**: multinomial logistic regression.
[a] Confusion index.

2016). In an extensive study on the use of resampling techniques on several imbalanced datasets, Loyola-González et al. (2016) reported accuracy improvement after the use of oversampling and under-sampling for contrast pattern based classifiers. In the field of medicine, protein classification has improved with the help of oversampling technique (Ahmad et al., 2017). In soil science, oversampling was reported to improve classification results using Markov chain random fields models for mapping soil type classes (Sharififar et al., 2019). However, there is still need for more research on the effect of different balancing techniques on different classifiers, particularly in the field of digital soil mapping.

Comparing the most probable maps produced by the three models (Fig. 4) shows that class H (Torriorthents, coded in blue) are absolutely lost from the map created by DT model using imbalanced data. This area, although small on the map, covers around 40 ha, which requires fairly different management for crop growth or rangeland utilization. Likewise, class C (Natrigypsids) was omitted in the MNLR map when the imbalanced classes were used (coded green in the map). This area also covers approximately 10 ha of the study area lands, as it contains high amount of sodium, it requires careful land management strategies different from those suitable for other parts of the area. The map produced by the RF model using imbalanced data shows that the majority of the area is predicted as class A (Calcigypsids) (yellow color in the map) which is dominated by Calcium and behaves completely different from other soil classes. Comparing this map with other maps in Fig. 4 show that RF has overestimated the majority class (A). All these findings reveal that maps produced using imbalanced classes could be misleading for the users or decision makers of the final produced maps.

Overall, data treatment with over- and under-sampling helped overcome the issue of modeling imbalanced class observations by improving the predictive models' results, in the sense of maintaining the minority class or classes in the calibration and validation tests to a reasonable extent. Although mostly producer's and user's accuracy of the minority classes were found to be zero in the validation test, but yet, the minority classes have been maintained in the modeling process and are not missed out, compared to when using imbalanced datasets with no prediction at all for them. Here in the validation, it should be clarified that zero means no correct prediction in the validation dataset, which could be due to the small size of our dataset for validation, particularly, in the minor soil classes that have very few number of observations. Nevertheless, this is still an achievement, as the minority classes were often completely lost in the mapping process, even at the calibration stage (e.g. DT model; Table 3), when data pretreatment was not applied. When the data were not treated for balancing, the outcome of the producer's and user's accuracy tests in some cases for the minority classes were found to be *not a number* (*NaN*), which is mathematically interpreted in such a way that these classes did not exist in the models output, as the ratio denominator in the user's accuracy test (refer to the Methods section) was zero for these classes. Beside the accuracy assessment results, uncertainty assessment also revealed a better performance of the models after the treatment.

Under-sampling decreases the number of observations in the majority classes (such as class A in our case) to balance the classes distribution, hence, it can be useful to prevent overfitting and overestimation of such a class. A major issue with the RF model is overestimation of the dominant soil class A when the imbalanced data were used (Fig. 4). In comparison, oversampling of the minority classes has helped in maintenance of these classes without bringing any artificial data into the dataset and mapping process, rather only duplicating some of the observations with the same coordinates. Altogether, these techniques resulted in noticeable improvement in models' performance by decreasing the uncertainties and prevention of losing the minority classes in the produced maps. However, RF model did not improve in terms of minority classes maintenance using the validation dataset.

A problem in dealing with the imbalanced soil classes with minority classes that have very few numbers of observations is that one might face a very small size of data for validation. As can be seen from the results of this research, individual classes accuracies in the validation dataset was zero for the minority classes, but this does not mean that the performance is not improved; rather it is the small size of the validation dataset that might not be capable to show real number of correct predictions for these classes.

5. Conclusions

This study brought some insight into soil type mapping with imbalanced and balanced number of observations using 3 well-utilised models within the approach of digital soil mapping with the help of data pre-treatment. Imbalanced distribution of class observation resulted in uncertain maps with minority classes being lost and relatively poor accuracies.

After data treatment, with over- and under-sampling, decision trees and multinomial logistic regression models showed significant improvement in maintaining the minority classes, in both calibration and validation evaluations. While data treatment can cause a slight decrease in the overall accuracy on the validation dataset, it decreased the uncertainty of all models. Zero correct prediction of the minority classes after data treatment in the validation set is mainly due to the small size of our validation dataset.

Comparing the 3 models, decision trees showed the highest calibration (Kappa and overall tests) and validation (Kappa) results with and without data treatment. RF has an issue of overestimation of some of the majority classes. According to the results, decision trees model was found to perform best in response to data resampling, compared to MNLR and RF models.

Acknowledgment

Ardebil Water Organization of Iranian ministry of energy is acknowledged for funding the soil sampling and laboratorial analyses.

References

Abdi, L., Hashemi, S., 2016. To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans. Knowl. Data Eng.(1).
Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214, 101–113.
Ahmad, J., Javed, F., Hayat, M., 2017. Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods. Artif. Intell. Med. 78, 14–22.

Baker, J.B., Fonnesbeck, B.B., Boettinger, J.L., 2016. Modeling rare endemic shrub habitat in the Uinta Basin using soil, spectral, and topographic data. Soil Sci. Soc. Am. J. 80, 395–408.
Breiman, L., 2006. randomForest: Breiman and Cutler's Random Forests for Classification and Regression.
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, 68–83.
Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115–135.
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.
Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46.
Costantini, E.A.C., L'Abate, G., 2016. Beyond the concept of dominant soil: preserving pedodiversity in upscaling soil maps. Geoderma 271, 243–253.
Dal Pozzolo, A., Caelen, O., Bontempi, G., 2015. Unbalanced: Racing for Unbalanced Methods Selection. (R Packag. Version 2).
de Martonne, E., 1926. Une nouvelle function climatologique: L'indice d'aridité. Meteorologie 2, 449–459.
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2012. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev) 42, 463–484.
Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207.
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.
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.
Kuhn, M., Weston, S., Culp, M., Coulter, N., Quinlan, R., 2018. Package 'C50.'
López, V., Fernández, 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.
Loyola-González, O., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., García-Borroto, M., 2016. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. Neurocomputing 175, 935–947.
Ma, Y.X., Minasny, B., Malone, B.P., McBratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70, 216–235.
Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using R for digital soil mapping. Springer.
McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Nayal, A., Jomaa, H., Awad, M., 2017. KerMinSVM for imbalanced datasets with a case study on arabic comics classification. Eng. Appl. Artif. Intell. 59, 159–169.
Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214, 91–100.
Quinlan, J.R., 1993. In: Kauffmann, Morgan (Ed.), C4. 5: Programming for machine learning. 38. pp. 48.
R Development Core Team, R, 2011. R: A Language and Environment for Statistical Computing.
Ripley, B., Venables, W., 2015. R-Package Nnet: Feed-forward Neural Networks and Multinomial Log-linear Models. Google Sch.
Sáez, J.A., Krawczyk, B., Woźniak, M., 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recogn. 57, 164–178.
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.
Siers, M.J., Islam, M.Z., 2018. Novel algorithms for cost-sensitive classification and knowledge discovery in class imbalanced datasets with an application to NASA software defects. Inf. Sci. (Ny). 459, 53–70.
Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253, 67–77.
USDA, N.R.C.S., 2010. Keys to soil taxonomy. Soil Surv. Staff Washington.
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.