Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines)

Emmanuel John M. Carranza, Alice G. Laborte

PII:  S0098-3004(14)00227-1
DOI:  http://dx.doi.org/10.1016/j.cageo.2014.10.004
Reference:  CAGEO3446

To appear in:  Computers and Geosciences

Received date:  5 July 2014
Revised date:  30 August 2014
Accepted date:  13 October 2014

Cite this article as: Emmanuel John M. Carranza and Alice G. Laborte, Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines), Computers and Geosciences, http://dx.doi.org/10.1016/j.cageo.2014.10.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Random Forest Predictive Modeling of Mineral Prospectivity with Small Number of Prospects and Data with Missing Values in Abra (Philippines)

Emmanuel John M. Carranza¹,², Alice G. Laborte²
¹James Cook University, Townsville 4811, Queensland, Australia
²International Rice Research Institute, Los Baños, Laguna, Philippines

Abstract

Machine learning methods that have been used in data-driven predictive modeling of mineral prospectivity (e.g., artificial neural networks) invariably require large number of training prospect/locations and are unable to handle missing values in certain evidential data. The Random Forests (RF) algorithm, which is a machine learning method, has recently been applied to data-driven predictive mapping of mineral prospectivity, and so it is instructive to further study its efficacy in this particular field. This case study, carried out using data from Abra (Philippines), examines (a) if RF modeling can be used for data-driven modeling of mineral prospectivity in areas with few (i.e., <20) mineral occurrences and (b) if RF modeling can handle evidential data with missing values. We found that RF modeling outperforms weights-of-evidence (WofE) modeling of porphyry-Cu prospectivity in the Abra area, where 12 porphyry-Cu prospects are known to exist. Moreover, just like WofE modeling, RF modeling allows analysis of the spatial associations of known prospects with individual layers of evidential data. Furthermore, RF modeling can handle missing values in evidential data through an RF-based imputation technique whereas in WofE modeling values are simply represented by zero weights. Therefore, the RF algorithm is potentially more useful than existing methods that are currently used for data-driven predictive mapping of mineral prospectivity. In particular, it is not a purely black-box method like artificial neural networks in the context of data-driven predictive modeling of mineral prospectivity. However, further testing of the method in other areas with few mineral occurrences is needed to fully investigate its usefulness in data-driven predictive modeling of mineral prospectivity.

Keywords:

Data-driven modeling; ensemble of regression trees; porphyry copper; GIS.

* Email for correspondence: john.carranza@jcu.edu.au
1. Introduction

Predictive modeling of mineral prospectivity entails the analysis and synthesis of various layers of spatial evidence derived from various relevant geoscience spatial datasets in order to delineate and rank areas that are prospective for exploration of mineral deposits of the type sought (Bonham-Carter, 1994; Carranza, 2009b, 2011b). Data-driven predictive modeling of mineral prospectivity is appropriate in moderately- to well-explored (or so-called ‘brownfields’) regions, where the goal is to define new targets for further exploration of mineral deposits of the type sought. In data-driven predictive modeling of mineral prospectivity, the weights assigned to every layer of spatial evidence are quantified spatial associations between training locations (i.e., discovered mineral deposits/prospects) and individual datasets used as evidence of mineral prospectivity. Data-driven predictive mapping of mineral prospectivity makes use of mathematical methods that involve either bivariate or multivariate analysis to quantify spatial associations between training locations and individual evidential datasets. Bivariate methods involve analysis of spatial association between a map of training locations and a map of an evidential data. There are two methods used for bivariate data-driven predictive modeling of mineral prospectivity: (i) weights-of-evidence (WofE) modeling (Bonham-Carter et al., 1988, 1989; Agterberg et al., 1990, 1993); and (ii) evidential belief modeling (An et al., 1994a,b; Carranza and Hale, 2003; Carranza et al., 2005, 2008a, 2008b, 2009; Carranza and Sadeghi, 2010; Carranza, 2009a, 2011a, 2014). In contrast, multivariate techniques involve simultaneous analysis of spatial associations between a map of training locations and maps of evidential datasets. The two methods most commonly used for multivariate data-driven predictive modeling of mineral prospectivity are (i) logistic regression (Chung, 1978, 1983; Chung and Agterberg, 1980, 1988; Carranza and Hale, 2001; Harris et al., 2001, 2006; Mejía-Herrera et al., 2014) and (ii) artificial neural networks (Singer and Kouda, 1996, 1999; Brown et al., 2000, 2003a,b; Harris et al., 2003; Porwal et al., 2003, 2004; Rigol-Sanchez et al., 2003; Skabar, 2005; Behnia, 2007; Nykänen, 2008; Oh and Lee, 2010; Abedi et al., 2012).

The success of data-driven predictive modeling of mineral prospectivity depends chiefly on the sufficiency of the number of training locations used in the analysis and on evidential data without missing values. However, no study has specified a minimum sufficient number of training locations for data-driven predictive modeling of mineral prospectivity. The smallest numbers of training locations that have been used successfully for data-driven modeling of mineral prospectivity are 19 epithermal-Au deposits in the Baguio district of the Philippines (Carranza and Hale, 2000, 2001), 15 volcanogenic massive sulfide deposits in the Chisel Lake – Anderson Lake area in Canada (Wright and Bonham-Carter, 1996), 13 hydrothermal vents on the East Pacific Rise (Agterberg et al., 1993; Agterberg and Cheng, 2003), and 12 porphyry-Cu deposits in Abra province of the Philippines (Carranza, 2004). These previous works have applied either WofE or logistic regression. Data-driven
evidential belief modeling has also been used successfully with 19 epithermal-Au deposits in the Baguio district of the Philippines (Carranza and Hale, 2003), 17 hydrothermal Cu-Au deposits in Catanduanes Island of the Philippines (Carranza, 2011a) and 12 porphyry-Cu prospects in Abra province of the Philippines (Carranza, 2014).

For applications of artificial neural networks, which are machine learning methods that usually require large training data, only few studies have used less than 20 training deposit locations for data-driven predictive modeling of mineral prospectivity. Pereira Leite and De Souza Filho (2009a,b) successfully applied artificial neural networks in two studies of data-driven mineral prospectivity mapping in the Carajás Mineral Province (Brazil) using (i) 17 Cu-Au deposits and (ii) two PGE (platinum group elements) deposits but each represented as four training cells. Alves Magalhães and De Souza Filho (2012) successfully applied artificial neural networks with 12 orogenic Au deposits in the central region of the Amapá State in the northernmost part of the Brazilian Amazon. However, unlike the WofE modeling of mineral prospectivity described in Bonham-Carter et al. (1989), Agterberg et al. (1990) and Carranza (2004), the artificial neural network modeling of mineral prospectivity described in Pereira Leite and De Souza Filho (2009a,b) and Alves Magalhães and De Souza Filho (2012) did not make use of evidential data with missing values.

Another machine learning method – Random Forests (Breiman, 2001) – is increasingly being used in predictive mapping of areas of interest according to certain 'suitability' criteria. For example, one of us has recently applied Random Forests (RF) to predict areas suitable for rice production (Laborte et al., 2012). The growing major application of RF is land-cover classification (e.g., Gislason et al., 2006; Grimm et al., 2008; Rodriguez-Galiano et al., 2012) and species distribution mapping (e.g., Prasad et al., 2006; Evans and Cushman, 2009; Bradter et al., 2013). Meanwhile, the application of RF to spatial data integration for geological mapping is also growing (e.g., Waske et al., 2009; Cracknell and Reading, 2013, 2014; Cracknell et al., 2013). However, RF is a sophisticated and elegant extension of a decision-tree algorithm known as CART (classification and regression trees) first introduced by Breiman et al. (1984). Reddy and Bonham-Carter (1991) demonstrated a decision-tree approach to map mineral prospectivity for base-metal deposits in the Snow Lake area in Manitoba (Canada). However, this approach was either not very easy to understand or computationally demanding at that time and, thus, it was rarely applied to mineral prospectivity mapping. Just recently, Rodriguez-Galiano et al. (2014) have demonstrated the applicability of RF to data-driven predictive modeling of mineral prospectivity using 49 training locations of epithermal Au deposits in the Rodalquilar district (Spain). In addition, Carranza and Laborte (2014) found that, in data-driven predictive mapping of gold prospectivity in the Baguio district (Philippines) based on 19 training locations for epithermal Au deposits, RF is not sensitive to different sets of non-deposit locations used for training together with known deposit locations.

In this paper, we demonstrate further that RF can be applied to data-driven modeling of mineral
prospectivity using less than 20 training locations. We also demonstrate that an imputation method based on and linked with RF makes it suitable for data-driven predictive modeling of mineral prospectivity using evidential data with missing values, just like WofE modeling. To achieve this objective, we use the same evidential datasets and 12 training prospect locations that were used by Carranza (2004) in WofE modeling of porphyry-Cu prospectivity in Abra area (Philippines).

2. RF algorithm

Random Forests are an ensemble of multiple decision trees, or a set of hierarchically organized restrictions or conditions, which are successively applied from a root (parent) node to a terminal (or child) node or leaf of a tree to make repeated predictions of the phenomenon represented by training data (Breiman, 2001). The decision trees can be either classification trees or regression trees (RTs). Each decision tree in RF uses a training subset that is randomly chosen and then replaced for a number of times equal to the number of trees in the ensemble. That means each decision tree uses bootstrap aggregation, referred to as bagging (Breiman, 1996), whereby approximately two-thirds of the training samples are employed for prediction (and referred to as bag samples) while the remaining roughly one-third of the training samples are employed for validation of prediction accuracy (and referred to as out-of-bag (OOB) samples). Meanwhile, for each node/split in a decision tree, a random selection of the predictor variables (or predictors) is made. The prediction output of RF is based on the average of the prediction of all the regression trees.

To run the analysis of the decision trees, recursive splitting and multiple classifications or regressions are carried out from the dataset (for details, e.g., flowchart of RF, see Rodriguez-Galiano et al. (2014)). In other words, the RF algorithm starts by splitting the target variable, or the parent node (root), into binary pieces, where the child nodes are ‘purer’ than the parent node. Through this process, the decision trees search through all candidate splits to find the optimal split that maximizes the ‘purity’ of the resulting tree. Whereas regression trees can be grown or pruned according to a specific condition, decision trees in RF can be grown to maximum 'purity'. The RF algorithm uses the Gini impurity index (Breiman et al., 1984) to calculate the information purity of child nodes compared to that of their parent node. Split thresholds are determined from the maximum reduction in purity (Breiman, 2001). From the root (parent) node, the data splitting process in each internal node of a restriction or condition of the tree is repeated until a pre-specified stop condition is reached. Each of the terminal (child) nodes, or leaves, has attached to it a simple regression model, which applies in that node only.

For data-driven predictive mapping of mineral prospectivity, based on training data of the target variable consisting of 1s (representing known deposit locations) and 0s (representing non-deposit locations), the RF consists of multiple regression trees (Rodriguez-Galiano et al., 2014). Therefore, the predictions are floating values ranging from 0 to 1 denoting likelihoods of mineral deposit
occurrence, which can be classified using a certain threshold value for mapping of prospective and non-prospective areas.

3. Application to Abra area (Philippines)

3.1 Geology and porphyry copper mineralization

The study area is located in the province of Abra in northwestern Philippines (Fig. 1). Its shape was defined by the joint Philippine-Japan geological and mineral exploration project (JICA, 1980). The area lies along the north-south trending Luzon Central Cordillera, which is a regional geanticlinal mountain range representing a magmatic arc linked to the eastward subduction of the South China Sea oceanic crust beneath the Philippine Plate along the Manila Trench. The magmatic arc is represented by older batholithic plutons, which intrude and form the core of the regional geanticline, and younger porphyry plutons. The area is cut by fault/fracture splays the form the northern sections of the sinistral strike-slip Philippine fault system (Balce et al., 1980), which is a response to east-west compressive stresses associated with opposing convergent systems on either side of the Philippine archipelago (Rangin, 1991). Hydrothermal mineral resources are abundant in the Luzon Central Cordillera (BMG, 1986). Outside the study area, for example, some major prospects (i.e., indicated/inferred resources) and deposits (i.e., measured resources) of porphyry-Cu are known (Sillitoe and Gappe, 1984; BMG, 1986) (Fig. 1). In the study area, several small prospects of porphyry-Cu are also known (Bureau of Mines, 1976; JICA, 1980; BMG, 1986).

The geology and porphyry-Cu mineralization in the study area are described by Bureau of Mines (1976), JICA (1980), BMG (1982) and several unpublished reports in public domain archives of Philippine Mines and Geosciences Bureau. The area is underlain mostly by the Eocene Licuan Group Formation II (Fig. 1), which is made up of pyroclastic rocks and andesitic lavas with intercalated limestone lenses. The Oligocene Tineg Formation, which unconformably overlies the Licuan Group Formation II, is made up of pyroclastic rocks and dacitic lavas with intercalations of sandstone, mudstone, and limestone. The Late Miocene to Pliocene Alava Formation, which unconformably overlies the Tineg Formation, consists of sandstone and tuffaceous siltstone in its lower portions, and limestone in its upper portions. Diorite and granodiorite, belonging to the Agno Batholith, have intruded the Licuan and Tineg Formations. The different plutons of the Agno Batholith were intruded in several phases, based on radiometric dating. The older (diorite) phases have an average age of 27 Ma whereas the younger (granodiorite) phases have an age range of 12-15 Ma (Wolfe, 1981). Stocks/dikes of Late Miocene quartz-diorite porphyry intruded into pre-existing volcano-sedimentary rocks and batholithic plutons. The area is cut mostly by north-south and northeast trending faults/fractures (Fig. 1). The fault/fracture system in the area forms part of horse-tail-like strike-slip structures at the northern sections of the Philippine fault system (Ringensch et al., 1990) (Fig. 1). The area has been explored mainly for porphyry-Cu deposits. Bureau of Mines (1976) reported 12
porphyry-Cu prospects whereas JICA (1980) mapped 12 mineralized zones (Fig. 2). Only four of the 12 prospects exist within the mapped mineralized zones, and three of the 12 prospects lie close to the boundary of a mapped mineralized zone (Fig. 1). Prospect ‘9’ in mineralized zone ‘G’ is known as Kilongolao prospect (Fig. 1), which was explored by diamond drilling in early 1970s by Marcopper Mining Corporation (Sillitoe and Gappe, 1984).

3.2 Conceptual model of porphyry-Cu mineralization

The following geological events, hypothesized by the Bureau of Mines (1976), JICA (1980) and several unpublished reports in the study area, have brought about porphyry-Cu mineralization in the area. Batholithic plutons (diorite and granodiorite) intruded into volcano-sedimentary rocks along north-south trending faults/fractures. Northeast trending faults/fractures were formed during or after the intrusion of batholithic plutons. Then, porphyry dikes/stocks (quartz diorite porphyry) intruded along north-south and northeast trending faults/fractures. Hydrothermal solutions related to porphyry plutons formed porphyry-Cu mineralizations. These hypotheses were based on field observations of Bureau of Mines (1976) and JICA (1980) that porphyry-style mineralization and/or hydrothermal alteration (a) occur near and along the margins of batholithic (mainly granodiorite) plutons where porphyry dikes/stocks intruded and (b) are related spatially with north-south and northeast trending faults/fractures.

3.3 Spatial datasets

The spatial datasets used include a geological map, map of faults/fractures, and locations of porphyry-Cu prospects (Fig. 1), which were digitized from paper maps in published literature (Bureau of Mines, 1976; JICA, 1980) and in unpublished maps/reports in the Mines and Geosciences Bureau of the Philippines. These are the same spatial datasets used by Carranza (2004) to map porphyry-Cu prospectivity in the study area using WofE modeling. The digital capture and map operations to prepare the spatial datasets were carried out using ILWIS (Integrated Land and Water Information Systems), a raster-based GIS software package (http://www.ilwis.org/).

We used a pixel (or cell) size of 100 m by 100 m in rasterizing the input vector maps to prepare the spatial datasets needed for the analysis. This pixel size was objectively determined based on the spatial pattern of the known porphyry-Cu prospects so that it is suitable to the scale of the analysis, that it adequately represents the spatial resolutions of the datasets, and that only one prospect is present in any given pixel (Carranza, 2009c); it is also a realistic lateral extent for a mineralized area. The same pixel size was used by Carranza (2004) in predictive modeling of porphyry-Cu prospectivity in the study area through WofE modeling.

3.3.1 Target variable

Mineral deposit occurrence is a dichotomous variable, which is represented by prospect and non-
prospect locations that are given scores of 1 and 0, respectively (Fig. 2). For prospect locations, we used the locations of 12 porphyry-Cu prospects. For non-prospect locations, we applied three selection criteria, as explained below (cf. Carranza et al., 2008a), to ensure that they have a strong likelihood of not being undiscovered prospect locations.

Firstly, the number of non-prospect locations is equal to the number of prospect locations, because mineralizations are rare events and because the use of equal number of ‘zeros’ (e.g., non-prospect locations) and ‘ones’ (e.g., prospect locations) in regression is optimal when the latter represents rare events (Breslow and Cain, 1988; Schill et al., 1993). In addition, in cases of rare ‘ones’, like the target variable in this study, King and Zeng (2001) aver that information content from predictors starts to diminish as the number of ‘zeros’ exceeds the number of ‘ones’. Secondly, non-prospect locations are distal to any known prospect location because locations proximal to existing mineral deposits are likely to have similar multivariate spatial data signatures as the prospect locations and thus preclude achievement of desired results. Thus, we applied point pattern analysis (Diggle, 1983; Boots and Getis, 1988) to find distance from any prospect location and corresponding probability that there is one prospect location situated next to it. In the study area, the distance from any porphyry-Cu prospect location in which there is 100% probability of a neighboring porphyry-Cu prospect location is 7 km. Hence, we selected 12 non-prospect locations from areas beyond 7 km of every known porphyry-Cu prospect location. Thirdly, in contrast to prospect locations, which tend to cluster and are thus non-random, non-prospect locations must be spatially randomly distributed. Thus, we applied point pattern analysis (Diggle, 1983; Boots and Getis, 1988) to evaluate the degrees of spatial randomness of every set of 12 non-prospect locations we have selected and then we used the set with strongest spatial randomness.

3.3.2 Predictor variables

Based on hypotheses about porphyry-Cu mineralization in the study area (section 3.2), it is instructive to quantify spatial associations of the target variable with (1) batholithic granodiorite margins, (2) batholithic diorite margins, (3) porphyry pluton centroids, (4) north-south trending faults/fractures and (5) northeast trending faults/fractures. To determine indirectly whether other structures in the study area are genetically linked to porphyry-Cu mineralization, it is instructive to quantify spatial associations of the target variable with (6) northwest trending faults/fractures and (7) east-west trending faults/fractures. It is also instructive to quantify spatial association of the target variable with (8) intersections of regional structures (i.e., north-south and northeast trending faults/fractures) because intersections of faults/fractures represent zones of high structural permeability that could favour hydrothermal fluid circulation. Quantifying the spatial associations of interest requires distances to pertinent geological features as predictors (Fig. 2), which are the same as those used by Carranza (2004) in WofE modeling.

In addition, we used the same two sets of geochemical variables (principal component or PC
scores) derived by Carranza (2004) through catchment basin analysis of stream sediment geochemical data (Carranza and Hale, 1997). The PC1 (Cu, Zn) and PC2 (Cu, -Zn) scores were attributed to sample catchment basins and were used as predictor variables (9) and (10), respectively (Fig. 2), in RF to determine which PC exhibits robust spatial association with and, thus, reflects meaningful drainage anomaly signature of porphyry-Cu mineralization in the area. Because the PC scores of stream sediment geochemical data are attributed sample catchment basins, but not interpolated, there are parts (~25%) of the study area with missing values for predictors 9 and 10 (Fig. 2).

3.4 RF analysis

In the randomForest package that we used, which is implemented in the R statistical environment (Liaw and Wiener, 2002; R Development Core, 2008), the parameters that need to be set are the number of trees ($k$) to be generated and the number of predictors ($m$) randomly sampled at each split. Rodriguez-Galiano et al. (2014) found in their study (with nine predictors and training samples of 46 deposit and 57 non-deposit locations) a minimum $k$ of 1,000 that resulted in the lowest prediction errors and the most stable predictions. Based on the study of Oshiro et al. (2012), the density of our datasets (with 10 predictors and training samples of 12 prospect and 12 non-prospect locations) is roughly dissimilar to that of Rodriguez-Galiano et al. (2014) and, therefore, we used a larger $k$ of 20,000 to obtain stable predictions. We did not apply pruning because, according to Breiman (1996) and Buhlmann and Yu (2002), the variance of the output error is reduced if several similar datasets (predictors and training samples) are created by bagging and if regression trees are grown without pruning and averaged. For the value of $m$, which is a fraction of the total number of predictors, the square root of the number of predictor variables can be used (Breiman et al., 1984). We used, however, the ‘tuneRF’ function in R, which determines the optimum value of $n$ that minimizes the OOB error (Liaw and Wiener, 2002). With these settings, the RF algorithm (i) finds a fit between the target variable and all the predictor variables of the training locations and (ii) then applies the RF-fit to compute probabilities at all locations (both training and non-training ones). However, like in regression analysis, training locations for target variables (both 1s and 0s) with missing values for any of the predictor variables are excluded in finding the RF-fit, and probabilities are not calculated for non-training locations with missing values for any of the predictor variables.

For predictors with missing values, Stekhoven and Bühlmann (2012) developed an imputation technique called 'missForest' that uses the RF algorithm whereby (i) a predictor variable with missing values is used as the pseudo-target variable and is fitted with all the other predictor variables without missing values and (ii) missing values in the pseudo-target variable are predicted or imputed using the RF-fit. Readers are referred to Stekhoven and Bühlmann (2012) for details of the missForest imputation algorithm. We used the missForest package implemented in the R statistical environment (R Development Core, 2008; Stekhoven, 2013). After imputation of missing values in a predictor layer, RF training and prediction of the real target variable follow.
4. Results and discussion

4.1 Spatial associations between target and predictor variables

The RF algorithm ranks the importance of input predictors according to a predictor's marginal effect on the target variable while holding all the other predictors constant. The marginal effect is, therefore, akin to bivariate measures of spatial associations between target and predictor variables. A graph of RF marginal effect on target variable versus values of a predictor variable (Figs. 3 and 4) is akin to a graph of WofE contrast versus values of a predictor variable. However, RF modeling, unlike WofE modeling, does not indicate whether spatial association is positive or negative and it is up to the analyst to interpret the results in that respect.

The results of RF modeling (Fig. 3) show that among all the predictors, either with missing values or with imputed values for predictors 9 and 10, the five most important predictors are: (3) distance to porphyry pluton centroids; (10) PC2 (Cu, -Zn) scores; (5) distance to northeast trending faults/fractures; (2) distance to batholithic diorite margins; and (8) distance to intersections of regional structures (i.e., north-south and northeast trending faults/fractures). The spatial association of the known porphyry-Cu prospects with porphyry pluton centroids is positive, meaning that the former are located proximal to the latter, and that the positive spatial association is optimal within 3 km of porphyry pluton centroids. The spatial association of the known porphyry-Cu prospects with PC2 scores of stream sediment geochemical data is positive, meaning that locations of the former are characterized by high PC2 scores. The spatial association of the known porphyry-Cu prospects with northeast trending faults/fractures is positive, but it is optimal within 10 km of northeast trending faults/fractures. The spatial association of the known porphyry-Cu prospects with batholithic diorite margins is negative (i.e., porphyry-Cu prospects are located distal to batholithic margins); therefore, proximity to batholithic diorite margins is negative evidence of porphyry-Cu prospectivity. The spatial association of the known porphyry-Cu prospects with intersections of north-south and northeast trending faults/fractures is positive, and it is optimal within 2 km of intersections of north-south and northeast trending faults/fractures. The results further show that the target variable has "flat" responses to the other five predictors (Fig. 3), indicating that each of the latter lacks spatial association with known porphyry-Cu prospects or that the geological or geochemical processes they represent are not genetically associated with porphyry-Cu mineralization in the study area.

In terms of mineralization, the results indicate that porphyry plutons are genetically associated with porphyry-Cu mineralization in the study area, but diorite plutons are unlikely genetically linked to porphyry-Cu mineralization in the study area. These interpretations are consistent with field observations of porphyry-Cu mineralizations in the study area (Bureau of Mines, 1976; JICA, 1980) and elsewhere (Sillitoe and Gappe, 1984) and with results and interpretations of spatial analysis of porphyry-Cu mineralizations elsewhere (Carranza and Hale, 2002b; Carranza et al., 2008b). The
results also indicate that porphyry-Cu mineralization in the study area is controlled by intersections of north-south and northeast trending faults/fractures but not by northeast trending faults/fractures. This interpretation is consistent with results and interpretations of spatial analysis of porphyry-Cu mineralizations elsewhere, which indicate that porphyry-Cu mineralization occurs in zones of discontinuities along strike-slip structures (Carranza and Hale, 2002b; Carranza et al., 2008b). The positive spatial associations of porphyry-Cu prospects in the study area with porphyry plutons and intersections of regional strike-slip faults are consistent with general characteristics of porphyry-Cu systems worldwide (Sillitoe, 2010). Therefore, based on interpretations of results of RF modeling, the most important and most meaningful predictors of porphyry-Cu prospectivity in the study area include only (a) proximity to porphyry pluton centroids, (b) proximity to intersections of north-south and northeast trending faults/fractures, and (c) PC2 (Cu, -Zn) scores. Carranza (2004) reached the same interpretations of results of WofE modeling.

The imputations of missing values for predictors 9 and 10 have an important consequence in the results of RF modeling as shown in Figure 3, and that is the increase in contrast of spatial association of the target variable with every predictor. For example, in Figure 3a the maximum and minimum marginal effects of predictor 3 on the target variable are 0.5 and -0.4 whereas in Figure 3b the maximum and minimum marginal effects of predictor 3 on the target variable are 0.5 and -0.5. The amount of increase in contrast of spatial association varies for every predictor (Fig. 3), but this consequence is important because it means that if predictors with missing values are used in RF modeling, or in any other predictive modeling techniques, the spatial associations obtained are non-optimal. This can be critical to the interpretation of which predictors are meaningful or not. However, for the study area, the quantified spatial associations of the target variable with the predictors are strongly similar when either missing values or imputed values for predictors 9 and 10 are used, probably because the missing values in either predictor 9 or 10 comprise only ~25% of the study area. However, if all the other less important and non-meaningful predictors are excluded, the marginal effects of each of the most important predictors on the target variable increase significantly (Fig. 4). For example, in Figure 4a the maximum and minimum marginal effects of predictor 3 on the target variable are 1.0 and -0.8 whereas in Figure 4b the maximum and minimum marginal effects of predictor 3 on the target variable are 1.0 and -1.1. The results demonstrate that, in predictive modeling of mineral prospectivity, it is important to (a) represent (e.g., by imputation) missing values in predictor layers and (b) use only evidential layers having positive spatial associations with known mineral deposit occurrences.

4.2 Accuracy of RF modeling

Because in RF modeling the predictions are floating values ranging from 0 to 1 denoting the likelihood of the presence of a mineral prospect, output values of ≤0.5 are classified as non-prospect and values of >0.5 are classified as prospect. With this classification scheme, we see in Table 1 that
the RF modeling using all predictors, with missing values for predictors 9 and 10, yielded classification error of 38% for the training non-prospect locations, which is about four times as much for that of the training prospect locations. The results of RF modeling using all predictors, with imputed values for predictor 10 yielded classification error of 25% for the training non-prospect locations, which is about three times as much for that of the training prospect locations. The differences in classification errors between the prospect and non-prospect locations are reduced when using only three most important and most meaningful predictors, with either missing or imputed values for predictors 9 and 10 (Table 1). In general, the classification errors for the training non-prospect location are reduced when missing values in predictors are imputed. However, the lower classification errors for the training prospect locations compared to training non-prospect locations actually make sense because we know for certain that porphyry-Cu prospects exist at those locations but the training non-prospect locations were only selected using the criteria discussed above but not verified in the field.

The results of the RF modeling also show (Table 1) that (a) OOB errors are reduced by at least 4% when missing values in predictors are imputed and (b) OOB errors are reduced by at least 4% when only the most important and most meaningful predictors are used. Moreover, the results of the RF modeling show (Table 1) that (a) over-all accuracies are increased by at least 4% (i.e., from 80% to 85% and from 82% to 86%) when missing values in predictors are imputed and (b) over-all accuracies are increased by at least 1% (i.e., from 80% to 82% and from 85% to 86%) when only the most important and most meaningful predictors are used. Furthermore, the results of the RF modeling show (Table 1) that (a) kappa coefficients (i.e., measures of agreement between classification result and reference data (Landis and Koch, 1977)) are increased by at least 0.09 (i.e., from 0.54 to 0.67, and from 0.66 to 0.75) when missing values in predictors are imputed and (b) over-all accuracies are increased by at least 0.08 (i.e., from 0.54 to 0.66, and from 0.67 to 0.75) when only the most important and most meaningful predictors are used. However, all experiments of RF modeling resulted in kappa coefficients indicating only moderate agreements between the predictors and the training prospect/non-prospect locations (cf. Landis and Koch, 1977), and that is likely due to small number of training locations although this needs further verification. Nevertheless, the results show that RF modeling can yield results with acceptable accuracy with only 12 training prospect locations and 12 non-prospect locations.

4.3 Performance of RF modeling

The accuracy assessments of RF modeling discussed above provide some indication of the predictive quality of the output only at the training prospect and non-prospect locations but not on a spatial context. To further examine the performance of RF modeling resulting in a predictive map with floating values ranging from 0 to 1 denoting the likelihood of the presence of a mineral prospect, we adopted the procedure of obtaining occurrence-area proportion plots described by Agterberg and
Bonham-Carter (2005). This involves a series of classifications of prospective pixels based on cut-off likelihood values at 5-percentile intervals of values in predictive map. The highest (100 percentile) cut-off likelihood value results in minimum proportion \([\approx 0]\) of prospective pixels in the study area, whereas the lowest (0 percentile) cut-off likelihood value results in maximum proportion \([\approx 1]\) of prospective pixels in the study area. For prospective pixels corresponding to each cut-off likelihood value, we determined the success-rate as the proportion of training porphyry-Cu prospects delineated by prospective pixels. We then created success-rate curves for the individual predictive maps obtained from using all predictors and using only the most important and most meaningful predictors.

The success-rates of training the RF model are better when missing values for predictors are replaced by imputed values (Fig. 5), although the success-rates of training the RF model when using all predictors and when using only the most important predictors are more or less similar. However, the success-rates of the RF models are better than that of the WofE model of Carranza (2004) using only the most important predictors and with weights for missing predictor data replaced by zeros. Note that in WofE, which is a log-linear model, weights in logit form are added to update posterior probability; thus, weights of zeros (representing lack of spatial association because of missing data) have no effect on prior probability. Therefore, the results show that RF modeling with missing values imputed provides better fit between target and predictor variables compared to WofE modeling.

However, the success-rate only represents an overall measure of adequacy (or goodness-of-fit) of all the analyses and interpretations of the datasets in order to ‘fit’ the predictors and the training prospect/non-prospects locations but it does not represent a measure of the chance of finding undiscovered prospects (i.e., prediction-rate). To quantify the prediction-rate of a predictive model of mineral prospectivity, the approach proposed by Fabbri and Chung (2008) for ‘blind testing’ of prediction models should be followed. This approach involves removing one prospect location from a training set of \(d\) number of prospect locations and using only \(d - 1\) prospect locations to create a map of prospect occurrence likelihood values, which is then cross-validated with the prospect location removed from the analysis. The blind testing is performed \(d\) times, each time using a different prospect location for blind testing. However, we did not follow this approach because then we would only have 11 prospect locations each time for training and it was not the objective of this study to determine if we could use RF for prospectivity modeling using less than 12 training prospect locations. Alternatively, we used the mineralized zones mapped by JICA (1980), shown in Fig. 1, as independent data for testing and then we followed the procedure of obtaining occurrence-area proportion plots described by Agterberg and Bonham-Carter (2005) to create prediction-rate curves for the individual predictive maps obtained from using all predictors and using only the most important and most meaningful predictors.

The prediction-rate curves of the RF models using all predictors and only the most important predictors, with missing values for predictors 9 and/or 10, are inferior to that of the WofE model of
Carranza (2004) using only the most important predictors and with weights for missing predictor data replaced by zeros (Fig. 6a). The prediction-rate curve of the RF model using all predictors, with imputed values for predictors 9 and/or 10, is similar to that of the WofE model of Carranza (2004) (Fig. 6b). However, the prediction-rate curve of the RF model using only the most important predictors, with imputed values for predictor 10, is better than that of the WofE model of Carranza (2004) (Fig. 6b). These illustrate that RF modeling of mineral prospectivity using predictors with imputed values for missing data can outperform WofE modeling.

Based on the success-rate and prediction-rate curves (Figs. 5 and 6), the RF model obtained from the most important predictors, with imputed values for predictor 10 (Fig 4b), is the best predictive model of porphyry-Cu prospectivity because it outlines 30% of the study area as prospective with 100% success-rate (Fig. 5b) and 70% prediction-rate (Fig. 6b). This model (Fig. 7) can be used to guide further exploration of porphyry-Cu prospects in the study area. Apparently, mineralized zones A, B, E, F and G should be prioritized, but not necessarily in that order, for porphyry-Cu exploration.

5. Conclusion

Machine learning methods, which have been used in data-driven predictive modeling of mineral prospectivity, such as artificial neural networks (e.g., Porwal et al., 2003; 2004) and support vector machines (e.g., Zuo and Carranza, 2011; Abedi et al., 2012), usually require a large number of training prospect/deposit locations (e.g., >20) and are unsuitable for evidential data with missing values. However, as shown in this study, the Random Forests algorithm, which is also a machine learning method, can be used in data-driven predictive modeling of mineral prospectivity in the case of few training prospect/deposit locations. Nevertheless, this proposition needs further verification by testing RF modeling in other areas and with few (i.e., <20) occurrences of the same or different deposit-types. An advantage of Random Forests over artificial neural networks and support vector machines is the linked imputation technique for representation of missing values in evidential data. This is an important advantage because evidential data with missing values is a common feature of areas with few mineral occurrences. Another important advantage of Random Forests over artificial neural networks and support vector machines is that it allows analysis of the target variable's response to each predictor variable, which makes Random Forests a non-black-box but rather an exploratory method, just like weights-of-evidence and evidential belief modeling, of data-driven predictive modeling of mineral prospectivity.

Acknowledgment

We thank Dr. Graeme Bonham-Carter and two anonymous reviewers for their constructive comments on the initial version of this paper. We also thank Dr. Jef Caers, Editor-in-Chief of Computers & Geosciences, for expert handling of our manuscript.
References


Nykanen, V., 2008. Radial basis function link nets used as a prospectivity mapping tool for orogenic gold deposits within the Central Lapland Greenstone Belt, northern Fennoscandian Shield. Natural Resources Research 17, 29–48.


Table 1. Accuracy of RF modeling using all predictors and only the three most important predictors having positive spatial associations with the target variable.

<table>
<thead>
<tr>
<th></th>
<th>Using all predictors, with missing values for predictors 9 and 10</th>
<th>Using only the three most important and most meaningful predictors, with missing values for predictor 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1 (prospect location) error</td>
<td>10% (no. of 1s = 10)</td>
<td>10% (no. of 1s = 10)</td>
</tr>
<tr>
<td>Class 0 (non-prospect location) error</td>
<td>38% (no. of 0s = 8)</td>
<td>25% (no. of 0s = 8)</td>
</tr>
<tr>
<td>OOB error</td>
<td>22%</td>
<td>17%</td>
</tr>
<tr>
<td>Over-all accuracy</td>
<td>80%</td>
<td>82%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.54</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Using all predictors, with imputed values for predictors 9 and 10</th>
<th>Using only the three most important and most meaningful predictors, with imputed values for predictor 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1 (prospect location) error</td>
<td>8% (no. of 1s = 12)</td>
<td>8% (no. of 1s = 12)</td>
</tr>
<tr>
<td>Class 0 (non-prospect location) error</td>
<td>25% (no. of 0s = 12)</td>
<td>17% (no. of 0s = 12)</td>
</tr>
<tr>
<td>OOB error</td>
<td>17%</td>
<td>13%</td>
</tr>
<tr>
<td>Over-all accuracy</td>
<td>85%</td>
<td>86%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.67</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Fig. 1. (a) Regional geotectonic map of Philippines. (b) Location of the case study area in Abra province, locations of major porphyry-Cu deposits/prospects (black dots) in and around province of Abra (Sillitoe and Gappe, 1984; BMG, 1986) and locations of small porphyry-Cu prospects (squares) in the study area (Bureau of Mines, 1976). Map coordinates in meters (UTM projection, zone 51). (c) Simplified geological map of study area (modified from JICA, 1980). Think black lines are mapped/interpreted faults/fractures. Number-identified circles are locations of porphyry-Cu prospects (Bureau of Mines, 1976). Letter-identified polygons are mapped mineralized zones (JICA, 1980).
Fig. 2. Variables used in RF modeling. The target variable comprises values of 1 (existing prospects) and 0 (non-prospect locations). Predictor variables 1 to 8 are distances to (1) batholithic granodiorite margins, (2) batholithic diorite margins, (3) porphyry pluton centroids, (4) north-south trending faults/fractures, (5) northeast trending faults/fractures, (6) northwest trending faults/fractures, (7) east-west trending faults/fractures, (8) intersections of regional structures (i.e., north-south and northeast trending faults/fractures). The last two predictor variables, derived from stream sediment geochemical data, are (9) PC1 (Cu, Zn) scores and (10) PC2 (Cu, -Zn) scores. Values of the predictor variables in the grey-scale maps vary from minimum (white) to maximum (black), although for predictors 9 and 10 missing values are white and minimum values are lightest grey.
Fig. 3. Predictors of porphyry-Cu occurrence arranged in decreasing order of importance (from left to right and top to bottom) based on marginal effect: (a) with missing values in predictors 9 and 10 (see Fig. 2); (b) with imputed values in predictors 9 and 10 (denoted by 9i and 10i, respectively). Predictor variables are: 1 = distance to batholithic granodiorite margins; 2 = distance to batholithic diorite margins; 3 = distance to porphyry pluton centroids; 4 = distance to north-south trending faults/fractures; 5 = distance to northeast trending faults/fractures; 6 = distance to northwest trending faults/fractures; 7 = distance to east-west trending faults/fractures; 8 = distance to intersections of regional structures (i.e., north-south and northeast trending faults/fractures); 9 = PC1 (Cu, Zn) scores; and 10 = PC2 (Cu, Zn) scores. Values of predictors 1 to 8 are in kilometers, whereas values of predictors 9 and 10 are unitless (i.e., PC scores). Inner hash marks along the x-axis of each graph are deciles.
Fig. 4. Most important and most meaningful predictors of porphyry-Cu occurrence arranged in decreasing order of importance (from left to right) based on marginal effect: (a) with missing values in predictor 10 (see Fig. 2); (b) with imputed values in predictor 10 (denoted by 10i). Predictor variables are: 3 = distance to porphyry pluton centroids; 8 = distance to intersections of regional structures (i.e., north-south and northeast trending faults/fractures); and 10 = PC2 (Cu, -Zn) scores. Values of predictors 3 and 8 are in kilometers, whereas values of predictor 10 are unitless (i.e., PC scores). Inner hash marks along the x-axis of each graph are deciles.
Fig. 5. Success-rate (goodness-of-fit) curves of RF-based predictive models of porphyry-Cu prospectivity using predictors (a) with missing values in predictors 9 and/or 10 and (b) with imputed values in predictors 9 and/or 10.
Fig. 6. Prediction-rate curves of RF-based predictive models of porphyry-Cu prospectivity using predictors (a) with missing values in predictors 9 and/or 10 and (b) with imputed values in predictors 9 and/or 10.
Fig. 7. Predictive map (a) of likelihood values of porphyry-Cu prospectivity obtained by using 12 training prospect locations and only the most important and most meaning predictors (Fig. 4b), with imputed values for predictor 10, and the corresponding binary prospectivity map (b) with prospective areas covering 30% of the test area. Number-identified circles are locations of porphyry-Cu prospects (Bureau of Mines, 1976). Letter-identified polygons are mapped mineralized zones (JICA, 1980).
Highlights:

- The Random Forest (RF) algorithm is tested data-driven modeling of mineral prospectivity.
- The RF algorithm can be used in areas with few (i.e., <20) mineral occurrences.
- The RF algorithm can handle evidential data with missing values.
- The RF algorithm allows analysis of spatial associations of prospects with every evidence layer.