Journal of Mining and Environment (JME) Vol. 12, No. 4, 2021, 987-1001

Shahrood University of Technology

Iranian Society of Mining Engineering (IRSME)

Journal of Mining and Environment (JME)

journal homepage: www.jme.shahroodut.ac.ir

THOMSON REUTERS

A Comparative Study on Machine Learning Algorithms for Geochemical Prediction Using Sentinel-2 Reflectance Spectroscopy

Muhammad Ahsan Mahboob[1,2], Turgay Celik[3,4] and Bekir Genc[1]

1- School of Mining Engineering, University of the Witwatersrand, Johannesburg, South Africa
2- Sibanye-Stillwater Digital Mining Laboratory (DigiMine), Wits Mining Institute (WMI), University of the Witwatersrand, Johannesburg, South Africa
3- School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa
4- Wits Institute of Data Science, University of the Witwatersrand, Johannesburg, South Africa

Article Info

Received 29 July 2021
Received in Revised form 27 October 2021
Accepted 16 November 2021
Published online 16 November 2021

DOI: 10.22044/jme.2021.11041.2080

Keywords

Ore potential
Machine Learning
Geochemical Stream Sedimentation
Remote Sensing
Satellite Spectral Reflectance

Abstract

The distribution of stream sediments is usually considered as an important and very useful tool for the early-stage exploration of mineralization at the regional scale. The collection of stream samples is not only time-consuming but also very costly. However, the advancements in space remote sensing has made it a suitable alternative for mapping of the geochemical elements using satellite spectral reflectance. In this research work, 407 surface stream sediment samples of the zinc (Zn) and lead (Pb) elements are collected from Central Wales. Five machine learning models, namely the Support Vector Regression (SVR), Generalized Linear Model (GLM), Deep Neural Network (DNN), Decision Tree (DT), and Random Forest (RF) regression, are applied for prediction of the Zn and Pb concentrations using the Sentinel-2 satellite multi-spectral images. The results obtained based on the 10 m spatial resolution show that Zn is best predicted with RF with significant R2 values of 0.74 ($p < 0.01$) and 0.7 ($p < 0.01$) during training and testing. However, for Pb, the best prediction is made by SVR with significant R2 values of 0.72 ($p < 0.01$) and 0.64 ($p < 0.01$) for training and testing, respectively. Overall, the performance of SVR and RF outperforms the other machine learning models with the highest testing R2 values.

1. Introduction

The identification and mapping of potential mineralization through the geochemical exploration of stream sediments has been successfully done over the past several decades [1, 2]. These geochemical observations at the surface of earth usually represent the diverse effects of the primary and secondary geological processes inside the earth [3]. The field-based sampling of stream sedimentation is not only time-consuming but also expensive; usually the samples are only collected from the downstream areas with less slope and easy accessibility. However, reflectance spectroscopy of satellite-based remote sensing data provides a unique edge to map the geochemically enriched areas using stream sedimentation not only at a larger scale but is also time- and cost-effective [4-6]. In the research work conducted by Martinez et al. [7], the MODIS satellite data was used in order to quantify the Amazon River sedimentation. It was concluded that sedimentation was assessed through satellite, and the field datasets showed a very good agreement with a mean difference lower than 1%. Another research work conducted by Abedi and Norouzi [8] tested the ASTER and LANDSAT satellite datasets along with the geochemical and geological data for mineral exploration using the TOPSIS method. The research work concluded that the proposed methodology based on satellite and field data was satisfactory for mapping of the porphyry copper

Corresponding author: Mahsan.Mahboob@wits.ac.za (M.A. Mahboob).

deposit. The research work conducted by Afzal et al. [9] also used the ASTER satellite data for exploration and mapping of Cu mineralization in Iran using the stream sedimentation data. A good correlation was found between the satellite reflectance data and field-based sedimentation values, and concluded that the satellite multispectral data could be used for mineral exploration using stream estimations. However, one of the key components in using the satellite data is the spatial auto-correlation of satellite data and the field measurements [10], a phenomenon usually used to map the features in a region based on the systematic spatial similarities between the satellite reflectance and laboratory measurements. Usually the spatial interpolation models like inverse distance weighted and kriging are used in order to predict the presence or absence of a geochemical value of a variable at the non-surveyed regions using the surveyed datasets. However, these interpolation models usually incorporate the estimation errors such as the smoothing effects in the prediction of geochemical variables [11]. Besides the smoothing effects, the complexity, non-linearity, and spatial variability in the geochemical datasets makes it challenging to predict the values at unknown locations based on the known location datasets.

The advancements in machine learning (ML) have made it a strong alternative choice to overcome the classical interpolation challenges and to develop the predictive models by analysing and learning from the data by incorporating the spatial autocorrelation. Several researchers [12-14] have applied ML as a tool for modelling the geospatial phenomenon by building the models capable of identifying the patterns in geospatial data and predicting from these models. The applications of ML in geochemical modelling are very limited, and there are very few publications on this topic. However, the ML-based predictive models are potentially powerful for geochemical mapping and mineral explorations [15-18]. This research work aims to construct and compare the ML models for prediction of the Pb and Zn concentrations associated with the potential mineralization and mining activities at Central Wales (in the United Kingdom) using the satellite-based remotely-sensed spectral reflectance data.

2. Materials and methods

2.1. Geology setting

The studied area for this research work is in Central Wales of Great Britain (Fig. 1). It has a prolonged history of gold mining, with the main centre of activity being the "Dolgellau Gold Belt", where the Clogau and Gwynfynydd mines are very active gold mines, particularly towards the South, in the Welsh Basin, where gold has been mined since the Roman times. Besides gold, several other secondary products such as lead, copper, zinc, iron, and nickel sulphides are present throughout the Wales area. The general geology of Central Wales is consisted of Ordovician and Silurian marine sedimentary rocks. The rocks settled at the bottom of the Welsh Basin are dominated by series of sandstone, siltstone, and mudstone. Many of these classifications are referred to as turbidites because they were settled from violent, sediment-laden undersea flows, which flooded off the shallower shelf areas onto the deep floor of the basin [19]. There has been extensive mining of Pb and Zn as Ball and TK; [20] has reported that the area is enriched with these two minerals, and they are most commonly found within the interbedded sequences of sandstone, siltstone, and mudstone.

3. Machine learning models

Machine learning models have been applied, tested, and proved in several real-world applications. The main objective of this research work is to predict Pb and Zn in Central Wales using satellite spectral reflectance data through following machine learning models and their comparative evaluation in terms of predictions. Regardless of the machine learning model type applied, the data was split into the training and testing purposes at 70% and 30%, respectively [22].

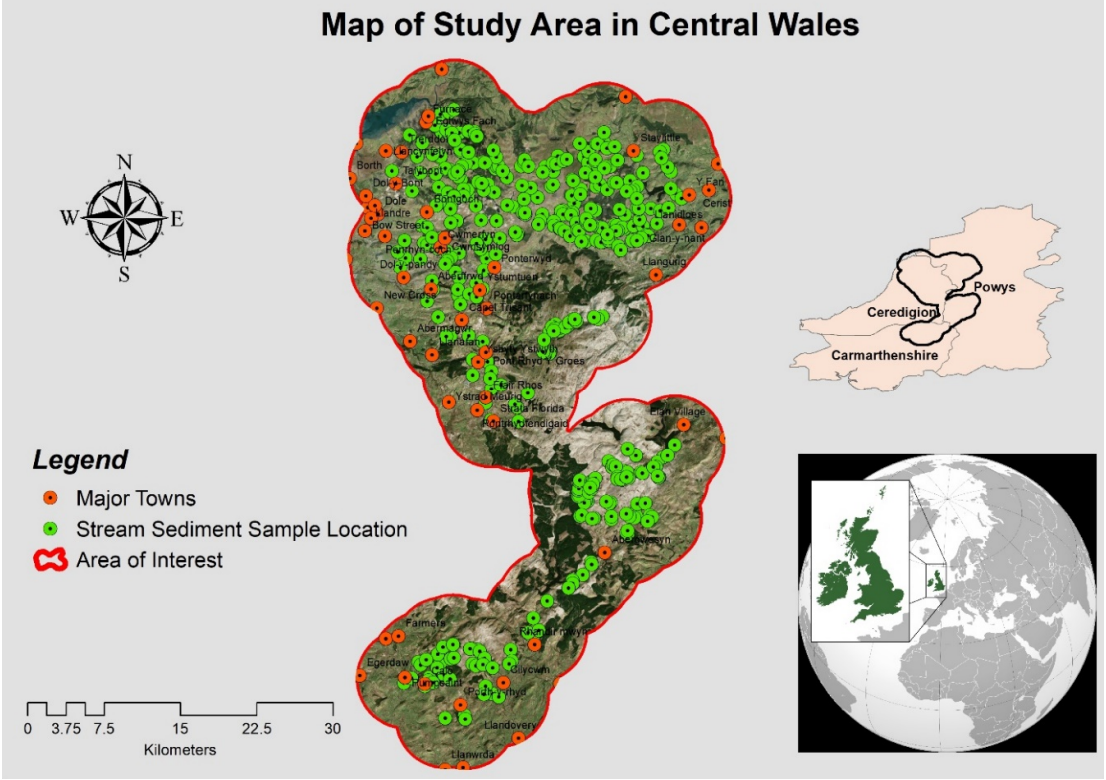


Figure 1. Geographical map of the studied area highlighting the major towns and locations of the collected samples [21]

3.1. Support Vector Regression (SVR)

SVR was initially introduced by Drucker et al. [23], and has been successfully applied for hidden pattern identifications in the data due to its strong predictive capability, flexibility, and robustness [24]. SVR is more suitable and applicable to the classification problems; however the SVR concepts can be generalized to become applicable to the regression problems. Several researchers have applied SVR for urban land cover mapping [25], hotspot detection [26], predication of organic carbon content in soils [27], landslide modelling and mapping [28], and in mining applications [29, 30]. In the SVR algorithm, all the data is plotted as points in an n-dimensional space (where n is the total of features in the data), and then the classification is performed on the n-dimensional space with the main objective to find the hyper-plane that can differentiate the two feature classes distinctively.

For the regression problems, rather than finding a hyperplane that can mainly distinguish the training points, SVR uses an ε-insensitive loss function for the computation of the hyperplane so that the predicted response values of the training points have at least an ε deviation from their actual response values. The hyperplane along with ε will define an ε-insensitive band (decision boundary) for the regression. For a set of training data $\{(x_1, y_1), \ldots (x_i, y_i)\}$, where x_i is the input data and y_i is the target output, following Equation 1 can be considered for general SVR.

$$y = f(x) = w, x + b = w^T x + b \quad (1)$$

where w, x denotes the dot product of the input data x and the weight vector w. The function f can be calculated through an ε-insensitive band as flat as possible, which is usually mentioned as flatness in order to seek a small w. The approximation of f can be described as Equations 2 and 3.

$$\min_w \frac{1}{2} ||w||^2 \quad (2)$$

$$\text{subject to} \begin{cases} y_i - w^T x_i - b \leq \varepsilon \\ w^T x_i + b - y_i \leq \varepsilon \end{cases} \quad (3)$$

For the above equation, it can be seen that SVR is to perform a linear regression with an ε-insensitive loss function that can be linear or quadratic, as given in Equations 4 and 5, respectively:

for linear

$$L\left(y, f\left(x\right)\right) = \begin{cases} 0 & \text{if } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases} \quad (4)$$

for quadratic

$$L\left(y, f\left(x\right)\right) = \begin{cases} 0 & \text{if } |y - f(x)| \leq \varepsilon \\ (|y - f(x)| - \varepsilon)^2 & \text{otherwise} \end{cases} \quad (5)$$

3.2. Generalized Linear Model (GLM)

GLM is used to focus on the arbitrary distributions (non-normal) response of the linear regression models. The main aim of a GLM is not to model-dependent factor as a linear combination of independent factors but to model a function of dependent factors as a linear combination of dependent factors. Many research studies have applied GLM for predictive modelling, e.g. Youssef et al. [31] have applied different machine learning models including GLM for landslides prediction, concluding that GLM could be used efficiently for landslide susceptibility mapping with a reasonable accuracy of 76.9%. Another research work conducted by Miller and Franklin [32] has applied GLM along with other classification trees for an effective mapping of vegetation alliances. For a GLM model, the three main components are a family function f, link function k, and parameters required to train the model. The family function can be Gaussian, binomial, fractional-binomial, ordinal, quasi-binomial, multinomial, Poisson, gamma, Tweedie, and negative-binomial. In this research work, the Tweedie family function has been applied, which further consist of gamma, normal, poisson, and their combinations, also because it is highly recommended, if the dataset has the positive continuous and exact zero responses. The variation in the Tweedie function is directly proportional to the p^{th} power of the mean-variance var $(y_i) = \delta\varepsilon_i^p$, where δ is the distribution factor and p is the power of variation. The Tweedie distribution is the characterized by power of variation p, while δ is an unknown constant. The values of p will be defined as per Equation 6:

$$\begin{aligned} &p = 0: \text{for Normal case} \\ &p = 1: \text{for Poisson case} \\ &p \in (1,2): \text{for Compound Poisson,} \\ &\text{non} - \text{negetive with zeros case} \\ &p = 2: \text{for Gamma case} \\ &p = 3: \text{for Inverse} - \text{Gaussian case} \\ &p > 2: \text{for positive reals case} \end{aligned} \quad (6)$$

The following maximum likelihood equation (7) is used to fit the model:

$$\sum_{i=1}^{N} \log\left(\alpha(y_i, \delta)\right) + \begin{cases} \frac{1}{\delta}\left(y_i \log(\varepsilon_i) - \frac{\varepsilon_i^{2-p}}{2-p}\right), & p = 1 \\ \frac{1}{\delta}\left(y_i \frac{\varepsilon_i^{1-p}}{1-p} - \log(\varepsilon_i)\right), & p = 2 \\ \frac{1}{\delta}\left(y_i \frac{\varepsilon_i^{1-p}}{1-p} - \frac{\varepsilon_i^{2-p}}{2-p}\right), & p \neq 1,\ p \neq 2 \end{cases} \quad (7)$$

where the function $\alpha(y_i, \delta)$ is assessed using an unbounded sequence increase, and will not have a logical explanation. However, because δ is an unknown constant, $\sum_{i=1}^{N} \log\left(\alpha(y_i, \delta)\right)$ can be considered as a constant as well and can be ignored. Therefore, the final function to minimize with the penalty will be as per the following Equation 8:

$$\min_{\alpha,\alpha_0} \mu \left(\beta||\alpha||_1 + \frac{1}{2}(1-\beta)||\alpha||_2\right) \begin{cases} \left(y_i \log(\varepsilon_i) - \frac{\varepsilon_i^{2-p}}{2-p}\right), & p = 1 \\ \left(y_i \frac{\varepsilon_i^{1-p}}{1-p} - \log(\varepsilon_i)\right), & p = 2 \\ \left(y_i \frac{\varepsilon_i^{1-p}}{1-p} - \frac{\varepsilon_i^{2-p}}{2-p}\right), & p \neq 1,\ p \neq 2 \end{cases} \quad (8)$$

Usually, the following function (Equation 9) is always used in the GLM model for the Tweedie function:

$$g(\varepsilon) = \begin{cases} \varepsilon^{q} = \omega = X\alpha, q \neq 0 \\ \log(\varepsilon) = \omega = X\alpha, q = 0 \\ q = 1 - p \end{cases} \quad (9)$$

The link power q can also be set to some other values as per the value of p. The resultant deviation will be as per the following Equation 10:

$$D = 2 \times \begin{cases} \sum_{i=1}^{N} y_i \log\left(\frac{y_i}{\varepsilon_i}\right) - \frac{\left(y_i^{2-p} - \varepsilon_i^{2-p}\right)}{2-p}, & p = 1 \\ \sum_{i=1}^{N} \frac{y_i}{1-p}\left(y_i^{1-p} - \varepsilon_i^{1-p}\right) - \log\left(\frac{y_i}{\varepsilon_i}\right), & p = 2 \\ \sum_{i=1}^{N} \frac{y_i\left(y_i^{1-p} - \varepsilon_i^{1-p}\right)}{1-p} - \frac{\left(y_i^{2-p} - \varepsilon_i^{2-p}\right)}{2-p}, & p \neq 1,\ p \neq 2 \end{cases} \quad (10)$$

3.3. Deep Neural Network (DNN)

Multi-layer Feedforward Artificial Neural Network, also known as DNN, is used in deep learning, and is trained with stochastic gradient descent using back-propagation. DNN comprises multiple levels of non-linear processes like neural nets with many hidden layers, as shown in Figure 2, and is also suitable for tabular datasets like the geochemical samples.

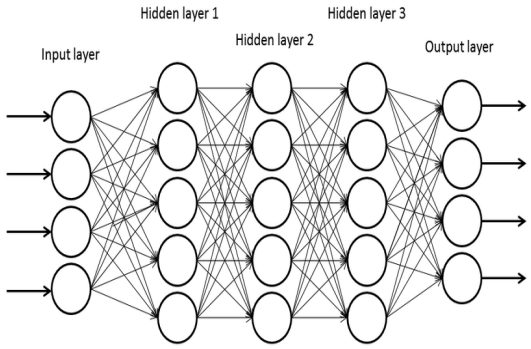


Figure 2. General architecture of deep neural network.

A classical DNN with a single hidden layer can be mathematically stated with these equations (11):

$$u_j = \sum_{i=1}^{N_{imp}} X_i W_{ij} + W_{0j}$$
$$H_j = f(u_j) \quad (11)$$
$$V_k = \sum_{i=1}^{N_{hid}} L_j m_{jk} + m_{0k}$$
$$O_k = g(V_k)$$

The results of the hidden layer (H_j) will be acquired by summing the products of the inputs (X_i) and the weight vectors W_{ij} in addition to a hidden layer's bias term W_{0j}, and then translating this sum using a transfer function f. The commonly used transfer functions are logistic and hyperbolic tangent [33]. Similarly, the outcome of the output layer O_k is obtained by summing the products of hidden layer's outputs L_j and weight vectors m_{jk} and output layer's bias term (m_{0k}), and converting the sum using the transfer function f.

3.4. Decision Tree (DT)

DTs or regression trees are non-parametric techniques that can explain the response of a dependent variable. The algorithm maps the whole dataset by representing different variables as internal nodes as the inputs and the leaf nodes as the outputs. Usually the decision trees do not require the data normalization and other data preparation requirements before its application, and can be applied to the data with outliers, and this algorithm is also known as a white-box algorithm so the behavior of the model along with the structure of predictions can be analyzed through the visual and instinctive interpretation of the results.

3.5. Random Forest (RF) Regression

The RF regression algorithm can be the improved form of the decision tree algorithm, and was first introduced by Breiman in 2001 [34]. In RF, several

small decision trees are generated from the random subsets of the dataset. Usually it categorizes each input vector into the random trees in order to build a forest and decide the output of each class based on the vote majority [35]. The final model is a voting model of all the generated random trees. Each random tree predicts each subset of the dataset by following the branches of the tree in line with the splitting rules and assessing the leaf. The importance of each effective factor can be estimated based on the mean decrease accuracy received at the end of the model. As all single predictions are taken with the same importance, and are based on the subsets of the dataset, the final prediction inclines to show a less variability than the single predictions. The concept of pruning can reduce the complexity of the model by replacing sub-trees that provide little predictive power only with leaves.

4. Preparation of input datasets

The dataset used in this work was obtained from the British Geological Survey (BGS). The stream sediment baseline geochemistry data was estimated from the samples collected across the central Wales region during the late 90s by BGS in Figure 1. The concentration values of the stream sediments are usually the representation of the parent geological material in the region and is developed over millions of years. Sampling was based on the collection of heavy minerals accumulated from the 1st and 2nd order streams. An active stream sediment was moved through a 2 mm sieve, collected in a wooden pan (about 3-4 kg), and condensed by panning to about 60 g. This process was repeated using an additional sediment from the same site, and the two concentrates combined were inspected on-site for heavy minerals and collected in a Kraft bag. A total of 407 samples were collected and analyzed each for Zn and Pb. The geochemical sample points were split into the training and the testing group with 80% (325 samples) and 20% (82 samples), respectively. The criterion to split the data was that the sample points located in the homogenous geological group and on the same streams should be selected and separated from the training data as the test data (Fig. 3).

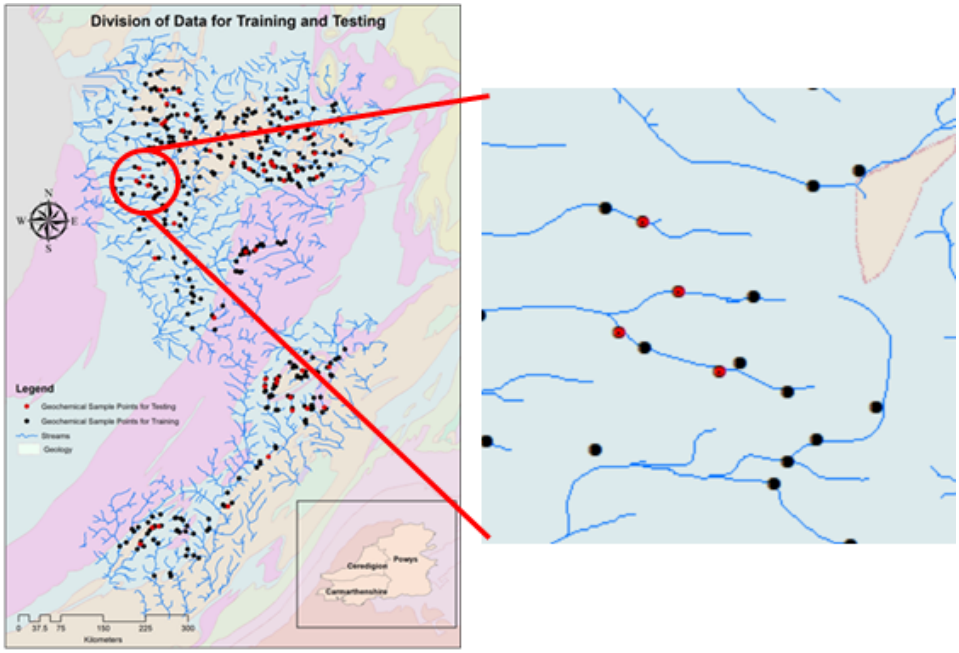


Figure 3. Splitting of geochemical sample points into train (black dots) and test (red dots) group (N = 407).

Prediction of Pb and Zn in the stream sediments was made using the five machine learning algorithms at stream levels using the MultiSpectral Instrument (MSI) Sentinel-2 satellite data of 2015. Although there is a time difference between the filed sampling and the satellite data, it will take a very long for the parent material to be changed. The time required for developing a geological profile is usually altered by the significant tectonic activities (earthquakes, tsunami, landslides), earth's inner temperature, gravity, climate, and weathering. Hence, the time difference between the collection of samples and the satellite data cannot be significant compared to the geological timelines, and therefore, the spectral profiles of the region will merely be changed. The Sentinel-2 satellite

has thirteen spectral bands, which range from the Visible and Near Infrared (NIR) to the Short-Wave Infrared (SWIR) of the electromagnetic spectrum. The spectral and spatial profile of Sentinal-2 satellite is given in Table 1.

Sentinel-2 has a 12-bit radiometric resolution, which means that the sensor can differentiate 4096 grey-levels in each spectral band. Due to its high spatial, spectral, and radiometric resolutions, the Sentinel-2 data is usually called the super-spectral satellite data. Many research works have applied, tested, and validated the Sentinel-2 dataset for several applications, i.e. geology [36], hydrology and hydrogeology [37], mining [38], and mineral exploration [39]. Similarly, the dataset has also been successfully used for stream sedimentation mapping [40] and geochemical mapping [41]. Usually the stream sediments have a huge capacity to hold the traces of different elements along with their chemical compositions that can be a good indicator of related mineralization. The Sentinel-2 satellite data was divided into four input subsets (Table 2) based on the spatial resolution of each band; however, the topographic elevation extracted from the ASTER Global Digital Elevation Model (GDEM) and the slope was also considered a part of each subset.

Table 1. Data characteristics of Sentinel-2 satellite.

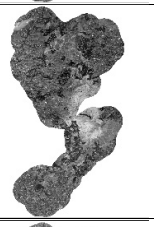
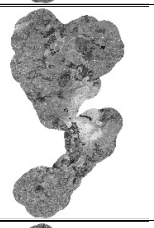
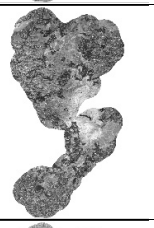
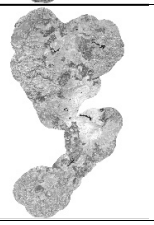
Band No	Spectral band	Central wavelength (nm)	Bandwidth (nm)	Spatial resolution	Satellite image of spectral band of the studied area
B1	Coastal aerosol	442.7	21	60	
B2	Blue	492.4	66	10	
B3	Green	559.8	36	10	
B4	Red	664.6	31	10	
B5	Red edge–1	704.1	15	20	

Table 1. Continuous of Table 1.

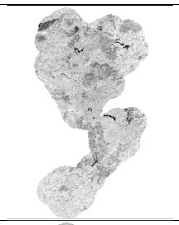
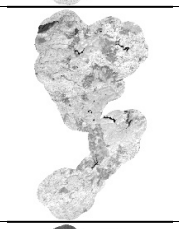
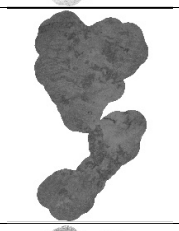
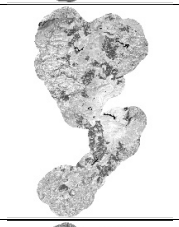
B6	Red edge–2	740.5	15	20	
B7	Red edge–2	782.8	20	20	
B8	NIR	832.8	106	10	
B8A	Narrow NIR	864.7	21	20	
B9	Water vapour	945.1	20	60	
B10	SWIR–Cirrus	1373.5	31	60	
B11	SWIR–1	1613.7	91	20	
B12	SWIR–2	2202.4	175	20	

Table 2. Input datasets composed of Sentinel-2 satellite and topographic variables.

Input subset	Parameters	Properties
DS-1	B2, B3, B4, B8, elevation, and slope	All the spectral bands of Sentinel-2 satellite with 10 m spatial resolution along with topographic elevation and slope
DS-2	B5, B6, B7, B8A, B11, B12, elevation, and slope	All the spectral bands of Sentinel-2 satellite with 20 m spatial resolution along with topographic elevation and slope
DS-3	B1, B9, B10, elevation, and slope	All the spectral bands of Sentinel-2 satellite with 60 m spatial resolution along with topographic elevation and slope
DS-4	B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B10, B11, B12, elevation, and slope	All the spectral bands of Sentinel-2 satellite along with topographic elevation and slope

All the five machine learning models were trained and validated on the input subsets of DS-1 to DS-4 for prediction of the Zn and Pb stream sediments, as shown in summary Figure 4.

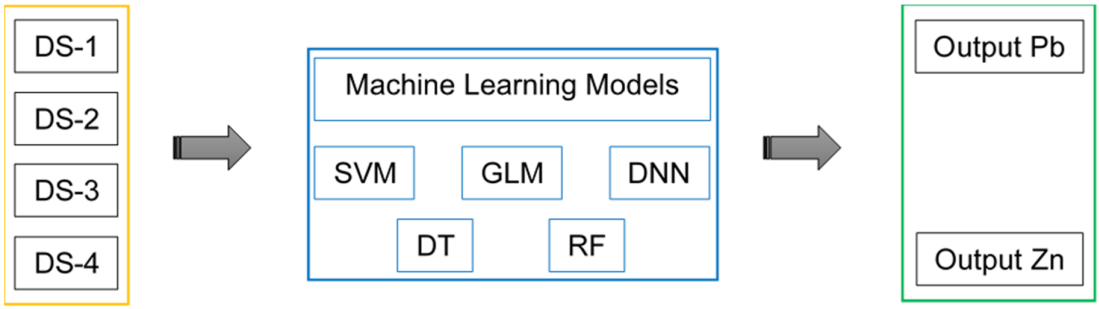


Figure 4. Summary of the four input sets and machine learning models used for predictions of Pb and Zn.

Performing the models were evaluated based on the coefficient of determination (R^2) as per Equation 12, which represents the total variation in the predictions made by the model as given in the following equation, and its value ranges between 0 (poor) and 1 (perfect) [42]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (12)$$

where y in the equation is an estimation of the average predictions, and N is the number of stream sample points.

5. Results and discussion

The results obtained on the prediction of Pb and Zn by applying through various machine learning applications are given in this section. The satellite image of the studied area was overlaid with stream sampling locations, and the value of each spectral band was extracted against each sample location and prepared in a separate database to be further used in the machine learning models.

The results based on DS-1 showed that Zn was best predicted with the RF model with the R^2 values of 0.74 and 0.7 during training and testing. However, for Pb, the best prediction was made by SVR with the R^2 values of 0.72 and 0.64 for training and testing, respectively. In RF, the model determined the highest and the lowest weight, and was assigned to the near-infrared (NIR) spectral band slope, respectively. The research work conducted by Cozzolino and Moron [43] also concluded that the Zn mineral ore could be separated with waste and other materials using the NIR reflectance spectroscopy. Pb also showed a good correlation with the NIR region along with SWIR due to its high reflectance in these two spectral bands, as also discussed by Hauff [44]. Using the DS-2 subset of the input dataset, Zn was best predicted through the SVR model with the R^2 values of 0.73 and 0.63 during training and testing. However, for Pb, the best prediction was made by DNN with the R^2 values of 0.72 and 0.61 for training and testing, respectively. DS-2 has eight input parameters with six Sentinel-2 satellite spectral bands including the NIR edge and SWIR.

These spectral bands are highly suitable for the mapping of Zn and Pb on the surface of the earth and in the outcrops regions [44-46]. The third input subset DS-3 showed that Zn and Pb both were best predicted through the SVR model with the R^2 values of 0.61 and 0.55 for Zn and 0.49 and 0.39 for Pb during training and testing, respectively. DS-3 has five input parameters with three Sentinel-2 satellite spectral bands and the R^2 value for DS-3 was less than the first two input subsets. This might be related to the spectral bands of DS-3, which are usually suitable for monitoring the atmospheric conditions and are less suitable for the mapping of features on the surface of the earth [47]. The other main reason could be the spatial resolution of the spectral bands, which is 60 m, and is coarser than DS-1 and DS-2. The fourth input subset DS-4 showed that Zn and Pb both were best predicted through the RF model with the R^2 values of 0.74 and 0.64 for Zn and 0.73 and 0.63 for Pb during training and testing, respectively. DS-4 was composed on 15 parameters including all the spectral bands of Sentinel-2 satellite ranged from visible, NIR, and SWIR along with topographic elevation and slope.

Overall, the performance of SVR and RF outperformed the other machine learning models (i.e. GLM, DNN, and DT) with the highest testing R^2 value. Similarly, the input subset DN-1 can be categorized as the most suitable for prediction of Zn and Pb consisting of visible and near-infrared spectral bands along with topographic elevation and slope of the studied area. The second-best input subset was DS-4, which consisted of all the spectral bands, the topographic elevation, and the slope of the studied area. The R^2 values of all the ML models were significant at $p < 0.01$, both for testing and training, as given in Table 3.

Table 3. Coefficients of determination as obtained for Zn and Pb prediction using machine learning models.

			SVR	GLM	DNN	DT	RF
DS-1	Zn	Training	0.71	0.52	0.73	0.67	0.74
		Testing	0.69	0.5	0.69	0.63	0.7
	Pb	Training	0.72	0.6	0.64	0.44	0.64
		Testing	0.64	0.51	0.55	0.35	0.54
DS-2	Zn	Training	0.73	0.7	0.72	0.56	0.7
		Testing	0.63	0.59	0.6	0.46	0.59
	Pb	Training	0.7	0.7	0.72	0.52	0.71
		Testing	0.58	0.59	0.61	0.42	0.62
DS-3	Zn	Training	0.61	0.59	0.6	0.41	0.57
		Testing	0.55	0.52	0.52	0.34	0.47
	Pb	Training	0.49	0.47	0.48	0.25	0.41
		Testing	0.39	0.35	0.36	0.16	0.28
DS-4	Zn	Training	0.56	0.54	0.72	0.72	0.74
		Testing	0.44	0.44	0.59	0.59	0.64
	Pb	Training	0.6	0.6	0.62	0.52	0.73
		Testing	0.48	0.49	0.52	0.4	0.63

The research work conducted by Perez et al. [48] and Abbaszadeh et al. [49] also concluded that SVR was better for mineral resource exploration and estimations, and Sheng et al. [50] concluded that RF was better for mineral resource exploration and estimations. The spatial distribution of the original and predicted Zn and Pb stream sediment contents (Fig. 5) shows significant degrees of association.

The predicted Zn and Pb geochemical surfaces are also under the bedrock permeability and geological fault line map of the region as well (Fig. 6). Permeability is the porosity of a rock; a higher permeability means a higher movement of liquid through the bed rock geology, and vice versa.

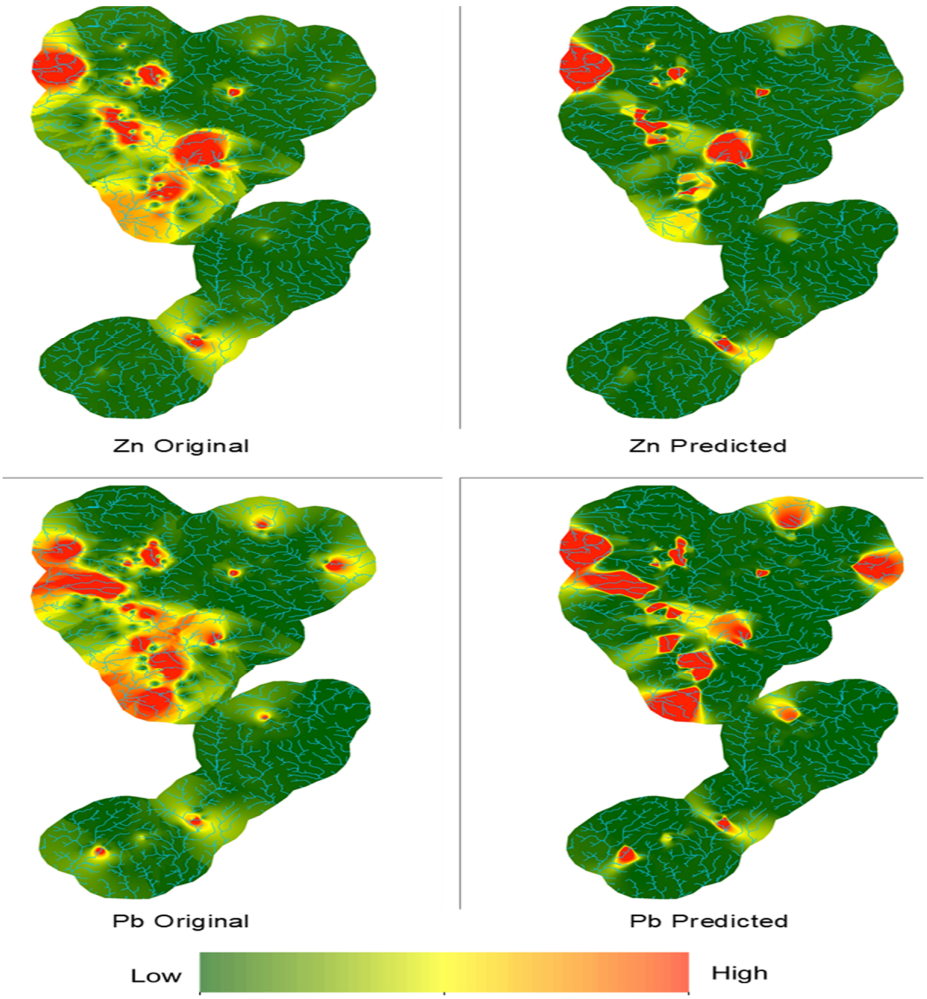


Figure 5. Spatial distribution of original and predicted Zn and Pb concentrations in the study area.

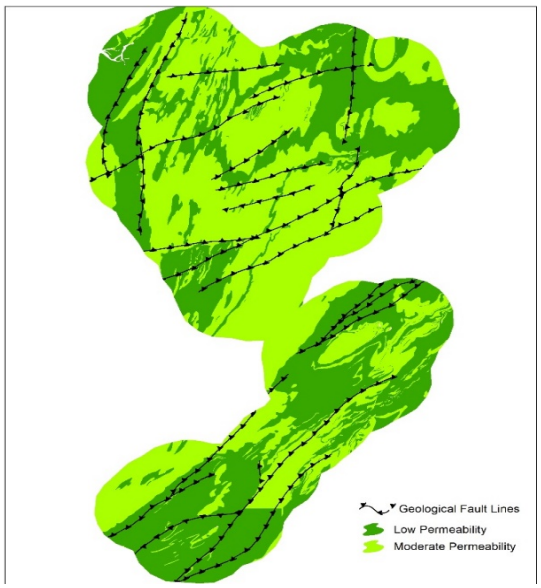


Figure 6. Bedrock permeability and geological fault line map of the Central Wales.

There are only two, i.e. low and moderate, zones of bedrock permeability present in the studied area, and high concentration pockets of Zn and Pb are present in the low zones. The research work conducted by Freedman [51], Bouabdellah and Sangster [52], and Gao et al. [53] also concluded that the higher concentrations of Zn and Pb were associated with the lower permeability of the bedrock. The other important observation is that the predicted high concentration values of Zn and Pb are in a close vicinity of the geological fault lines, which also indicate this mineralization in the studied area.

Although the values of the predicted stream sediment concentrations are higher than the original, overall, it is under the spatial location of the original dataset. The higher Zn and Pb content distribution can also be related to the ongoing mining activities of the same elements in the studied area. However, other areas in the central

Wales are enriched in these elements as well, and could be the future exploitation zones.

6. Conclusions

Five machine learning algorithms including support vector regression, generalized linear model, deep neural network, decision tree, and random forest regression were trained and tested for prediction of the Zn and Pb stream sediments using the reflectance spectroscopy of Sentinel-2 satellite data along with topographic elevation and slope. For a better prediction and in order to check the impacts of different spectral channels and spatial resolution of the data, the input datasets were divided into four subsets, and the accuracy was measured using the coefficient of determination (R^2). The prediction of stream sediments based on the DS-1 input dataset was the best with the SVR for Zn and RF for Pb as compared to the other models and input datasets with the measured R^2 values of 0.69 (for Zn) and 0.64 (for Pb). This could be related to the NIR band in the dataset in association with the high spatial resolution of 10 m. The input dataset DS-4 could be concluded as the second-best for the predictions of the Zn and Pb stream sediments using the RF machine learning model with the R^2 values of 0.64 (for Zn) and 0.63 (for Pb). DS-4 consisted of 15 input parameters including 13 Sentinel-2 spectral bands, the topographic elevation, and the slope of the area. The spatial locations of the predicted concentration values are not only under the original dataset but also with the existing mines present in the studied area. The proposed methodology for the prediction of the concentration of the stream sediments is useful for the mapping of the mineral enriched zones and future mineral resource exploration in the studied area. However, applying this methodology to the new studied area requires the new training and testing of the models so that the model can adapt to the new data for predictions of particular stream sediments. In the future research works, the field-based reflectance spectroscopy, topographic factor (aspect, surface curvature), and hydro-meteorological factors (streams flow, rainfall) can be added to the machine learning models along with the hyper-spectral satellite-based reflectance datasets for more accurate concentration predictions of the stream sediments.

Conflict of Interest

The authors ensure that there are no conflicts of interest.

Acknowledgements

The work presented here is part of a PhD research work in the School of Mining Engineering at the University of the Witwatersrand. The authors wish to acknowledge the administrative and financial support provided by the Sibanye-Stillwater Digital Mining Laboratory (DigiMine), Wits Mining Institute (WMI), University of the Witwatersrand, Johannesburg, South Africa.

References

[1]. Yousefi, M., Kamkar-Rouhani, A. and Carranza, E.J.M. (2012). Geochemical mineralization probability index (GMPI): a new approach to generate enhanced stream sediment geochemical evidential map for increasing probability of success in mineral potential mapping. Journal of Geochemical Exploration. 115: 24-35.

[2]. Lin, X., Hu, Y., Meng, G. and Zhang, M. (2020). Geochemical patterns of Cu, Au, Pb and Zn in stream sediments from Tongling of East China: Compositional and geostatistical insights. Journal of Geochemical Exploration, 210, 106457.

[3]. Kirkwood, C., Everett, P., Ferreira, A. and Lister, B. (2016). Stream sediment geochemistry as a tool for enhancing geological understanding: An overview of new data from south west England. Journal of Geochemical Exploration, 163, 28-40.

[4]. Choe, E., van der Meer, F., van Ruitenbeek, F., van der Werff, H., de Smeth, B. and Kim, K.W. (2008). Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: A case study of the Rodalquilar mining area, SE Spain. Remote Sensing of Environment. 112 (7): 3222-3233.

[5]. Cyples, N.N., Ielpi, A. and Dirszowsky, R.W. (2020). Planform and stratigraphic signature of proximal braided streams: remote-sensing and ground-penetrating-radar analysis of the Kicking Horse River, Canadian Rocky Mountains. Journal of Sedimentary Research, 90(1), 131-149.

[6]. Wang, Q., Li, F., Jiang, X., Wu, S. and Xu, M. (2020). On-stream mineral identification of tailing slurries of tungsten via NIR and XRF data fusion measurement techniques. Analytical Methods, 12(25), 3296-3307.

[7]. Martinez, J.M., Guyot, J.L., Filizola, N. and Sondag, F. (2009). Increase in suspended sediment discharge of the Amazon River assessed by monitoring network and satellite data. Catena, 79(3), 257-264.

[8]. Abedi, M. and Norouzi, G.H. (2016). A general framework of TOPSIS method for integration of airborne geophysics, satellite imagery, geochemical and geological data. International journal of applied earth observation and geoinformation. 46: 31-44.

[9]. Afzal, P., Asl, R.A., Adib, A. and Yasrebi, A.B. (2015). Application of fractal modelling for Cu mineralisation reconnaissance by ASTER multispectral and stream sediment data in Khoshname area, NW Iran. Journal of the Indian Society of Remote Sensing. 43 (1): 121-132.

[10]. Mondini, A.C. (2017). Measures of spatial autocorrelation changes in multitemporal SAR images for event landslides detection. Remote Sensing. 9 (6): 554.

[11]. Yousefi, M. (2017). Analysis of zoning pattern of geochemical indicators for targeting of porphyry-Cu mineralization: a pixel-based mapping approach. Natural Resources Research. 26 (4): 429-441.

[12]. Tehrany, M.S., Jones, S., Shabani, F., Martínez-Álvarez, F. and Bui, D.T. (2019). A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. Theoretical and Applied Climatology. 137 (1): 637-653.

[13]. Ahmed, N., Firoze, A. and Rahman, R.M. (2020). Machine learning for predicting landslide risk of Rohingya refugee camp infrastructure. Journal of Information and Telecommunication. 4 (2): 175-198.

[14]. Dornan, T., O'Sullivan, G., O'Riain, N., Stueeken, E. and Goodhue, R. (2020). The application of machine learning methods to aggregate geochemistry predicts quarry source location: an example from Ireland. Computers & Geosciences, 140, 104495.

[15]. Coimbra, R., Rodriguez-Galiano, V., Olóriz, F. and Chica-Olmo, M. (2014). Regression trees for modeling geochemical data An application to Late Jurassic carbonates (Ammonitico Rosso). Computers & Geosciences. 73: 198-207.

[16]. Zuo, R. and Xiong, Y. (2018). Big data analytics of identifying geochemical anomalies supported by machine learning methods. Natural Resources Research. 27 (1): 5-13.

[17]. Chen, Y. and Wu, W. (2017). Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. Geochemistry: Exploration, Environment, Analysis. 17 (3): 231-238.

[18]. Wang, Z., Zuo, R. and Dong, Y. (2019). Mapping geochemical anomalies through integrating random forest and metric learning methods. Natural Resources Research. 28 (4): 1285-1298.

[19]. Toghill, P. (2011). The geology of Britain: an introduction. Crowood.

[20]. Ball, T.K. and TK, B. (1976). PRELIMINARY MINERAL RECONNAISSANCE OF CENTRAL WALES.

[21]. Mahboob, M.A., Celik, T. and Genc, B. (2020). Predictive modeling and comparative evaluation of geostatistical models for geochemical exploration through stream sediments. Arabian Journal of Geosciences. 13 (20): 1-21.

[22]. Chen, W., Pourghasemi, H.R., Kornejady, A. and Zhang, N. (2017). Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. Geoderma. 305: 314-327.

[23]. Drucker, H., Burges, C.J., Kaufman, L., Smola, A. and Vapnik, V. (1997). Support vector regression machines. Advances in neural information processing systems. 9: 155-161.

[24]. Lim, E.P., Foo, S., Khoo, C., Chen, H., Fox, E., Shalini, U. and Thanos, C. (Eds.). (2002). Digital Libraries: People, Knowledge, and Technology: 5th International Conference on Asian Digital Libraries, ICADL 2002, Singapore, December 11-14, 2002, Proceedings (Vol. 2555). Springer Science & Business Media.

[25]. Okujeni, A., van der Linden, S., Tits, L., Somers, B. and Hostert, P. (2013). Support vector regression and synthetically mixed training data for quantifying urban land cover. Remote Sensing of Environment. 137: 184-197.

[26]. Pozdnoukhov, A. and Kanevski, M. (2007). Multi-scale support vector regression for hotspot detection and modeling.

[27]. Tan, M., Song, X., Yang, X. and Wu, Q. (2015). Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study. Journal of Natural Gas Science and Engineering. 26: 792-802.

[28]. Miao, F., Wu, Y., Xie, Y. and Li, Y. (2018). Prediction of landslide displacement with step-like behavior based on multialgorithm optimization and a support vector regression model. Landslides. 15 (3): 475-488.

[29]. Nourali, H. and Osanloo, M. (2019). Mining capital cost estimation using Support Vector Regression (SVR). Resources Policy. 62: 527-540.

[30]. X. Ding, M. Hasanipanah, H. N. Rad, and W. Zhou. (2020). "Predicting the blast-induced vibration velocity using a bagged support vector regression optimized with firefly algorithm," Engineering with Computers, pp. 1-12.

[31]. Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S. and Al-Katheeri, M.M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides. 13 (5): 839-856.

[32]. Miller, J. and Franklin, J. (2002). Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with

spatial dependence. Ecological Modelling. 157 (2-3): 227-247.

[33]. Hussain, F. and Jeong, J. (2015, March). Exploiting deep neural networks for digital image compression. In 2015 2nd world symposium on web applications and networking (WSWAN) (pp. 1-6). IEEE.

[34]. Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R. (2006). Random forests for land cover classification. Pattern recognition letters. 27 (4): 294-300.

[35]. K. Fawagreh, M.M. Gaber, and E. Elyan. (2014). "Random forests: from early developments to recent advancements," Systems Science Control Engineering: An Open Access Journal, Vol. 2, No. 1, pp. 602-609.

[36]. Van der Meer, F.D., Van der Werff, H. M.A. and Van Ruitenbeek, F.J.A. (2014). Potential of ESA's Sentinel-2 for geological applications. Remote sensing of environment. 148: 124-133.

[37]. M. Karaman, E. Özelkan, and S. Tasdelen. (2018) "Influence of basin hydrogeology in the detectability of narrow rivers by Sentinel2-A satellite images: A case study in Karamenderes (Çanakkale)," Journal of Natural Hazards Environment, Vol. 4, pp. 140-155.

[38]. Lobo, F.D.L., Souza-Filho, P.W.M., Novo, E.M.L.D.M., Carlos, F.M. and Barbosa, C.C.F. (2018). Mapping mining areas in the brazilian amazon using msi/sentinel-2 imagery (2017). Remote Sensing. 10 (8): 1178.

[39]. Mielke, C., Boesche, N.K., Rogass, C., Segl, K. and Kaufmann, H. (2014, June). Multi-and hyperspectral satellite sensors for mineral exploration, new applications to the sentinel-2 and enmap mission. In Proceedings of the 34th EARSeL Symposium, Poland, Warsaw (pp. 16-20).

[40]. Karim, M., Maanan, M., Maanan, M., Rhinane, H., Rueff, H. and Baidder, L. (2019). Assessment of water body change and sedimentation rate in Moulay Bousselham wetland, Morocco, using geospatial technologies. International journal of sediment research. 34 (1): 65-72.

[41]. Cardoso-Fernandes, J., Lima, A. and Teodoro, A.C. (2018). Potential of Sentinel-2 data in the detection of lithium (Li)-bearing pegmatites: a study case. In Earth resources and environmental remote sensing/GIS applications IX (Vol. 10790, p. 107900T). International Society for Optics and Photonics.

[42]. Piepho, H.P. (2019). A coefficient of determination (R2) for generalized linear mixed models. Biometrical Journal. 61 (4): 860-872.

[43]. Cozzolino, D. and Moron, A. (2004). Exploring the use of near infrared reflectance spectroscopy (NIRS) to predict trace minerals in legumes. Animal Feed Science and Technology. 111 (1-4): 161-173.

[44]. Hauff, P. (2008). An overview of VIS-NIR-SWIR field spectroscopy as applied to precious metals exploration. Spectral International Inc, 80001, 303-403.

[45]. Hunt, G. R. (1977). Spectral signatures of particulate minerals in the visible and near infrared. Geophysics. 42 (3): 501-513.

[46]. Chattoraj, S.L., Sharma, R.U., Kumar, C. and Sengar, V. (2020). Identification and characterization of hydrothermally altered minerals using surface and space-based reflectance spectroscopy, in parts of south-eastern Rajasthan, India. SN Applied Sciences. 2 (4): 1-9.

[47]. Vanhellemont, Q. (2019). Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the Landsat and Sentinel-2 archives. Remote Sensing of Environment. 225: 175-192.

[48]. Perez, C.A., Estévez, P.A., Vera, P.A., Castillo, L.E., Aravena, C.M., Schulz, D.A. and Medina, L.E. (2011). Ore grade estimation by feature selection and voting using boundary detection in digital image analysis. International Journal of Mineral Processing. 101 (1-4): 28-36.

[49]. Abbaszadeh, M., Hezarkhani, A. and Soltani-Mohammadi, S. (2013). An SVM-based machine learning method for the separation of alteration zones in Sungun porphyry copper deposit. Geochemistry. 73 (4): 545-554.

[50]. Sheng, L., Zhang, T., Niu, G., Wang, K., Tang, H., Duan, Y. and Li, H. (2015). Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF). Journal of Analytical Atomic Spectrometry. 30 (2): 453-458.

[51]. Freedman, J. (1972). Geochemical prospecting for zinc, lead, copper, and silver, Lancaster Valley, southeastern Pennsylvania (No. 1314). US Government Printing Office.

[52]. Bouabdellah, M. and Sangster, D.F. (2016). Geology, geochemistry, and current genetic models for major Mississippi valley-type Pb–Zn deposits of Morocco. In Mineral Deposits of North Africa (pp. 463-495). Springer, Cham.

[53]. Gao, R., Xue, C., Zhao, X., Chen, X., Li, Z. and Symons, D. (2019). Source and possible leaching process of ore metals in the Uragen sandstone-hosted Zn-Pb deposit, Xinjiang, China: Constraints from lead isotopes and rare earth elements geochemistry. Ore Geology Reviews. 106: 56-78.

مطالعه تطبیقی الگوریتم‌های یادگیری ماشین برای پیش‌بینی ژئوشیمیایی با استفاده از طیف‌سنجی بازتابی Sentinel-2

محمد احسن محبوب[۱،۲*]، تورگای چلیک[۳،۴] و بکیر گِنج[۱]

۱- دانشکده مهندسی معدن، دانشگاه ویت واترزراند، ژوهانسبورگ، آفریقای جنوبی

۲- آزمایشگاه معدن دیجیتال Sibanye-Stillwater (DigiMine)، موسسه معدنی Wits (WMI)، دانشگاه ویت واترزراند، ژوهانسبورگ، آفریقای جنوبی

۳- دانشکده مهندسی برق و اطلاعات، دانشگاه ویت واترزراند، ژوهانسبورگ، آفریقای جنوبی

۴- مؤسسه علوم داده Wits، دانشگاه ویت واترزراند ، ژوهانسبورگ، آفریقای جنوبی

ارسال ۲۰۲۱/۰۶/۲۹ پذیرش ۲۰۲۱/۱۱/۱۶

* نویسنده مسئول مکاتبات: Mahsan.Mahboob@wits.ac.za

چکیده:

توزیع رسوبات رودخانه معمولاً به عنوان یک ابزار مهم و بسیار مفید در مراحل اولیه برای اکتشاف کانی‌ها در مقیاس منطقه‌ای در نظر گرفته می‌شود. جمع‌آوری نمونه‌های رودخانه‌ای تنها زمان بر نبوده و بسیار پرهزینه است. با این حال، پیشرفت در روش سنجش از دور آن را به یک جایگزین مناسب برای نقشه برداری از عناصر ژئوشیمیایی با استفاده از بازتاب طیفی ماهواره‌ای تبدیل کرده است. در این کار تحقیقاتی، ۴۰۷ نمونه رسوب جریان سطحی عناصر روی (Zn) و سرب (Pb) از ولز مرکزی جمع‌آوری شد. پنج مدل یادگیری ماشین، یعنی رگرسیون بردار پشتیبان (SVR)، مدل خطی تعمیم یافته (GLM)، شبکه عصبی عمیق (DNN)، درخت تصمیم (DT) و رگرسیون جنگل تصادفی (RF) برای پیش‌بینی غلظت و تمرکز سرب و روی با استفاده از تصاویر چند طیفی ماهواره ای Sentinel-2 اعمال شد. نتایج به‌دست‌آمده بر اساس تفکیک فضایی ۱۰ متر نشان می‌دهد که روی با استفاده از روش RF با مقادیر R2 قابل قبول ۰/۷۴ ($p<0.01$) و ۰/۷ ($p<0.01$) در طول آموزش و آزمایش بهتر پیش‌بینی می‌شود. با این حال برای سرب، بهترین پیش‌بینی توسط روش SVR با مقادیر R2 قابل قبول ۰/۷۲ ($p<0.01$) و ۰/۶۴ ($p<0.01$) برای داده‌های آموزش و آزمایش انجام شد. به طور کلی، عملکرد روش‌های SVR و RF از سایر مدل‌های یادگیری ماشین با بالاترین مقادیر تست شده R2 بهتر است.

کلمات کلیدی: امکان‌سنجی ماده معدنی، یادگیری ماشین، رسوبات ژئوشیمیایی رودخانه‌ای، سنجش از دور، بازتاب طیفی ماهواره‌ای.