International Journal of Geographical Information Science, 2014
Vol. 28, No. 7, 1336–1354, http://dx.doi.org/10.1080/13658816.2014.885527

Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain

V.F. Rodriguez-Galiano[a]*, M. Chica-Olmo[a] and M. Chica-Rivas[b]

[a]*Departamento de Geodinámica, Universidad de Granada, Granada, Spain;* [b]*Departamento de Análisis Matemático, Universidad de Granada, Granada, Spain*

(*Received 6 April 2013; final version received 14 January 2014*)

Mineral exploration activities require robust predictive models that result in accurate mapping of the probability that mineral deposits can be found at a certain location. Random forest (RF) is a powerful machine data-driven predictive method that is unknown in mineral potential mapping. In this paper, performance of RF regression for the likelihood of gold deposits in the Rodalquilar mining district is explored. The RF model was developed using a comprehensive exploration GIS database composed of: gravimetric and magnetic survey, a lithogeochemical survey of 59 elements, lithology and fracture maps, a Landsat 5 Thematic Mapper image and gold occurrence locations. The results of this study indicate that the use of RF for the integration of large multisource data sets used in mineral exploration and for prediction of mineral deposit occurrences offers several advantages over existing methods. Key advantages of RF include: (1) the simplicity of parameter setting; (2) an internal unbiased estimate of the prediction error; (3) the ability to handle complex data of different statistical distributions, responding to nonlinear relationships between variables; (4) the capability to use categorical predictors; and (5) the capability to determine variable importance. Additionally, variables that RF identified as most important coincide with well-known geologic expectations. To validate and assess the effectiveness of the RF method, gold prospectivity maps are also prepared using the logistic regression (LR) method. Statistical measures of map quality indicate that the RF method performs better than LR, with mean square errors equal to 0.12 and 0.19, respectively. The efficiency of RF is also better, achieving an optimum success rate when half of the area predicted by LR is considered.

Keywords: mineral potential mapping; data-driven models; random forest; mineral prospectivity; mineral exploration

1. Introduction

The demand of mineral resources and, hence, the need to find new deposits and map them have increased significantly in recent years. This is a consequence of the industrial development of emerging countries, but it is particularly a consequence of the development of China, India and Brazil (Moon and Evans 2006, Varet 2012). The exploration and mapping of potential mineral deposits require the application of diverse methods and techniques, based on the geological and mining knowledge of the region under investigation and on the application of predictive models of mineral potential. This term, mineral potential, refers to the probability that mineral deposits of the type sought can be found at

*Corresponding author. Email: vrgaliano@ugr.es a certain location (Carranza 2008, 2011). It is obvious that a large part of the mineral resources available near the Earth's surface have already been found. Therefore, new searches must be more complex in principle, and hence will require the application of even more sophisticated spatial data analysis techniques than those available in the present literature (Moon and Evans 2006).

Within the context of mining exploration, GIS has become an essential tool to support decision-making in mineral exploration (Bonham-Carter 1994, Carranza 2008). With the aim of deciding which areas are favourable for exploration, large volumes of data that have been collected cannot be effectively analysed without an adequate and efficient spatial data management system, i.e., without a GIS. In this sense, GIS can be used to capture, store, organize, consult, handle, transform, analyse and integrate geospatial information from different sources (Carranza 2008). From all these GIS functions, information analysis and integration are paramount, as the final aim is to elaborate predictive spatial models that allow for the incorporation and combination of relevant variables related to the occurrence of mineral deposits.

Several numerical methods have been devised for the district-scale mapping of mineral potential. These can be categorized into knowledge-driven and data-driven types, depending on the nature of the inference procedure used. Knowledge-driven models use subjective evidence based on expert knowledge of processes that might have led to formation of mineral deposits in a given geological setting, where no or very few mineral deposits are known to occur (Carranza 2008, Abedi *et al*. 2013). Data-driven models use objective evidence based on the associations between evidential features and known deposit locations (Carranza 2011). For a deeper review of the different mineral mapping methods, please refer to Carranza (2011).

Within data-driven models, statistical multivariate methods exist, such as the discriminant analysis (Chung 1977, Harris 2003, Carranza 2008) and logistic regression (LR) methods (Chung 1978, Carranza and Hale 2001, Fallon *et al*. 2010, Porwal *et al*. 2010a, Cuihua *et al*. 2011), and a set of methods known as artificial intelligence or machine learning (Singer and Kouda 1996, Porwal *et al*. 2003, Rigol-Sanchez *et al*. 2003, Pereira Leite *et al*. 2009a, 2009b, Lewkowski *et al*. 2010, Oh and Lee 2010, Porwal *et al*. 2010b, Zuo and Carranza 2011). Multivariate statistical techniques, together with a machine learning technique, called neural networks, have probably been the most widely used in this context. The main reason that this set of techniques has been more widely used is its greater accessibility, as these techniques are included within different packages for ArcView and ArcGIS software (ESRI, Redlands, CA, USA) such as ArcWofE (Kemp *et al*. 1999), MI-SDM v2.51 (Avantra Geosystems 2006) and ArcSDM9.3 and ArcSDM10 (Sawatzky *et al*. 2009). However, these techniques show a variety of problems, such as their sensibility towards outlier values of LR (Hastie *et al*. 2009) and the opacity of neural networks (Carranza 2011).

New machine learning methods have been proposed in recent years to solve some of the problems described above. For instance, ensembles of trees have recently emerged and are receiving highlighted interest in other fields of knowledge (Hansen and Salamon 1990, Krogh and Vedelsby 1995, Friedl *et al*. 1999, Steele 2000, Gislason *et al*. 2006, Sesnie 2008, Ghimire *et al*. 2010). Ensemble learning algorithms use the same base algorithm to produce repeated multiple predictions, which are averaged in order to produce a unique model (Friedl *et al*. 1999, Breiman 2001). Ensemble learning techniques are supposed to have higher accuracy than other machine learning algorithms, because the set of predictions performs more accurately than any single one, and utilizes the strengths of the individual set of regressions, while, at the same time, the weaknesses are circumvented (Kotsiantis and Pintelas 2004, Ghimire *et al*. 2010). An ensemble learning technique called random forests (RFs) is increasingly being applied in land-cover classification from remotely sensed data (Pal 2005, Lawrence *et al.* 2006, Chan and Paelinckx 2008, Sesnie *et al.* 2008, Ghimire *et al.* 2010, Peters 2011) and other fields related to GIS (Vincenzi 2011, Huang *et al.* 2012, Laborte *et al.* 2012, McGinnis and Kerans 2013). RF offers a new approach to the problem of mineral prospectivity mapping, as it is relatively robust to outliers and it can overcome the 'black-box' limitations of artificial neural networks, assessing the relative importance of the variables. At the same time, the parametrization of RF is very simple and it is computationally lighter than other machine learning methods (neural networks or support vector machines) (Rodriguez-Galiano and Chica-Rivas 2012). Although RF is currently being used as a remote-sensing data classifier (Rodriguez-Galiano *et al.* 2012a, 2012b, Rodriguez-Galiano and Chica-Olmo 2012), its potential as a spatial modelling tool is still underexplored due to its novelty.

An RF model was conceived for gold potential mapping in the Rodalquilar gold mining district (Spain) using a comprehensive exploration GIS database. This locality is favourable to pilot studies, given the abundant information and the previous published works that make a reasonable database for comparison of results and robustness of the methodology. Chica-Olmo *et al.* (2002) developed a mineral exploration decision support system for gold potential mapping in the Rodalquilar–San Jose districts. Rigol-Sanchez *et al.* (2003) proposed an artificial neural network model for gold prospectivity mapping in the Rodalquilar district. Carranza *et al.* (2008) proposed a new hybrid model based on evidential belief functions. Debba *et al.* (2009) demonstrated a new methodology for deriving optimal exploration target zones in the Rodalquilar district. The potential of RF for generating a gold potential map is assessed considering multiple criteria related to variations in the algorithm parameters and the accuracy of the gold potential maps. The performance of the RF is also evaluated in comparison to the LR method.

2. Random forest

RF is an ensemble method which combines multiple decision tree algorithms to produce repeated predictions of the same phenomenon. Decision trees can be divided into classification trees and regression trees (RTs). The main objective of this study is to predict gold potential, so only regression mode will be presented in this section. A RT represents a set of restrictions or conditions which are hierarchically organized, and which are successively applied from a root to a terminal node or leaf of the tree (Breiman 1984, Quinlan 1993). In order to induce the RT, recursive partitioning and multiple regressions are carried out from the data set. From the root node, the data splitting process in each internal node of a rule of the tree is repeated until a stop condition previously specified is reached. Each of the terminal nodes, or leaves, has attached to it a simple regression model which applies in that node only. Once the tree's induction process is finished, pruning can be applied with the aim of improving the tree's generalization capacity by reducing its structural complexity. The number of cases in nodes can be taken as pruning criteria.

As described by Breiman *et al.* (1984), the induction of the RT involves first selecting optimal splitting measurement vectors. The process starts by splitting the dependent variable, or the parent node (root), into binary pieces, where the child nodes are 'purer' than the parent node. Through this process, the RTs search through all candidate splits to find the optimal split, s^*, that maximizes the 'purity' of the resulting tree, as defined by the largest decrease in the impurity.

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (1)$$

In this equation, s is the candidate split at node t, and the node t is divided by s into the left child node t_L with a proportion of p_L, and right child node t_R with a proportion of p_R. $i(t)$ is a measure of impurity before splitting, $i(t_L)$ and $i(t_R)$ are measures of impurity after splitting, and $\Delta i(s, t)$ measures the decrease in impurity from split s.

There are many approximations for measuring impurity. Some of the most frequent ones are gain-ratio (Quinlan 1993), Gini Index (Breiman *et al*. 1984) and Chi-square (Mingers 1989). The most common measure is the Gini index. The Gini index used in this research measures $i(t)$ as follows:

$$I_G\left(t_{X(x_i)}\right) = 1 - \sum_{j=1}^{m} f\left(t_{X(x_i)}, j\right)^2 \tag{2}$$

where $f\left(t_{X(x_i)}, j\right)$ is the proportion of samples with the value x_i belonging to leaf j as node t. The decision tree splitting criterion is based on choosing the attribute with the lowest Gini impurity index (I_G).

RF was proposed by Breiman (2001) as a new development of decision trees. It combines the predictions of every single RT algorithm using a rule-based approach. For regression, RF builds a number K of RTs and averages the results. After K such trees $\{T(x)\}_1^k$ are grown, the RF regression predictor is:

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K} T(x) \tag{3}$$

To avoid the correlation of the different trees, RF increases the diversity of the trees by making them grow from different training data subsets created through a procedure called bagging. Bagging is a technique used for training data creation by resampling randomly the original data set with replacement, i.e., with no deletion of the data selected from the input sample for generating the next subset $\{h(\mathbf{x}, \mathbf{\Theta_k}), k = 1, \ldots, K\}$, where $\{\mathbf{\Theta_k}\}$ are independent random vectors with the same distribution. Hence, some data may be used more than once in the training of classifiers, while others might never be used. Thus, greater stability is achieved, as it makes it more robust when facing slight variations in input data and, at the same time, it increases prediction accuracy (Breiman 2001). When the RF makes a tree grow, it uses the best variable within a subset of variables which has been selected randomly from the overall set of input variables. Therefore, this can decrease the strength of every single tree, but it reduces the correlation between the trees, reducing the generalization error (Breiman 2001). Another characteristic of interest is that the trees of an RF algorithm grow with no pruning, which makes them light, from a computational perspective.

Additionally, the samples not selected for the training of the k-th tree in the bagging process are included as part of another subset called out-of-bag (oob). These oob elements can be used by the k-th-tree to evaluate performance (Peters *et al*. 2007). This way RF can compute an unbiased estimation of the generalization error without using an external text data subset (Breiman 2001). A flow chart of RF is given in Figure 1. The generalization error converges as the number of trees increases; therefore, the RF does not overfit the data. RF also provides an assessment of the relative importance of the different variables. This aspect is useful for multisource studies, where data dimensionality is very high, and it is important to know how each predictive variable influences the prediction model to be able to select the

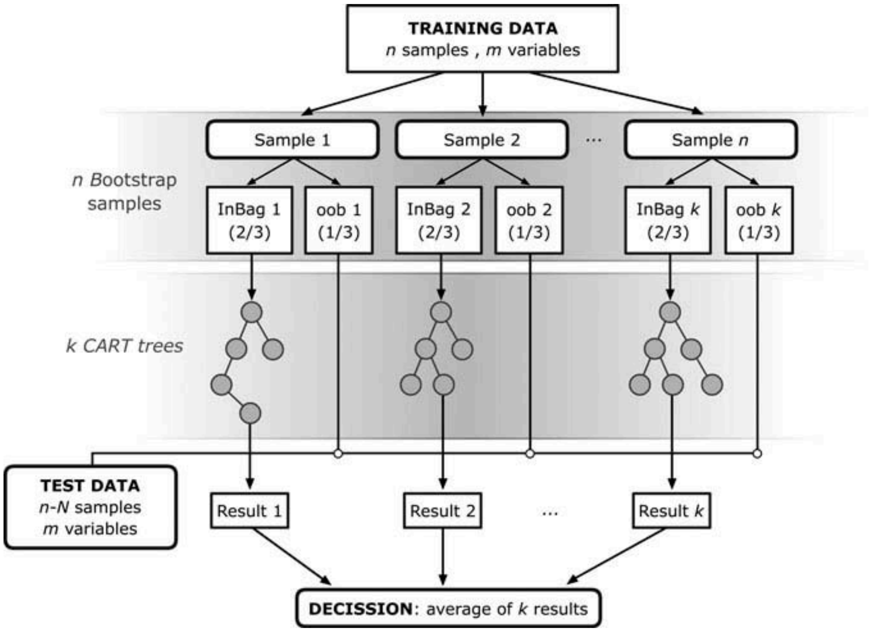

Figure 1. The flowchart of RF for regression (adapted from Guo 2011)).

best variables (Gislason *et al*. 2004, Ham 2005, Pal 2005, Gislason *et al*. 2006, Ghimire *et al*. 2010). To assess the importance of each variable (e.g. GIS layer), the RF switches one of the input variables while keeping the rest constant, and it measures the decrease in accuracy which has taken place by means of the oob error estimation (Breiman 2001).

3. Study area and data

Rodalquilar was chosen for this pilot study to test the application of RF to mineral potential mapping, because it contains a sufficiently large number of gold occurrences (46) to provide training data for the application of this methodology.

The GIS database for this study corresponds to the Rodalquilar mining district, which is located in the far southeast of the Iberian Peninsula, within the province of Almería. This mining district covers an area of 150 km^2 (Figure 2). From a geological perspective, the study area is located in the south-eastern portion of the Betic Mountain Range. This area mostly coincides with the volcanic field from the Miocene epoch of Cape of Gata, which makes up a mountain range of the same name, which goes along the coast from the Cape of Gata. Volcanic rocks range in composition from pyroxene andesite to rhyolite and in age from about 15 to 7 million years. This area is characterized by epithermal quartz–alunite gold deposits associated with felsic-to-intermediate tertiary volcanic rocks showing fracturing and pervasive hydrothermal alteration (Rytuba 1990, Demoustier *et al.* 1999). Mineralization within the Rodalquilar caldera complex consists of low-sulphidation Pb–Zn–(Cu–Ag–Au) quartz veins and high-sulphidation Au-alunita-(Cu–Te–Sn) epithermal deposits (Arribas *et al.* 1995). The gold epithermal deposits are associated with intensely hydrothermally altered rocks. These alterations resulted from the reaction of volcanic rocks and extremely acidified fluids. These fluids contained sulphides from a dioritic magma in depth and, very likely, from the sea.

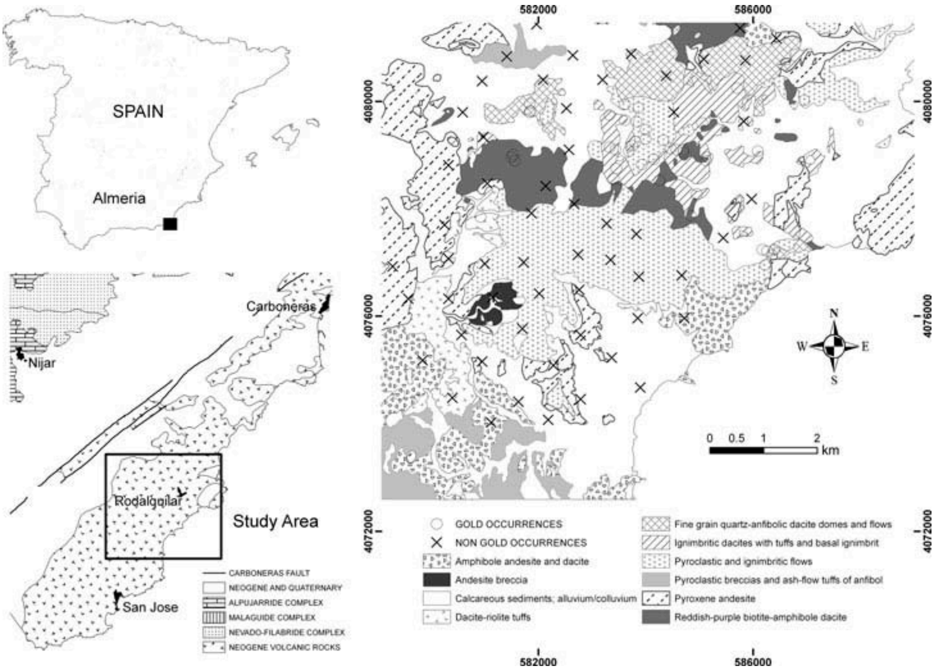

Figure 2. Location of the study area (left panels) and the distribution of Neogene volcanic rocks and locations of epithermal deposits (right panel). Map coordinates are in meters (UTM project, zone 30N, International 1924 ellipsoid, European 1950 datum).

From a climatic point of view, this region is characterized by its dryness, showing a semi-desert kind of climate. Unusual and intense precipitations, together with scarce vegetation, result in a strong run-off with flooding and important land erosion. Regarding land cover, there is an abundance of bare soils with very dispersed and scarce vegetation. This scarce vegetation, together with its lithological/geological characteristics, make this area a favourable sector for remote-sensing studies, as shown by diverse pilot studies carried out in the area in recent years (García 2008, Bedini *et al.* 2009, Escribano 2010).

On the basis of the deposit model for the district outlined by Rytuba *et al.* (1990), a database was created comprising the following thematic layers (Figure 2): (a–c) lithogeochemical survey (59 elements, 372 locations); (d, e) gravity and magnetic surveys (330 ground stations); (f) fracture map; (g, h) Landsat 5 Thematic Mapper (TM) image acquired in July 1991; (i) lithology, and (a–i) gold occurrence database containing 46 locations (Figure 3). Each of these locations corresponds to exploited deposits and known mineralized structures. Data sets and GIS database were gathered in the context of the DARSTIMEX European Project (Chica-Olmo *et al.* 2002).

4. Methods

4.1. Data processing

Landsat TM data processing was focused on identifying hydrothermally altered areas. Landsat TM band ratios have been shown to be very useful in delineating rock alteration in the area (Crósta and Moore 1989). Landsat TM 5/7 and 3/1 band ratios were computed because of their ability to discriminate ore-related hydroxyl and iron oxide alteration,

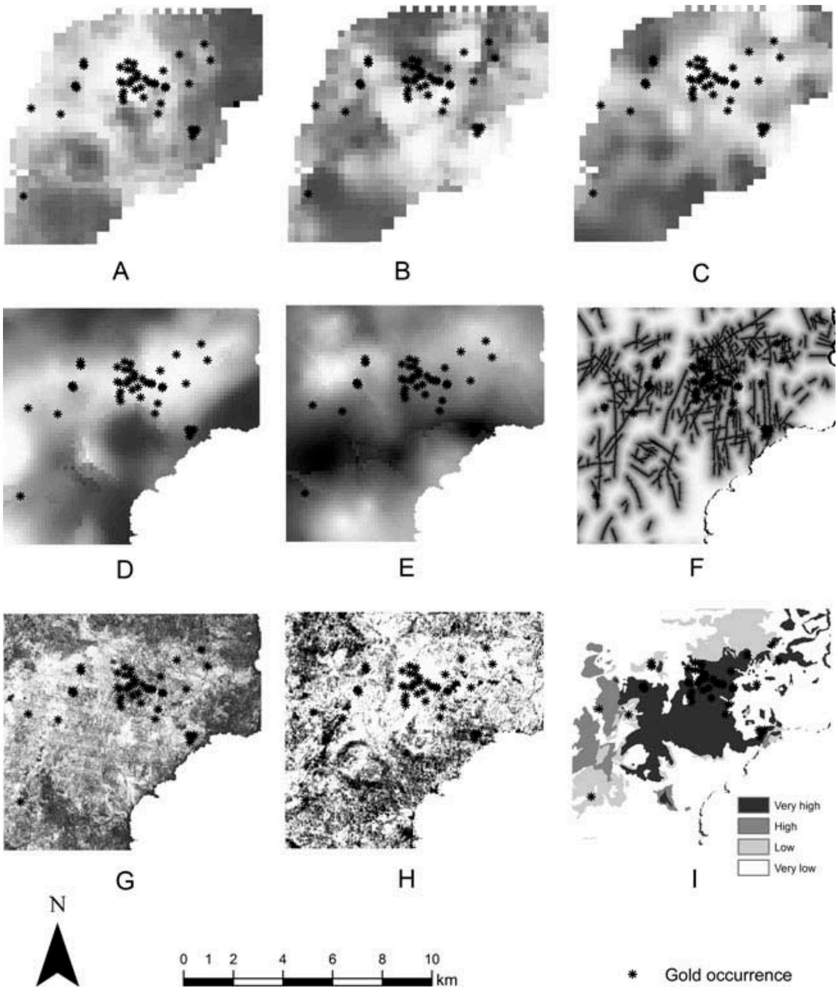

Figure 3. Layers used in mapping prospectivity for Au deposits: (A–C) geochemical survey (59 elements, 372 locations); (D, E) gravity and magnetic survey (330 ground stations); (F) fracture map; (G, H) Landsat 5 Thematic Mapper (TM) image acquired in July 1991; (I) lithology, and (A–I) gold occurrence database containing 46 locations. Lighter tones indicate higher values.

respectively. Geochemical data were processed by performing principal component analysis (PCA) on 59 selected mineralization-related elements. PC1, PC2 and PC3, indicating lithology (Si-Al versus Ca) and Au–Ag–As–S association, respectively, were chosen for further modelling. Au mineralization can be found in hydrothermal (epithermal) alteration zones of volcanic rocks situated in central areas. It is mainly associated with volcanic breccia of highly silicificated rocks (Si), and with argillite or clay (Al) minerals with presence of sulphides–sulphates. In this argillite hydrothermal alternation area, there is impoverishment in Ca, Mg, Na, Al, K, Ti, Mn, etc.

Continuous layers were created by Kriging (Chica-Olmo *et al*. 2002) from geochemical PCs to minimize estimation error. Gravity and magnetic residual values were also interpolated by Kriging to generate residual anomaly maps indicating the presence of potential ore-related buried anomalous bodies (Chica-Olmo *et al*. 2002). A distance-to-nearest-fracture

Table 1. Description of lithological gold favourability categories.

Class Id.	Category	Description of the category
1	Very favourable	Pyroclastic and ignimbritic flows, reddish-purple biotite-amphibole dacite and ignimbritic dacites with tuffs and basal ignimbrites.
2	Favourable	Dacite-riolite tuffs and pyroxene andesite.
3	Little favourable	Fine grain quartz-anfibolic dacite domes and flows, pyroclastic breccias and ash-flow tuffs of anfibol, amphibole andesite and dacite.
4	Non-favourable	Calcareous sediments; alluvium/colluvium and andesite breccia.

map was derived from the fracture map using GIS analysis functions. The lithology map of the area was reclassified into four classes: very favourable, favourable, less favourable and non-favourable (Table 1).

All seven input maps were stored as raster layers with 25 m spatial resolution in the exploration GIS. The thematic layers in the Rodalquilar GIS database were combined into a set of input feature vectors at each cell location in the set of grids (Figure 3a). These vectors formed the input to the RF algorithm and are known as input-feature vectors. Known deposit locations were used as a response variable for the training of the algorithms. Training patterns were created by recording the input feature vector values at each of the 46 locations of the gold occurrence database. The training data set was completed by adding 57 sterile locations scattered over the district, selected by means of stratified random sampling. Each training pattern consisted of an input feature vector paired together with a binary target value (target values used in the training data were 1 for gold occurrences and 0 for non-gold occurrences) (Figure 4). Hence, the output of the algorithm will be a floating value ranging from 0 to 1, representing the probability of mineral deposits.

Predictive modelling was performed outside the GIS using a series of computer programs written in R statistical software. Data processing in the RF method consists of two main stages:

(1) training and processing the entire data set after training; and

(2) post-processing requiring conversion of the output values to a map.

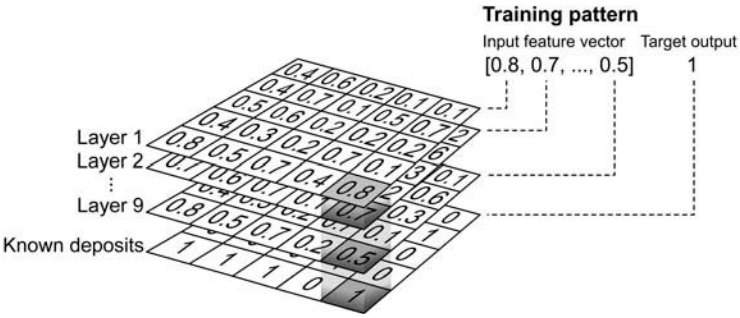

Figure 4. Relationship between GIS thematic layers and feature vectors used as input to random forest and logistic regression. For each cell location on the Rodalquilar database grid, the components of the input feature vector for that location are set to the values stored in the thematic layers. Each pattern in the data set used to train the models consists of an input feature vector paired together with the gold occurrences.

4.2. Random forest modelling of Au deposits

RF only needs two parameters to be set for generating a prediction model: the number of classification trees and the number of predictive variables (m), which are used in each node to make decision trees grow (Rodriguez-Galiano *et al.* 2012c). Breiman (1996) demonstrated that by increasing the number of trees, the generalization error always converges; hence, overtraining is not a problem and the number of trees can be fixed once the error has converged. However, the selection of the m parameter is very important. In each node, a subset of m predictive variables, from 1 to the maximum has to be specified. This value of m remains constant while the tree is growing, and variable selection is made randomly. Hence, the definition of this parameter affects both the correlation and strength of each individual tree and, therefore, it also affects the generalization error and predictive model accuracy (Breiman 2001).

In order to set the value of k from which the error converges and which also makes estimation more reliable, models made up of from 10 to 10,000 trees were generated (Rodriguez-Galiano *et al.* 2012c). Once the optimum number of threes was determined, for which the error was minimum and stable, all prediction models were created (see Section 5.1). The m parameter was optimized by varying the number of variables between 1 and the maximum number of variables (9 when all the variable subsets were considered). The resulting models were evaluated using the oob error estimation (see Section 2). For the selection of the most accurate model, we determined the one with the lowest oob error.

There are several commercial and open source implementations for RF model development (Liaw and Wiener 2002, Breiman and Cutler 2004, Witten and Frank 2005). In this study, the 'randomForest' package within the statistical software R 2.10.1 was used. The R implementation of RF (Liaw and Wiener 2002) contains a function called 'tuneRF' which automatically selects the optimal value of m with respect to oob correct classification rates. We did not use this function because there is no research as yet to assess the effects of choosing RF parameters such as m to optimize oob error rates on the generalization error rates for RF using this kind of data. The performance of RF was compared to LR using the 'glm' function within the statistical software R2.10.1.

4.3. Evaluating mineral potential maps

In order to quantitatively compare the quality of gold potential maps, verification of the results using some form of ground truth is required. Training data or an independent test data subset can be used as ground truth. The evaluation of the models considering the same patterns (known gold occurrences) used in their parametrization is not recommendable. The model is biased towards the patterns used in its construction and will be prone to overestimate its accuracy. On the other hand, the problem of data scarcity can exist. In many mineral potential mapping applications, the data to build data-driven models are scarce and to refuse using a part of them in the training of the model can lead to a minor learning. In this sense, RF incorporates a non-biased evaluation which allows it to generate an independent test (oob) without losing information for the calibration of the overall method (see Section 2). The oob estimate for the prediction error is the mean squared error (mse) computed between the prediction of the model and the original values for the oob subset (considering both positive and negative both occurrences). It is demonstrated in studies carried out by Breiman (2001, 1996) that the oob error is a good estimator of the prediction error when the number of trees is large enough. This measure based on the estimation of the error from the oob subset was used to select the RF model which generates the most accurate gold potential maps.

The performance of RF as a new mineral prospectivity mapping method was evaluated with respect to the LR method: (1) reclassifying the gold potential maps according to different thresholds of areal percentages of prospective zones and calculating the success rate of those prospective zones against the known gold occurrences; and (2) calculating the goodness of fit of the maps using an independent test (Agterberg and Bonham-Carter 2005). The success rate is the percentage of training deposits delineated correctly in prospective zones. Model performance curves are then created by plotting percentage of prospective zones versus success rates. For the comparison of the two methods (RF and LR) using an independent test, a variant of the *k*-fold cross-validation method was used (Webb 2010). The training set was randomly divided into 10 folds of equal size. Different predictive models were then trained 10 times, each time leaving out one of the subsets from training, but using only the omitted subset to assess performance.

5. Results and discussion

5.1. Random forest performance

Figure 5 shows the relationship between the number of trees (*k*) and split variables (*m*), which were used for training each RF regression model. The relationship between *k* and the mse is inversely proportional until a certain number of trees (*k*) is achieved from which the variance in the mse is low. Breiman (2001) demonstrated that RFs do not overfit. Thus, a limiting value of the generalization error is obtained as more trees are added until a threshold value is reached. However, the value of *m*, which is constant during forest growth, affects both the correlations between the trees and the strength of the individual trees. Reducing *m* reduces correlation and strength; increasing *m* increases both. As seen in forests made up of more than 1000 trees, the mse is minimal and remains constant. The addition of more trees neither increases nor decreases the prediction error. Very small values of *k* resulted in lower prediction performance; larger values of *k* resulted in more stable predictions (low variability in accuracy) and variable importance measures. The results of this analysis show that the differences in stability after 1000 trees are very small and the computation time increases for larger ranges of possible values of *k*. Hence, it is not advisable to adopt those values of the *k* parameter which are well over this threshold, in order to reduce the computational cost (Rodriguez-Galiano *et al*. 2012c).

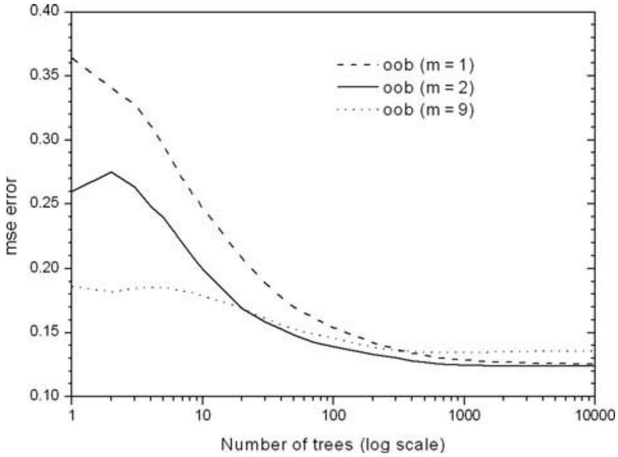

Figure 5. Effect of number of trees (*k*) and random split variables (*m*) on mse.

Table 2. Mean square errors of the RF models depending on the number of split variables used (*m*).

Number of trees (*k*)	Number of split variables (*m*)	mse
1000	1	0.129
1000	2	0.123
1000	3	0.128
1000	4	0.129
1000	5	0.129
1000	6	0.131
1000	7	0.133
1000	8	0.135
1000	9	0.134

To assess the optimal value of *m*, numerous RF models were created, each of them made up of 1000 trees for the different possible values of *m* to divide nodes (from 1 to 9). Table 2 shows the error depending on the number of split variables. The mse is minimum (0.123) when three random variables are used in the prediction (m = 2). The error increases as one moves away from this optimal value of *m*, reaching its maximum when all predictive variables were used (m = 9).

The average values of the absolute differences of the oob estimation between the minimum and maximum of the possible random variables (m = 1 and m = 9) and the optimum number of variables (m = 2) are equal to 0.013 and 0.015, respectively. However, from tree number 1000, these differences were lower, being equal to 0.002 and 0.11, respectively, from which it can be inferred that an RF is less sensitive to the value of *m* once the point has been reached at which the error converges. The results of this study agree with Breiman's (2001) recommendations, to use values of *m* close to a third of the total number of variables. Taking this into account, it is preferable to use a large number of trees (*k*) and a number of split variables (*m*) roughly equal to one-third of the overall number of split variables to reduce the generalization error and the correlation between trees.

5.1.1. Importance of variables

The layout of the split variables in a RT provides information about the importance of the variables in the prediction model. However, it is almost impossible to carry out this interpretation with classifier ensembles based on multiple decision trees. An RF allows for assessment of the importance of the variables by means of the oob subset (Section 2). Thus, if variables are too many, the RF can be applied only to variables which have been identified as most important in the first application of the RF.

Models with 1000 trees and two variables were considered to calculate the importance of the contribution of each variable to the general prediction model (Breiman 2001). Figure 6 shows the contribution of each variable to the regression model generated by considering all the GIS layers. According to percentage of the mse increasing (computed from the oob subset), the layer with the highest contribution to the RF regression model is the third principal component of the geochemical data followed by the distance to faults and lithology, with values of 24, 22 and 19, respectively. The information derived from remote sensing also played an important role in the mineral potential mapping. As ratio TM5/TM7 (ore-related hydroxyl) produced an 18% improvement in the estimation of the model. The altered volcanic rocks of the study area contain clay minerals (alunite, pirofilite, caolinite, illite, etc.), which are not present or are scarce in rocks with no hydrothermal alteration.

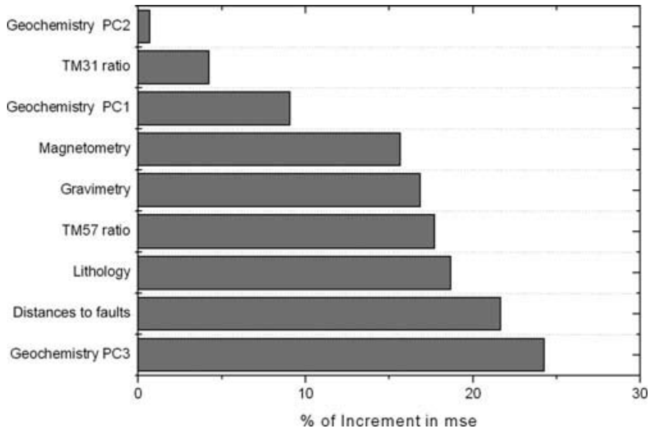

Figure 6. Variable importance contribution of different layers to the prospective model.

These minerals have characteristic absorption peaks in spectral curves in the TM7 band, and have high reflectivity in the TM5 band. For this reason, the TM5/TM7 ratio has been of special interest for the detection of hydrothermal alteration areas. However, there are other types of satellite sensors that could contribute additional radiometric information for mineral potential mapping, e.g. ASTER (Galvão *et al.* 2005, Carranza *et al.* 2008). These four variables chosen by RF as most important support the criteria established by Carranza *et al.* (2008) for the prospective modelling of gold epithermal deposits: (1) volcanic rocks, (2) short distances to fractures/faults, and (3) intense hydrothermal alteration. Criterion number 1 is given by the lithology and geochemistry layers; criterion number 2 is given by the distance layer to fractures and faults and, lastly, criterion number 3 is given by TM5/TM7 ratio and the third component of geochemistry (see Section 3).

Geophysics was of a slightly smaller importance, being more relevant gravimetry than magnetometry, with values equal to 17 and 16, respectively. Lastly, the rest of the main components of geochemistry and the ratio of TM3/TM1 bands were the least important layers in prospective modelling.

5.2. Comparison of RF and LR prospectivity models

Gold potential maps produced using the RF and LR are shown together with gold occurrence points in Figure 7. Areas with higher gold potential are located mainly in the central part of the study area and around certain fracturing and faults. Similarity between maps stems from the fact that for both methods, the distance variable to faults or fractures had an outstanding importance. As seen in Section 5.1.1, RF identified high potential areas for gold as those pixels with high values in the PC3 of geochemical data,

Table 3. LR coefficients.

Dist. Faults	Litho. C2	Litho. C3	Litho. C4	GPC1	GPC2
−66.17	−3.30	−0.80	−1.07	0.57	−2.48
GPC3	Grav.	Mag.	TM31	TM57	Intercept
2.56	20.40	−2.69	−19.75	32.18	−16.18

Note: Dist. Faults, distance to faults; Litho. C2, lithology class 2; Litho. C3, lithology class 3; Litho. C4, lithology class 4; GPC1, geochemical principal component 1; GPC2, geochemical principal component 2; GPC3, geochemical principal component 3.

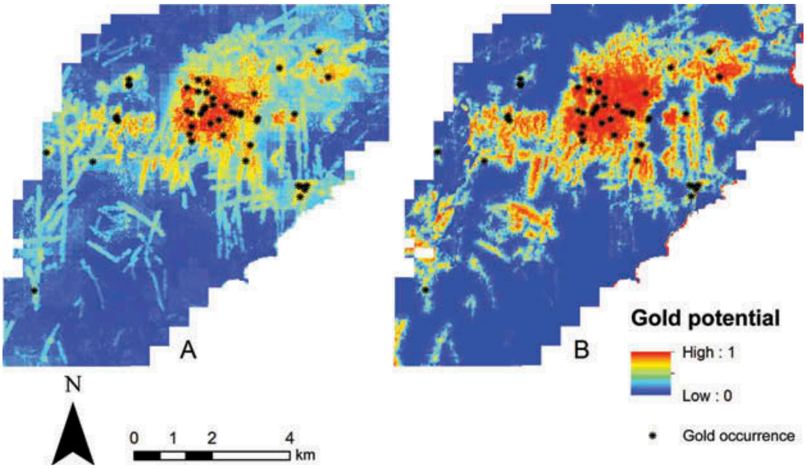

Figure 7. Gold potential maps generated using RF (A) and LR (B). Available in colour online.

corresponding to volcanic rocks linked to high fracturing and hydrothermal alteration areas. Table 3 shows that, in the case of LR, the layers with stronger influence on the gold potentiality predictive model were the distance to faults and fractures and, to a lesser extent, the TM5/TM7 and TM3/TM1 ratios and gravimetry. Unlike RF, which uses a large number of variables, LR bases prediction mainly on the distance to faults and fractures. The greater importance assigned by LR to fracture and fault areas, in relation to the rest of the layers, caused that the LR method identified all the zones close to faults or fractures as of high potentiality. However, very relevant variables in the RF predictive model, such as geochemistry and lithology, had a marginal role in the case of LR.

The area defined as highly prospective is significantly smaller in the RF map compared to the LR map. Hence, in order to reach a success rate similar to RF, LR needs to delimit larger prospective areas. This overestimation is not desirable, for it entails a higher economic cost both for mineral prospection and exploitation. Figure 8 shows the success rate in the estimation of the known gold deposits according to different percentages of prospective areas (see Section 4.3). The RF model has a better success rate than LR model. Thus, for percentage threshold values of prospective areas over 10%, the success rate of RF is over 90%. Nevertheless, in the case of LR, this success rate was reached considering percentage thresholds of prospective areas which tripled the previous case (over 30%). Likewise, RF managed to predict 100% of the known gold occurrences correctly, considering prospective areas greater than 35%, while in the case of LR, this success rate was only achieved when most of the study area was considered as prospective (65%). This difference in the performance of both methods is especially relevant in mining exploration, where the economic costs associated with the extension of the exploration area must be taken into account (Figure 9).

Finally, the value of the mse calculated from the cross validation procedure described in Section 4.3 was noticeably better for RF than for LR, with values of 0.12 and 0.19, respectively. It should be noted that the error estimation of the RF model obtained from the *k*-fold cross validation is equal to that obtained from the oob subset (see Section 5.1). Apart from those aspects referred to in previous sections, the best RF performance, with relation to LR, is linked to its non-parametric nature, i.e., it does not need the layers included in the GIS to follow a normal distribution.

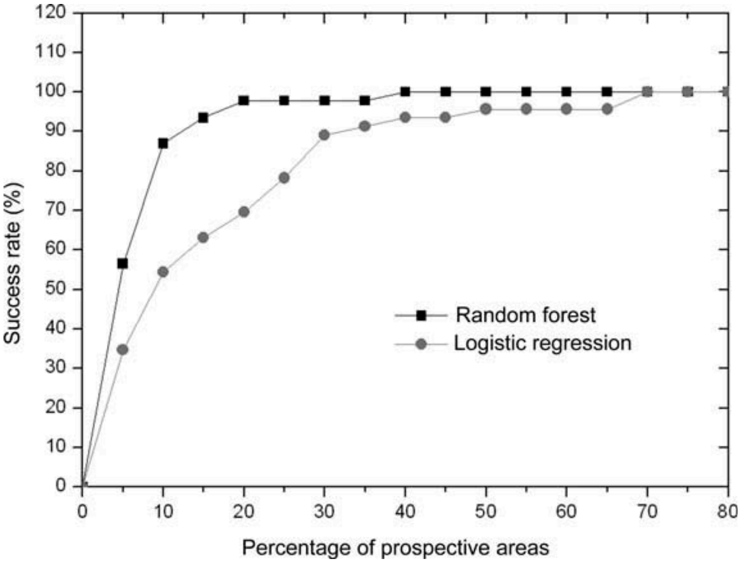

Figure 8. Mineral prospectivity success of the gold potential maps.

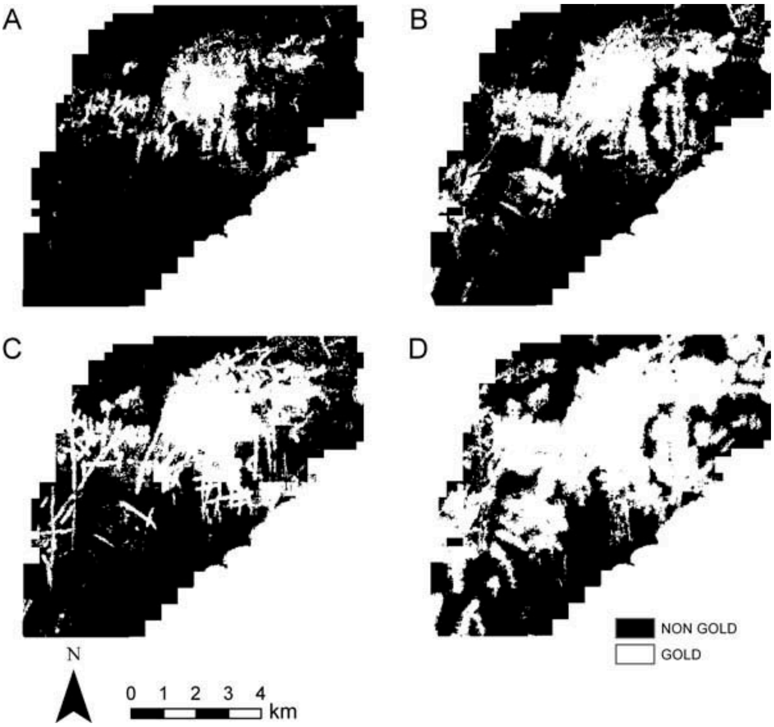

Figure 9. Prospectivity maps for success rates equal to 90% (A and B) and 100% (C and D). RF maps are given on the left and LR on the right.

6. Conclusions

This study aimed to evaluate the performance of the RF method for prospectivity mapping of a very well-known mining area: Rodalquilar (Southern Spain). RF performed well in the context of gold potential mapping from a suite of geochemical, geophysical, geological and remotely sensed GIS data.

The RF algorithm generates an internal unbiased estimation of the mapping error (oob error), so it is not necessary to split data into training and testing subsets or resort to cross-validation. The configuration of the design parameters of the RF models is simple. The number of trees is directly proportional to the classifier's accuracy until reaching a state in which the error converges (1000 trees). Once the error has converged, the number of random variables (*m*) only alters the map's accuracy slightly, so the RF is a very operative technique and therefore could be a valuable tool for GIS software.

Furthermore, the algorithm can estimate the importance of variables (GIS layers). These important measures are consistent with geological knowledge about the gold mineralization in the Rodalquilar area. Although feature selection was not carried out in this study, this estimation of importance could be used for variable selection in the study of complex areas, where it is mandatory to use large multisource data sets with a large number of variables.

Apart from the simplicity of its application and result interpretability, the RF method can handle complex data of different statistical distributions, responding to non-linear relationships between variables.

Statistical measures used to compare map quality indicate that the RF method performs better than LR method. The gold potentiality mapping average error is lower for RF and success rate is higher. Once the error has converged, the number of random variables (*m*) only alters the map's accuracy slightly, so RF is a very operative technique and therefore could be a valuable tool for GIS software.

Acknowledgements

We are grateful for the financial support given by the Spanish MICINN (Project CGL2010-17629) and Junta de Andalucía (Group RNM122). We would like to thank the reviewers for their constructive criticism.

References

Abedi, M., Norouzi, G.-H., and Fathianpour, N., 2013. Fuzzy outranking approach: a knowledge-driven method for mineral prospectivity mapping. *International Journal of Applied Earth Observation and Geoinformation*, 21, 556–567. doi:10.1016/j.jag.2012.07.012

Agterberg, F.P. and Bonham-Carter, G.F., 2005. Measuring the performance of mineral-potential maps. *Natural Resources Research*, 14 (1), 1–17. doi:10.1007/s11053-005-4674-0

Arribas, A. Jr., *et al.*, 1995. Geology, geochronology, fluid inclusions, and isotope geochemistry of the Rodalquilar gold alunite deposit, Spain. *Economic Geology*, 90 (4), 795–822. doi:10.2113/gsecongeo.90.4.795

Avantra Geosystems, 2006. *MI-SDM (MapInfo Spatial Data Modeller) v2.51* [online]. Available from: http://www.avantra.com.au/mi-sdm.htm [Accessed December 2013].

Bedini, E., Van Der Meer, F., and Van Ruitenbeek, F., 2009. Use of HyMap imaging spectrometer data to map mineralogy in the Rodalquilar caldera, southeast Spain. *International Journal of Remote Sensing*, 30 (2), 327–348. doi:10.1080/01431160802282854

Bonham-Carter, G.F., 1994. *Geographic information systems for geoscientists: modelling with GIS*. Tarrytown, NY: Pergamon.

Breiman, L., 1984. *Classification and regression trees*. London: Chapman & Hall/CRC.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24 (2), 123–140. doi:10.1007/BF00058655

Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32. doi:10.1023/A:1010933404324

Breiman, L. and Cutler, A., 2004. *Random forest* [online]. Available from: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Accessed December 2013].

Breiman, L., *et al*, 1984. *Classification and regression trees.* 1st ed. Belmont, CA: Chapman and Hall/CRC.

Carranza, E.J.M., 2008. *Geochemical anomaly and mineral prospectivity mapping in GIS*. Amsterdam: Elsevier.

Carranza, E.J.M., 2011. Geocomputation of mineral exploration targets. *Computers & Geosciences*, 37 (12), 1907–1916. doi:10.1016/j.cageo.2011.11.009

Carranza, E.J.M. and Hale, M., 2001. Logistic regression for geologically constrained mapping of gold potential, Baguio district, Philippines. *Exploration and Mining Geology*, 10 (3), 165–175. doi:10.2113/0100165

Carranza, E.J.M., *et al*., 2008. Knowledge-guided data-driven evidential belief modeling of mineral prospectivity in Cabo de Gata, SE Spain. *International Journal of Applied Earth Observation and Geoinformation*, 10 (3), 374–387. doi:10.1016/j.jag.2008.02.008

Chan, J.C.-W. and Paelinckx, D., 2008. Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112 (6), 2999–3011. doi:10.1016/j.rse.2008.02.011

Chica-Olmo, M., Abarca, F., and Rigol, J.P., 2002. Development of a decision support system based on remote sensing and GIS techniques for gold-rich area identification in SE Spain. *International Journal of Remote Sensing*, 23 (22), 4801–4814. doi:10.1080/01431160110104656

Chung, C.F., 1977. Application of discriminant analysis for the evaluation of mineral potential. *In*: RVR, ed. *APCOM Symposium*, New York: Society of Mining Engineers of American Institute of Mining, Metallurgical, and Petroleum Engineers, 299–311.

Chung, C.F., 1978. *Computer program for the logistic model to estimate the probability of occurrence of discrete events*. Paper-Geological Survey of Canada 78–12. Ottawa: Minister of Supply and Services Canada, 23.

Crósta, A.P. and Moore, J.M., 1989. Geological mapping using landsat thematic mapper imagery in Almeria province, south-east Spain. *International Journal of Remote Sensing*, 10 (3), 505–514. doi:10.1080/01431168908903888

Cuihua, C., *et al*., 2011. Mineral prospectivity mapping integrating multi-source geology spatial data sets and logistic regression modeling. *In: Proceedings of the 2011 IEEE international conference on spatial data mining and geographical knowledge Services (ICSDM 2011)*, 29 June–1 July, Fuzhou: IEEE, 214–217.

Debba, P., *et al.*, 2009. Deriving optimal exploration target zones on mineral prospectivity maps. *Mathematical Geosciences*, 41 (4), 421–446. doi:10.1007/s11004-008-9181-5

Demoustier, A., Charlet, J.M., and Castroviejo, R., 1999. Characterization of epithermal quartz veins from the volcanic area of Cabo de Gata (Almeria Province, southeastern Spain) by low-temperature thermoluminescence; relation with petrographic textures and fluid inclusions. *Caracterisation des quartz filoniens epithermaux de la zone volcanique de Cabo de Gata (province d'Almeria, Espagne) par thermoluminescence basse temperature; relation avec les textures petrographiques et les inclusions fluides*, 328 (8), 521–528.

Escribano, P., *et al.*, 2010. Spectral properties and sources of variability of ecosystem components in a Mediterranean semiarid environment. *Journal of Arid Environments*, 74 (9), 1041–1051. doi:10.1016/j.jaridenv.2010.02.001

Fallon, M., Porwal, A., and Guj, P., 2010. Prospectivity analysis of the Plutonic Marymia Greenstone Belt, Western Australia. *Ore Geology Reviews*, 38 (3), 208–218. doi:10.1016/j.oregeorev.2010.03.009

Friedl, M.A., Brodley, C.E., and Strahler, A.H., 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience and Remote Sensing*, 37 (2), 969–977. doi:10.1109/36.752215

Galvão, L.S., Almeida-Filho, R., and Vitorello, Í., 2005. Spectral discrimination of hydrothermally altered materials using ASTER short-wave infrared bands: evaluation in a tropical savannah environment. *International Journal of Applied Earth Observation and Geoinformation*, 7 (2), 107–114. doi:10.1016/j.jag.2004.12.003

García, M., *et al.*, 2008. Monitoring land degradation risk using ASTER data: the non-evaporative fraction as an indicator of ecosystem function. *Remote Sensing of Environment*, 112 (9), 3720–3736. doi:10.1016/j.rse.2008.05.011

Ghimire, B., Rogan, J., and Miller, J., 2010. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. *Remote Sensing Letters*, 1, 45–54. doi:10.1080/01431160903252327

Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R., 2004. Random forest classification of multisource remote sensing and geographic data. *In*: *IGARSS 2004: IEEE international geoscience and remote sensing symposium*, 20–24 September, Anchorage, AK, 1049–1052.

Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R., 2006. Random forests for land cover classification. *Pattern Recognition Letters*, 27 (4), 294–300. doi:10.1016/j.patrec.2005.08.011

Guo, L., *et al*., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (1), 56–66. doi:10.1016/j.isprsjprs.2010.08.007

Ham, J., *et al.*, 2005. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (3), 492–501. doi:10.1109/TGRS.2004.842481

Hansen, L.K. and Salamon, P., 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 (10), 993–1001. doi:10.1109/34.58871

Harris, D., *et al*., 2003. A comparative analysis of favorability mappings by weights of evidence, probabilistic neural networks, discriminant analysis, and logistic regression. *Natural Resources Research*, 12 (4), 241–255. doi:10.1023/B:NARR.0000007804.27450.e8

Hastie, T., Tibshirani, R., and Friedman, J., 2009. Linear methods for classification. *The Elements of Statistical Learning*. New York: Springer, 101–137. doi:10.1007/978-0-387-84858-7_4

Huang, Z., *et al*., 2012. Predictive modelling of seabed sediment parameters using multibeam acoustic data: a case study on the Carnarvon Shelf, Western Australia. *International Journal of Geographical Information Science*, 26 (2), 283–307. doi:10.1080/13658816.2011.590139

Kemp, L.D., Bonham-Carter, G.F., and Raines, G.L., 1999. *Arc-WofE: Arcview extension for weights of evidence mapping* [online]. Available from: http://www.ige.unicamp.br/wofe [Accessed January 2013].

Kotsiantis, S. and Pintelas, P., 2004. Combining bagging and boosting. *International Journal of Computational Intelligence*, 1 (4), 324–333.

Krogh, A. and Vedelsby, J., 1995. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 7, 231–238.

Laborte, A.G., Maunahan, A.A., and Hijmans, R.J., 2012. Opportunities for expanding paddy rice production in Laos: spatial predictive modeling using random forest. *Journal of Land Use Science*, 7 (1), 21–33. doi:10.1080/1747423X.2010.519788

Lawrence, R., Wood, S., and Sheley, R., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). *Remote Sensing of Environment*, 100 (3), 356–362. doi:10.1016/j.rse.2005.10.014

Lewkowski, C., Porwal, A., and González-Álvarez, I., 2010. Genetic programming applied to base-metal prospectivity mapping in the Aravalli Province, India. *In*: *EGU general assembly 2010.* 2–7 May, Vienna: EGU.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R News*, 2/3, 18–22.

Mcginnis, S. and Kerans, B.L., 2013. Land use and host community characteristics as predictors of disease risk. *Landscape Ecology*, 28 (1), 29–44.

Mingers, J., 1989. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3 (4), 319–342. doi:10.1007/BF00116837

Moon, C.J. and Evans, A.M., 2006. Ore, mineral economics and mineral exploration. *In*: C.J. Moon, M.K.G. Whateley, and A.M. Evans, eds. *Introduction to mineral exploration*. 2nd ed. Oxford: Blackwell Publishing, 3–18.

Oh, H.J. and Lee, S., 2010. Application of artificial neural network for gold-silver deposits potential mapping: a case study of Korea. *Natural Resources Research*, 19 (2), 103–124. doi:10.1007/s11053-010-9112-2

Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26 (1), 217–222. doi:10.1080/01431160412331269698

Pereira Leite, E. and De Souza Filho, C.R., 2009a. Artificial neural networks applied to mineral potential mapping for copper-gold mineralizations in the Carajás Mineral Province, Brazil. *Geophysical Prospecting*, 57 (6), 1049–1065. doi:10.1111/j.1365-2478.2008.00779.x

Pereira Leite, E. and De Souza Filho, C.R., 2009b. Probabilistic neural networks applied to mineral potential mapping for platinum group elements in the Serra Leste region, Carajás Mineral Province, Brazil. *Computers and Geosciences*, 35 (3), 675–687. doi:10.1016/j.cageo.2008.05.003

Peters, J., *et al.*, 2007. Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, 207 (2–4), 304–318.

Peters, J., *et al.*, 2011. Synergy of very high resolution optical and radar data for object-based olive grove mapping. *International Journal of Geographical Information Science*, 25 (6), 971–989. doi:10.1080/13658816.2010.515946

Porwal, A., *et al.*, 2010a. Weights-of-evidence and logistic regression modeling of magmatic nickel sulfide prospectivity in the Yilgarn Craton, Western Australia. *Ore Geology Reviews*, 38 (3), 184–196. doi:10.1016/j.oregeorev.2010.04.002

Porwal, A., Carranza, E.J.M., and Hale, M., 2003. Artificial neural networks for mineral-potential mapping: a case study from Aravalli Province, Western India. *Natural Resources Research*, 12 (3), 155–171. doi:10.1023/A:1025171803637

Porwal, A., Yu, L., and Gessner, K., 2010b. SVM-based base-metal prospectivity modeling of the Aravalli Orogen, northwestern India. *In*: *EGU general assembly 2010*. 2–7 May, Vienna: EGU.

Quinlan, J.R., 1993. *C4.5 programs for machine learning*. 1st ed. San Mateo, CA: Morgan Kaurmann.

Rigol-Sanchez, J.P., Chica-Olmo, M., and Abarca-Hernandez, F., 2003. Artificial neural networks as a tool for mineral potential mapping with GIS. *International Journal of Remote Sensing*, 24 (5), 1151–1156. doi:10.1080/0143116021000031791

Rodriguez-Galiano, V. and Chica-Olmo, M., 2012. Land cover change analysis of a Mediterranean area in Spain using different sources of data: multi-seasonal Landsat images, land Surface temperature, digital terrain models and texture. *Applied Geography*, 35, 208–218. doi:10.1016/j.apgeog.2012.06.014

Rodriguez-Galiano, V.F. and Chica-Rivas, M., 2012. Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. *International Journal of Digital Earth*. doi:10.1080/17538947.2012.748848

Rodriguez-Galiano, V.F., *et al.*, 2012a. Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, 121, 93–107. doi:10.1016/j.rse.2011.12.003

Rodriguez-Galiano, V.F., *et al.*, 2012b. Incorporating the downscaled Landsat TM thermal band in land-cover classification using random forest. *Photogrammetric Engineering and Remote Sensing*, 78 (2), 129–137. doi:10.14358/PERS.78.2.129

Rodriguez-Galiano, V.F., *et al.*, 2012c. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104. doi:10.1016/j.isprsjprs.2011.11.002

Rytuba, J.J., *et al.*, 1990. Mineralized and unmineralized calderas in Spain; Part II, evolution of the Rodalquilar caldera complex and associated gold-alunite deposits. *Mineralium Deposita*, 25 (S1), SS29–SS35. doi:10.1007/BF00205247

Sawatzky, D.L., *et al.*, 2009. *Spatial Data Modeller (SDM): ArcMAP 9.3 geoprocessing tools for spatial data modelling using weights of evidence, logistic regression, fuzzy logic and neural networks* [online]. Available from: http://arcscripts.esri.com/details.asp?dbid=15341 [Accessed January 2013].

Sesnie, S., *et al.*, 2008. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment*, 112 (5), 2145–2159. doi:10.1016/j.rse.2007.08.025

Singer, D.A. and Kouda, R., 1996. Application of a feedforward neural network in the search for kuroko deposits in the hokuroku district, Japan. *Mathematical Geology*, 28 (8), 1017–1023. doi:10.1007/BF02068587

Steele, B.M., 2000. Combining multiple classifiers an application using spatial and remotely sensed information for land cover type mapping. *Remote Sensing of Environment*, 74 (3), 545–556. doi:10.1016/S0034-4257(00)00145-0

Varet, J., 2012. Mineral resources: An overall assessment. *Ressources minérales: Un état des lieux*, (381), 29–54.

Vincenzi, S., *et al*., 2011. Application of a random forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecological Modelling*, 222 (8), 1471–1478. doi:10.1016/j.ecolmodel.2011.02.007

Webb, G.I., 2010. Cross-validation. *In*: C. Sammut and G.I. Webb, eds. *Encyclopedia of machine learning*. New York: Springer US, 249.

Witten, I.H. and Frank, E., 2005. *Data mining: practical machine learning tools and techniques.* 2nd ed. San Francisco, CA: Morgan Kaufmann.

Zuo, R. and Carranza, E.J.M., 2011. Support vector machine: a tool for mapping mineral prospectivity. *Computers & Geosciences*, 37 (12), 1967–1975. doi:10.1016/j.cageo.2010.09.014