University of Tehran Press

Pollution

https://jpoll.ut.ac.ir/

Print ISSN: 2383-451X
Online ISSN: 2383-4501

Regularization Learning of Trace Element Contamination Stemmed from Tailings Dam-Break

Bulent Tutmez[1]✉ | **Osamu Komori**[2]

1. Department of Mining Engineering, Inonu University, Malatya, Türkiye
2. Department of Computer and Information Science, Seikei University, Tokyo, Japan.

Article Info

Article type:
Research Article

Article history:
Received: 12 Oct 2022
Revised: 29 Jan 2023
Accepted: 17 May 2023

Keywords:
Contamination
Tailings Dam Failure
Regularization
U-Lasso

ABSTRACT

An important practice in environmental risk management is assessing the consequences of heavy metal concentrations resulting from a mine dam tailing failure on soil, water, and trees. To appraise the extent of pollution, an effective classification is essential. In this study, trace element contamination is handled as a two-group classification problem and examined the performance of supervised regularization algorithms as spatial classifiers using imbalanced uncertain data. In addition to conventional shrinkage algorithms such as Ridge, the Lasso and Elastic-Net, the generalized t-statistic-based U-Lasso classifiers have been introduced and tested for mitigating such imbalances and adjusting weights for class distributions. The feature interpretation studies underlined that the most important indicator of the models is Zinc (Zn). The experimental studies revealed that the Ridge classifier (l_2 *penalty*) outperforms the other models. Statistically, the U-Lasso models exhibited notable explanation capacity and their performances recorded close to the conventional shrinkage algorithms. The use of statistical learning-based classification approach to appraise geo-environmental contamination under the conditions of natural variability and spatial uncertainty provides useful meta-data and reliable classification models.

Cite this article: Tutmez, B., and Komori, O. (2023). Regularization Learning of Trace Element Contamination Stemmed from Tailings Dam-Break. *Pollution*, 9 (3), 1082-1097.
https://doi.org/10.22059/poll.2023.349809.1657

DOI: https://doi.org/10.22059/poll.2023.349809.1657

INTRODUCTION

Mining aims to extract ores and solid substances recoverable at a profit. In general, the run-of-mine extracted from the earth needs further mineral processing such as the leaching (Revuelta, 2018). As a matter of course, mining operations cover various natural uncertainties and application risks. To minimize the natural and operational risks, first of all environmental impact assessment is required (Palma et al., 2019). However, either uncontrolled factors like earthquakes or incompatible design and working practices like unsafe waste storage induce environmental problems like soil and water contamination as well as health risks (Akoto et al., 2018).

Extractive mining has crucial importance for providing some fundamental materials used by modern society such as construction materials, ceramics, metals and glass. However, the mining industry generates egregious volumes of waste. A tailing dam is employed to store the mining waste then by separating the concentrate (valuable material) from the gangue (worthless material). As tailings can be liquid, solid, or slurry of fine particles, tailings have mostly toxic properties. Therefore, building tailing dams to collect and store the wastes is a common risky operation in mining and ore processing industries. These dams are usually established by

*Corresponding Author Email: *bulent.tutmez@inonu.edu.tr* available local materials and concrete (Kossoff et al., 2014). The potential risks of the dam-break have been discussed in (Gildeh et al., 2021).

Heavy metal contamination from tailings has become a major concern on account of its toxicity (Li et al., 2015). For this reason, many assessments have been made on tailings dam failures. The main reasons for the failures have been reported as: weak foundation, slope instability, mine subsidence, unexpected rain and snow, and seismic liquefaction (Rico et al., 2008; Lyu et al., 2019). In practice, trace elements in tailings create contamination both for water and soil. Ultimately, river, lake and groundwater and aquifer are generally affected because of high concentration of these elements. The influence of tailings dam breaks on plant life, fish and terrestrial animals can be harsh (Hudson-Edwards, 2003).

The impacts of the tailings dam failures on the environment have recently been a hot topic in literature. The effects of the Pb-Zn mine tailing dam-break on the degree of environmental response from soil properties were assessed (Jin et al., 2015). One of the detailed studies, water quality impacts and river system recovery following the mine tailings dam spill in Canada were appraised (Byrne et al., 2018). For this purpose, a conceptual model for analysing chemical mobilization has been developed. Similarly, the level of heavy metals contamination resourced from gold mine tailings in Ghana has been handled by (Sey and Belford, 2019). Different factors and indexes have been considered to evaluate the levels of trace element concentrations in the sampling sites. In another study, the effect of Radon concentration on human health in a domain dominated by failure mine tailings dams in South Africa has been investigated (Moshupya et al., 2019). Environmental risk assessment for geochemical distribution was made for a failure antimony mine in Serbia by (Randelovic et al., 2019). In this study, multivariate statistical analyses have been utilized for analysing the distribution. Most recently, heavy metals released by the dam collapse realized in Brazil have been considered (Davila et al., 2019). In this study, devastating contamination recorded from four years later when the accident was made has been appraised by hierarchical cluster analysis and correlation matrices.

Statistically, evaluation of trace element contamination in an environmental site addresses a case-control study based on imbalanced data and in which it is challenging to achieve high classification accuracy (Kumar et al 2017). This difficulty is referred to as "class imbalance problem" (Komori and Eguchi 2019). Because the appraisal of the impact of the heavy metal contamination resulting from tailings dam corruptions by artificial intelligence-based data analytics is a requirement for literature, instead of conventional and shrinkage classification methods, a new learning algorithm based on mitigating such imbalances and adjusting weights for class distributions is focused in this study. For this purpose, the capacity of generalized t-statistics for two group classifications is examined in this study. The t-statistic uses the standardized difference between the means of the two distributions (Eguchi and Copas, 2002). Komori et al. (2015) introduced the generalized t-statistic as an alternative solution for two-group classification based on asymptotic consistency and normality.

If the U-function is properly determined and some assumptions about residual vectors of the estimated linear classifier are satisfied, then the estimation of the parameters is best in terms of asymptotical efficiency.

The heavy metal contamination problem has been handled as a two-group (two different type trees) classification problem and cause-effect relationships are identified for providing a reliable balance differently from the conventional (logistic regression, discriminant analysis) and shrinkage-based (Ridge, the Lasso, Elastic Net) classifiers (Igual and Segui, 2017). By identifying the difference between the exposing levels of two group trees using a semi-parametric U-Lasso algorithm, this study aims to make two main contributions to literature. At first, the advanced multivariate statistical learning algorithms are enabled for handling the devastating environmental problem and the levels of heavy metal contamination in the trees have been manifested in this manner. Thus, it has been recorded that use of statistical learning-based classification algorithms covers a big capacity for evaluating this problem. The learning algorithms' feature interpretation highlighted the significance of the trace elements needed for environmental impact studies. Second, introducing and use of t-statistic-based classification algorithm U-Lasso in environmental problems involve a methodological contribution to ecological informatics. The performance indicators and experimental results demonstrated the notable capacity of the U-Lasso algorithms.

MATERIALS AND METHODS

The mining works like ore dressing produce vast amounts of waste. As the outcomes of crushed rocks and processing fluids, tailings remain at the site as a result of the extractive mining. The chemical texture of tailings relies on the mineralogy and the content of the processing fluids for extracting the economic minerals. Toxicity of trace elements in tailings disturbs nature and living in the sense of ecological, nutritional, and environmental sources. Permanent contamination in soil and plants is the main outcome of the trace elements in the site. As reported by Kossoff et al. (2014), sulphide tailings oxidation includes a potential for metal mobilization and acidification. Finally, destructive effects for vegetation, crops and trees are recorded. Evaluation of the impact of the trace elements over the trees can be determined using the concentrations measured in leafs and soil.

A cause-effect mapping based on in-situ environmental measurements requires cutting-edge multivariate analyses of simultaneous independent relationships (Jafarzadeh et al., 2020; Tutmez, 2020). Since machine learning provides effective time-tested toolkits for evaluating the impacts on ecological components such as water, vegetation, soil and trees (Hsieh, 2009), it can be underlined that the identification of mine-based contamination problems using machine learning includes a potential both for regression and classification purposes. This study aims to conduct a binary classification for inspecting the trace element contamination on the trees by shrinkage models. For analysing contamination intensity and heterogeneity in the site, probabilistic density functions and standardized differences can be utilized from a statistical learning perspective.

Statistically, two-group classification considered in this study comprises input variables (trace elements) and response variable (tree type). Inputs are composed of numerical concentration values and the response is two populations $p_0(x)$and $p_1(x)$ having the density functions. Providing good discriminant scores and successful classification both in training and testing is objected to.

Regularization methods for classification

As a statistical learning approach, regularization obtains a tool to constrain the predicted coefficients, which can decrease the variance and reduce out of sample error (Boehmke and Greenwell, 2020). The approach is also against overfitting and it corresponds to a shrinkage analysis. Three regularization methods such as Ridge, the Lasso and Elastic Net are used both for regression and classification purposes (James and Witten, 2013).

Ridge classifier

The general matrix solution employed for exhibiting multivariate input-output relationship addresses the ridge estimation by an additive small constant:

$$\hat{\beta}_R = \left(X^T X + \lambda I_p\right)^{-1} X^T y \tag{1}$$

In Eq (1), as a preliminary invariant, λ is employed as a tuning parameter. $\hat{\beta}_R$ is calculated to minimize the penalized sum of squares:

$$\sum_{i=1}^{n}\left(y_i\text{-}\beta_0\text{-}\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{2}$$

In Eq (2), RSS represents the residual sum of squares. The term $\lambda\sum_{j=1}^{p}\beta_j^2$ is expressed as shrinkage penalty. When coefficients are close to zero, this penalty decreases. The relative importance of the terms is expressed via the constant λ (Dorugade, 2018).

The Lasso

In the ridge model structure, all the independent variables are used in the resulting structure. This approach results in limited generality. To overcome limited generality and provide the model interpretability, the Lasso path was recommended (Hastie and Tibshirani, 2015).

To perform a feature selection, The Lasso uses simultaneous regularization and it shrinks down the model coefficients. The Lasso employs an l_1 norm penalty in place of an l_2. The coefficients $\hat{\beta_L}$ can be expressed as:

$$\sum_{i=1}^{n}\left(y_i\text{-}\beta_0\text{-}\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{3}$$

Elastic-Net

In comparison with the advantages of the Lasso regularization, it may involve some objections. High number of variables compared with observations and high correlations among the variables cause some limitations for the Lasso-based analysis (Zou and Hastie, 2005). As a combination of the Lasso and ridge, elastic net provides a new basis in sparsity (Khan et al., 2019). Elastic-net permits efficient regularization by ridge structure with the feature selection properties of the Lasso (Mokhtia et al., 2020). The integration can be formulated as follows:

$$L(\lambda_1,\lambda_2,\beta) = |y\text{-}X\beta|^2+\lambda_2|\beta|^2+\lambda_1|\beta|_1 \tag{4}$$

where, λ_1,λ_2 are fixed and non-negative. For $\alpha\epsilon[0,1)$ the elastic-net penalty is

$$\alpha = \lambda_2/(\lambda_1+\lambda_2) \tag{5}$$

$$(1-\alpha)|\beta|_1+\alpha|\beta|^2 \tag{6}$$

Based on two-step regularization, the following minimization addresses the elastic-net model structure (Boehmke and Greenwell, 2020):

$$RSS+\lambda_1\sum_{j=1}^{p}\beta_j^2+\lambda_2\sum_{j=1}^{p}\left|\beta_j\right| \tag{7}$$

The l_1 norm of the penalty structures a sparse model. In other respects, the quadratic part of the penalty conducts the l_1 component more stable. In the first step of the two-stage regularization for every stationary λ_2, the ridge regression coefficients are specified. After that, the shrinkage along the Lasso coefficient estimation path is conducted.

U-Lasso: Generalized t-statistics-based regularization

In two-group classification, maximization of the standardized difference between the means of the two distributions can be performed by t-statistics $\mathbb{L}_1(\beta)$. Following statistics is an extended version of the univariate t-test:

$$\mathbb{L}_1(\beta)=\frac{\beta^T(\mu_2-\mu_1)}{\left(\beta^T\sum_1\beta\right)^{1/2}} \tag{8}$$

The critical points of the multi-group classification performed by t-statistics are symmetry of distributions and variance-based optimality. As discussed in (Komori and Eguchi, 2019), providing a generalized version of t-statistics is necessary for ensuring optimality, consistency and normality.

Suppose two populations including pollution measurements $\{x_{1i}: i=1,\dots,n_1\}$ are trees of the first group (adult) and $\{x_{2j}: j=1,\dots,n_2\}$ are trees of the second group (sapling). The main motivation is to specify the generalized t-statistic as follows:

$$L_U(\beta)=\frac{1}{n_2}\sum_{j=1}^{n_2}U\left(\frac{\beta^T\left(x_{2j}-\bar{x}_1\right)}{\left(\beta^T S_1\beta\right)^{1/2}}\right) \tag{9}$$

In Eq. (9), U denotes a generator function. $\bar{x}_y$ and S_y represent the sample mean and sample variance, respectively. Finally, expectation of $L_U(\beta)$ can be adjusted as

$$\mathbb{L}_U(\beta)=E_2\left\{U\left(\frac{\beta^T(x-\mu_1)}{\left(\beta^T\sum_1\beta\right)^{1/2}}\right)\right\}. \tag{10}$$

In Eq. (10), $\sum y$, μ_y and E_y represents the variance, mean, and conditional expectation, respectively.

The Lasso and its variants have gained popularity as they automatically specify the effective variables via L$_1$-regularization. As a novel classifier, the U-Lasso has been suggested by adding L1-regularization term to generalized t-statistic as follows (Komori et al., 2015):

$$L_U^{\lambda}(\beta)=L_U(\beta)-\lambda\sum_{k=1}^{p}|\beta_k|, \tag{11}$$

where $\beta=\left(\beta_1,\dots,\beta_p\right)^T$, $L_U(\beta)$ is determined by Eq. (9). λ is a nonnegative component that manages the shrinkage of β. As is in the case with the Lasso, the absolute value of β_k procures smaller as λ increases.

RESULTS and DISCUSSION

The application site illustrated in Figure 1, the Guadiamar River Valley is located near the Mediterranean Sea and one of Europe's greatest sulphide mineralizations is found in this region. As reported in Domínguez et al. (2008), the collapse of a mine tailing dam in Seville released about 4 million m^3 of trace element-contaminated sludge into the Guadiamar River in 1998. The tailing dam failure had catastrophic ecological and socio-economic results (Grimalt et al., 1999). A large-scale cleaning and remediation program was performed in the Guadiamar site

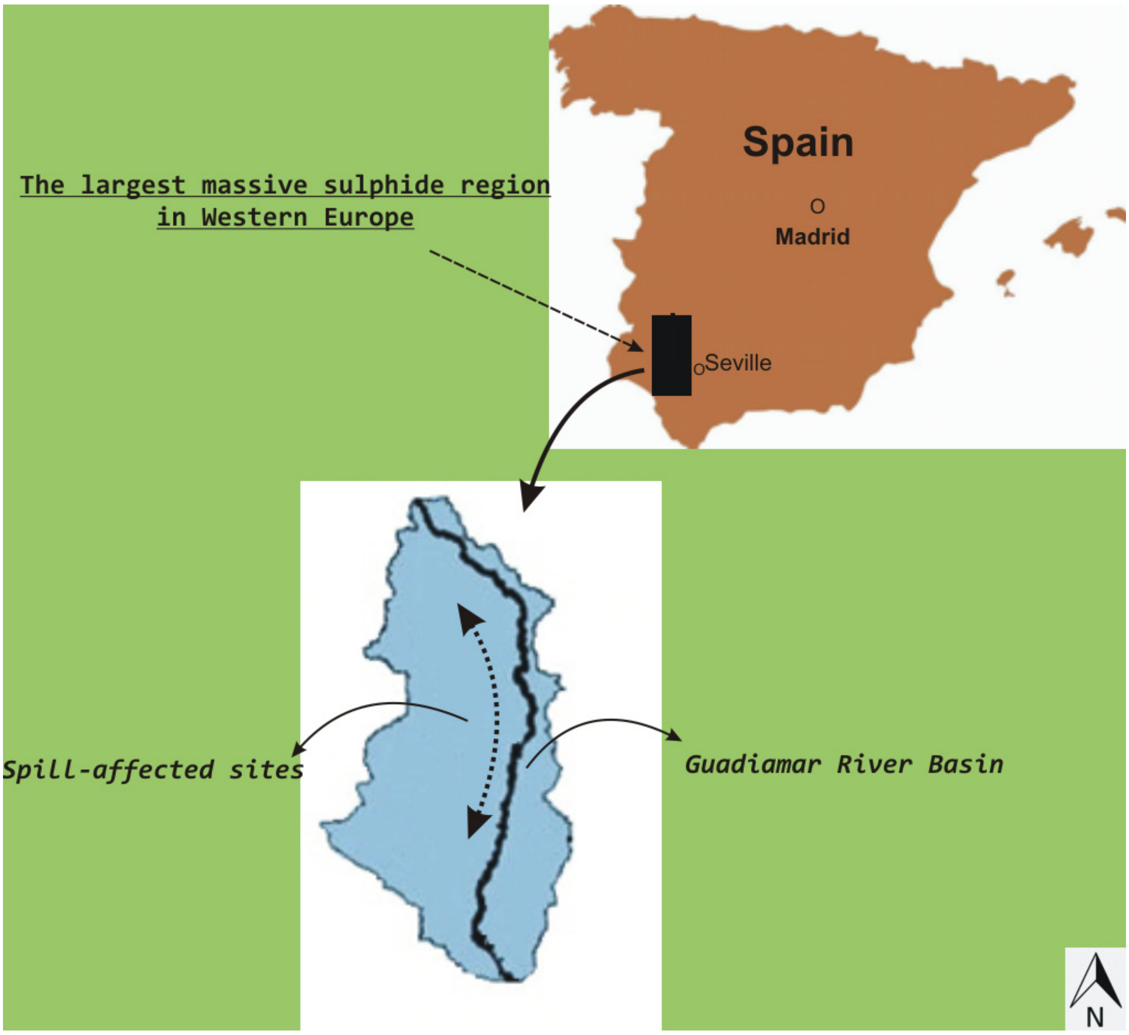

Fig. 1. Site location map for contaminated province.

affected by a mine-spill in 1998 (Madejón et al., 2018a).

Structure Identification and Implementation

As a result of the floods, many contaminants and trace-elements such as Cd, As, Cu, Zn, Pb accumulated in the agricultural areas and water. To appraise the destructive effects on soil, water and plant, many experimental studies were performed and data sets were generated and assessed (Domínguez et al., 2008; Grimalt et al., 1999; Domínguez, 2010; Madejón et al., 2018b).

In this study, 116 samples including 11 indicator variables (5 heavy metal concentrations measured both in soil and plant leaf as well as pH measurements) and response variable (tree type) have been considered. Based on the impacts on two tree types such as adult (coded as 0) and sapling (coded as 1), a series of classification experiments have been conducted for identification. The analyses were performed based on an experimental and algorithmic framework (Friedman et al., 2010). It should be noticed that since this study focuses on providing a reliable statistical framework based on environmental data, developing a mathematical model for the contaminant transport related to some physical problem is beyond the extent of the study.

To discover the relationships among the pollution indicator variables and tree types, bag plots which are the scatter plot variants, have been constructed. In Figure 2, 'bags' account for the box in a box and whisker plot to illustrate outliers. The inner polygon is called a bag, which corresponds to the box of a box and whisker plot, representing an area where 50% observations exist. The outer polygon and boundary are called a loop and a fence, respectively, which distinguish observations from outliers. As seen in Figure 2, there are no outliers for all indicator

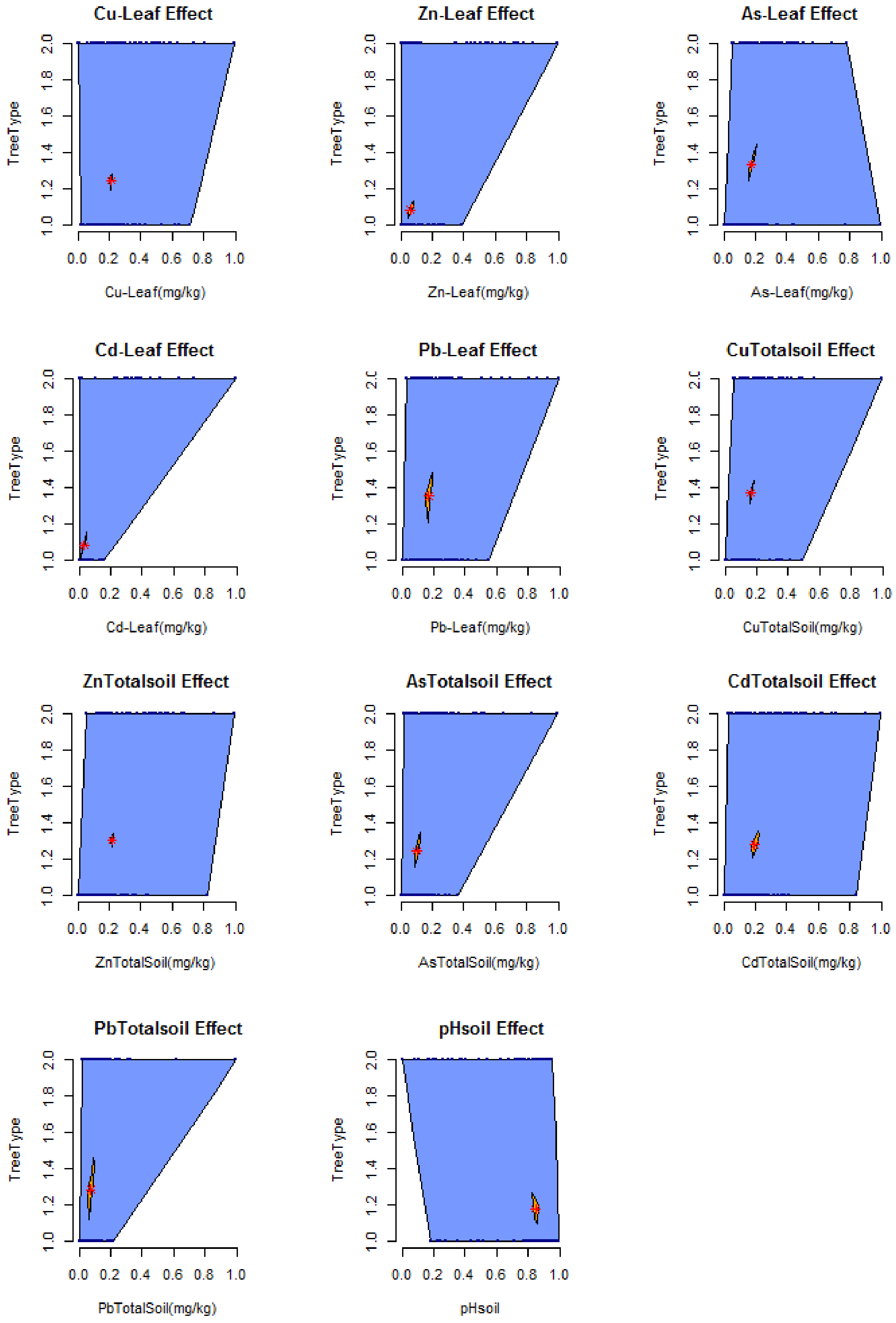

Fig. 2. Relationships between indicators and tree type.

variables and the areas of bags are concentrated around the centres (medians) marked in red, suggesting relatively small variations and spreads in the data. The shapes of bag and loop show correlation between the tree type and the variables. The shapes having upward slopes such as Zn-leaf, Cd-Leaf and Cu-Total Soil indicate positive correlations; that having downward slope such as pH-Soil indicates negative correlation.

To perform the experiments using the site data, the first data set was divided into two parts: 75% (Training) and 25% (Testing). Both conventional methods (Ridge, the Lasso, Elastic-net) and the suggested method (U-Lasso) have been applied in the same manner of approach. To provide stability and make reliable analyses independent from the data heterogeneity, the implementations have been conducted 10 times using separately sampled data groups.

Meta-data analysis and feature interpretation

Figure 3 exhibits the optimized model coefficients obtained by the λ grid values for Ridge, the Lasso and Elastic-Net models, the first implementations. From a statistical perspective, the solution paths by Ridge show gradual shrinkage as λ takes large values due to the property l_2 penalty. The next two path plots by Lasso and Elastic-Net show high similarity, indicating the tuning parameter $\alpha\epsilon[0,1)$ of Elastic-Net is estimated to be around 0. As a whole several variables seem to be informative to characterize tree types (adult or sapling).

There is a close relationship between good uncertainty diagnosis and successful discrimination scores. The U-Lasso considered distributional t-statistic and obtained optimality using different function types such as linear-U, auc-U and quadratic-U as well as optimal-U. Figure 4 illustrates the resultant path plots provided by these function structures for the first algorithmic experiment. From the perspective of generalized-t statistic with a linear U-function, Cd-Leaf, As-leaf and Cu total soil turned out to be informative. When AUC function is used in the generalized-t statistic, As-Leaf, Pb total soil, As-Total Soil and Cu-Total Soil remain active even if the penalty term λ is large. From the perspective of Fisher discriminant analysis (quadratic U function), Pb total soil, Cd-Leaf, Zn total soil and As total soil are selected to be informative. The solution paths generated by the optimal-U show very few informative variables except for Cd-Leaf.

To make a reliable feature interpretation, first variable importance for the shrinkage models have been explored. The importance has been designated using the magnitude of the standardized coefficients. Figs. 5 and 6 illustrate the feature importance plots for conventional (Ridge, the Lasso, Elastic Net) and U-Lasso-based regularization (Linear, Quadratic, Auc, Optimal). The largest absolute coefficients of the conventional shrinkage models have been recorded as Ridge (Cd-Leaf), the Lasso (Cd-Leaf, Zn-Soil), Elastic Net (Zn-Soil, Cd-Leaf). The U-Lasso models underlined the parameters: U-Linear (As-Soil, Zn-Leaf), U-Quadratic (Zn-Soil, As-Soil, Zn-Leaf), U-Auc (As-Soil, Zn-Soil), U-Optimal (As-Soil, Zn-Leaf).

The magnitude of the indicators provided by the conventional algorithms indicated that Cd and Zn are the principal parameters of the contamination for Leaf and Soil, respectively. There is a consensus of all algorithms that "Zn" is the one of the important heavy metals for both soil and leaf. With the difference of conventional algorithms, U-Lasso models underline the trace element "As" instead of "Cd" as the major indicator variable. As discussed in different studies, these heavy metals can be potential threats to a wide range of biota, soil as well as human health (Norini et al., 2019; Holtra and Zamorska-Wojdyla, 2020).

Performance evaluation-based benchmarking

To evaluate the performance of the classification models, one of the effective performance measures, AUC (Area Under the ROC curve) was utilized. Because AUC has a common insensitivity and also adaptivity to machine learning classification algorithms, it has been preferred. For the discriminant function F(x) and threshold value c, AUC can be stated as follows:

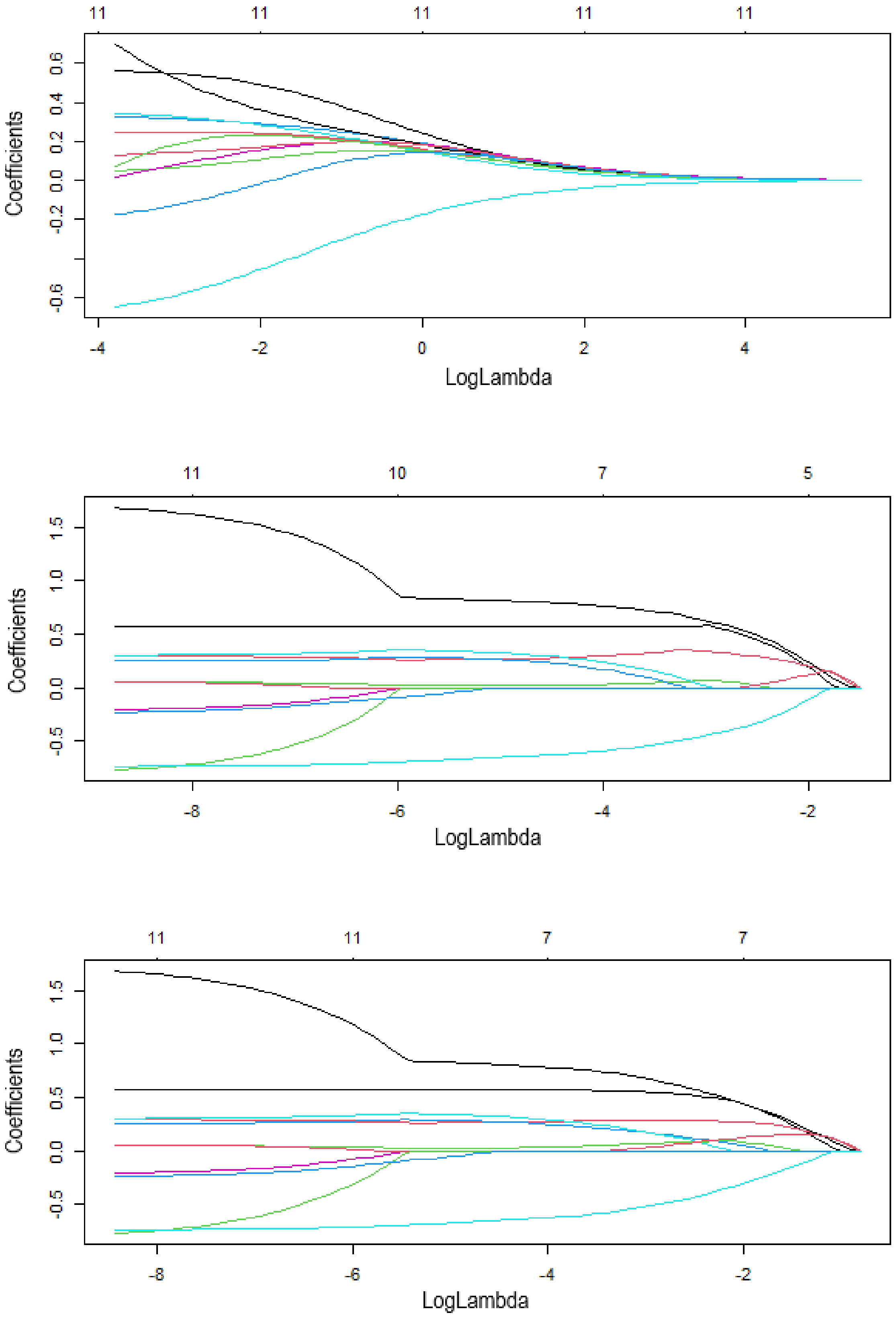

Fig. 3. Solution paths for conventional regularization algorithms.

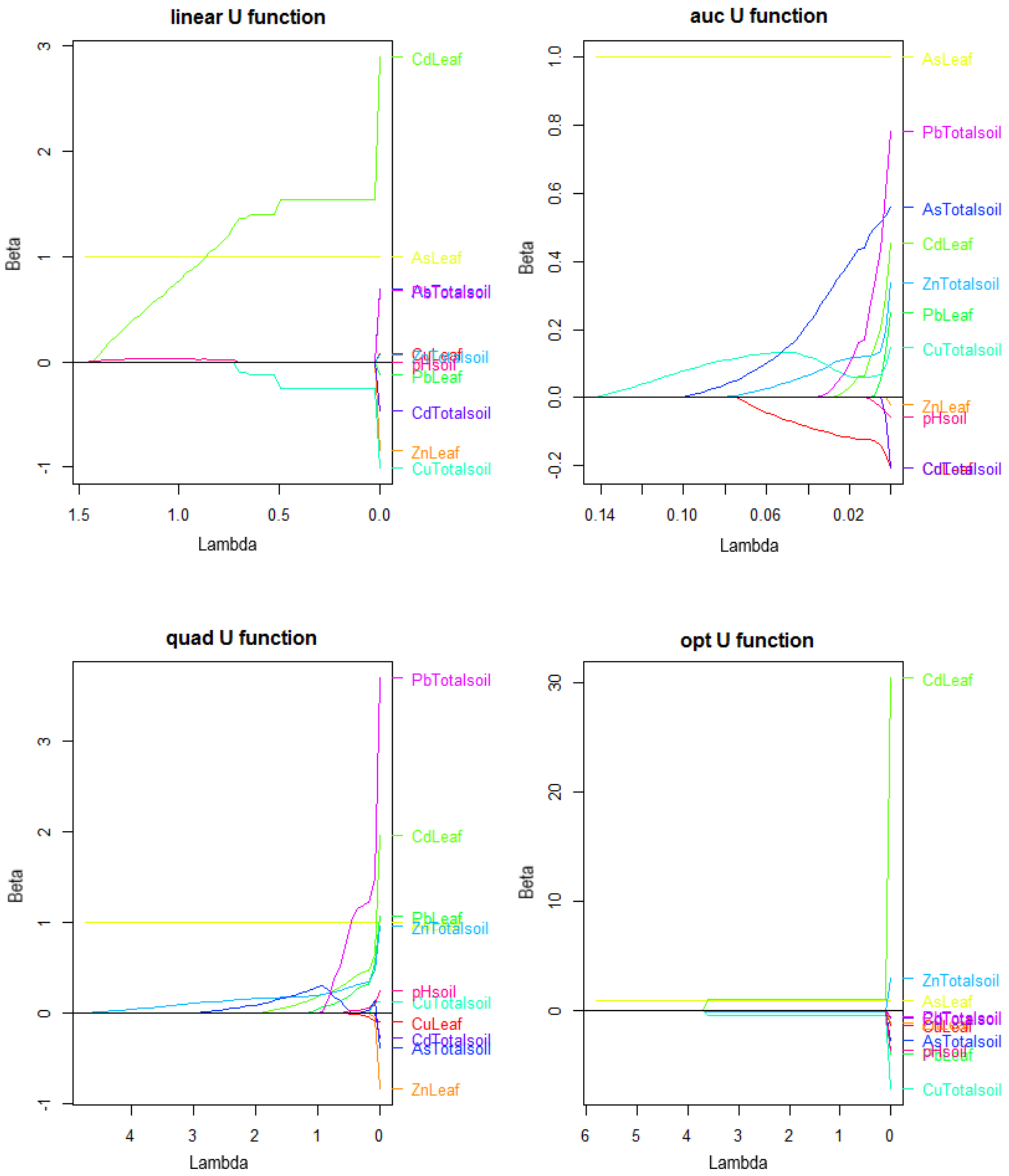

Fig. 4. Solution paths for U-Lasso algorithms.

$$\overline{AUC}(F)=\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}H\left(F\left(x_{2j}\right)-F\left(x_{1i}\right)\right), \qquad (12)$$

where, *H(.)* denotes the Heaviside function: $H\left(z\right)=1 \, if \; z \geq 0$ and 0 otherwise. The samples, $\left\{x_{11},\ldots,x_{1n_1}\right\}$ and $\left\{x_{21},\ldots,x_{2n_2}\right\}$ correspond for *y=0* and *y=1*, respectively.

AUC represented probabilities and it obtained an aggregate measure of performance across all possible classification thresholds. Table 1 summarizes 10 experimental AUC results of both conventional and U-Lasso-based classification models. The outcome of the performances with variability is presented by Figure 7.

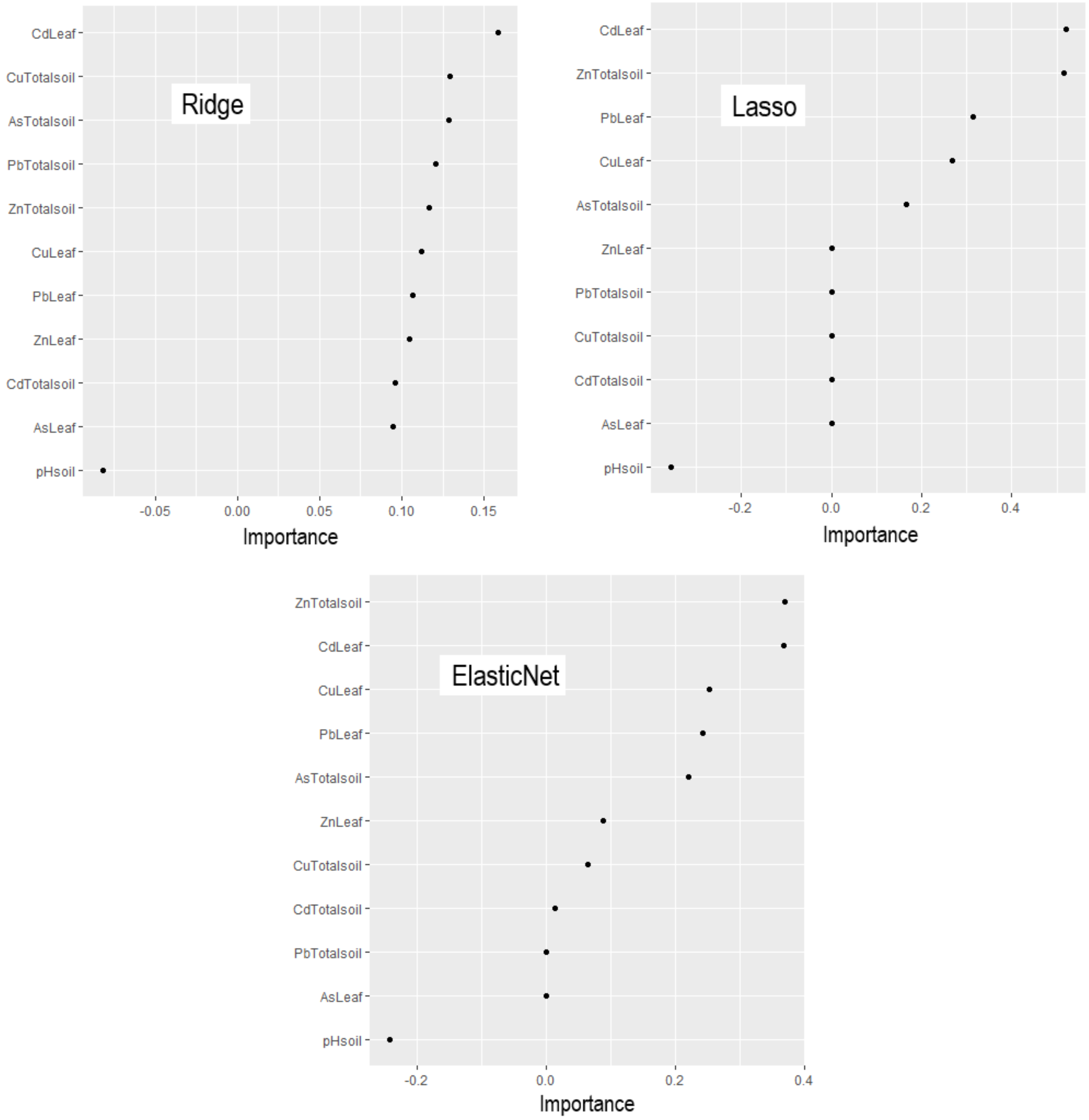

Fig. 5. Variable importance for conventional regularization models.

The numerical outcomes exhibited that all the classification models produce identical biases. This may be explained by natural variability of site and also spatial heterogeneity. Although the Lasso-based models provide simple and interpretable models that contain only a subset of the indicators, the results indicate that the Ridge model outperforms the other models. One of the main reasons for better accuracy provided by l_2 penalty can be explained by the slightly lower variance of the Ridge structure. The achievement obtained for classification accuracy is also rooted in the bias-variance trade-off. Even though the flexibility of the Ridge model reduces with high λ, it also provides decreased variability.

It should be noticed that the generalized t-statistic-based classifiers, the U-Lasso models have remarkable capacity and their performances close to the well-known shrinkage algorithms. Practically, if we choose a quadratic U-function in U-Lasso, it is equivalent to Fisher linear classification with l_1 penalty. If we choose the standard normal distribution function as

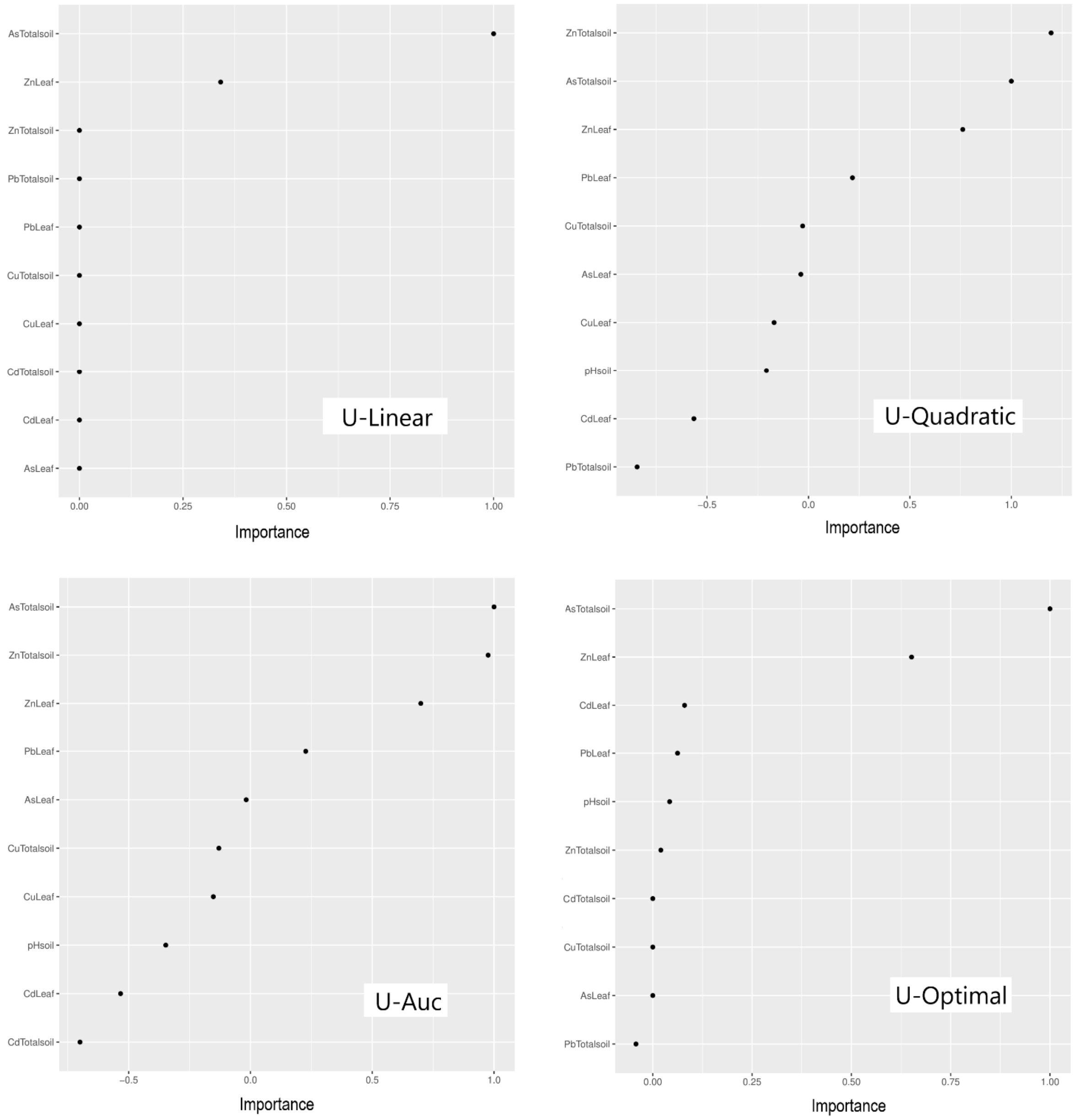

Fig. 6. Variable importance for U-Lasso models.

Table 1. Performance measure for classification models.

			Testing Performance - AUC				
No	**Ridge**	**Lasso**	**Elastic-Net**	**U-Lin.**	**U-Auc**	**U-Quad.**	**U-Opt.**
1	0.902	0.858	0.858	0.73	0.848	0.848	0.725
2	0.923	0.893	0.893	0.833	0.833	0.833	0.833
3	0.926	0.868	0.916	0.689	0.768	0.758	0.689
4	0.914	0.862	0.876	0.948	0.952	0.943	0.938
5	0.784	0.712	0.755	0.716	0.712	0.731	0.712
6	0.874	0.808	0.843	0.773	0.864	0.828	0.758
7	0.869	0.753	0.803	0.641	0.641	0.641	0.646
8	0.808	0.793	0.798	0.651	0.779	0.774	0.683
9	0.922	0.811	0.872	0.606	0.789	0.822	0.667
10	0.848	0.71	0.814	0.7	0.695	0.695	0.69
Mean	**0.877**	**0.806**	**0.842**	**0.728**	**0.788**	**0.787**	**0.734**

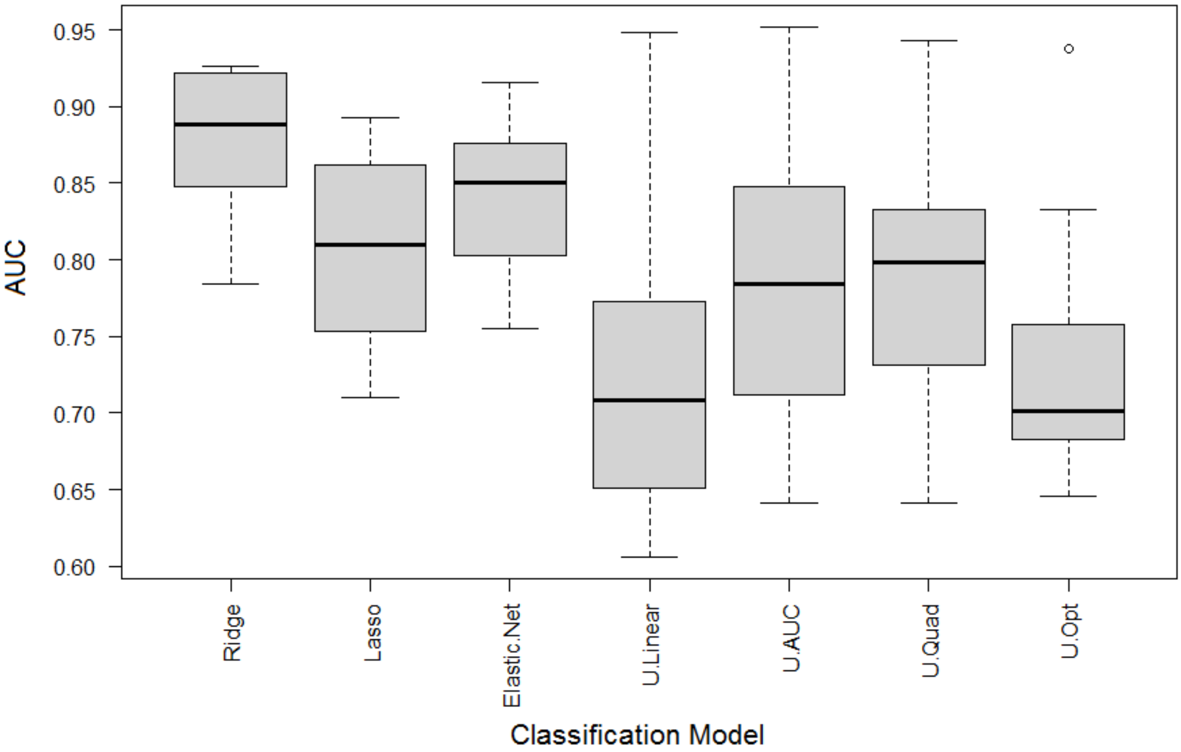

Fig. 7. AUCs for classifiers.

U-function, it is equivalent to maximize the AUC with l_1 penalty. These properties are useful in various data analysis. Theoretically, the estimation of the parameters is most asymptotically efficient if the U-function is correctly calculated and certain residual vector assumptions of the estimated linear classifier are met. Also the semiparametric efficiency of the parameters in a more general framework is established in Baek et al. (2018).

CONCLUSIONS

Mineral processing and beneficiation embody contamination potential due to natural and practical risks. As a result of which, heavy metal contamination is recorded by an accident like industrial-scale mine waste storage dam failure. Based on the advanced regularization models, the pollution problem originated from this mining action has been appraised. The statistical evaluation of the trace element contamination stemming from dam tailings was considered as a two-group machine learning classification problem and a new supervised regularization algorithm U-Lasso model examined and tested as a solution tool. The abilities of both conventional shrinkage and the U-Lasso classifiers to distinguish between classes have been inspected.

The feature interpretations and the benchmarking analyses on the classification models showed that the shrinkage approach is a useful tool to identify and understand the relationships and effects in the contaminated site. The studies on feature interpretation revealed that Zn is the most influential indicator for the regularization-based effect analysis. Among the shrinkage algorithms, the Ridge classifier and l_2 penalty become prominent. However, all of the models exhibited at least 70% accuracy and this level can be acceptable for a real geosciences model due to natural and sampling uncertainties. The classification studies along with meta-data analysis demonstrated that as an alternative classier, generalized t-statistic based classification (U-Lasso) can provide additional information on the ground of feature interpretation and semi-parametric statistical learning. Thus, the U-Lasso approach suggests different types of classification functions and also additional information for exploring the bounds of the pollution on soil, water as well as trees.

ACKNOWLEDGEMENT

The authors extend their appreciations to the Editor-in-Chief and anonymous reviewers for the insightful comments.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Akoto, A., Bortey-Sam, N., Nakayama, S.M.M., Ikenaka, Y., Baidoo, E., Apau, J., Marfo, J.T. & Ishizuka, M. (2018). Characterization, Spatial Variation & Risk Assessment of Heavy Metals & a Metalloid in Surface Soils in Obuasi, Ghana. *Journal of Health & Pollution*, 8(19); 180902. https://doi.org/10.5696/2156-9614-8.19.180902.

Baek, S., Komori, O. & Ma, Y. (2018) An optimal semiparametric method for two-group classification. *Scandinavian Journal of Statistics*, 45; 806-846. https://doi.org/10.1111/sjos.12323.

Boehmke, B. & Greenwell, B. (2020). Hands-on machine learning with R. (Boca Raton: CRC Press).

Byrne, P., Hudson-Edwards, K.A., Bird, G., Macklin, M.G., Brewer, P.A., Williams, R.D. & Jamieson, H.E. (2018). Water quality impacts & river system recovery following the 2014 Mount Polley mine tailings dam spill, British Columbia. Canada. *Applied Geochemistry*, 91; 64-74. https://doi.org/10.1016/j.apgeochem.2018.01.012.

Davila, R.B., Fontes, M.P.F., Pacheco, A.A. & Ferreira, M.D.S. (2019). Heavy metals in iron ore tailings & floodplain soils affected by the Samarco dam collapse in Brazil. *Science of the Total Environment*, 709; 136151. https://doi.org/10.1016/j.scitotenv.2019.136151.

Domínguez, M.T., Marañón, T., Murillo, J.M., Schulin,R. & Robinson, B.H. (2008). Trace elements accumulation in woody plants of the Guadiamar Valley, SW Spain: a largescale phytomanagement case study. *Environ. Pollut.*, 152; 50–59. https://doi.org /10.1016/j.envpol.2007.05.021.

Domínguez, M.T., Madejón, P., Marañón, T. & Murillo, J.M. (2010). Afforestation of a trace element polluted area in SW Spain: woody plant performance & trace element accumulation. *Eur. J. For. Res.*, 129; 47–59. https://doi.org /10.1007/s10342-008-0253-3.

Dorugade, A.V. (2014). New ridge parameters for ridge regression. *Journal the Association of Arab Universities for Basic & Applied Sciences*, 15(1); 94-99. https://doi.org/10.1016/j.jaubas.2013.03.005

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *J Stat Softw.*, 33(1); 1-22.

Eguchi, S. & Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika*, 89(1); 1-22. https://doi.org/10.1093/biomet/89.1.1.

Gildeh, H.K., Halliday, A., Arenas, A. & Zhang, H. (2021). Tailings dam breach analysis: a review of methods, practices, & uncertainties. *Mine Water & the Environment*, 40; 128-150. DOI:10.1007/s10230-020-00718-2

Grimalt, J.O., Ferrer, M. & Macpherson, E. (1999). The environmental impact of the mine tailing accident in Aznalcóllar. *Sci. Total Environ.*, 242 (1-3); 3-11. doi: 10.1016/s0048-9697(99)00372-1

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning, (New York: Springer).

Jin, Z.J., Li, Z.Y., Li, Q., Hu, Q.J., Yang, R.M., Tang, H.F., Li, M., Huang, B.F., Zhang, J.Y. & Li, G.W. (2015). Canonical correspondence analysis of soil heavy metal pollution, microflora & enzyme activities in the Pb-Zn mine tailing dam collapse area of Sidi village, SW China. *Environmental Earth Science*, 73(1); 267-274. https://doi.org/10.1007/s12665-014-3421-4.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). Statistical learning with sparsity. (CRC Press).

Holtra, A. & Zamorska-Wojdyla, D. (2020). The pollution indices of trace elements in soils & plants close to the copper & zinc smelting works in Poland's Lower Silesia. *Environmental Science & Pollution Research*, 27; 16086-16099, https://doi.org/10.1007/s11356-020-08072-0.

Hsieh, W.W. (2009). Machine learning methods in the environmental sciences. (Cambridge: Cambridge University Press).

Hudson-Edwards, K.A., Macklin, M.G., Jamieson, H.E., Brewer, P., Coulthard, T.J., Howard, A.J. & Turner, J. (2003). The impact of tailings dam spills & clean-up operations on sediment & water quality in river systems: the Rios Agrio-Guadiamar, Aznalcollar, Spain. *Applied Geochemistry*, 18; 221–239. https://doi.org/10.1016/S0883-2927(02)00122-1

Igual, L. & Segui, S. (2017). Introduction to data science, a Python approach to concepts, techniques & applications. (Switzerland: Springer).

Jafarzadeh, A.A., Pal, M., Servati, M., Fazelifard, M.H. & Ghorbani, M.A. (2016). Comparative analysis of support vector machine & artificial neural network models for soil cation exchange capacity prediction. *Int. J. Environ. Sci. Technol.*, 13; 87–96. https://doi.org/10.1007/s13762-015-0856-4

Khan, M.H.R., Anamika, B. & Tamanna, H. (2019). Stability selection for Lasso, ridge & elastic net implemented with AFT models. *Statistical Applications in Genetics & Molecular Biology*, 18(5). https://doi.org /10.1515/sagmb-2017-0001.

Komori, O., Eguchi, S. & Copas, J.B. (2015). Generalized t-statistic for two-group classification. *Biometrics*, 71; 404-416. https://doi.org/10.1111/biom.12265

Komori, O. & Eguchi, S. (2019). Statistical methods for imbalanced data in ecological & biological studies. (Japan: Springer).

Kossoff, D., Dubbin, W.E., Alfredsson, M., Edwards, S.J., Macklin, M.G. & Hudson-Edwards, K.A. (2014). Mine tailings dams: Characteristics, failure, environmental impacts. & remediation. *Applied Geochemistry*, 51; 229-245. http://dx.doi.org/10.1016/j.apgeochem.2014.09.010

Kumar, M., Ramanathan, A.L., Tripathi, R., Farswan, S., Kumar, D. & Bhattacharya, P. (2017). A study of trace element contamination using multivariate statistical techniques & health risk assessment in groundwater of Chhaprola Industrial Area, Gautam Buddha Nagar, Uttar Pradesh, India. *Chemosphere*, 166; 135-145. https://doi.org/10.1016/j.chemosphere.2016.09.086.

Li, P., Qian, H., Howard, KWF. & Wu, J. (2015). Heavy metal contamination of Yellow River alluvial sediments, Northwest China. *Environ Earth Sci*, 73; 3403–3415. https://doi.org /10.1007/s12665-014-3628-4.

Lyu, Z.J., Chai, J.R., Xu, Z.G., Qin, Y. & Cao, J. (2019). A comprehensive review on reasons for tailings dam failures based on case history. *Advances in Civil Engineering*, ID:159306. https://doi.org/10.1155/2019/4159306.

Madejón, P., Domínguez, M.T., Gil-Martinez, M., Navarro-Fernandez, C.M., Montiel-Rozas, M.M., Madejón, E., Murillo, J.M., Cabrera,F. & Marañón, T. (2018a). Evaluation of amendment addition & tree planting as measures to remediate contaminated soils: The Guadiamar case study (SW Spain). *Catena*, 166; 34-43. https://doi.org/10.1016/j.catena.2018.03.016.

Madejón, P., Domínguez, M.T., Madejón, E., Cabrera,F., Marañón, T. & Murillo, J.M. (2018b). Soil-plant relationships & contamination by trace elements: A review of twenty years of experimentation & monitoring after the Aznalcóllar (SW Spain) mine accident. *Science of the Total Environment*, 625; 50-63. https://doi.org/10.1016/j.scitotenv.2017.12.277.

Mokhtia, M., Eftekhari, M. & Saberi-Movahed, F. (2020). Feature selection based on regularization of sparsity based regression models by hesitant fuzzy correlation. *Applied Soft Computing*, 91, ttps://doi.org/10.1016/j.asoc.2020.106255.

Moshupya, P., Abiye, T., Mouri, H., Levin, M., Strauss, M. & Strydom, R. (2019). Assessment of Radon concentration & impact on human health in a region dominated by abandoned Gold mine tailings dams: a case from the West Rand Region, South Africa. *Geosciences*, 9(11); 466. https://doi.org /10.3390/geosciences9110466.

Norini, M.P., Thouin, H., Miard, F., Battaglia-Brunet, F., Gautret, P., Guégan, R., LeForestier, L., Morabito, D., Bourgerie, S. & Motelica-Heino, M. (2019). Mobility of Pb, Zn, Ba, As & Cd toward soil pore water & plants (willow & ryegrass) from a mine soil amended with biochar. *Journal of Environmental Management*, 232; 117-130. https://doi.org 10.1016/j.jenvman.2018.11.021.

Palma, P., Lopez-Orozco, R., Mourinha, C., Oropesa, A.L., Novais, M.H. & Alvarenga, P. (2019). Assessment of the environmental impact of an abandoned mine using an integrative approach: A case-study of the Las Musas mine (Extremadura, Spain), *Science of the Total Environment*, 659; 84-94. https://doi.org/10.1016/j.scitotenv.2018.12.321

Randelovic, D., Mutic, J., Marjanovic, P., Dordevic, T. & Kasanin-Grubin, M. (2019). Geochemical distribution of selected elements in flotation tailings & soils/sediments from the dam spill at the abandoned antimony mine Stolice, Serbia. *Environmental Science Pollution Research*, 27; 6253-6258. https://doi.org/10.1007/s11356-019-07348-4

Revuelta, M.B. (2018). Mineral resources-from exploration to sustainability assessment, (Cham:Springer)

Rico, M., Benito, G., Salgueiro, A.R., Diez-Herrero, A. & Pereira, H.G. (2008). Reported tailings dam failures: a review of the European incidents in the worldwide context. *J. Hazard. Mater.*, 152; 846–852. https://doi.org /10.1016/j.jhazmat.2007.07.050

Sey, E. & Belford, E.J.D. (2019). Levels of heavy metals & contamination status of a decommissioned tailings dam in Ghana. *EQA - International Journal of Environmental Quality*, 35; 33-50. DOI: 10.6092/issn.2281-4485/9060 33.

Tutmez, B. (2020). Air quality assessment by statistical learning-based regularization. *Çukurova University Journal of the Faculty of Engineering & Architecture*, 35(2); 271-278.

Zou, H. & Hastie, T. (2005). Regularization & variable selection via the elastic net. *Journal of the Royal Statistical Society*, Series B:301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x