Application of random forest classification and remotely sensed data in geological mapping on the Jebel Meloussi area (Tunisia)

Gáspár Albert¹ · Seif Ammar²

Received: 3 April 2021 / Accepted: 22 September 2021 / Published online: 29 October 2021
© The Author(s) 2021

Abstract

Remotely sensed data such as satellite photos and radar images can be used to produce geological maps on arid regions, where the vegetation coverage does not have a significant effect. In central Tunisia, the Jebel Meloussi area has unique geological features and characteristic morphology (i.e. flat areas with dune fields in contrast with hills of folded and eroded stratigraphic sequences), which makes it an ideal area for testing new methods of automatic terrain classification. For this, data from the Sentinel 2 satellite sensor and the SRTM-based MERIT DEM (digital elevation model) were used in the present study. Using R scripts and the random forest classification method, modelling was performed on four lithological variables—derived from the different bands of the Sentinel 2 images—and two morphometric parameters for the area of the 1:50,000 geological map sheet no. 103. The four lithological variables were chosen to highlight the iron-bearing minerals since the spectral parameters of the Sentinel 2 sensors are especially useful for this purpose. The training areas of the classification were selected on the geological map. The results of the modelling identified Eocene and Cretaceous evaporite-bearing sedimentary series (such as the Jebbs and the Bouhedma Formations) with the highest producer accuracy (> 60% of the predicted pixels match with the map). The pyritic argillites of the Sidi Khalif Formation were also recognized with the same accuracy, and the Quaternary sebkhas and dunes were also well predicted. The study concludes that the classification-based geological map is useful for field geologist prior to field surveys.

Keywords Geological mapping · Sentinel 2 · Remote sensing · Random forest classification · Machine learning

Introduction

The compilation of the 1:50,000 geological map series of Tunisia started in 1945, but only 80% of the sheets have been completed and published to date (ONM 2021). The Jebel Meloussi sheet (no. 103) is one of the latest map sheets in the 1:50,000 map series, which provides up to date geological information. It was published by the Tunisian National Mining office in 2005, preceded by fieldwork and aerial

stereotopographic surveys (Mahjoub 2005). In the southern part of the country, the geological map covering is planned to be fulfilled with 1:100,000 maps, but only 55% of the sheets are finished (ONM 2021). Completing the map covering of a country is usually a national and strategic goal since the lack of detailed geological information makes it difficult for industrial firms to acquire the necessary knowledge in order to initiate new mineral exploration projects. Accelerating the compilation of maps is therefore desirable. One of the possibilities for quick geological mapmaking is to automate the process and use the data that is already available.

The automation of geological mapping based on remotely sensed data intrigues geoscientists for a long time, it speeds up the reconnaissance of an area, makes it cheaper and may lead to new discoveries (Bedini 2009; Carneiro et al. 2012; He et al. 2015). Sensors can collect detailed information over large areas in very short time, and surveying geologists nowadays concentrate on data classification and validation processes rather than fieldwork (Drury 2001). The automatic method of making a geological map using satellite imagery

Communicated by Biswajeet Pradhan.

This paper was selected from the 3rd Conference of the Arabian Journal of Geosciences (CAJG), Tunisia 2020

✉ Gáspár Albert
albert@ludens.elte.hu

¹ ELTE Eötvös Loránd University, Budapest 1117, Hungary

² ELTE Doctorate School of Earth Sciences, Budapest 1117, Hungary

works well if no vegetation coverage is present in the area, but the presence of paved surfaces and buildings increases the difficulty of automation (Grebby et al. 2011). Consequently, barren and uninhabited areas are ideal candidates to be used for processing when using satellite imagery as a base data.

In the case of geological mapping, the spectral parameters of the most characteristic minerals of the surveyed area play an important role in the selection of the remotely sensed data for modelling (Kruse et al. 2002). Also, open-access data from NASA (Landsat), ESA (Sentinel) and other state-sponsored space programmes are often preferred over private and/or premium data, especially if the mapping is aimed at reconnaissance or testing. The remotely sensed (RS) data is processed with different image classification methods, and the results are “classes” which—after evaluation—can be considered as geological features (e.g. lithological types).

The classification of multi-band satellite images is well studied, and several methods were tried for geological purposes as well (Cracknell and Reading 2014). For example, the K-means and the SOM (self-organizing maps) are applied for unsupervised, and the image segmentation, the convolutional neural networks and the decision trees are often used methods in supervised classifications (Kotsiantis et al. 2007; Carneiro et al. 2012). The RFC (random forest classification) is a machine learning technique that is based on decision trees to perform predictions from a set of data (Breiman 2001). The trees branch (mtry) as many times as much predictor variables are used (tree height). The no. of trees (ntree) is the no. of random sampling points from the learning areas. The RFC is reported to be useful in earth sciences especially for qualitative geological mapping (Cracknell and Reading 2014).

While the unsupervised classification is less labour-intensive, the creation of classes is not controlled (only their numbers) and the resulting map may contain categories which do not necessarily match with existing quality classes, such as geological formations. Supervised classification provides this control and the classes will represent the same categories that were used to train the algorithm (Kotsiantis et al. 2007).

In this study, we aim to examine the effectiveness of the random forest classification method on the Jebel Meloussi area (Fig. 1) using the existing map as training data. Our aim is to simulate the situation where the field surveying of the area is not yet complete but a geological map is being prepared. By documenting our work in detail, we hope to help accelerate the process of geological mapping in the adjacent areas, where the geology, morphology and land cover are similar.

In this paper, we first describe the geological and morphological characteristics of the area, which played an important role in the selection of the variables used in the modelling.

Fig. 1 The Jebel Meloussi study area is located in central Tunisia and covers ~ 650 km²

Then, the basic data used and the geoinformatics and parameterization methods applied to them are described, followed by the results. Finally, the advantages and disadvantages of the method used are discussed. Our method uses predictor variables that have not, to our knowledge, been tried for a similar purpose.

Geology and morphology of the area

The most characteristic morphological feature of the area is the Jebel Meloussi (626 m), which is the eastern continuation of the Jebel Majoura (874 m) forming a southwest-northeast trending mountain chain that turns west-east in the Meloussi range. The Jebel Zebbeus (461 m) to the south-east forms a north-south trending chain of hills with the Jebel Gouleub (Fig. 2), which is part of the north-south trending structural unit.

Triassic evaporitic mudstones, Cretaceous and Eocene carbonates and clastic sediments, Oligo-Miocene siliciclastic and Quaternary terrestrial sediments are present (Burolet 1956; Mahjoub 2005; Trabelsi et al. 2006). Outcrops of the Triassic are present in Jebel Zebbeus (Jebbs Hill), where it is best exposed, but the majority of the outcrops of the region is mainly made up of Cretaceous to Quaternary geological strata (Trabelsi et al. 2006). The Eocene sedimentary succession is representative in the Jebel Zebbeus area in an

Fig. 2 The Gouleb Hill seen from the South is mainly composed of folded Turonien-Coniacien mudstone and evaporite layers

elongated north–south trending synform structure which formed partly due to the halokinetics of older evaporites (Zaïer et al. 1998).

The Jebel Meloussi is an anticline with an eroded core. Due to the tilted axial plain, the flanks are not symmetrical. Two different structural directions could be distinguished in the area: the first follows the E–W trend and the second chopped structure has an N–S axis (Boutib and Zargouni 1998). The structures are contemporaneous with the folded nappe systems and the deformation of the Atlas Mountains, which was most active during the Neogene (Burollet 1991).

Morphology is rough in the older, folded and flat in the younger terrestrial formations. The plain in between the hills is on average 280- to 300-m altitude and is characterized by a generally gentle and fairly regular slope towards the periodic streams. Dune fields are also present in the northern parts of the study area, but their extent is small. Although dune migration rates can be as high as 10–50 m/year in areas free of vegetation (Lorenz et al. 2013), in the study area, most of the dune fields are covered by plantations in order to stop sand mobilization.

The selection of the Meloussi sheet in our study was based mainly on the low rate of vegetation, anthropogenic coverage and the characteristic structural settings of the area. The aim of the presented study was to reconstruct the geological map of Jebel Meloussi using only freely accessible remotely sensed data sources, by means of GIS and modelling, and to evaluate the possibility of automatic geological mapping.

Materials and methods

Supervised methods in machine learning-based classifications usually require a labour-intensive initial stage when a set of data is created with labels of different quality (Kotsiantis et al. 2007). In the present study, the labelling was done using the results (map) of a previous geological survey (Mahjoub 2005). The spectral and morphological parameters of the labelled areas were then analyzed using

remotely sensed data to identify the characteristics of each geological type. Using these data sources, a complex raster classification was performed to reconstruct the geological map of Jebel Meloussi.

Remotely sensed data

The Copernicus Programme provides several types of remotely sensed data for free since 2015 (EU-EOP 2021). Due to the spectral band setting of the sensors, the multi-spectral optical images of Sentinel 2 satellites are especially useful for mapping iron-rich lithological types in arid regions (Van der Werff and Van der Meer 2016). The ground resolution of these images is 10 m in the case of spectral bands 2, 3, 4 and 8, 20 m in the case of 7, 8A, 11 and 12 and 60 m in the case of spectral bands 1, 9 and 10. The selected satellite data for the target area was collected on 10 July 2018. For the subsequent analysis, the 2A level data were used,¹ to which the atmospheric correction has already been applied previously (Main-Korn et al. 2017). In the present study, the year of satellite data acquisition should not be necessarily the same as the year of the geological survey since the geology did not change in such a short time.

For mapping, topographic elevation data are also essential. Freely accessible data sources for digital surface models—such as the SRTM and ASTER—are available and widely used in the earth sciences (e.g. Hirt et al. 2010; Yamazaki et al. 2017). However, a DSM contains the elevation of man-made objects and the vegetation, and using these data for morphometric analysis may lead to false conclusions. The MERIT DEM is also a freely accessible post processed SRTM3 dataset which can be utilized for calculating morphometric variables, as the vegetation cover and the buildings were removed from the original dataset (Yamazaki et al. 2017). In the present study, we used the MERIT DEM to calculate morphometric variables.

¹ The ID of the data is: subset_0_of_S2B_MSIL2A_20180710T100029_N0208_R122_T32SND_20180710T143733.

Fig. 3 Mosaic of the satellite image (Google Earth photo) and the geological map (Mahjoub 2005) of the Jebel Hill

Cartographic materials

The map of the Jebel Meloussi was compiled in the Carthage Nord System (EPSG: 22391). The geological map depicts formations in the case of older than Quaternary rocks or uses merged symbols by geological age. Quaternary formations are represented by their genetics. Lithological descriptions of the formations were used to identify the iron-rich rock types.

Iron-rich sediments in the area are usually red coloured due to oxidation. These geological formations characterize the landscape of the area (Fig. 3) and are represented on the map in the form of several stratigraphic units. The Triassic successions are composed of evaporites and red mudstone layers, where the red colour comes from the iron oxide (Burllet 1991; Mahjoub 2005). The red mudstones of the Cretaceous Bouhedma Formation reportedly contain iron oxides in 5–9% (Boussen et al. 2016). During the Late Eocene, the area was uplifted and was covered by evaporite- and phosphate-rich sedimentary layers (Jebel Formation) and later overlain by marls containing oolitic iron (Zaïer et al. 1998).

Image processing

The resolution of the satellite images was spatially harmonized (scaled up) to make it possible to carry out index calculations with band ratios. After the upscaling, the common ground resolution was 10 by 10 m for each band of the used satellite image stack. The bicubic method was used for the reclassification of the originally 20-m resolution bands. The bands of the Sentinel 2A dataset were processed separately using the SAGA GIS program. After the reclassification, the bands were used for calculations to create the NDVI (normalized difference vegetation index) and those variables (geological indices), which are appropriate for detecting iron-rich materials.

The NDVI was calculated by using the $NDVI = (NIR - RED) / (NIR + RED)$ formula, where NIR is near-infrared reflectance and RED is red reflectance (Rouse et al. 1974). In the case of the Sentinel 2 dataset, band 8 (NIR) and band 4 (RED) are usually used. The calculated values ranged from -0.03 to 0.75 , but 94.6% of the study area was under 0.2 , and the mean value for the whole area was 0.12 (st. dev. 0.05). Based on the classification of

Table 1 Geological indices and the band ratio formulas for calculations

Feature	Band ratio–Sentinel 2B
Ferric iron, Fe3 +	Band 4/band 3
Ferrous iron, Fe2 +	Band 12/band 8 + band 3/band 4
Laterite/alterations	Band 11/band 12
Ferric oxides	Band 11/band 8

the NDVI values for semi-arid regions (e.g. Aquino et al. 2018), the 0–0.2 value range represents bare soil or very low vegetation. The areas where the vegetation was low ($0.2 < \text{NDVI} \leq 0.4$), moderately low ($0.4 < \text{NDVI} \leq 0.6$) or moderately high ($0.6 < \text{NDVI}$) are found on the low lying plains, where mainly quaternary sediments are present.

Four geological indices were calculated, concentrating on iron-containing minerals (FeO, Fe3 +, Fe2 + ions and laterite). The lithological indices were calculated from the spectral bands of the Sentinel 2A image based on published formulas (Van der Werff and Van der Meer 2016). The indices and band calculations are shown in Table 1.

Morphometric derivatives

For the classification, two morphometric variables (topographic wetness index—TWI and topographic ruggedness index—TRI) were created from the MERIT DEM. The MERIT DEM was processed with the SAGA module of QGIS to calculate the indices. The morphometric indices were calculated as raster layers with a slightly better spatial resolution (50 by 50 m pixels) than the original MERIT DEM. The physical parameters of the modelling area are shown in Table 2. All the raster-type data was transformed and cropped to this extent during the modelling processes.

Vector processing

After scanning, the map sheet no. 103 was georeferenced in the Carthage Nord System and the content was digitized using the QGIS programme. The geological indices and the lithological descriptions of the map units were also recorded in a geodatabase. In the case of those categories, where the formations were subdivided into different facies (e.g. the Cenomanian calcarenites and dolomites “Cce-a” with the

dolomitic gypsum facies “Cce-b”), the subcategories were digitized separately, but in the training process only the general, solo-indexed (e.g. “Cce” = dolomite with rudist bivalves and ammonites), polygons were used as training areas for the classification.

A total of 203 training areas were selected from 26 geological types (typically 5–8 areas/types) for the supervised classification (Fig. 4). The training areas were exported as polygon shape features into a separate geodatabase containing numbered indices (1–26) as attributes. The areas were selected to simulate the results of a quick field survey, which typically involves only a few sites per formation, but were otherwise randomly selected from the geological map. Also, the areas having higher than 0.2 NDVI value were avoided during the selection.

Classification

The method is based on applying a random forest classification (RFC) on the 50-m resolution multi-band image and using training areas. The process included the downscaling of the finer resolution rasters (10 and 20 m down to 50 m) and the upscaling of the 60-m resolution satellite image bands. In the first case, the weighted moving average method was used, while in the second case bicubic interpolation was applied. The training areas were selected on the geological map sheet no. 103. To avoid imbalances in the sampling, the 26 geological types were sampled with the aim to have approximately equal numbers in each category; however, the small areal distribution of some of the geological formations (e.g. the Middle Miocene Ain Grab Formation) made this difficult. The average size of the training areas was 36,780 m², and all training areas covered at least one 50 m by 50 m pixel on the multi-band raster image that was used for the classification. Most of the samples were from Quaternary dunes, sebhkas and alluvial sediments, as well as the Pleistocene residual and slope sediments (gravel, sand and silt), providing approximately half of the training pixels. The remaining half was distributed between the Triassic (~2%), Cretaceous (~31%), Paleogene (~7%) and Neogene (~5%) formations.

The RFC was executed in Rstudio (a freely accessible program), and the rasters (Sentinel-2 bands, geological and morphometric indexes) were put together in multi-band Geo-Tiff-files, which had the same physical parameters. Several

Table 2 The physical parameters of the rasters, which were used for the classification

CRS	EPSG: 32632—WGS 84/UTM zone 32 N					
	Min. easting (m)	Min. northing (m)	Max. easting (m)	Max. northing (m)	Width (px)	Height (px)
Extent	529,475	3,839,075	563,325	3,862,425	677	467

Fig. 4 Distribution of the training areas (orange polygons). The background shows the 1:50,000 scale geological map sheet (no. 103) with the original colouring

models were tested during the process, aiming at the right choice of predictors and parameters.

The first step was selecting the stronger predictor variables from the available data, excluding those that might contain bias. The B1 and B9 bands for example indicate the presence of aerosol particles and water vapor, respectively (EU-EOP 2021). As these are not related to the geological properties of the solid surface, and especially as they can change rapidly over time, they have been excluded from the classification. Using the remaining 10 bands of the Sentinel data package, as well as the lithological indices and morphometric indices (4 + 2 variables), and also using elevation as a baseline, a preliminary classification was performed. In this model, the RFC parameters were set to 100 trees and to the default $mtry = \sqrt{n}$ value, where n was the number of classes. In our case $n = 26$, so the $mtry = 5$. The strength of the variables was tested by examining the role of the variables in the $mtry$ events using the MDG (Mean Decrease Gini) values (Fig. 5). Although elevation was the strongest predictor, the maps produced using it showed misleading patterns. Some formations appeared only at a certain altitude and followed the morphology, giving a false impression that there was a rock outcrop (Fig. 6A). Numerically, this

appeared as the imbalance of the producer and the user accuracy of the 26 class variables. The accuracy test comparing the raster resulting from the classification with the geological map showed that classes with good user accuracy have poorer producer accuracy, and vice versa (Fig. 6B).

Of the remaining variables, the two morphometric indices were the strongest, along with the lithological variables red iron (Fe^{3+}) and laterite. Two of the lithological indices, black iron (Fe^{2+}) and iron oxide, were also selected as they were considered important for raw material research.

By examining the values of the predictor variables in the different geological categories, we can determine how well a variable characterizes a given geological category (Table 3). The average standard deviations of the variables relative to the mean values were relatively small ($> 4\%$) for TWI, FeO , Fe^{2+} , Fe^{3+} and limonite, while the standard deviations of TRI were generally large.

A multi-band raster of the six variables was created and then used for modelling. The classifications were performed by modifying three parameters: $mtry$, $ntree$ and the type of sampling, which could be done with or without replacement. The $mtry$ parameter was set to 26 (all variables), 5 (the square root of the number of variables) and 2. The $ntree$

Fig. 5 Importance of the possible predictor variables determined by the random forest analysis using Mean Decrease in Gini (MDG)

was set to 100, 250, 500 and 750, and the type of the sampling at the mtry events was done either with or without replacement of the randomly selected (in bag) pixel values. Our aim was to specify the parameters that would yield the highest estimation probability. In each model, the probability of estimation for the entire area was calculated, using the out of bag (OOB) method (Li 2013). For this, the probability grids were calculated for each of the 26 classes in all models and aggregated in one grid; the mean value of the aggregated grid was the selection criteria for the best model parameters. The model settings and results are shown in Table 4. The finally selected parameters were mtry = 2, tree count = 500 with using sampling without replacement method during the mtry events, which is reported to decrease the bias in OOB error calculations (Mitchel 2011). The prediction probability for the whole study area using the mentioned parameters was 0.86 (with 0.2 st. dev).

The cross-correlation of the variables and their strengths were also checked (Table 5). The cross-correlation refers to the strength of TRI and Fe3+ and the weakness of the FeO and Fe2+ variables. The Laterite and the TWI are also weaker predictors as they cross-correlate with the other variables in 3/5 cases.

Results

The overall distribution of the 26 different geological types in the model was very similar to the existing geological map (Fig. 7). Also, the elongated concentric pattern of the folded Cretaceous sequence of Jebel el Barda el Hamra and the Triassic evaporites of the Jebel Zebbeus (Jebs) at the Northern Meknassy region was well outlined in the model (Fig. 8).

The results of the modelling identified Eocene and Cretaceous evaporite-bearing sedimentary series (such as the Jebs and the Bouhedma Formations) with the highest accuracy (> 60% of the predicted pixels match with the map). The pyritic argillites of the Sidi Khalif Formation were also recognized with the same accuracy, and the Quaternary sebkhas and dunes are also well predicted (Table 6).

The RFC produced the best results when the ratio of training and the testing area was 90/10 (the selection was random and pixel based). In this model, the mtry accuracy was 0.62 and the Kappa 0.58, which indicates a good-to-moderate agreement. We performed a pixel-wise analysis of the producer accuracy to compare the true positive results with the false positives on pixel level (Table 6). Using the digitized geological map of the Jebel Meloussi area, a polygon-wise accuracy test was also performed on the whole area. The model was compared to the original digitized polygon map to see whether it matches the type or not. If the majority of the pixels within a polygon matched, the model prediction was considered “fit”; otherwise, it was considered to be a “misfit”. From the proportion of fitting polygons to all polygons of a given type, the polygon-wise producer accuracy was calculated (Table 6).

Discussion

The method and the training areas

The RFC is a powerful machine learning technique for multivariate classification that uses several decision trees, each providing a prediction. However, the internal structure of the data—such as the stratigraphic sequence of the

Fig. 6 Excerpts of the classification results and accuracy plots representing the premier and the final model settings. **A** Classification with the premier settings: 17 predictor variables, mtry=5, ntree=100; **B**

accuracy plots of the 26 classes using the premier settings; **C** classification with the final settings: 6 predictor variables, mtry=2, ntree=500; **D** accuracy plots of the 26 classes using the final settings

Table 3 List of the 26 geological categories, their pixel-counts and statistical characteristics for the 6 predictor variables

Category	Symbol	Fromation	CELLS	TWI_MEAN	TWI_STDDEV	TRI_MEAN	TRI_STDDEV	FeO_MEAN	FeO_STDDEV	Fe2+_MEAN	Fe2+_STDDEV	Fe3+_MEAN	Fe3+_STDDEV	Lim_MEAN	Lim_STDDEV
1 a		Alluvial sediments	17489	13.72	0.44	1.41	1.28	1.35	0.08	1.90	0.09	1.40	0.07	1.14	0.05
2 dQ		Eolian dunes	8914	13.70	0.16	0.60	0.35	1.29	0.08	1.86	0.10	1.45	0.08	1.11	0.05
3 sQ		Sebkhas	7483	13.77	0.16	0.40	0.44	1.29	0.05	1.90	0.06	1.38	0.05	1.11	0.04
4 aQp-bQp		Terrasse deposits	2919	13.63	0.25	0.88	0.90	1.33	0.07	1.91	0.08	1.46	0.07	1.09	0.03
5 Qp-a-b-c		Pleistocene silts & clays	137971	13.74	0.35	0.91	0.95	1.31	0.09	1.88	0.10	1.42	0.08	1.12	0.05
6 M-PI		Fm. Segui	5492	13.11	0.64	3.11	2.14	1.37	0.05	1.87	0.05	1.43	0.06	1.17	0.03
7 M2-3		Fm. Beglia	299	13.45	0.46	3.19	2.61	1.38	0.04	1.81	0.07	1.53	0.10	1.20	0.02
8 M2		Fm. Ain Ghrab	50	13.16	0.62	3.11	1.14	1.39	0.03	1.80	0.04	1.58	0.09	1.20	0.02
9 O-Ma		Fm. Fortuna	638	13.15	0.53	3.36	1.75	1.39	0.05	1.87	0.06	1.42	0.07	1.19	0.03
10 E2		Fm. Jebes	2827	12.66	0.68	3.96	2.53	1.28	0.08	1.75	0.10	1.30	0.06	1.33	0.10
11 E1		Fm. Metlaoui	362	12.69	0.69	4.03	2.76	1.29	0.08	1.78	0.11	1.32	0.05	1.28	0.10
12 Cm-p		Fm. El Haria	3970	13.60	0.56	1.88	2.63	1.31	0.06	1.86	0.06	1.37	0.04	1.16	0.05
13 Cm		Fm. Abiod upper member	2166	12.82	0.83	5.29	4.94	1.35	0.08	1.89	0.05	1.34	0.05	1.18	0.04
14 Cca-m		Fm. Abiod	644	13.19	0.85	4.22	4.69	1.35	0.08	1.90	0.05	1.35	0.05	1.16	0.05
15 Cca		Fm. Abiod lower member	1531	12.65	0.95	6.57	5.52	1.36	0.08	1.89	0.05	1.36	0.05	1.18	0.04
16 Cs		Santonian general	4589	13.02	0.64	3.63	3.51	1.37	0.05	1.90	0.05	1.38	0.05	1.17	0.03
17 Ct-co		Turonian and Coniacian general	6489	12.30	0.70	9.98	5.41	1.42	0.07	1.92	0.07	1.38	0.04	1.19	0.04
18 Cce-a-b		Cenomanian general	15173	12.91	0.72	5.40	5.08	1.36	0.08	1.89	0.08	1.42	0.06	1.16	0.07
19 Cal		Albian general	4902	12.23	0.60	9.22	4.66	1.39	0.05	1.90	0.05	1.38	0.04	1.18	0.03
20 Cap		Fm. Orbata	3210	12.01	0.56	9.70	4.70	1.40	0.05	1.92	0.04	1.36	0.03	1.18	0.02
21 Cba		Fm. Sidi Aich	1311	12.29	0.64	10.02	6.69	1.39	0.05	1.93	0.05	1.37	0.04	1.15	0.03
22 Ch-ba		Fm. Bouhedma	7015	12.88	0.62	4.85	4.18	1.39	0.05	1.97	0.06	1.36	0.05	1.13	0.03
23 Ch		Fm. Boudinar	5894	13.28	0.62	3.43	2.64	1.43	0.06	1.99	0.05	1.38	0.06	1.13	0.03
24 Cb		Fm. Sidi Khalif	13370	12.38	0.67	8.13	4.79	1.48	0.06	2.01	0.07	1.40	0.07	1.14	0.02
25 Cb-v		Fm. Meloussi	44	12.79	0.34	3.16	0.82	1.45	0.03	1.98	0.03	1.42	0.04	1.14	0.01
26 T		Triassic general	982	12.81	0.54	3.48	2.52	1.33	0.06	1.85	0.08	1.34	0.04	1.21	0.09
		Average value	9836	13.00	0.57	4.38	3.06	1.36	0.06	1.89	0.07	1.39	0.06	1.17	0.04

Table 4 The model settings and the prediction probabilities of the models calculated for the study area

Model no.	Mtry	Ntree	Replace*	Prediction probability	St.dev.
1	26	100	True	0.4524	0.2010
2	26	250	True	0.4502	0.2006
3	26	500	True	0.4483	0.2008
4	26	750	True	0.4491	0.2012
5	26	100	False	0.4927	0.2117
6	26	250	False	0.4883	0.2114
7	26	500	False	0.4882	0.2111
8	26	750	False	0.4861	0.2107
9	5	100	True	0.5215	0.2183
10	5	250	True	0.5198	0.2177
11	5	500	True	0.5197	0.2179
12	5	750	True	0.5181	0.2182
13	5	100	False	0.8531	0.2026
14	5	250	False	0.8507	0.2056
15	5	500	False	0.8510	0.2054
16	5	750	False	0.8501	0.2067
17	2	100	True	0.5396	0.2186
18	2	250	True	0.5269	0.2147
19	2	500	True	0.5360	0.2153
20	2	750	True	0.5270	0.2154
21	2	100	False	0.8544	0.2042
22	2	250	False	0.8599	0.2064
23	2	500	False	0.8603	0.2042
24	2	750	False	0.8541	0.2024

*Sampling method of the in-bag samples at the mtry events with replacement

Values in bold indicate the best model based on the prediction probability

formations—cannot be implemented in the modelling process. For this reason, classified maps can often contain pixels adjacent to each other whose categories are stratigraphically distant from each other in the layer sequence. Because of this, the manual identification of lithological boundaries after the RF classification is usually done in order to create a geological interpretation of the area (Radford et al. 2018). In our case, the geological map was already compiled, so this step was not necessary.

The variables

The predictor variables plays important role in the accuracy of the classification. Selecting the proper variable cannot depend only on the selection of the strongest predictor variable, but also on the results of the user and producer accuracy tests. In our case, altitude initially seemed to be a very strong

predictor variable, and its strength caused that it played a decisive role in the classification even when it should not have done so. Because of this, each of the classes was “pulled apart” on both sides of the accuracy plot diagonal towards the user and producer accuracy by the model (see Fig. 6B and D). This meant that although a large proportion of polygons in a given class was correctly classified by the model, in many other places, the model would output the class incorrectly as a result (e.g. tied to a particular height). After removing the elevation from the predictor variables, the accuracy test of the classes showed a higher balance.

The MERIT DEM excludes artificial artefacts and vegetation but the spatial resolution is poor (~90 m at the equator). The RS data was scaled down, and the MERIT was scaled up to 50×50-m spatial resolution. With better (~10 m) resolution, the accuracy of the RFC may have increased. The terrain indices of TRI and TWI were amongst the stronger predictor variables (Fig. 5). The successful use of TRI has been reported in other studies aimed at classifying geology using the random forest method (Bachri et al. 2020). However, while the standard deviation of TWI for the 26 geological categories was small, that of TRI was relatively large (Table 3). For this reason, TWI is considered a more reliable predictor variable than TRI.

Lithological indices suitable to detect iron-containing minerals were selected because these minerals have spectral properties covered by the Sentinel 2 sensors (Van der Werff and Van der Meer 2016). When combined with the terrain indices, the spectral features generally worked well to correctly classify the pixels in the higher iron-mineral content formations, but in some cases, they produced false results. The Qp category (Quaternary slope debris with soil), for example has similar spectral parameters to the source rock of the rock debris (e.g. the carbonates and red clay beds of the Bouhedma Formation) and often located in a rough terrain, so the classification results sometimes incorrectly predicted Bouhedma Fm for the pixels in these areas. This is reflected in the relatively poor user accuracy of this category (Table 6). However, one of the well-predicted formations was the Cretaceous Sidi Khalif Formation composed of green argillites and carbonates according to the lithological descriptions of the geological map (Mahjoub 2005). It has the highest mean FeO and Fe₂+ value (Table 3) referring to a high iron content.

The Fe₂+ (ferrous iron) correlates positively with FeO (ferric oxide) and negatively with alterations/laterite. This type of iron is also present in green plants, though plants can also absorb ferric iron (Fe₃+), but less effectively (Olsen and Miller 1986). The vegetation-covered areas were thus more pronounced in the modelling and were sometimes falsely predicted to be one of the iron-bearing formations. This type of error can be mostly ruled out with the use of the NDVI classification map masking those areas where

Table 5 Cross-correlation of the 6 predictor variables examined on the pixel level, $N=1000$

	TWI	TRI	FeO	Fe2 +	Laterite	Fe3 +
TWI	1	−0.86	−0.38	−0.02	−0.36	0.113
TRI		1	0.339	0.048	0.274	−0.07
FeO			1	0.789	−0.17	0.105
Fe2 +				1	−0.73	−0.37
Laterite					1	−0.37
Fe3 +						1

Fig. 7 The anticline of Jebel el Barda el Hamra on the manually compiled map (left) and on the automatically generated one (right). Coordinates are in UTM 32 N (EPSG: 32632) **Fig. 8** The RFC method resulted in maps, which have very similar appearance as the original geological map

Table 6 List of the formations that were predicted with better than 50% producer accuracy and the result of the accuracy tests (ACC)

Categories		Polygon-wise producer ACC [polygons]	Pixel-wise producer ACC	Pixel-wise user ACC	Overall ACC
dQ	Sand dunes	71% [14]	63%	9%	36%
sQ	Sebhkas	100% [1]	53%	17%	35%
E2	Fm. Jebes	17% [18]	62%	60%	61%
Ch-ba	Fm. Bouhedma	40% [53]	53%	27%	40%
Cb	Fm. Sidi Khalif	50% [2]	60%	74%	67%
Total avg		27% [500]	30%	17%	24%

vegetation is present. The correlation results suggest that ferrous iron (Fe^{2+}) can be substituted with the FeO and vice versa, and omitting one of them as a predictor variable would decrease the chance of false predictions. Examining the characteristic values for the 26 geological classes, FeO and Fe^{2+} show a very similar pattern (Table 3) when looking at the minimum and maximum values and the standard deviation values for the classes. This also confirms that the two predictor variables are interchangeable.

The results

Industrial uses of the raw materials in this region are also reported and are often associated with the formations that are referred to in the present study as predictable classes. For example, the Bouhedma formation contains red clays which are reported as a raw material for the ceramics industry (Boussen et al. 2016). This implies that the method has potential in mineral exploration as well.

The geological map of Jebel Meloussi was compiled by a team of geologists, who surveyed the area to produce a generalized map of the surface geology at a scale of 1:50,000. This generalization inevitably leads to a loss of some information regarding the lithological types of the surface. Small details or geologically less important geological formations are suppressed, while structural features and characteristic formations are emphasized. For example, in the area of the Barda el Hamra hill, in the anticlinal axis, there are tributary valleys containing young valley-filling sediments, which are not shown on the geological map (Fig. 7A), but are indicated in the classification results (Fig. 7B). The detected error in the categorization of the RFC is partly the result of this, but its exact extent is unknown.

For category Qd (dunes), it is important to note that several years have passed between the time the map was produced and the time the satellite image was processed, and the boundary of the dune area may have changed during this time despite the intention to stop sand migration by plantations.

Conclusions

This study aimed to test the capabilities of RFC in geological mapping using only freely accessible remotely sensed data. By documenting the process in detail, we provide new

insights into the application of machine learning in geological mapping.

We showed the usefulness of focusing on iron-bearing minerals and combining their spectral indices with morphometric indicators when classifying them. Other researchers also report successful application of morphometric derivatives (including the TRI) for RFC as predictor variables for geology (Bachri et al. 2020), but the TWI as predictor variable was not demonstrated previously according to our knowledge. We theorize that using higher resolution DEM (e.g. 20-m spatial resolution), and other morphometric indices (e.g. the slope steepness), the accuracy of the model can be increased.

We also demonstrated that using only the freely accessible Sentinel 2 satellite images and DEM, a good prediction of the geological pattern can be made with RFC in the sparsely vegetated and uninhabited regions. Based on the results of the accuracy tests, the reconstruction of the Jebel Meloussi sheet was rather successful in the case of the iron-bearing formations, sebkhas and dunes, but balanced chanced or unsuccessful in the case of other lithological types. However, the results show that the pattern of the geological formations well reflects reality, despite the fact that the numerical results do not support this for many formations. In our opinion, this may be explained to a considerable extent by the fact that the map used for validation is a generalized representation of the surface formations.

The model used for the prediction was justified by the results of the statistical analysis (mtry accuracy was 0.62 and the Kappa 0.58) referring to a good-to-moderate agreement. Based on modelling, we identified geological and morphometric indicators that could potentially be useful predictors in the classifications. The TRI (topographic ruggedness index), the TWI (topographic wetness index), the Fe^{3+} (ferric iron) and the laterite seem to be powerful predictors when combined but individually the two topographic indices are strong, while the two lithological were intermediate predictors. The TWI, however, is considered a better predictor than the TRI because of its smaller standard deviation within geological categories. Concerning the two other lithological predictor variables, Fe^{2+} (ferrous iron) was found to be correlated positively with FeO (ferric oxides). Though the area was sparsely vegetated, the false predictions of iron-bearing formations concentrated

in the vegetated areas. Since ferrous iron is present in green plants, it is more useful for detecting the vegetation rather than lithology. We suggest omitting ferrous iron as a predictor variable for geological mapping purposes and using FeO instead. Amongst the Sentinel 2 spectral bands, the B8A and the B11 were intermediate predictors.

We conclude that using RFC and the already mapped information some geological formations can be predicted with good accuracy. The best-predicted formations were the Eocene Jebb Formation and the Cretaceous Sidi Khalif Formation with 61% and 67% overall accuracy, respectively. These formations can be predicted in other adjacent areas using the Meloussi sheet no. 103 as a training area. Other formations in the concerning area were generally below 50% overall accuracy and with the here presented predictor variables and data parameters, the adjacent areas cannot be classified based on sheet no. 103.

Though the RFC is not appropriate to predict geological structures, such as the sequence of stratigraphic units outcropping in proper order, it can be used for preliminary analysis prior to doing fieldwork. A manual identification of geological boundaries should be done after the classification. Applying this method in the process of geological map production, we assume that the compilation of the 1:50,000 and 1:100,000 geological map series of Tunisia can be quickened.

Acknowledgements The study was partly supported by the National Research Development and Innovation (NRDI) Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme.

Funding Open access funding provided by Eötvös Loránd University. The study was partly supported by the National Research Development and Innovation (NRDI) Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme.

Data availability The study used freely accessible remotely sensed data.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aquino DDN, Rocha OCD, Moreira MA, Teixeira ADS, Andrade EMD (2018) Use of remote sensing to identify areas at risk of degradation in the semi-arid region. *Rev Ciênc Agron* 49(3):420–429. <https://doi.org/10.5935/1806-6690.20180047>
- Bachri I, Hakdaoui M, Raji M, Benbouziane A (2020) Geological mapping using random forests applied to remote sensing data: a demonstration study from Msaidira-Souk Al Had, Sidi Ifni inlier (Western Anti-Atlas, Morocco). 2020 IEEE International conference of Moroccan Geomatics (Morgeo), 11–13 May 2020, <https://doi.org/10.1109/Morgeo49228.2020.9121888>
- Bedini E (2009) Mapping lithology of the Sarfartoq carbonatite complex, southern West Greenland, using HyMap imaging spectrometer data. *Remote Sens Environ* 113:1208–1219. <https://doi.org/10.1016/j.rse.2009.02.007>
- Bouri S, Gasmi M, Jaouadi M, Souissi I, Mimi AL, Dhia HB (2007) Etude intégrée des données de surface et de subsurface pour la prospection des bassins hydrogéothermiques: cas du bassin de Maknassy (Tunisie centrale)/Integrated study of surface and subsurface data for prospecting hydrogeothermal basins: case of the Maknassy basin (central Tunisia). *Hydrolog Sci J* 52(6):1298–1315. <https://doi.org/10.1623/hysj.52.6.1298>
- Boussen S, Sghaier D, Chaabani F, Jamoussi B, Bennour A (2016) Characteristics and industrial application of the Lower Cretaceous clay deposits (Bouhedma Formation), Southeast Tunisia: potential use for the manufacturing of ceramic tiles and bricks. *Appl Clay Sci* 123:210–221. <https://doi.org/10.1016/j.clay.2016.01.027>
- Boutib L, Zargouni F (1998) Disposition et géométrie des plis de l'Atlas centro-méridional de Tunisie: découpage et cisaillement en lanières tectoniques. *Comptes Rendus De L'académie Des Sciences - Series IIA - Earth and Planetary Science* 326(4):261–265
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Burollet PF (1956) Contribution à l'étude stratigraphique de la Tunisie centrale. *Ann Mines Géol Tunisie* 18:195–203
- Burollet PF (1991) Structures and tectonics of Tunisia. *Tectonophysics* 195(2–4):359–369. [https://doi.org/10.1016/0040-1951\(91\)90221-D](https://doi.org/10.1016/0040-1951(91)90221-D)
- Carneiro CDC, Fraser SJ, Crósta AP, Silva AM, Barros CEDM (2012) Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. *Geophysics* 77(4):17–24. <https://doi.org/10.1190/geo2011-0302.1>
- Cracknell MJ, Reading AM (2014) Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput Geosci* 63:22–33. <https://doi.org/10.1016/j.cageo.2013.10.008>
- Drury F (2001) *Image Interpretation in Geology*. 3rd edn. Nelson Thornes, Cheltenham, UK
- EU-EOP (2021) Copernicus in Brief. The European Union Earth Observation programme, <https://www.copernicus.eu/>. Accessed 20 March 2021
- Grebby S, Naden J, Cunningham D, Tansey K (2011) Integrating airborne multispectral imagery and airborne LiDAR data for enhanced lithological mapping in vegetated terrain. *Remote Sens Environ* 115(1):214–226. <https://doi.org/10.1016/j.rse.2010.08.019>
- He J, Harris JR, Sawada M, Behnia P (2015) A comparison of classification algorithms using Landsat-7 and Landsat-8 data for mapping lithology in Canada's Arctic. *Int J Remote Sens* 36(8):2252–2276. <https://doi.org/10.1080/01431161.2015.1035410>
- Hirt C, Filmer MS, Featherstone WE (2010) Comparison and validation of the recent freely available ASTER-GDEM ver1, SRTM ver4. 1 and GEODATA DEM-9S ver3 digital elevation models

- over Australia. *Aust J Earth Sci* 57(3):337–347. <https://doi.org/10.1080/08120091003677553>
- Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. In: Maglogiannis I et al (eds) *Emerging artificial intelligence applications in computer engineering*. Ios Press, Amsterdam, pp 3–24
- Kruse FA, Boardman JW, Huntington JF, Mason P, Quigley MA (2002) Evaluation and validation of EO-1 Hyperion for geologic mapping. *IEEE International Geoscience and Remote Sensing Symposium*, 24–28 June 2002, Toronto, Canada, 1:593–595. <https://doi.org/10.1109/IGARSS.2002.1025115>
- Li C (2013) Probability estimation in random forests. All Graduate Plan B and other Reports, Utah State University. <https://digitalcommons.usu.edu/gradreports/312>. Accessed 02 August 2021
- Lorenz RD, Gasmi N, Radebaugh J, Barnes JW, Ori GG (2013) Dunes on planet Tatooine: Observation of barchan migration at the Star Wars film set in Tunisia. *Geomorphology* 201:264–271. <https://doi.org/10.1016/j.geomorph.2013.06.026>
- Mahjoub K (2005) *Jebel Meloussi – Carte Géologique Tunisia 1:50 000*. No 103, Service Géologique de Tunisie Office Nationale des Mines, Tunis
- Main-Knorn M, Pflug B, Louis J, Debaecker V, Müller-Wilm U, Gascon F (2017) Sen2Cor for Sentinel-2. In Bruzzone L (ed) *Image and signal processing for remote sensing XXIII*. International Society for Optics and Photonics, p 1042704. <https://doi.org/10.1117/12.2278218>
- Mitchell MW (2011) Bias of the random forest out-of-bag (OOB) error for certain input parameters. *Open J Stat* 1(03):205. <https://doi.org/10.4236/ojs.2011.13024>
- Olsen RA, Miller RO (1986) Absorption of ferric iron by plants. *J Plant Nutr* 9(3–7):751–757
- ONM (2021) Basic geological cartography. National Office of Mines, <http://www.onm.nat.tn/en>. Accessed 21 March 2021
- Radford DD, Cracknell MJ, Roach MJ, Cumming GV (2018) Geological mapping in western Tasmania using radar and random forests. *IEEE J Sel Top App* 11(9):3075–3087. <https://doi.org/10.1109/JSTARS.2018.2855207>
- Rouse JW, Haas RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publication* 351:309
- Trabelsi H, Gasmi M, Kadri A, Jaouadi M, Balti H, Hachani F (2006) Contribution de la géophysique (Sismique Réflexion) à la caractérisation d'une tectonique compressive aptienne en Tunisie Centrale (Plaine de Maknassy). 3. Colloque Maghreb de Géophysique Appliquée, 11 May 2006, Oujda, Morocco pp 42–46
- Van der Werff H, Van der Meer F (2016) Sentinel-2A MSI and Landsat 8 OLI provide data continuity for geological remote sensing. *Remote Sens-Basel* 8(11):883. <https://doi.org/10.3390/rs8110883>
- Yamazaki D, Ikeshima D, Tawatari R, Yamaguchi T, O'Loughlin F, Neal JC, Sampson CC, Kanae S, Bates PD (2017) A high-accuracy map of global terrain elevations. *Geophys Res Lett* 44(11):5844–5853. <https://doi.org/10.1002/2017GL072874>
- Zaïer A, Beji-Sassi A, Sassi S, Moody RTJ (1998) Basin evolution and deposition during the Early Paleogene in Tunisia. *Geological Society, London, Special Publications*, 132(1):375–393. <https://doi.org/10.1144/GSL.SP.1998.132.01.21>