Ore Geology Reviews 71 (2015) 777–787

Contents lists available at ScienceDirect

Ore Geology Reviews

journal homepage: www.elsevier.com/locate/oregeorev

Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm

Emmanuel John M. Carranza [a,*], Alice G. Laborte [b]

[a] School of Earth and Oceans, James Cook University, Townsville 4811, Queensland, Australia
[b] International Rice Research Institute, Los Baños 4030, Laguna, Philippines

ARTICLE INFO

Article history:
Received 5 June 2014
Received in revised form 12 August 2014
Accepted 23 August 2014
Available online 29 August 2014

Keywords:
Mineral prospectivity mapping
Ensemble of regression trees
Epithermal Au
Spatial correlation

ABSTRACT

The Random Forests (RF) algorithm has recently become a fledgling method for data-driven predictive mapping of mineral prospectivity, and so it is instructive to further study its efficacy in this particular field. This study, carried out using Baguio gold district (Philippines), examines (a) the sensitivity of the RF algorithm to different sets of deposit and non-deposit locations as training data and (b) the performance of RF modeling compared to established methods for data-driven predictive mapping of mineral prospectivity. We found that RF modeling with different training sets of deposit/non-deposit locations is stable and reproducible, and it accurately captures the spatial relationships between the predictor variables and the training deposit/non-deposit locations. For data-driven predictive mapping of epithermal Au prospectivity in the Baguio district, we found that (a) the success-rates of RF modeling are superior to those of weights-of-evidence, evidential belief and logistic regression modeling and (b) the prediction-rate of RF modeling is superior to that of weights-of-evidence modeling but approximately equal to those of evidential belief and logistic regression modeling. Therefore, the RF algorithm is potentially much more useful than existing methods that are currently used for data-driven predictive mapping of mineral prospectivity. However, further testing of the method in other areas is needed to fully explore its usefulness in data-driven predictive mapping of mineral prospectivity.

1. Introduction

Predictive mapping of mineral prospectivity involves the analysis and synthesis of various layers of spatial evidence derived from various pertinent geoscience spatial data sets in order to outline and prioritize areas that are prospective for exploration of undiscovered mineral deposits of the type sought (Bonham-Carter, 1994; Carranza, 2009b). Predictive mapping of mineral prospectivity can either be knowledge-driven or data-driven. Knowledge-driven predictive mapping of mineral prospectivity is suitable in less-explored (or so-called 'greenfields') geologically permissive regions where no or very few mineral deposits are known to occur (e.g., Lusty et al., 2012). In knowledge-driven predictive mapping of mineral prospectivity, the weights assigned to every layer of spatial evidence reflect one's 'expert' judgment of its spatial association with mineral deposits of the type sought. In contrast, data-driven predictive mapping of mineral prospectivity is suitable in moderately- to well-explored (or so-called 'brownfields') geologically permissive regions, where the objective is to demarcate new targets for further exploration of undiscovered mineral deposits of the type sought (e.g., Mejía-Herrera et al., 2014). In data-driven predictive mapping of mineral prospectivity, the weights assigned to individual layers of spatial evidence are quantified spatial associations between discovered mineral deposits and individual data sets used to represent prospectivity recognition criteria.

Data-driven predictive mapping of mineral prospectivity makes use of mathematical methods that involve either bivariate or multivariate analysis. Bivariate techniques involve pair-wise analysis of spatial association between a map of mineral deposit occurrence and a map of an evidential dataset. There are two bivariate techniques commonly used for data-driven predictive mapping of mineral prospectivity: (i) weights-of-evidence modeling (Agterberg, 2011; Agterberg et al., 1990, 1993; Bonham-Carter et al., 1988, 1989); and (ii) evidential belief modeling (Carranza, 2009a, 2011a, 2014; Carranza and Hale, 2003; Carranza and Sadeghi, 2010; Carranza et al., 2005, 2008a,b,c, 2009). In contrast, multivariate techniques involve simultaneous analysis of spatial associations between a map of mineral deposit occurrence and maps of various evidential datasets. The multivariate techniques commonly used for predictive mapping of mineral prospectivity include: (i) logistic regression (Chung, 1978, 1983; Chung and Agterberg, 1980, 1988; Carranza and Hale, 2001; Harris et al., 2001, 2006); and (ii) artificial neural networks (Behnia, 2007; Harris et al., 2003; Nykänen, 2008;

* Corresponding author.
E-mail address: john.carranza@jcu.edu.au (E.J.M. Carranza).

http://dx.doi.org/10.1016/j.oregeorev.2014.08.010
0169-1368/

Porwal et al., 2003, 2004; Rigol-Sanchez et al., 2003; Singer and Kouda, 1996, 1999).

The number of multivariate techniques used for data-driven predictive mapping of mineral prospectivity is greater than the number of bivariate techniques used for the same purpose (Carranza, 2009b, 2011b). This indicates that, because of the highly complex, and probably non-linear, nature of spatial associations between mineral deposits and geological features, it is usually more desirable to develop and/or apply multivariate rather than bivariate techniques for data-driven predictive mapping of mineral prospectivity. Nevertheless, multivariate techniques, like bivariate techniques, show a variety of problems that can undermine the accuracy of predictive mapping of mineral prospectivity in different cases. For example, logistic regression assumes that the distribution of the data is known and follows a stochastic data model, and that variables are independent. These assumptions rarely apply, however, to data-driven predictive mapping of mineral prospectivity whereby data distributions are usually not known a-priori and evidential data represent geological processes involved in mineral deposit formation that are not independent of each other. Moreover, logistic regression is quite sensitive to data outliers. On the other hand, artificial neural networks, like other data-driven techniques, require a large number of known locations of mineral deposits of the type sought for quantifying spatial associations with multiple layers of evidential data. More importantly, unlike the parameters (i.e., coefficients representing degrees of spatial associations between mineral deposits and evidential data) in bivariate techniques and other multivariate techniques such as logistic regression, the parameters in artificial neural networks are not interpretable in terms of relative importance of predictor maps. In other words, artificial neural networks do not provide insights into the inter-play of geologic controls on mineralization.

In the last two decades or so, various multivariate methods have been developed in various fields to overcome some of the problems discussed above. From the field of machine learning, ensemble methods like bagging (Breiman, 1996) and Random Forests (Breiman, 2001) are 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., Bradter et al., 2013; Evans and Cushman, 2009; Prasad et al., 2006). Meanwhile, the application of RF to spatial data integration for geological mapping is also growing (e.g., Cracknell and Reading, 2013, 2014; Cracknell et al., 2013; Waske et al., 2009). Just recently, Rodriguez-Galiano et al. (2014) have demonstrated the applicability of RF to data-driven predictive mapping of mineral prospectivity for epithermal Au deposits in the Rodalquilar district (Spain).

As the method of RF has now become a fledgling multivariate technique for data-driven predictive mapping of mineral prospectivity, various questions regarding its applicability as well as efficacy need to be answered. Rodriguez-Galiano et al. (2014) have investigated the optimum number of trees and optimum number of random variables needed for proper training in RF modeling (see Section 2). Because training in RF modeling requires both training samples for deposit and non-deposit locations and because there can be an infinite number of sets of non-deposit locations, we investigate in this paper the following questions: (a) is RF modeling sensitive to different training sets of non-deposit locations? and (b) is RF modeling better than weights-of-evidence, logistic regression and evidential belief modeling in terms of predictive ability to delineate prospective areas for mineral deposits of the type sought? The latter three techniques have been applied by Carranza (2002) and Carranza and Hale (2000, 2001, 2003) to predict prospective areas for epithermal Au deposits in the Baguio gold district (Philippines). Therefore, to answer the second question, we also applied the RF algorithm using the same datasets for predictive mapping of prospectivity for epithermal Au in the same district.

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, 1984, 2001). The decision trees can be either classification trees or regression trees (RTs). Every decision tree in RF employs a training subset that is randomly chosen as much times with replacement as the number of trees in the ensemble. That means every decision tree employs bootstrap aggregation, referred to as bagging (Breiman, 1996), whereby roughly two-thirds of the training samples are used to create a prediction (and referred to as bag samples) while the remaining roughly one-third of the training samples are used to validate the accuracy of the prediction (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 final prediction output of RF (in regression) is the average of the prediction of all the regression trees.

To induce the decision trees, recursive splitting and multiple classifications or regressions are carried out from the data set. From the root (parent) node, the process of data splitting in every internal node of a restriction or condition of the tree is repeated until a pre-specified stop condition is achieved. Each of the terminal (child) nodes, or leaves, has attached to it a simple regression model, which applies in that node only. 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 pruned or grown until a specific condition is achieved, decision trees in RF can be grown to maximum 'purity'. The RF algorithm uses the Gini impurity index (Breiman, 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).

For data-driven predictive mapping of mineral prospectivity, based on the training data of the target variable consisting of 1 s (representing deposit locations) and 0 s (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 test area: the Baguio gold district

3.1. Geology and epithermal Au mineralization

Five lithologic formations underlie the Baguio gold district (Fig. 1). The oldest formation — Pugo Formation of Cretaceous to Eocene age — comprises a sequence of metasedimentary and metavolcanic rocks. Unconformably overlying the Pugo Formation is the Zigzag Formation, which, according to Balce et al. (1980) is made up mostly of marine sedimentary of Early to Middle Miocene age. However, andesite porphyries, which have been dated 15.0 ± 1.6 Ma (Wolfe, 1981) or pre-Middle Miocene, have intruded into the Zigzag Formation. Therefore, Mitchell and Leach (1991) consider the Zigzag Formation to be largely Late Eocene although it may include rocks of Early Miocene age. Unconformably overlying the Zigzag Formation is the Kennon Formation of Middle Miocene age (Balce et al., 1980), which consists of limestones that outcrop in a discontinuous north-trending belt west of the district. Unconformably overlying all of the above-mentioned formations is the Klondyke Formation of Late Miocene age (Balce et al., 1980; Mitchell and Leach, 1991; Wolfe, 1988), which is composed mainly of clastic rocks that are very largely or entirely andesitic in composition. The