ARTICLES

Mediterranean Botany
ISSNe 2603-9109

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/mbot.67609

Use of machine learning to establish limits in the classification of hyperaccumulator plants growing on serpentine, gypsum and dolomite soils

Marina Mota-Merlo[1] & Vanessa Martos[2]

Received: 7 February 2020 / Accepted: 18 September 2020 / Published online: 8 March 2021

Abstract. So-called hyperaccumulator plants can store heavy metals in quantities a hundred or a thousand times higher than typical plants, making hyperaccumulators very useful in phytoremediation and phytomining. Among these, there are many serpentinophytes, i.e., plants that grow exclusively on ultramafic rocks, which produce soils with a great proportion of heavy metals. Even though there are multiple classifications, the lack of consensus regarding which parameters should be used to determine if a plant is a hyperaccumulator and the arbitrariness of established thresholds elicits the need to propose more objective criteria. Therefore, this work aims to refine the existing classification. To this end, plant mineral composition data from different vegetal species were analyzed using machine learning techniques. Three complementary approaches were established. Firstly, plants were classified into three types of soils: dolomite, gypsum, and serpentine. Secondly, data about normal and hyperaccumulator plant Ni composition were analyzed with machine learning to find differentiated subgroups. Lastly, association studies were carried out using data about the mineral composition and soil type. Results in the classification task reached a success rate of over 75%. The clustering of plants by Ni concentration in parts per million (ppm) resulted in four groups with cut-off points in 2.25, 100 (accumulators) and 3000 ppm (hyperaccumulators). Associations with a confidence level above 90% were found between high Ni levels and serpentine soils, as well as between high Ni and Zn levels and the same type of soil. Overall, this work demonstrates the potential of machine learning to analyze plant mineral composition data. Finally, after consulting the IUCN's red list as well as those of countries with high richness in hyperaccumulator species, it is evident that a greater effort should be made to establish the conservation status for this type of flora.

Keywords: hyperaccumulators; serpentinophytes; nickel; phytoremediation; phytomining; artificial intelligence.

How to cite: Mota-Merlo, M. & Martos, V. 2021. Use of machine learning to establish limits in the classification of hyperaccumulator plants growing on serpentine, gypsum and dolomite soils. Mediterr. Bot. 42, e67609. https://dx.doi.org/10.5209/mbot.67609

Introduction

Nowadays, research on the use of certain organisms to eliminate heavy metals and other pollutants is of great interest. Due to various factors, both natural (geogenic activities) and anthropogenic (mining, electroplating, smelting operations, domestic and agro-allied industries), these compounds are present in soils, water or air in concentrations that can be toxic for most living beings, including humans (Okereafor *et al*., 2020) farmlands, plants, livestock and subsequently humans through the food chain. Most of the toxic metal cases in Africa and other developing nations are a result of industrialization coupled with poor effluent disposal and management. Due to widespread mining activities in South Africa, pollution is a common site with devastating consequences on the health of animals and humans likewise. In recent years, talks on toxic metal pollution had taken center stage in most scientific symposiums as a serious health concern. Very high levels of toxic metals have been reported in most parts of South African soils, plants, animals and water bodies due to pollution. Toxic metals such as Zinc (Zn.

Heavy metals are mostly transition metals that can be both essential, such as iron (Fe), zinc (Zn), manganese (Mn), copper (Cu), nickel (Ni), molybdenum (Mo), and cobalt (Co, nitrogen fixation in legumes), and non-essential (Marschner, 2016); for example, chrome (Cr), cadmium (Cd), mercury (Hg) and lead (Pb). In the case of soil, high concentrations of heavy metals may be found in the rhizosphere solution, as ions or as part of complexes with organic or inorganic compounds (Rostami & Azhdarpoor, 2019). However, the heavy metal deficit is more common than toxicity, especially in crops (Alloway, 2013). Their presence in soils can be the result of natural processes such as the weathering of parent rocks (Cd, Co, Ni, Pb), high soil acidity (Mn, Zn), lack of proper oxygenation, for example, due to inundation (Fe, Mn), etc. Moreover, as previously stated, heavy metals (Cd, Co, Cr, Cu, Mo, Ni, Pb, Sr, Zn) can appear as the result of human activities, such as irrigation and land clearing (Shabala, 2013).

[1] Bioinformatics MSc Student, Uppsala University. E-mail: marina.mota.merlo@gmail.com
[2] Department of Plant Physiology, Sciences Faculty, University of Granada. E-18071, Granada, Spain

There are three relevant concepts regarding nutrient availability in plants: deficit zone, deficiency critical concentration, and critical toxicity concentration. In the deficit zone, element concentration in the soil hinders optimal plant growth. In severe cases, this can lead to plant death. The critical deficiency concentration is the minimal concentration in a tissue allowing for 90% of its maximum yield. Lastly, when element concentration is over toxicity critical concentration, the yield is reduced >10% compared to the maximum. Toxicity critical concentration is considered, in some cases, as the threshold concentration for hyperaccumulation (Marschner, 2016).

Critical concentrations depend on sample tissue and soil solute composition. They can differ among, as well as within, species. It probably results from their adaptation to ancestral habitats and the development of special ecological strategies. The capability of some plants to accumulate heavy metals in concentrations that are toxic for the majority of living beings has a significant genetic component related to the entrance of the element into the cell and its subsequent processing by the plant, an insight of great use in biotechnology (Khan *et al*., 2015). In the case of soils with long-lasting contamination, the strength of the response ranges from acclimation mechanisms, consisting of blocking entrance and transport of heavy metals inside the plant, to tolerance mechanisms, including the aforementioned heavy metal detoxification. The complexity and abundance of said mechanisms result from the adaptive responses' evolution for hundreds, thousands, or even millions of years (Shabala, 2013).

Heavy metals may be found in acid, alkaline, saline, sodic and calcareous soils. Endemic plants are common among hyperaccumulator species since these are closely linked to very unusual substrates, such as serpentine. This is a particular type of substrate where numerous heavy metal hyperaccumulator species grow. Due to their narrow distribution and habitat specificity, said endemism presents a high probability of being endangered. This circumstance may be aggravated if there is also mining activity in their habitats. Therefore, these species' conservation status is also of great interest according to the IUCN and national red lists, especially those of the countries where these species are frequent.

Hyperaccumulator species are of great interest in phytoremediation and phytomining because heavy metals, despite being often necessary for plant growth in low concentrations, become toxic for non-accumulator plants in higher concentrations. The heavy metal whose hyperaccumulation is most widely documented is nickel since soils, where hyperaccumulator plants are found are usually rich in this element (Reeves *et al*., 2017). In low concentrations, nickel acts as a cofactor for some plant enzymes; thus its deficit has negative effects on plant growth. However, it causes toxicity (seen as growth inhibition and photosynthetic activity decay) in high concentrations (Reeves *et al*., 2017; Batool, 2018). It is important to take into account that, even though some hyperaccumulator species are already well-known and there are databases for these plants (such as the Global Hyperaccumulator Database, GHD: http://hyperaccumulators.smi.uq.edu.au/collection/), the use of autochthonous species in phytoremediation and phytomining must be favored because of the risks involved in the introduction of invasive species. Moreover, the characterization and study of hyperaccumulators allows for further investigation about hyperaccumulation-related genes and the use of deeper knowledge about these genes in genetic engineering, aiming to obtain more efficient plants for phytoremediation and phytomining. On the other hand, plant productivity is also important in phytoremediation because a high productivity could compensate for a lower accumulation in contrast to hyperaccumulators whose biomass production is low (Al Chami *et al*., 2015).

The existence of heavy metal hyperaccumulator plants has been known for a long, but the idea of cultivating plants to extract soil pollutants (natural phytoextraction) is relatively recent (Shah & Daverey, 2020). This technique has numerous limitations that prevent it from decontaminating soils in a short time. However, investigations up until today have had promising results (Corzo Remigio *et al*., 2020)while achieving monetary gain. Phytoextraction can be applied to a limited number of elements depending on the existence of hyperaccumulator plants with suitable characteristics. Although phytoextraction has been trialled in experimental settings, it requires testing at field scale to assess commercial broad-scale potential. Scope: The novelty and purported environmental benefits of phytoextraction have attracted substantial scientific inquiry. The main limitation of phytoextraction with hyperaccumulators is the number of suitable plants with a high accumulation capacity for a target element. We outline the main considerations for applying phytoextraction using selected elemental case studies in which key characteristics of the element, hyperaccumulation and economic considerations are evaluated. Conclusions: The metals cobalt, cadmium, thallium and rhenium and the metalloids arsenic and selenium are present in many types of minerals wastes, especially base metal mining tailings, at concentrations amenable for economic phytoextraction. Phytoextraction should focus on the most toxic elements (arsenic, cadmium, and thallium. Studies about the augmentation in productivity for these species and their metal accumulation capability should be noted by selecting and reproducing improved cultures and optimizing soil management practices. Simultaneously, progress in understanding soil-plant-microorganism relationships will soon allow for the modification of rhizosphere conditions for hyperaccumulator plants to increase metal absorption and translocation. Kidd *et al*. (2007) documented clear examples of these breakthroughs in numerous works about different species in the genus *Alyssum*, the one that includes the greater number of hyperaccumulator species. Many of these species are local endemisms. According to the GHD, around 100 species (13.9% of the total) are local endemisms. Since restricted geographical distribution is a threat factor, many hyperaccumulators are likely endangered (e.g., Faucon *et al*., 2010), natural outcrops

of copper-rich rocks are colonised by highly original plant communities. A number of plant species have been proposed as possibly endemic to those sites. Here we revise the taxonomic, phytogeographic and conservational status of these plants. Methods - Almost all the herbarium materials of supposed Cuendemics available in BR and BRLU have been revised and all relevant taxonomic revisions have been consulted. Literature and herbarium data have been supplemented by original observations in the field. Conservational status was established using IUCN criteria based on current and projected variation of population size and number. Key results - Thirty-two taxa are identified as strict endemics of Curich soil in Katanga, i.e. absolute metallophytes. Twenty-four of these are known from one to five localities only. Twenty-three other taxa are identified as broad endemics, i.e. with > 75% of occurrence on Curich soil. Fifty-seven other names formerly used for supposed endemics are rejected either for nomenclatural or phytogeographic reasons. A number of species formerly regarded as endemics have been discovered off copperenriched substrates due to progress in the botanical exploration of Katanga. The taxonomic value of a number of proposed endemics is still uncertain and requires further research. For a number of taxa, local geographic distribution still remains insufficiently known. The low proportion of endemics (c. 5%). However, this information is not present in the GHD.

According to Buscaroli *et al*. (2017) and Mganga *et al*. (2011), a line transect of 700m long was established opposite the gold mine wastes. A total of eight sampling points were systematically established each after every 100m in that transect. Fifteen plant species were sampled; at least one species per sampling point. Approximately 5g of the root and shoot portions of the plants were separately collected from each plant. Three soil samples were also collected at each sampling point where vegetations were previously sampled. The soils and vegetations were analyzed for heavy metals (copper, lead, chromium, zinc, cadmium and nickel, there are four criteria to define nickel (and other heavy metal) hyperaccumulator plants:

1. Hyperaccumulator plants have a metal content equal to or greater than 1000 ppm in their leaves (dry weight, DW).
2. Hyperaccumulator plants have a mean nickel content in their above-ground part (even though sometimes the whole plant or different parts are used instead) greater than the total mean content of said mineral in the soil.
3. Hyperaccumulator plants can store a metal quantity 10-500 times greater than that for “normal” plants.
4. Heavy metal level in the shoot (the above-ground part of the plant includes stem and leaves) is greater than that in the root.

The disadvantage for the first classification lies in the fact that, apart from being based on an arbitrary limit, it does not establish any nickel concentration threshold for plants deficient in this element; it only establishes a minimum of 100 ppm for accumulators and 1000 ppm for hyperaccumulators (Brooks *et al*., 1977). Regarding the other classifications, one of their main limitations is the disagreement concerning which part of the plant should be used since total element concentration is not a real measure of available minerals (Buscaroli, 2017).

These problems can be solved by resorting to data mining, which is the extraction of implicit, previously unknown, and potentially useful information in data (Witten *et al*., 2017). This multidisciplinary field combines works in machine learning, statistics, pattern recognition, and artificial intelligence (Han & Kamber, 2006). Machine learning provides a technical framework for data mining (Witten *et al*., 2017). It is used to extract information from raw data in databases such as the GHD mentioned above. The process is based on abstraction: data are collected, with all their defects, and the underlying structure is inferred (Witten *et al*., 2017).

Therefore, it is at this level that machine learning techniques can be applied: algorithms that are commonly used in the scientific sphere may be applied to plant heavy metal accumulation data to address this problem involving the management of a large amount of information. These algorithms can be applied *a priori* independently from the problem to address. Because they have already been used to deal with different issues, said algorithms offer an objective data analysis framework. Consequently, this study aims to evaluate the potential of machine learning methods to refine existing classifications of plants growing in extreme edaphic environments, with a special emphasis on serpentine soils.

Materials and Methods

1. Data sources

It was attempted to look for mineral composition data in specialized literature, yet, due to data heterogeneity, the Global Hyperaccumulator Database (http://hyperaccumulators.smi.uq.edu.au/collection/) was chosen as the preferential data source. It contains data for elements such as cobalt, copper, nickel, manganese, selenium or zinc, expressed as parts per million (ppm). If there were more than a single concentration value for a plant or mineral concentration were expressed as an interval, values would be averaged. In total, this database contains information about 721 plant species from all around the world, only two of which are found in Spain: *Alyssum malacitanum* and *Thlaspi* (*Noccaea*) *stenopterum*. Besides, other published data on the subject were used, consisting of mineral composition values for Spanish flora of dolomite, serpentine and gypsum soils (Martínez-Hernández, 2013; Medina-Cazorla, 2015). In total, this data source included 96 taxa (Appendix 1), mainly distributed throughout Spain in the case of gypsum and throughout the Baetic Mountains in the case of dolomite or serpentine soils.

2. Classification according to soil type

To carry out the first analysis, data were used from Martínez-Hernández *et al*. (2013) and Medina-Cazorla *et al*. (2015). Elements Mg, Fe, Mn, K, Cu, Ni, and Zn were selected, as well as soil type data for every plant: dolomite (D), gypsum (G), or serpentine (S). Only those plants with available data for every chosen element and the type of soil where they grew were used in the analysis; thus when some piece of information was lacking, the rest was excluded.

Soil type was used as the target variable for prediction. In contrast, the seven mineral composition attributes were the variables used to discern to which group (D, G, or S) the plant belonged. Therefore, this first approach was based on supervised learning. This means that the actual classification (the type of soil) was known beforehand, in contrast to unsupervised learning (Alphy & Sharma, 2020). The most commonly used algorithms were applied to these data because the aim is to assess machine learning performance when applied to this particular problem. To apply machine learning algorithms, Weka software, version 3.8.3, was used (Witten *et al*., 2017). Among the different available techniques, NaiveBayes (Naïve Bayes, NB), SMO (SVM), IBk (kNN), and J48 (C4.5) methods were selected to perform the analysis. These methods were chosen since they are the predominant state-of-the-art techniques in these different paradigms: statistical (NB), linear models with kernel (SVM), instance-based learning (kNN) and decision trees (C4.5) (Rooney *et al*., 2004). Naïve Bayes is a statistic classifier that predicts the probability of belonging to a certain class. It is based on Bayes theorem and on the assumption that the effect of one attribute's value on a specific class is independent of the values of other attributes (Witten *et al*., 2017). SVM is an algorithm that uses non-linear mapping to convert original data into a higher dimension. In this new space, SVM looks for the optimal hyperplane to divide data (i.e., the decision limit which separates one class from another). The algorithm finds this hyperplane by using support vectors (training tuples) and margins defined by said vectors (Han & Kamber, 2006; Xue *et al*., 2020). The kNN (k Nearest Neighbours) algorithm is based on learning by analogy. It compares test tuples to training tuples. N attributes define training tuples; thus, each one represents a point in an n-dimensional space. When a problem tuple is provided, the classifier looks for the nearest training tuples in the n-dimensional space (nearest neighbors). Lastly, the problem tuple is assigned to the most common class among its nearest neighbors (Han & Kamber, 2006). C4.5 is a decision tree based on the gain of information. That means that the tree ramifies in nodes so that each node has the highest possible percentage of data from every class (Quinlan, 1993).

For each algorithm, two kinds of experiments were carried out. Firstly, Weka's default parameters were used in order to obtain preliminary results. Secondly, different parameters were chosen to analyze how to optimize the method's predictive capabilities. Supervised discretization based on the standard method (Fayyad and Irani's MDL method) was used for NaiveBayes. This discretization transforms a range of numeric attributes into nominal attributes. For SMO, kernel, the function which transforms data into a higher-dimensional space, was changed. For IBk, what was optimized was the number of nearest neighbors used to calculate distances. Lastly, for J48, it was the confidence factor (Witten *et al*., 2017). A lower confidence factor results in an increment of the error attributed to each tree node. Consequently, nodes with a higher error value are discarded and the tree is simplified, which is 'pruned.' If the confidence factor were too high, there would be overfitting (Drazin & Montag, 2012)and sifts through to remove statistically insignificant nodes. Working from the bottom up, the probability (or relative frequency.

3. Clustering

GHD and published data for nickel (Martínez-Hernández, 2013; Medina-Cazorla, 2015) was used to discuss classification for Ni accumulator (between 100 and 1000 ppm) and hyperaccumulator (> 1000 ppm) plants (Brooks *et al*., 1977). It was compiled in a single file to be inputted in Weka. When there were several different data for a single species in the GHD or when data were expressed as intervals, the average was used. Additionally, before applying clustering algorithms, data were transformed by the function ln(x+1), were x is the original data, in order to correct data exponential distribution. Then, Ni data were divided into four clusters: low Ni content, normal content, high content (accumulators) and very high content (hyperaccumulator), according to Brooks *et al*. (1977). To this end, three algorithms based on different principles were applied to validate results: EM (Expectation Maximisation), HierarchicalClusterer (HC), and SimpleKMeans. EM assigns a probability distribution to every instance, which indicates the probability of the instance belonging to each cluster (Jung *et al*., 2014; Witten *et al*., 2017). EM can decide how many clusters to use by cross-validation (Witten *et al*., 2017), yet the number was fixed to four in this case, both in EM and in the rest of the algorithms used. Although fixing the number of clusters introduces statistical bias, this decision is justified by the aim of the study, which is not to establish a new classification for hyperaccumulator plants, but to refine the criteria in Brooks *et al*. (1977). HC implements classic hierarchical clustering agglomerative methods. It generates two initial clusters and evaluates whether it is worth dividing them again, which results in a hierarchy. K-means clustering works so that k initial points are chosen to represent the cluster center (centroid). Then, every data point is assigned to the nearest initial point, and the cluster point mean value becomes the new cluster centroid. This iteration is repeated until there are no changes (Jung *et al*., 2014; Witten *et al*., 2017). Two hierarchical clustering methods were used, SINGLE and WARD. SINGLE takes into account the minimal distance between any pair of data from different clusters, whereas WARD progressively merges clusters and computes the sum of squares (it begins at zero). The method tries to

maintain the growth of the sum of squares as small as possible (Witten *et al*., 2017).

4. Associations among minerals

In this case, published data for Mg, Fe, Mn, K, Cu, Ni and Zn and the type of soil where the plant grew (Martínez-Hernández, 2013; Medina-Cazorla, 2015) were used to search for associations between concentrations of different minerals or between these concentrations and the type of soil. The algorithm used in Weka was "Apriori", which transforms items into association rules based on coverage (number of items that fulfill the rules) and accuracy, i.e., this number is expressed as a proportion of total items (Witten *et al*., 2017). Accordingly, as there was more dolomite than serpentine or gypsum data, pre-processing was necessary to balance the three categories' influence. To this end, supervised filters "Resample" and "SpreadSubsample" were applied. "Resample" generates a random subsample using sampling with and without replacement. The supervised version was chosen because data have a nominal class attribute (the type of soil), and this attribute needed to be that determining the division of numerical data into discrete groups. Class distribution can be maintained or modified. "SpreadSubsample" produces a random subsample too, and it allows the user to control the frequency difference between the most common class and the rarest one. In addition, the amount of data to take from the original sample can also be specified (Witten *et al*., 2017). By using "Resample", class weight (distribution) was just made equal; in contrast, by using "SpreadSubsample", only 39 data points from every class were taken into account (the rest were discarded in the analysis) because the class with a lower number of data (serpentine) only had them for 39 different species in the file that was analysed. Accordingly, class weight also became equal for all three classes. In addition, different subsets of 39 data for dolomite and gypsum were used to assess whether the associations found were real, not merely resulting from the random data selection.

Lastly, since the algorithm did not work for numeric attributes, discretizing data was necessary to divide them into intervals. The unsupervised Discretize algorithm was used, which divides a range of attributes into a predetermined number of groups based on training data distribution (Witten *et al*., 2017). Cluster number was fixed to four to match the Ni existing classification, where plants were separated according to low, normal, high, and very high Ni levels.

5. Conservation status of hyperaccumulator flora

To gather information on the hyperaccumulator species contemplated in this study and to offer a global perspective on their threat degree, both the IUCN Red List (https://www.iucnredlist.org/) and National Red Lists were consulted, especially in the case of countries with rich hyperaccumulator flora (Appendix S1) such as New Caledonia (Wulff *et al*., 2013)we have made a first-pass quantitative assessment of the distribution of Narrow Endemic Species (NES, Cuba (González Torres *et al*., 2016), Turkey (Ekim *et al*., 1989) or Australia (http://anpsa.org.au/atrisk3.html).

Results and Discussion

1. Data sources: comments and gaps

The exploration of existing literature evidenced the need for a database to gather plant mineral composition data. In databases such as the GHD, there is only information about a specific mineral for each species. In specialized literature, there was a huge variation regarding the minerals which were determined, the part of the plant that was analyzed, the type of soil where it grew, and, in the case of soil data, whether the mineral composition was referred to as total or extractable elements (Buscaroli, 2017). The element for which there was more information, taking into account that the bibliography focused on serpentine plants, was Ni, both in specialized literature and in the GHD. This is logical because serpentine soils are characterised by their high Ni content (Rajakaruna *et al.*, 2009).

Other databases with large amounts of data, such as Watanabe *et al*. (2007), containing information about 2228 foliar samples from 670 species, barely include values for Ni and other heavy metals. As pointed out, heterogeneity and the fragmentary nature of available data hinder their analysis because there is a lack of analysis for plants that are not considered either hyperaccumulator or normal. To sum up, the initial information is strongly biased despite the efforts to include plants that grow on special soils, such as serpentine, even if the species are not considered hyperaccumulators.

The literature review also evidenced the lack of information regarding a plant's ability to accumulate radionuclides. In this case, information is also insufficient, and, for example, plant Sr content data are hardly ever available. This mineral is one of the most abundant in the earth's crust, and, concretely, its radioisotope ^{90}Sr, a subproduct of the rain following nuclear explosions, has a semi-disintegration period of 28.78 years. Said radioisotope represents an important health risk because it easily replaces bone calcium, hindering its removal. In the Chernobyl area, ^{90}Sr contributes significantly to radioactive contamination (Guillén *et al*., 2011). Phytoremediation also addresses this important aspect, but the approach seems different from that for heavy metals (Burger & Lichtscheidl, 2018).

Finally, regarding the information review carried out, a small number of works take account of plant productivity for different species when evaluating plant potential for phytoremediation. Productivity is the amount of biomass produced per unit of time (Monson, 2014). It is relevant since, for example, *Amaranthus retroflexus* (amaranth) is capable of extracting higher ^{90}Sr levels, although its bioconcentration factor is lower than the one for *Brassica juncea* (Indian mustard) or *Phaseolus acutifolius* (tepary bean), as Wang et al. (2017) pointed out. It means that *A. retroflexus*

extracts higher quantities of ^{90}Sr because of its faster growth, irrespective of whether the concentration of ^{90}Sr is higher in *B. juncea* or *P. acutifolius*. As a result, *A. retroflexus* might be a better candidate to be used for phytoremediation. Although there are some exceptions, such as the one mentioned before, the lack of works discussing this is remarkable (Al Chami *et al*., 2015).

2. Classification according to the type of soil

After applying algorithms with default Weka settings (Table 1), the percentage of success was 42.36% for NaiveBayes, 64.99% for SMO, 75.05% for IBk and 80.46% for J48. Table 1, which represents confusion matrices obtained for every algorithm, correctly classified plant data in the central diagonal (in red in the first one). Therefore, the success rate is the proportion of correct predictions over the total. However, the success rate does not determine how good the algorithm's performance has been. It is important to read confusion matrices as well, particularly the results for serpentine soils. For example, in the case of SMO, 64.80% of soil data are dolomite. Consequently, even though the success rate is above 50%, the algorithm's performance is poor.

Table 1. Confusion matrices after applying a, NaiveBayes; b, SMO; c, IBk; d, J48 algorithms. Columns indicate the classification made by the algorithm, whereas rows correspond to the real classification. Correctly classified instances are those that occupy the central diagonal. Abbreviations are: D, dolomite; G, gypsum; S, serpentine.

a) NaiveBayes				b) SMO			
a	b	c	← classified as	a	b	c	← classified as
71	259	5	a = D	333	1	1	a = D
13	129	1	b = G	143	0	0	b = G
12	8	19	c = S	36	0	3	c = S
c) IBk				d) J48			
a	b	c	← classified as	a	b	c	← classified as
279	50	6	a = D	299	31	5	a = D
58	84	1	b = G	58	84	1	b = G
10	4	25	c = S	2	4	33	c = S

In Table 1, as the success rate points out, the NaiveBayes classification did not perform well, but it made a distinction between serpentine and other groups better than SMO because it correctly identified 48.71% of serpentine soil data and 90.21% of gypsum soil data; in contrast, SMO did not correctly identify any gypsum and only 0.77% of serpentine data. Conversely, SMO correctly classified 99.40% of data from plants that grow on dolomite soils, whereas NaiveBayes only has a 21.19% of success rate for this kind.

IBk and J48 yielded good results, both regarding success rate and confusion matrices. IBk correctly classified 64.10% of serpentine data, which is an improvement in comparison with NaiveBayes. J48 yielded an even higher success rate, 84.62% for serpentine soils. For both NaiveBayes and J48, the worst success rates were obtained for gypsum soils (58.74%). In fact, the best results for plants that grow on serpentine soils are those obtained with NaiveBayes. However, it must be considered that this classifier assigned 259 dolomite data to the gypsum class. Because of the above, the best results are those obtained with J48, regarding both success rate and real ability to discriminate between plants depending on soil type.

As results were poor for NaiveBayes and SMO, some parameters were modified for all of the classifiers in order to achieve improved outcomes. For NaiveBayes, the option for supervised discretisation was selected. Thus, a success rate of 75.05% was achieved; however, the success rate for gypsum soils decreased from 90% to 48.25%. For SMO, kernel Puk was used instead of the default one, and a success rate of 77.76% was achieved. In this case, the success rate for dolomite soils remained high (94.33%), and, in turn, success rates for gypsum and serpentine increased. In the case of the IBk algorithm, the highest success rate (77.37%) was achieved when using a number of 9 k-nearest neighbors (KNN); however, after optimizing this parameter, the success rate for serpentine soils decreased to 28.21% and the percentage for gypsum decreased as well; consequently, even though the success rate increased, there was not a real classification improvement. Lastly, the best results for J48 were achieved using a confidence factor of 0.06, which led to a success rate of 81.62%.

As can be deduced from Table 2, for J48 the success rate increased and the ability to discern serpentine and gypsum soils remained unaffected; in fact, the only negative difference is that three serpentine data points were classified as dolomite, compared to two before lowering the confidence factor. This improvement in the results suggests that, by reducing the confidence factor, the effect of overfitting has been corrected.

A decision tree (Appendix S2) was obtained after applying the J48 algorithm. This tree shows classification rules. The number of data correctly classified are to be found inside the grey squares to the left, whereas exceptions are on the right side of the square. Those parameters which determine the first tree forks are the most relevant for the classification, whereas those that appear in subsequent bifurcations are less relevant. According to Appendix S2, the most relevant parameter is Ni concentration, in such a way that all serpentine data are associated with high

nickel levels, even though these levels are well below what is considered as accumulation or hyperaccumulation. Although Salmerón-Sánchez *et al.* (2014) obtained similar data for *Jurinea pinnata* in dolomite (with a foliar composition unusually high in Mg) and gypsum (a foliar composition unusually high in S), this phenomenon has not been reported in serpentine plants because "serpentinomics" has only addressed hyperaccumulation to date (Wright & Wettberg, 2009). However, some investigations have revealed the risks that Ni accumulation in plants, depending on soil composition, may pose to human health (Beygi & Jalali, 2019).

Table 2. Confusion matrices for a) NaiveBayes with supervised discretisation; b) SMO with Puk kernel; c) IBk with 9 KNN; d) J48 with a confidence factor of 0.06. Abbreviations are: D, dolomite; G, gypsum; S, serpentine. Columns indicate the classification made by the algorithm, whereas rows correspond to the real classification.

a) NaiveBayes				a) NaiveBayes			
a	b	c	← classified as	a	b	c	← classified as
294	29	12	a = D	316	13	6	a = D
73	69	1	b = G	83	59	1	b = G
10	4	25	c = S	8	4	27	c = S
c) IBk				d) J48			
a	b	c	← classified as	a	b	c	← classified as
322	11	2	a = D	304	27	4	a = D
76	67	0	b = G	57	85	1	b = G
22	6	11	c = S	3	3	33	c = S

From a methodological perspective, this second analysis also yielded better results with J48, which was able to balance the prediction for the three concepts to be learned (good performance for dolomite, gypsum, and serpentine). Moreover, it has an advantage over the other three algorithms: it shows the rules it uses to classify (in this case, rules can be found in Appendix S2 tree).

3. Clustering

Four different clusters were generated after applying the EM clustering and SimpleKMeans algorithms. In the case of EM, the first cluster (cluster 0 in Figure 1) ranges from the minimum Ni value (0.1 ppm) to 2.25 ppm, both included. Because plants in this group had the lowest Ni content, they were classified as plants with low Ni content. The following cluster ranges from 2.27 ppm to 103.23 ppm, also including both. Plants in this cluster (cluster 3 in Figure 1) have been considered normal plants regarding their Ni content. The next cluster (cluster 2 in Figure 1) comprises a Ni content from 208.75 to 3080 ppm. Plants in this group are considered to have high Ni levels, which makes them accumulators by analogy with Brooks *et al.* (1997) classification, according to which accumulator plants were those with more than 100 Ni ppm. Since the first datum below 208.75 ppm is 103.23 Ni ppm, the accumulator plant cluster result is close to the criteria established by the classification mentioned above. Lastly, the group ranges from 3131 ppm to the highest value, 65800 ppm (cluster 1 in Figure 1) comprises plants that accumulate very high nickel levels, namely hyperaccumulator plants for this element. In this case, the threshold above which a plant may be considered a hyperaccumulator is higher for clustering (>3000) than for Brooks' *et al.* (1977) classification, which proposes a level above 1000 ppm. Yet, both are in the same order of magnitude.

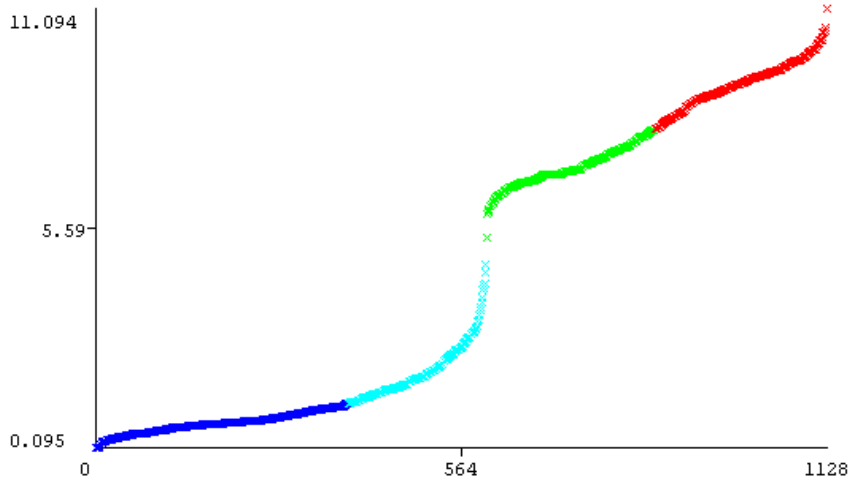

Figure 1. ln(x+1) clustering by EM, where x is Ni concentration data expressed in ppm. Four clusters can be observed: cluster 0 (blue), cluster 3 (cyan), cluster 2 (green) and cluster 1 (red). Weka gives the number of the cluster; it has no biological meaning. Y-axis shows the value of ln(x+1), whereas X-axis shows the number of the instance (the order of Ni datum in the original data file, where values were ordered from lowest to highest).

After applying the SimpleKMeans algorithm, similar results were obtained compared to EM: in this case, the first cluster, which corresponds to an interval from 0.1 to 3.7 Ni ppm, has a higher upper limit than that for EM. It corresponds to cluster 3 in Figure 2 and, as in the previous case, plants in this group were considered to contain low Ni levels. The following group, cluster 1 in Figure 2, contains values from 3.72 to 103.23 ppm; consequently, this group's upper limit is the same as that for EM. Cluster 1 comprises plants with normal Ni levels. The accumulator plant cluster is cluster 0 in Figure 2 which ranges from 208.75 ppm to 3573 ppm. Therefore, the upper limit is also slightly higher than the one for EM. However, because interval limits are similar for both SimpleKMeans and EM, this strengthens previously obtained results.

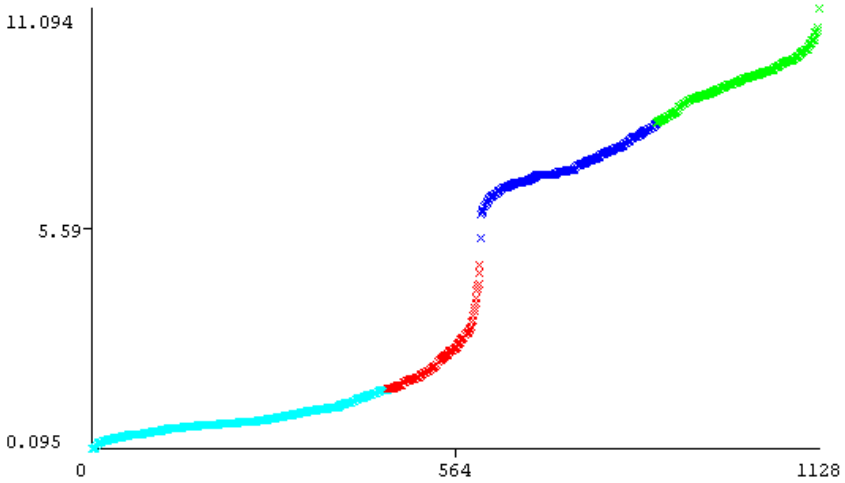

Figure 2. ln(x+1) clustering by SimpleKMeans, where x is Ni concentration data expressed in ppm. Four clusters can be observed: cluster 3 (cyan), cluster 1 (red), cluster 0 (blue) and cluster 2 (green). Y-axis shows the value of ln(x+1), whereas X-axis shows the number of the instance (the order of Ni datum in the original data file, where values were ordered from lowest to highest).

Lastly, the HierarchicalClusterer algorithm was applied. In this case, only two groups were obtained when using default settings, one from 0.1 ppm to 103.23 ppm (cluster 3 in Figure 3a) and another one from 380 ppm to 40875 ppm (cluster 2 in Figure 3a). Even though the turning point between the two groups (208.75) matches the result for EM and SimpleKMeans, that is to say, the threshold to set apart plants with a normal Ni content and hyperaccumulators, no more groups are formed because said turning point has been assigned as an independent cluster (cluster 1 in Figure 3a). Likewise, the maximum value was also assigned as a differentiated cluster (cluster 3 in Figure 3a).

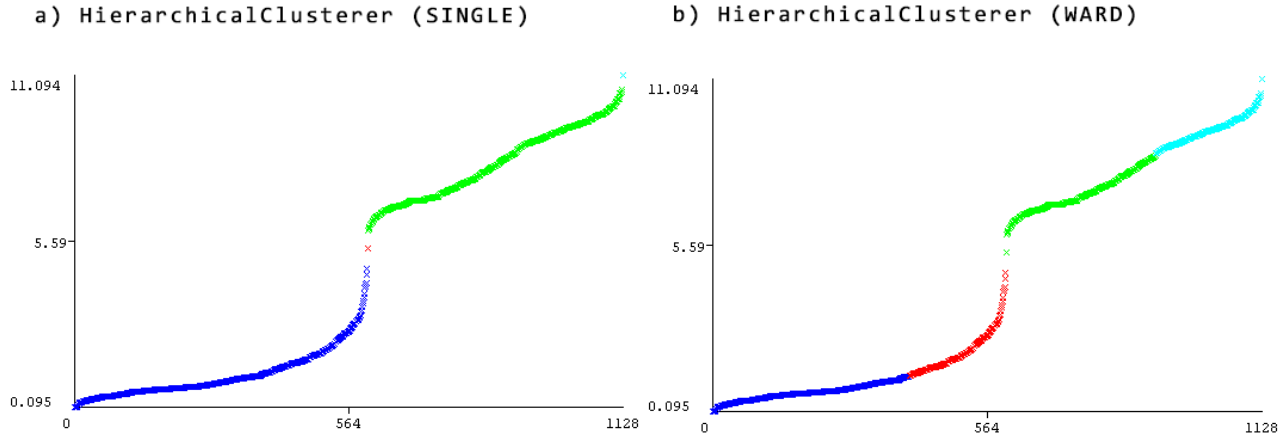

Figure 3. ln(x+1) clustering by HierarchicalClusterer (x=Ni concentration data expressed in ppm). Four clusters are observed: cluster 0 (blue), cluster 1 (red), cluster 2 (green) and cluster 3 (cyan). Y-axis shows the value of ln(x+1), whereas X-axis shows the number of the instance (the order of Ni datum in the original data file, where values were ordered from lowest to highest).

As there were two clear clusters instead of four, several methods were assessed in order to measure distance (dissimilarity). By default, Weka uses SINGLE method. The best results were attained with the WARD method: the first cluster (cluster 0 in Figure 3b), which comprises plants with low Ni levels, encloses the range from 0.1 to 2.46 ppm. The following cluster (cluster 1 in Figure 3b) includes the range from 2.51 to 103.23 ppm,

which corresponds to plants with normal Ni levels. As the figure shows, the turning point between plants with normal Ni levels and those with high levels does not vary depending on the classifier; in fact, the division between these two groups is visible at a glance in data graphs (Figures 1-3) because there are few data in the range from 103.23 ppm to 380 ppm. The interval for plants with high Ni levels encompasses values from 208.75 ppm to 5113 ppm (cluster 2 in Figure 3b). Therefore, in this case, the upper limit is above those obtained before, which were around 3000 ppm, but has the same order of magnitude as this previous result and Brooks' *et al*. (1997) classification (1000 ppm).

4. Associations among minerals

By applying "SpreadSubsample", a correlation was identified between intermediate/low Mg levels, high Ni levels, and the plant growing on serpentine soils. However, this association is not found when using "Resample". In addition, it is not consistent with the literature because, even though dolomite, serpentine and gypsum soils have a high Mg content, the highest Mg levels are found in serpentine soils (Berazain, 1999). Both with "SpreadSubsample" and "Resample", an association between high Zn and Ni levels and plants growing in serpentine soils is found. This correlation is indeed consistent with literature as serpentine soils are rich in Ni and in other minerals, among which is Zn (Rajkumar *et al*., 2009; Mohseni *et al*., 2019). A concurrence is also found between the highest Ni levels and serpentine soils. It is the most apparent association in data after discretizing and equalling variable weights (Figure 4). Lastly, there is a relation between the highest Mg levels and the highest K levels. Even though cations can compete when their concentrations are high, K can accumulate elements against concentration gradients (Taiz & Zeiger, 2010; Marschner, 2016).

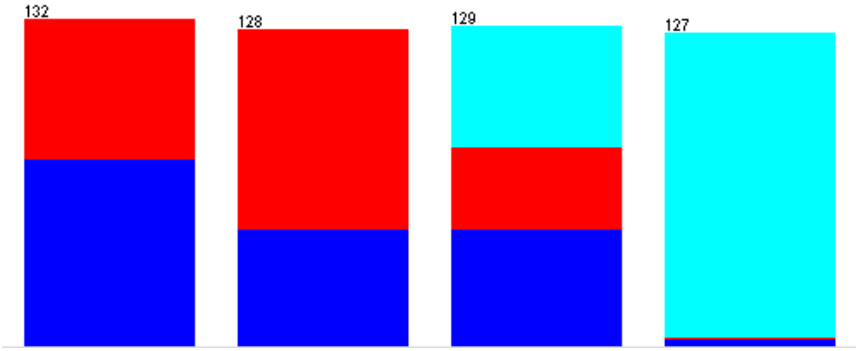

Figure 4. Visualization of Ni data after dividing them into four intervals with a similar number of data per interval (127−132), where Resample has equalized the weight for every class. The interval that corresponds to the lowest Ni content appears first from the left; intervals are ordered from left to right (Ni content increases to the right). This figure shows that, in the interval where nickel levels are the highest, most data correspond to serpentine soils (cyan), whereas only a few of them are from dolomite (blue), and the proportion of gypsum soils (red) is even lower.

5. Conservation status of hyperaccumulator flora

The recently created hyper-accumulating plant database represents an important advance in the study of this type of plant, not only from a scientific but also from an applied point of view. This database analysis shows that most known hyperaccumulator plants are concentrated in 6 or 7 regions of the planet (Prance & Brooks, 1988; Mengoni, Schat, & Vangronsveld, 2010). However, when these mega-diverse regions for hyperaccumulator flora are compared with the distribution of ultramafic soils (Echevarria, 2018), and is highly biased towards Ni hyperaccumulators. This is mainly due to the existence of a reagent paper test that is only specific to nickel (based on dimethylglyoxime, it is clear that many territories remain to be prospected. It is also true that recent research, for example, in Mexico, has not detected new hyperaccumulator taxa (Navarrete Gutiérrez *et al*., 2018). On the contrary, the Central American zone has found new hyperaccumulator species of the genus *Psychotria* (McCartha *et al*., 2019). Be that as it may, the information about these species on foliar contents in Ni and other heavy metals, accumulators, or not provides valuable insight that will help establish hyperaccumulation thresholds in a more objective way.

A relevant aspect of hyperaccumulator species that is not included in the database is related to their threat degree. Very few of the database species are listed in the IUCN red list (https://www.iucnredlist.org/). Thus, none of the European species appear on that list or are only considered as DD, which shows that a greater effort should be made to ascertain their conservation status. This is the case for *Alyssum akamasicum* and *A. pintasilvae* (Bilz *et al*., 2011). A review of other information sources, such as Cuba's red list, shows that some of these species are seriously endangered. In fact, the xeromorphic thickets on the serpentine house, together with the montane rainforest (González Torres *et al*., 2016), have the greatest number of threatened species, among which more than a hundred are critically endangered (CR). Out of the 129 species featured in the GHD, 14 appear in the Cuban red list.

Regarding Australia, the GHD features 14 species, two of which belong to the genus *Gossia* (*G.*

fragrantissima and *G. gonoclada*) and are considered EN (DSEWPC, 2009). In Turkey, one of the countries with the highest number of species in the GHD (Appendix 1), ten species face some threat degree: 1 CR, 6 EN and 4 VU. In other countries such as New Caledonia, the richest in terms of known hyperaccumulator species (Appendix 1), the red list is being updated (http://endemia.nc/page/la-liste-rouge). However, most species listed as hyperaccumulators are endemisms, of which Wulff *et al*. (2013)we have made a first-pass quantitative assessment of the distribution of Narrow Endemic Species (NES gather 35 narrow endemism present in 1, 2, or 3 localities. Analyzing data from New Caledonia, Cuba, Australia, and Turkey together, the proportion of hyperaccumulator flora that is endangered ranges between 10 and 20%. In New Caledonia, Ni mining affects many serpentinicolous plants that are not protected by local legislation (Wulff *et al*., 2013)we have made a first-pass quantitative assessment of the distribution of Narrow Endemic Species (NES. This datum is relevant to urge conservation of these species, and it accords that at least 14.9% of known hyperaccumulator flora is represented by local endemisms (Reeves *et al*., 2017).

Conclusions

A thorough literature review revealed that available foliar composition data are fragmentary and incomplete, not suitable for conducting global analyses. Even though economics is becoming increasingly important, different approaches (agronomical, ecological, functional, phytoremediation-focused, etc.) that focus on concrete aspects while disregarding others prevail. The case of elements with radioactive isotopes, such as Sr, is key to advancing the field of phytoremediation, taking into account the relevance of the last nuclear accidents (Chernobyl, Fukushima), with an almost global influence scope.

Machine learning algorithms are useful tools to analyze mineral composition data, as this first approach to the problem has proven. However, databases compiling multivariate plant composition data, as well as soil type, are needed in order to carry out more complex analyses. Analysing mineral content, soil type and plant nutritional strategies (hyperaccumulator, accumulator or normal) will elicit interesting results, given that there are certain types of soil, such as serpentine, where there is a higher proportion of hyperaccumulator plants. Therefore, confirming whether these associations are found in plants is interesting since soil mineral content affects plant mineral composition and, as this work has proven, plants can be classified according to their mineral composition depending on the soil where they grow.

On another note, a binary classification into accumulator and non-accumulator plants based only on plant mineral composition has its limitations; for example, it does not consider the effect of soil mineral content. The prevailing classification, which defines accumulators as plants which contain more than 100 Ni ppm and hyperaccumulators as plants which contain more than 1000 Ni ppm, is consistent with results obtained in this work, although said results suggest that a higher limit (3000-5000 Ni ppm) should be established to define hyperaccumulator plants. However, because the number of clusters was fixed according to the existing classification, optimizing it is the next step to devise a new classification or further refine established ones.

The appliance of machine learning techniques to multivariate plant mineral composition data, taking into account all three aspects (classification, clustering, and association) combined could provide relevant information which is not explicit in the dataset, and also improve hyperaccumulator plant classification, not only taking account of leaf mineral content, but also the criteria established by the four definitions of a hyperaccumulator, both for Ni and other relevant minerals. Thus, more hyperaccumulator plants could be identified to be used in phytoremediation and phytomining, with consequent biotechnological impact.

Information on the threat degree of hyperaccumulator species is incomplete even in territories that have been widely studied from a botanical point of view, such as Europe. Although this information is not available, the fact that this type of flora is often associated with very restricted habitats and territories suggests that they are endangered species in many cases. According to the authors' estimations, at least 10% of these plants are endangered. The applications that hyperaccumulators may have in phytoremediation or even in the research on other aspects of plant nutrition encourage greater efforts to evaluate their conservation status.

References

Alloway, B.J. 2013. Heavy Metals in Soils. In: Alloway, B.J. & Trevors, J.T. (Eds.). Environmental Pollution, vol. 22, 3rd ed. Pp. 195–209. Springer, Dordrecht.

Alphy, M. & Sharma, A. 2020. A literature review on different types of machine learning methods in web mining. Int. J. Psychosoc. Rehabil. 24(1): 1761–1769. doi: 10.37200/IJPR/V24I1/PR200276.

Batool, S. 2018. Effect of nickel toxicity on growth, photosynthetic pigments and dry matter yield of Cicer arietinum L. varieties. https://www.asianjab.com/wp-content/uploads/2018/06/2.-OK_Effect-of-nickel-toxicity-on-growth-photosynthetic-pigments.pdf.

Berazain, R. 1999. Estudios en plantas acumuladoras e hiperacumuladoras de níquel en el Caribe. Rev del Jardín Botánico Nac. 20:17–30. doi: 10.2307/42597044.

Beygi, M. & Jalali, M. 2019. Assessment of trace elements (Cd, Cu, Ni, Zn) fractionation and bioavailability in vineyard soils from the Hamedan, Iran. Geoderma. 337: 1009–1020. doi: 10.1016/j.geoderma.2018.11.009.

Bilz, M., Kell, S.P., Maxted, N. & Lansdown, R.V. 2011. European Red List of Vascular Plants. [accessed 2020 Feb 7]. www.tasamim.net.

Brooks, R.R., Lee, J., Reeves, R.D. & Jaffre, T. 1977. Detection of nickeliferous rocks by analysis of herbarium specimens of indicator plants. J.

Geochem. Explor. 7(C): 49–57. doi: 10.1016/0375-6742(77)90074-7.

Burger, A. & Lichtscheidl, I. 2018. Strontium in the environment: Review about reactions of plants towards stable and radioactive strontium isotopes. Sci. Total Environ. 653:1458–1512. doi: 10.1016/j.scitotenv.2018.10.312.

Buscaroli, A. 2017. An overview of indexes to evaluate terrestrial plants for phytoremediation purposes (Review). Ecol. Indic. 82:367–380. doi:10.1016/j.ecolind.2017.07.003. doi: 10.1016/j.ecolind.2017.07.003.

Al Chami, Z., Amer, N., Al Bitar, L. & Cavoski, I. 2015. Potential use of Sorghum bicolor and Carthamus tinctorius in phytoremediation of nickel, lead and zinc. Int. J. Environ. Sci. Technol. 12(12): 3957–3970. doi: 10.1007/s13762-015-0823-0.

Corzo Remigio, A., Chaney, R.L., Baker, A.J.M., Edraki, M., Erskine, P.D., Echevarria, G. & van der Ent, A. 2020. Phytoextraction of high value elements and contaminants from mining and mineral wastes: opportunities and limitations. Plant Soil 449(1–2): 11–37. doi: 10.1007/s11104-020-04487-3.

Drazin, S. & Montag, M. 2012. Decision Tree Analysis using Weka. Project report. 3p.

Echevarria, G. 2018. Genesis and Behaviour of Ultramafic Soils and Consequences for Nickel Biogeochemistry. In: Van der Ent, A., Echevarria, G., Baker, A. & Morel, J. (Eds.). Agromining: Farming for Metals. Pp. 135–156. Mineral Resource Reviews. Springer, Cham.

Ekim, T., Koyuncu, M., Vural, M., Duman, H., Aytaç, Z. & Adigüzel, N. 1989. Red data book of Turkish plants. (Pteridophyta and Spermatophyta). Turkish Assoc. Conserv. Nature, Ankara.

Faucon, M.P., Meersseman, A., Shutcha, M.N., Mahy, G., Luhembwe, M.N., Malaisse, F. & Meerts, P. 2010. Copper endemism in the congolese flora: A database of copper affinity and conservational value of cuprophytes. Plant Ecol. Evol. 143(1): 5–18. doi: 10.5091/plecevo.2010.411.

González Torres, L.R., Palmarola, A., González Oliva, L., Bécquer, E.R., Testé, E. & Barrios, D. 2016. Lista Roja de la Flora de Cuba. Bissea. 10(1): 352. doi: 10.13140/RG.2.2.24056.65288.

Guillén, F.J., Baeza, A. & Salas, A. 2011. Strontium. In: Atwood, D. (Ed.). Radionuclides in the environment. Pp. 1-17. Encyclopedia of Inorganic and Bioinorganic Chemistry. John Wiley & Sons Ltd., Chichester. http: // dx.doi.org/10.1002/9781119951438.eibc0416.

Han, J., Kamber, M. & Pei, J. 2006. Data Mining: Concepts and Techniques, 3rd ed. Elsevier, Amsterdam.

Jung, Y.G., Kang, M.S. & Heo, J. 2014. Clustering performance comparison using K-means and expectation maximization algorithms. Biotechnol. Biotechnol. Equip. 28(1):S44–S48. doi:10.1080/13102818.2014.949045. doi: 10.1080/13102818.2014.949045.

Khan, A., Khan, S., Khan, M.A., Qamar, Z. & Waqas, M. 2015. The uptake and bioaccumulation of heavy metals by food plants, their effects on plants nutrients, and associated health risk: a review. Environ. Sci. Pollut. R. 22(18): 13772–13799. doi: 10.1007/s11356-015-4881-0.

Kidd, P.S., Becerra Castro, C., Garcia Lestón, M. & Monterroso, C. 2007. Aplicación de plantas hiperacumuladoras de níquel en la fitoextracción natural: el género Alyssum L. Ecosistemas 2(2): 1–18.

Marschner, H. 2016. Marschner's Mineral Nutrition of Higher Plants. 3rd ed. Academic Press.

Martínez-Hernández, F. 2013. Patrones biogeográficos de la flora gipsícola ibérica. Mem. Doc. (ined.). Universidad de Almería.

McCartha, G.L., Taylor, C.M., van der Ent, A., Echevarria, G., Navarrete Gutiérrez, D.M. & Pollard, A.J. 2019. Phylogenetic and geographic distribution of nickel hyperaccumulation in neotropical Psychotria. Am. J. Bot. 106(10): 1377–1385. doi: 10.1002/ajb2.1362.

Medina-Cazorla, J.M. 2015. Conservación y biogeografía de la flora dolomitófila bética. Mem. Doc. (ined.). Universidad de Almería.

Mengoni, A., Schat, H. & Vangronsveld, J. 2010. Plants as extreme environments? Ni-resistant bacteria and Ni-hyperaccumulators of serpentine flora. Plant Soil. 331(1): 5–16. doi: 10.1007/s11104-009-0242-4.

Mganga, N., Manoko, M. & Rulangaranga, Z. 2011. Classification of Plants According to Their Heavy Metal Content around North Mara Gold Mine, Tanzania: Implication for Phytoremediation. Tanzania J. Sci. 37(1): 109–119.

Mohseni, R., Ghaderian, S.M. & Schat, H. 2019. Nickel uptake mechanisms in two Iranian nickel hyperaccumulators, Odontarrhena bracteata and Odontarrhena inflata. Plant Soil 434(1–2): 263–269. doi: 10.1007/s11104-018-3814-3.

Monson, R.K. (Ed.). 2014. Ecology and the environment. Springer-Verlag, New York.

Navarrete Gutiérrez, D.M., Pons, M.N., Cuevas Sánchez, J.A. & Echevarria, G. 2018. Is metal hyperaccumulation occurring in ultramafic vegetation of central and southern Mexico? Ecol. Res. 33(3): 641–649. doi: 10.1007/s11284-018-1574-4.

Okereafor, U., Makhatha, M., Mekuto, L., Uche-Okereafor, N., Sebola, T. & Mavumengwana, V. 2020. Toxic metal implications on agricultural soils, plants, animals, aquatic life and human health. Int. J. Environ. Res. Pu. 17(7): 1–24. doi: 10.3390/ijerph17072204.

Prance, G.T. & Brooks, R.R. 1988. Serpentine and Its Vegetation. A Multidisciplinary Approach. Brittonia 40(3): 268. doi: 10.2307/2807470.

Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufmann Publishers.

Rajakaruna, N., Harris, T.B. & Alexander, E.B. 2009. Serpentine Geoecology of Eastern North America: A Review. Rhodora 111(945): 21–108. doi: 10.3119/07-23.1.

Rajkumar, M., Prasad, M.N.V., Freitas, H. & Ae, N. 2009. Biotechnological applications of serpentine soil bacteria for phytoremediation of trace metals. Crit. Rev. Biotechnol. 29(2): 120–130. doi: 10.1080/07388550902913772.

Reeves, R.D., Baker, A.J.M., Jaffré, T., Erskine, P.D., Echevarria, G. & van der Ent, A. 2017. A global database for plants that hyperaccumulate metal and

metalloid trace elements. New Phytol. 218(2): 407–411. doi: 10.1111/nph.14907.

Rooney, N., Patterson, D. & Galushka, M. 2004. A comprehensive review of recursive Naïve Bayes Classifiers. Intell. Data Anal. 8(6): 615–628. doi: 10.3233/ida-2004-8607.

Rostami, S. & Azhdarpoor, A. 2019. The application of plant growth regulators to improve phytoremediation of contaminated soils: A review. Chemosphere 220: 818–827. doi: 10.1016/j.chemosphere.2018.12.203.

Salmerón-Sánchez, E., Martínez-Nieto, M.I., Martínez-Hernández, F., Garrido-Becerra, J.A., Mendoza-Fernández, A.J., de Carrasco, C.G., Ramos-Miras, J.J., Lozano, R., Merlo, M.E. & Mota, J.F. 2014. Ecology, genetic diversity and phylogeography of the Iberian endemic plant Jurinea pinnata (Lag.) DC. (Compositae) on two special edaphic substrates: dolomite and gypsum. Plant Soil. 374(1–2): 233–250. doi: 10.1007/s11104-013-1857-z.

Shabala, S. 2013. Plant stress physiology. Choice Rev. Online 50(05): 50–2652. doi: 10.5860/choice.50-2652.

Shah, V. & Daverey, A. 2020. Phytoremediation: A multidisciplinary approach to clean up heavy metal contaminated soil. Environ. Technol. Innov. 18: 100774. doi: 10.1016/j.eti.2020.100774.

Taiz, L. & Zeiger, E. 2010. Plant physiology. Sinauer Associates, Sunderland.

Wang, X., Chen, C., Wang, J. 2017. Phytoremediation of strontium contaminated soil by Sorghum bicolor (L.) Moench and soil microbial community-level physiological profiles (CLPPs). Environ. Sci. Pollut. Res. 24(8): 7668–7678. doi: 10.1007/s11356-017-8432-8.

Watanabe, T., Broadley, M.R., Jansen, S., White, P.J., Takada, J., Satake, K., Takamatsu, T., Tuah, S.J. & Osaki, M. 2007. Evolutionary control of leaf element composition in plants: Rapid report. New Phytol. 174(3): 516–523. doi: 10.1111/j.1469-8137.2007.02078.x.

Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. 2017. Data Mining - Practical Machine Learning Tools and Techniques, 4th ed. Kaufmann ed. Elsevier Inc, Cambridge.

Wright, J.W. & Wettberg, E. von. 2009. "Serpentinomics" - An Emerging New Field of Study. Northeast Nat. 16(sp5): 285–296. doi: 10.1656/045.016.0521.

Wulff, A.S., Hollingsworth, P.M., Ahrends, A., Jaffré, T., Veillon, J.M., L'Huillier, L. & Fogliani B. 2013. Conservation Priorities in a Biodiversity Hotspot: Analysis of Narrow Endemic Plant Species in New Caledonia. PLoS One. 8(9). doi: 10.1371/journal.pone.0073371.

Xue, H., Xu, H., Chen, X. & Wang, Y. 2020. A primal perspective for indefinite kernel SVM problem. Front. Comput. Sci-Chi. 14(2): 349–363. doi: 10.1007/s11704-018-8148-z.

Websites

DSEWPC. 2009. EPBC Act List of Threatened Flora. SPRATT Profile. [accessed 2020 Jun 7]. http://www.environment.gov.au/cgi-bin/sprat/public/publicthreatenedlist.pl?wanted=flora#flora_critically_endangered.

Supplementary Material

Appendix S1a. Number of species per country in the GHD.

Appendix S1b. Number of species per family in the GHD.

Appendix S2. Decision tree obtained after applying J48 algorithm on data with a confidence factor of 0.06.

Appendix 1. List of species present in Martínez-Hernández (2013) and Medina-Cazorla (2015) plant mineral composition data.

Abies pinsapo Boiss.
Alyssum gadorense P. Küpfer
Alyssum serpyllifolium Desf.
Andryala agardhii Haens. & Boiss.
Anthyllis cytisoides L.
Anthyllis tejedensis Boiss. subsp. *plumosa* (Cullen ex E. Domínguez) Benedí
Anthyllis montana L.
Anthyllis tejedensis Boiss
Anthyllis terniflora (Lag.) Pau
Anthyllis vulneraria L.
Arctostaphylos uva-ursi (L.) Spreng
Boleum asperum Desv.
Brassica repanda (Willd.) DC.
Centaurea bombycina Boiss. subsp. *bombycina*
Centaurea granatensis DC
Centaurea hyssopifolia Vahl.
Cistus albidus L.
Cistus clusii Dunal in DC.
Cistus populifolius L.
Cistus salvifolius L.
Convolvulus boissieri Steud.
Coris hispanica Lange
Dittrichia viscosa (L.) Greuter
Echium albicans Lag. & Rodr. subsp. *albicans*
Erica scoparia L
Erodium boissieri Cosson
Glandora nitida (Ern) D.C. Thomas
Globularia spinosa L.
Gypsophila bermejoi G. López
Gypsophila struthium L. subsp. *hispanica* (Willk.) G. López
Gypsophila struthium L. subsp. struthium
Halimium atriplicifolium (Lam.) Spach. subsp. *atriplicifolium*
Hedysarum boveanum Bunge ex *Basiner* subsp. *palentinum* Valdés
Helianthemum alypoides Losa & Rivas Goday
Helianthemum apenninum subsp. *stoechadifolium* (Brot.) Samp
Helianthemum canum (L.) Hornem
Helianthemum marifolium subsp. *conquense* Borja & Rivas Goday ex G. López
Helianthemum pannosum Boiss
Helianthemum raynaudii Ortega Oliv., Romero García & C. Morales
Helianthemum squamatum (L.) Dum. Cours.
Jurinea humilis (Desf.) DC.
Jurinea pinnata (Lag. ex Pers.) DC
Krascheninnikovia ceratoides (L.) Gueldenst
Lavandula lanata Boiss.
Lavandula latifolia Medik.
Leontodon boryi Boiss. ex DC.
Lepidium subulatum L.
Leucanthemopsis pallida subsp. *spathulifolia* (J. Gay) Heywood
Lomelosia pulsatilloides (Boiss.) Greuter & Burdet subsp. *pulsatilloides*
Ononis tridentata L subsp. *tridentata*
Ononis tridentata subsp. *angustifolia* (Lange) Devesa & G. López
Ononis tridentata L. subsp. *crassifolia* (Dufour ex Boiss.) Nyman
Ononis tridentata f. *edentula* (Webb ex Willk.) Irj.
Pinus halepensis Miller
Pinus nigra Arnold
Pinus pinaster Aiton
Pinus sylvestris L.
Pterocephalus spathulatus (Lag.) Coult.
Quercus faginea Lam. subsp. *faginea*
Quercus rotundifolia Lam
Rosmarinus eriocalyx Jordan & Fourr.
Rosmarinus officinalis L.
Rothmaleria granatensis (DC.) Font Quer
Salvia lavandulifolia Vahl.
Santolina elegans DC
Satureja montana subsp. *montana*
Scorzonera albicans Cosson
Sedum album L.
Sedum sediforme (Jacq.) Pau
Sideritis spinulosa Barnades ex Asso subsp. *spinulosa*
Sideritis incana subsp. *virgata* (Desf.) Malag.
Staehelina baetica DC
Teucrium balthazaris Sennen
Teucrium carolipaui Vicioso subsp. *fontqueri* (Sennen) Rivas Mart.
Teucrium lepicephalum Pau
Teucrium libanitis Schreb.
Teucrium pumilum Loefl. ex L. Cent.
Teucrium turredanum Losa & Rivas Goday
Thymelaea tartonraira (L.) All. subsp. *valentina* (Pau) O. Bolòs & Vigo
Thymus funkii Coss. subsp. *sabulicola* (Coss.) R. Morales

Helianthemum syriacum (Jacq.) Dum. Cours.

Helictotrichon filifolium (Lag.) Henrard. subsp. *filifolium*

Helianthemum viscidulum Boiss

Herniaria fruticosa L.

Hypericum ericoides L.

Hippocrepis squamata (Cav.) Coss.

Hormathophylla lapeyrousiana (Jord.) P. Küpfer

Jacobaea auricula (Bourg. ex Coss.) Pelser

Jasione crispa (Pourr.) Samp. subsp. *segurensis* Mota, C. Díaz, Gómez Merc. & F. Valle

Thymus granatensis Boiss. subsp. *granatensis*

Thymus granatensis Boiss. subsp. *micranthus* (Willk.) O. Bolòs & Vigo

Thymus lacaitae Pau

Thymus mastichina L.

Trisetum velutinum Boiss.

Vella pseudocytisus L. subsp. *pseudocytisus*

Ziziphora hispanica L.

Jurinea humilis (Desf.) DC.

Jurinea pinnata (Lag. ex Pers.) DC