




# The spatial leave-pair-out cross-validation method for reliable AUC estimation of spatial classifiers

Antti Airola<sup>1</sup>  · Jonne Pohjankukka<sup>1</sup> · Johanna Torppa<sup>2</sup> · Maarit Middleton<sup>3</sup> · Vesa Nykänen<sup>3</sup> · Jukka Heikkonen<sup>1</sup> · Tapio Pahikkala<sup>1</sup>

Received: 8 November 2017 / Accepted: 26 September 2018 / Published online: 20 December 2018  
© The Author(s) 2018

## Abstract

Machine learning based classification methods are widely used in geoscience applications, including mineral prospectivity mapping. Typical characteristics of the data, such as small number of positive instances, imbalanced class distributions and lack of verified negative instances make ROC analysis and cross-validation natural choices for classifier evaluation. However, recent literature has identified two sources of bias, that can affect reliability of area under ROC curve estimation via cross-validation on spatial data. The pooling procedure performed by methods such as leave-one-out can introduce a substantial negative bias to results. At the same time, spatial dependencies leading to spatial autocorrelation can result in overoptimistic results, if not corrected for. In this work, we introduce the spatial leave-pair-out cross-validation method, that corrects for both of these biases simultaneously. The methodology is used to benchmark a number of classification methods on mineral prospectivity mapping data from the Central Lapland greenstone belt. The evaluation highlights the dangers of obtaining misleading results on spatial data and demonstrates how these problems can be avoided. Further, the results show the advantages of simple linear models for this classification task.

**Keywords** Area under ROC curve · Classifier evaluation · Cross-validation · Mineral prospectivity mapping · Spatial data mining

---

Responsible editor: Alípio Jorge, Rui L. Lopes, German Larrazabal.

---

This work was supported by the Academy of Finland (Grants 289903, 311273).

---

✉ Antti Airola  
Antti.Airola@utu.fi

Extended author information available on the last page of the article

## 1 Introduction

Mineral prospectivity mapping or mineral potential mapping (MPM) techniques are used to delineate areas favorable for mineral exploration (see e.g., Bonham-Garter 1994; Carranza 2008; Nykänen 2008). By integrating information derived from spatial geological, geophysical and geochemical datasets, the MPM methodology is used to quantify the likelihood of presence of a specific type of mineral occurrence within a study area. In supervised MPM learning techniques, the locations of known mineral deposits or occurrences are used to relate the occurrences to the mapped quantities that are indicative of the corresponding mineral deposit type. Known mineral occurrences can be also used for validating the models (Bonham-Garter 1994).

In this work, we consider the issue of supervised binary classification in spatial prediction problems. Here the goal is to train a classifier that can predict some property of a geographical area, such as the presence or absence of a mineral deposit. Training and evaluation of such classifiers is challenging because the available data is typically highly imbalanced since the amount of positive instances denoting known mineral occurrences is small. Further, instead of known negative instances, data sets usually contain only positive and unlabeled instances (see e.g., Nykänen 2008; Rigol-Sanchez et al. 2003); a setting known as positive-unlabeled (PU) learning (Elkan and Noto 2008). Works such as Bradley (1997), Fawcett (2006) and Huang and Ling (2005) have suggested the use of area under the ROC curve (AUC) for classifier evaluation on imbalanced data, as the criterion is insensitive to relative class distributions on the test set. Further, AUC has also been established as a recommended metric for PU-learning problems (Elkan and Noto 2008; Jain et al. 2017). Thus, AUC is a natural performance measure for MPM classifier evaluation, and studies such as Brown et al. (2003), Nykänen (2008), Nykänen et al. (2015) and Rodriguez-Galiano et al. (2015) have used AUC for evaluating MPM models. Further, since adequately large separate test data may not be available for MPM, cross-validation (CV) is necessary for validating the models (see e.g., Abedi et al. 2012; Rigol-Sanchez et al. 2003; Carranza 2008; Rodriguez-Galiano et al. 2015).

The prediction and success rate curves commonly used to evaluate the accuracy of predictive spatial models in geoscience applications (see Chung and Fabbri 2003; Fabbri and Chung 2008; Frattini et al. 2010; Carranza et al. 2015; Rodriguez-Galiano et al. 2015) correspond to such ROC curves, where the instances are weighted according to the area covered by them. Similarly to our study, the prediction rate curves are computed from cross-validated predictions, whereas success rate curves measure goodness-of-fit to training data. Thus while not specifically considered in the following, the area under a prediction rate curve could be also estimated using the methods considered in this work. Other metrics popularly used to evaluate spatial classifiers include prediction-area plots (Yousefi and Carranza 2015) and classification accuracy (Brown et al. 2003; Abedi et al. 2012; Rodriguez-Galiano et al. 2015) as well as various other statistics (Frattini et al. 2010), these however fall outside the scope of our work.

Based on recent literature we suggest that there are two major sources of bias that can affect results when using CV for estimating the AUC of spatial classification problems. First, standard CV methods such as leave-one-out (LOOCV) and K-fold

are often affected by a negative bias resulting from pooling together predictions from different folds for AUC computation, as shown by Airola et al. (2009), Airola et al. (2011), Forman and Scholz (2010), Parker et al. (2007) and Smith et al. (2014). Airola et al. (2009) and Airola et al. (2011) proposed a leave-pair-out CV (LPOCV) method for correcting such bias in AUC estimation. LPOCV is further validated by Smith et al. (2014) on clinical data. Second, spatial autocorrelation causes standard CV methods to produce optimistically biased prediction performance estimates for spatial data. This is caused by the fact that leave-one-out and K-fold rely on the assumption that the data is independent and identically distributed (i.i.d.)—an assumption violated by spatial data where close instances tend to be more similar than ones distant from each other. Recently, Pohjankukka et al. (2014, 2017) and Le Rest et al. (2014) have proposed spatial CV (SCV) methods for correcting this bias.

In this work, combining the leave-pair-out and spatial CV methods, we introduce the leave-pair-out spatial CV (LPO–SCV) method for evaluating MPM classifiers. As a case study, we use the approach to benchmark a number of machine learning methods on an orogenic gold MPM classification task. In our experiments, we first show that one can obtain completely misleading results if the spatial and pooling biases are not corrected for. At worst, one can obtain with standard CV methods close to perfect AUC values for classifiers, that are in reality not much better than random at making predictions for new data. We demonstrate how the LPO–SCV method corrects the pooling and spatial biases, allowing one to reliably estimate the AUC of spatial classifiers. Finally, in the LPO–SCV based classifier comparison, we show simple linear models to be surprisingly competitive on the MPM data.

## 2 Cross-validation for AUC estimation with spatial data

First, we present our mathematical notation. Let us assume a set of  $m$  instances, divided into the so-called positive and negative classes. Further, let  $\mathcal{I} = \{1, 2, \dots, m\}$  denote the index set of these instances, with  $\mathcal{I}_+ \subset \mathcal{I}$  and  $\mathcal{I}_- \subset \mathcal{I}$  denoting the indices of the positive and the negative instances, respectively. We refer to the instances only by their indices, since their other properties such as features are not required when defining AUC and the CV methods.

Further, let  $f : \mathcal{I} \rightarrow \mathbb{R}$  denote a classifier, that maps each instance to a real-value, representing how likely it is to belong to the positive class. We can use  $f$  to classify data, by assigning each  $f(i) > t$  to the positive class, and the rest to the negative class for some threshold  $t$ . Finally, when defining the cross-validation methods we refer by  $f_{\mathcal{J}}$ , where  $\mathcal{J} \subseteq \mathcal{I}$ , to a classifier trained with a machine learning method on the subset of the instances indexed by  $\mathcal{J}$ .

### 2.1 AUC

Area under the ROC curve (AUC) is a common criterion for evaluating the quality of a classifier. It estimates the probability, that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Hanley and McNeil

1982). AUC is invariant to prior class distributions, and does not require one to define class specific error costs or a threshold  $t$ . These advantages make it especially popular metric for classifier evaluation and comparison, especially in applications dealing with imbalanced data (Bradley 1997; Fawcett 2006; Huang and Ling 2005).

AUC can be computed based on the Wilcoxon–Mann–Whitney statistic (Bamber 1975) as

$$\frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} H(f(i) - f(j)), \quad (1)$$

where

$$H(a) = \begin{cases} 1, & \text{if } a > 0 \\ 0.5, & \text{if } a = 0 \\ 0, & \text{if } a < 0 \end{cases} \quad (2)$$

is the Heaviside step function,  $|\mathcal{S}|$  denotes the number of elements in set  $\mathcal{S}$ ,  $i$  and  $j$  denote the indices of a positive and a negative instance, and  $|\mathcal{I}_+|$  and  $|\mathcal{I}_-|$  denote the number of positive and negative instances, respectively. The statistic is calculated by comparing each positive instance with all the negative instances and counting the number of times the prediction value for that positive instance exceeds the negative instance (ties are counted as 0.5). Thus, the statistic represents the fraction of positive–negative pairs, where the positive has higher prediction than the negative. The AUC value is between 0 and 1, with 0.5 corresponding to a random classifier, or one that always predicts the same value.

## 2.2 Pooling bias and LPOCV

When dealing with data sets where at least one of the classes has only a small number of instances belonging to it, CV is typically used for computing the AUC. In CV, one repeatedly divides the training set into two disjoint parts. The first part is used for training a classifier, and the second for testing how accurately it predicts new data. The predictions made by the different classifiers are then combined together in order to estimate, how accurately a classifier trained on all the training data would predict on new data. One standard way of computing the AUC from CV results is known as *pooling*. In this approach, the predictions made on different rounds of CV are pooled together, and one single AUC value is computed as defined in Eq. (1). In an alternative approach known as *averaging* a separate AUC is computed for each round of CV and the mean of these is taken as the final AUC estimate (see Bradley 1997 for more thorough discussion of pooling and averaging).

However, previous work has shown that pooling based CV methods such as leave-one-out (LOO) and (pooled) K-fold CV can have a large negative bias, when used for computing AUC (Airola et al. 2009, 2011; Forman and Scholz 2010; Smith et al. 2014; Parker et al. 2007). We refer to this effect as *pooling bias*. While averaging can correct this bias, the standard averaged K-fold method has been shown to have unacceptably high variance on small imbalanced data sets (Airola et al. 2009, 2011) Recently, Airola et al. (2009), Airola et al. (2011) and Smith et al. (2014) have shown that the pooling bias can be eliminated by using an averaging-based method known as

leave-pair-out CV (LPO). In LPO, each possible positive–negative pair is left out of the training set in turn, and the classifier trained on the remaining instances. The LPO AUC estimate is then computed as the fraction of pairs, where the positive instance has a higher prediction than the negative one.

Formally, this can be defined as

$$\frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} H(f_{\mathcal{I} \setminus \{i,j\}}(i) - f_{\mathcal{I} \setminus \{i,j\}}(j)), \quad (3)$$

where  $f_{\mathcal{I} \setminus \{i,j\}}$  is the classifier trained without the  $i$ :th and  $j$ :th instances.

For an example of pooling bias, let us consider a trivial classifier  $f(i) = \frac{|\mathcal{I}_+|}{m}$ , that just predicts the fraction of positive instances in the training set. In leave-one-out, the classifier would always obtain AUC of 0, since it would predict  $\frac{|\mathcal{I}_+|}{m-1}$  when a negative instance is left out, and  $\frac{|\mathcal{I}_+|-1}{m-1}$  when a positive instance is left out. While this is an extreme example, the strong effect pooling bias can have has been established experimentally in several studies (Airola et al. 2009, 2011; Forman and Scholz 2010; Smith et al. 2014; Parker et al. 2007), and is further validated by our results. LPO avoids the pooling bias, as it is an averaging based method where predictions made on different rounds of CV are not pooled together when computing the AUC estimate. Rather, the predictions for each compared positive–negative pair come from the same round of CV.

### 2.3 Spatial bias and SCV

Most of the methodologies in statistical inference rely on the assumption that data samples are realizations from i.i.d. random variables. In cases where we are concerned with spatio-temporal data sets this assumption can have major drawbacks. Take for example geographical instances sampled from soil. We are given three instances  $i$ ,  $j$ , and  $k$  with  $i$  and  $j$  located geographically much closer to each other than  $k$  to both the previous two. Anyone could argue in this scenario that  $i$  and  $j$  are probably the most similar to each other among the three instances due to the small geographical distance between them. In 1970 Waldo R. Tobler stated in his work (Tobler 1970) the Tobler’s first law of geography: *Everything is related to everything else, but near things are more related than distant things.*

The relationship of being near versus being similar in spatial data analysis is called *spatial autocorrelation* (SAC). SAC in spatial data sets is usually measured quantitatively using variograms or Moran’s index (Cressie 2015; Longley et al. 2005). SAC tends to be naturally high for instances close to each other and small for instances more distant from each other. It is therefore clear that when we have a set of geographical data samples, they are most certainly not i.i.d., and this needs to be addressed in model evaluation and selection.

To estimate a model’s prediction performance where the effect of SAC has been reduced, Pohjankukka et al. (2014, 2017) and Le Rest et al. (2014) proposed *spatial cross validation* (SCV) to be used for this purpose. The idea in SCV is to estimate a

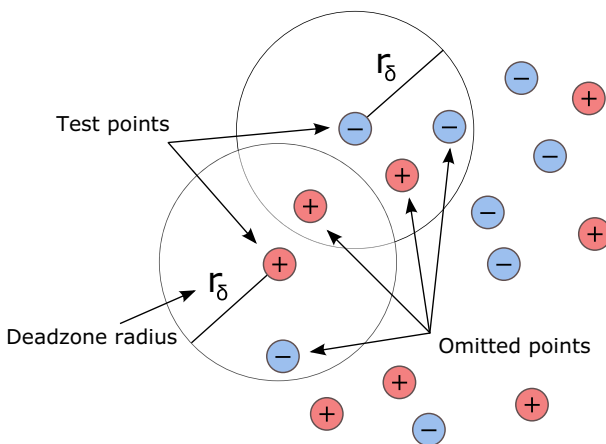
model’s prediction performance for a test point  $r$  units away from the closest known instances. This is done by altering the data in the CV procedure, so that a test point will always be at least  $r$  units away from the training data. Following Pohjankukka et al. (2017), we call this left out area the *dead zone*. SCV produces a prediction performance estimate of our model as a function of  $r$ , i.e. the distance of closest known data to the predicted instance. Thus SCV simulates the situation, where the trained model is used to make predictions for data that is further than  $r$  units of distance from the instances in the training data.

### 2.4 Spatial leave-pair-out CV

In order to eliminate both the biases caused by pooling and spatial autocorrelation simultaneously, we now introduce the LPO–SCV method, which combines the LPO and SCV methods. The method is illustrated in Fig. 1. In LPO–SCV, on each round of CV a positive–negative pair, and all the instances within  $r$  radius of these two points, are left out of the training set. The model is trained on the rest of the training set, and predictions are made for the left out positive and negative instance. The procedure is repeated for all possible positive–negative pairs. The AUC estimate is the relative fraction of pairs, for which the positive instance has a higher predicted value, than the negative one.

Formally, the estimate can be defined as follows. Let  $d(i, j)$  denote the geographical distance (e.g. Euclidean) between the  $i$ :th and  $j$ :th training instances. Further, let  $\mathcal{U}(i, j) = \{k \in \mathcal{I} | d(i, k) > r \wedge d(j, k) > r\}$  denote all training instances that have a larger distance than  $r$  from both  $i$ <sup>th</sup> and  $j$ <sup>th</sup> training instances. Then, the LPO–SCV is computed as

$$\frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} H(f_{\mathcal{U}(i,j)}(i) - f_{\mathcal{U}(i,j)}(j)), \tag{4}$$



**Fig. 1** Leave-pair-out spatial CV. On each round, a positive and negative instance are left out, as well as all the instances within the dead zone circles surrounding them. Thus the CV procedure simulates the setting, where the left out test pair is at least  $r$  distance away from nearest training instance

where  $f_{\mathcal{U}(i,j)}(i)$  is the classifier trained on all data outside circles of radius  $r$  around instances  $i$  and  $j$ .

Similarly to the ordinary LPO–CV, the approach corrects for pooling bias by ensuring that only predictions made on the same round of CV are ever compared. At the same time, the method corrects for spatial bias by excluding training data that is close to the selected pair of test instances. The dead zone ensures that the AUC result holds for data further than  $r$  units from the training instances, not just in the immediate neighborhood of the training data.

A downside of the approach is computational complexity, as full LPO–SCV requires training the classifier  $|\mathcal{Z}_+||\mathcal{Z}_-|$  times. When this is not computationally feasible, one may approximate full LPO–SCV by randomly sampling a subset of all the possible pairs. Further, for ridge regression classifiers, fast LPO–SCV can be implemented using the fast holdout algorithms (Pahikkala et al. 2012) implemented in the RLScore open source library (Pahikkala and Airola 2016).

### 3 Data

We chose to experiment with the LPO–SCV method for prospectivity modelling of orogenic gold occurrences in the Central Lapland greenstone belt (CLGB). As positive instances, we used the locations of known orogenic gold occurrences, and as negative instances, a random selection of locations in the study area. As evidence features, we used rasters derived from airborne and ground-based geophysics, till geochemistry and geological interpretations. Two datasets were generated: one with pixel size of  $200\text{ m} \times 200\text{ m}$  and another one with  $50\text{ m} \times 50\text{ m}$ . The coarser grid is a compromise between the resolutions of the original data, while the more accurate grid reveals the details in the geophysical data sets but is over accurate for geochemical and gravity data. Overall dimensions of the study area are approximately 170 km in the East–West and 110 km in the North–South direction, yielding 508,944 and 8,146,792 points for the 200 m and 50 m rasters, respectively. The study area is shown in Fig. 2, and the generated evidence features in Fig. 3. Geophysical data preprocessing was carried out using the Intrepid software. Interpolation, image filtering and fuzzy integration was carried out using ArcGIS and ERMMapper software.

#### 3.1 Evidence features

The evidence feature set was the same as the one generated by Nykänen (2008) and consists of typical mineral exploration related geoscientific spatial data that are derived from airborne geophysics (magnetic and electro-magnetic), ground geophysics (gravity), regional till geochemistry and a 1:200,000 scale digital geological map. Two evidence feature sets were generated with cell sizes of  $0.04\text{ km}^2$  and  $0.0025\text{ km}^2$ , while the resolution of the original measurements varies from 1 point/ $0.01\text{ km}^2$  to 1 point/ $4\text{ km}^2$ . Data preprocessing is briefly described below, while the geological basis and more detailed description of the preprocessing steps can be found in Nykänen (2008) and other references provided. Grid cell dimensions used by Nykänen (2008)



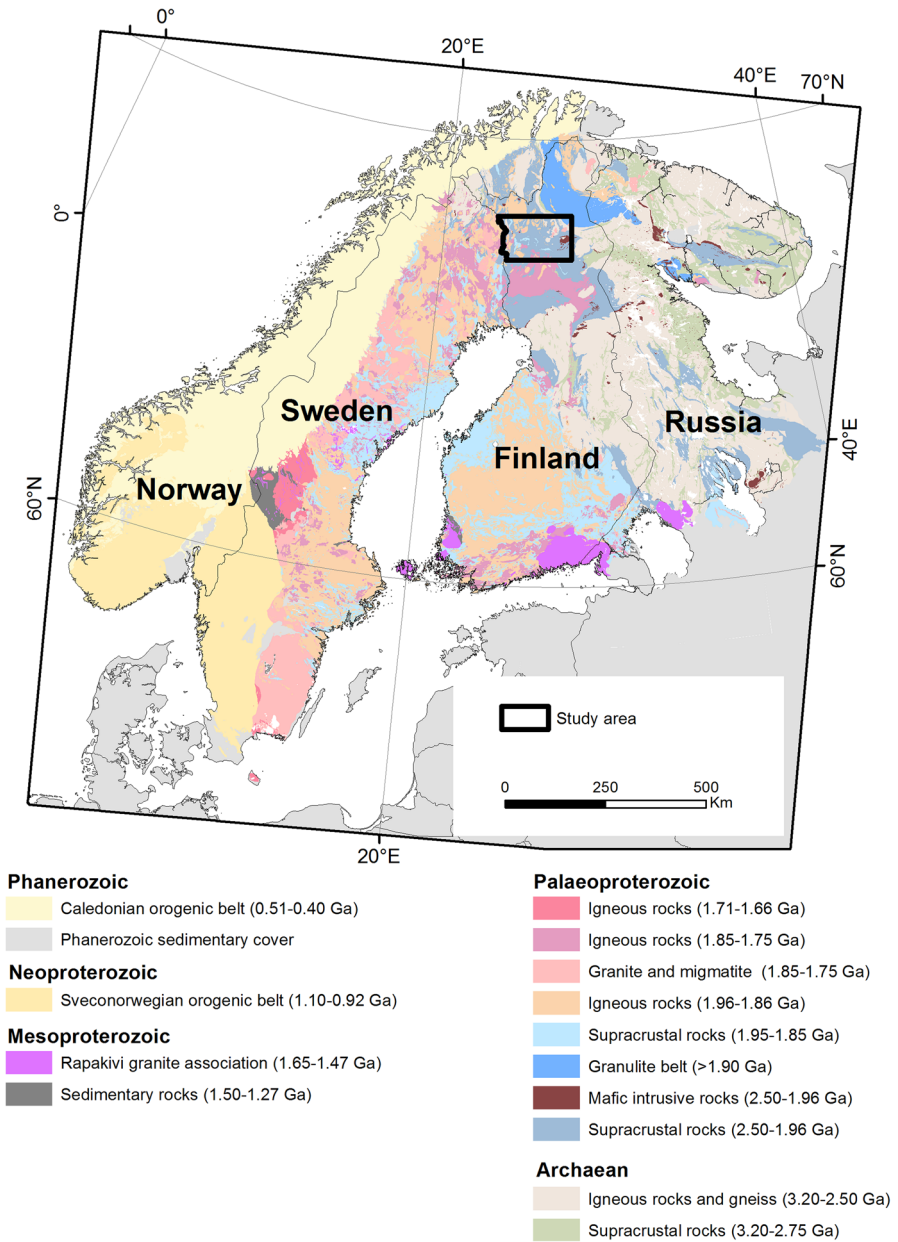
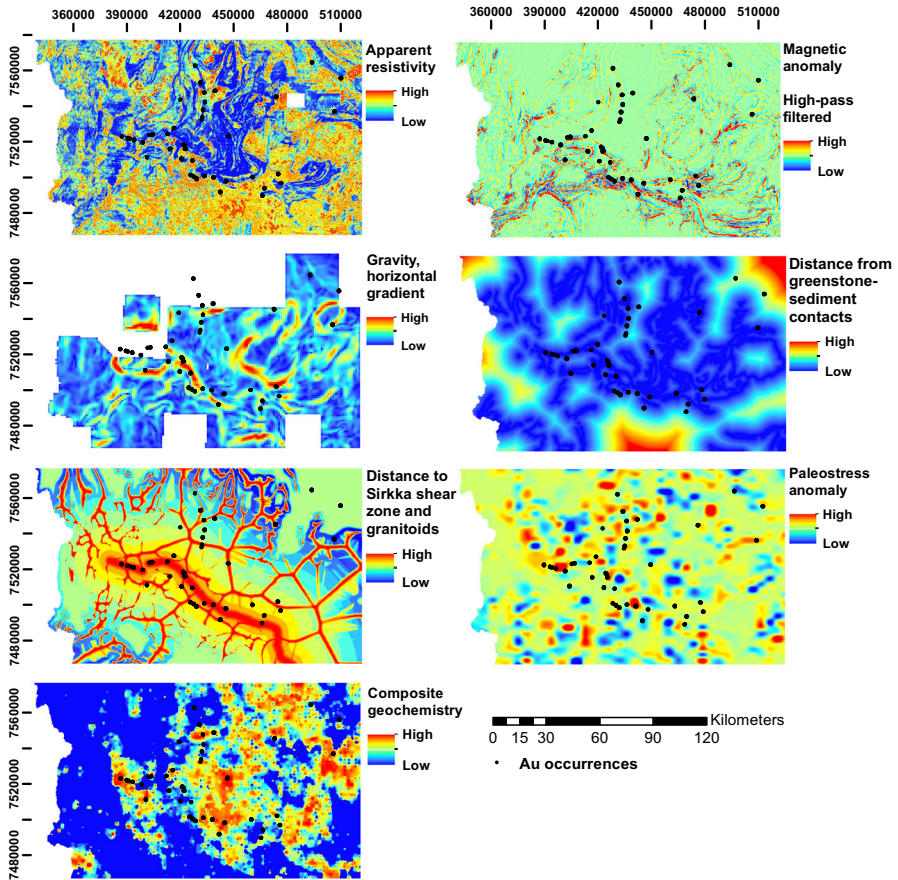


Fig. 2 Location of the study area. Generalized bedrock map is modified from Koistinen et al. (2001)

were 250 m × 250 m, and resampling of cell values for the 200 m × 200 m grid was done using the nearest neighbour method. Resampling for the 50 m × 50 m grid was done using a linear smoothing filter. All the features are standardized to zero mean and unit variance.





**Fig. 3** Maps showing the evidence features in the study area (EUREF-FIN coordinate system)

Airborne magnetic and electromagnetic data were derived from the nationwide airborne geophysical measurements collected by GTK in 1973–2007 (Airo 2005). Measurements were carried out with 200 m line spacing at a nominal 30–40 m altitude using a fixed-wing aircraft, with vertical coplanar coils (coaxial until 1979) for the electromagnetics (Hautaniemi et al. 2005). Magnetic data were interpolated to grids with a  $50\text{ m} \times 50\text{ m}$  cell size. The deviation from the Definitive Geomagnetic Reference Field was computed following Korhonen (2005). Further, deviation of each pixel value from the median of pixel values within a radius of 4 km was calculated by Nykänen (2008). Electromagnetic response was interpreted as apparent resistivity and interpolated to grids with  $50\text{ m} \times 50\text{ m}$  cell size following Suppala et al. (2005), and further resampled to  $250\text{ m} \times 250\text{ m}$  by Nykänen (2008).

The regional scale gravity map was derived from the ground-based gravity measurements collected by GTK and the Finnish Geospatial Research Institute (former Finnish Geodetic Institute) in 1990's (Kääriäinen and Mäkinen 1997) with 1 point/ $1\text{ km}^2$ . Gravity is the only evidence feature that does not cover the entire study area. Nykänen

(2008) computed the horizontal gradient of Bouguer anomaly derived from the gravity measurements.

Geochemical data were derived from GTK's national geochemical survey of glacial tills, conducted in 1970's and 80's (Salminen and Tarvainen 1995). Three to five samples, taken at a density of 1 sample/km<sup>2</sup>, were combined for analysis. Thus, the concentrations represent the average till concentration in an area of approximately 4 km<sup>2</sup>. Data for Au, As, Cu, Fe, Ni and Te were interpolated by Nykänen (2008) using inverse distance weighting with the weight decreasing as the square of the distance. Since the grid cell size was much smaller than the sampling density, anomalous average concentrations appear spot-like near the locations associated with the combined sample. Nykänen (2008) further combined the different element concentration grids by setting conditions such that Cu must always be elevated for a prospective area, at least one of As, Fe, Ni or Te must be elevated and the presence of Au increases prospectivity.

From the digital 1:200,000 scale bedrock map of northern Finland (Lehtonen et al. 1998), three evidence features were derived. The first feature is the paleostress model computed following Holyland and Ojala (1997) by geomechanical interpretation at 1:100,000 scale using faults and lithological contacts from the digital 1:200,000 scale bedrock maps and 1:100,000 scale geophysical maps. The second feature is the combination of proximity to granitoids in the Kittilä, Savukoski and Sodankylä Groups and distance to the Sirkka Shear Zone. The mean distance to granitoids within a 2500 × 2500 m square neighborhood is subtracted from the original proximity grid resulting in a grid which defines the midpoint between the granitoids within the greenstone belt, and this grid is combined with the proximity grid to the Sirkka Shear Zone. Values are discretized to 10 classes. The third feature derived from the bedrock map is the distance to contact zones between the greenstone belt lithological units and the overlying sedimentary units.

The geospatial data covers a 20,000 km<sup>2</sup> area centered on the Central Lapland Greenstone Belt (CLGB), located in the Northern Fennoscandian Shield. This area is a typical Paleoproterozoic greenstone belt composed of mafic to ultramafic volcanic successions and largely overlying sedimentary units surrounded and intruded by younger granitoids and mafic intrusions (Lehtonen et al. 1998). There has been noticeable amount of mineral exploration activity within the area during the recent years resulting more than 30 drill-defined gold occurrences and one currently operating gold mine. The majority of the gold occurrences within the CLGB are classified as orogenic gold deposits, as defined by Groves et al. (1998). Indirect age constraints suggest two separate gold mineralization events within the Fennoscandian Shield at 1.9–1.86 and 1.85–1.79 Ga (Weiheid et al. 2005). The assumption is that gold mineralization occurred late during orogenic events, enabling use of the current geometries on the bedrock map as a source of inputs for the spatial modeling because they approximate the geometries at the time of gold mineralization (Nykänen 2008).

### 3.2 Training data

Positive instances were extracted from the Geological survey of Finland's (GTK) database of mineral deposits and occurrences in Finland, and contain all the 27 gold deposits and other occurrences in CLGB that have been categorized as orogenic. Definition of the exact location of the occurrences is somewhat vague since they are not point-like. Usually orogenic gold deposits are no more than 100 m in width, but can extend hundreds of meters along structures. Defining whether an occurrence is a single one or consists of multiple separate occurrences is subject to interpretation. Here, the deposits with undefined extents are represented as single pixels in the coarser grid, and extended using a linear smoothing filter to cover a square area of 32 pixels ( $6 \times 6$  pixel square with corners omitted) in the  $50 \text{ m} \times 50 \text{ m}$  grid.

Negative instances are generated by randomly sampling pixels in the study area. Random sampling for the negatives is justified, since the vast majority of the study area can be considered unprospective. The first data set contains a total of 1000 instances, the second 2000. The area for which the gravity evidence feature was not available, was excluded when sampling the data. In the first data set there are 27 positive, and 973 negative instances. In the second data set, each deposit is represented by 32 pixels, leading to  $27 \times 32 = 864$  positive, and 1136 negative instances.

## 4 Experiments

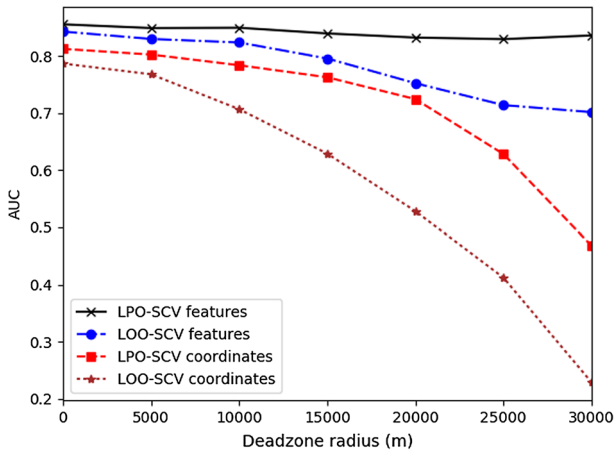
In the experiments, we demonstrate the effects of both pooling and spatial bias, and how LPO–SCV allows correcting for both of them. Then we proceed to benchmark a number of different classifiers on the prospectivity mapping data sets. We consider three linear methods, support vector machine (SVM), logistic regression and ridge regression, as well as two non-linear ones, k-NN and random forest (Hastie et al. 2001; Breiman 2001).

For ridge regression, we used the training and fast CV algorithms implemented in RLScore library (Pahikkala and Airola 2016). For the other methods, we used the scikit-learn library (Pedregosa et al. 2011), where the SVM implementation is based on the LIBLINEAR package (Fan et al. 2008). Example code for running LPO–SCV is made freely available at <https://github.com/jjepsuomi/LPO-SCV>.

For the  $200 \text{ m} \times 200 \text{ m}$  data we run full LPO–SCV, using all the  $27 \times 973 = 26,271$  positive–negative pairs. For the  $50 \text{ m} \times 50 \text{ m}$  resolution data, we select as test pairs for LPO–SCV a random subsample of 50,000 positive–negative pairs in order to speed up validation.

### 4.1 Pooling and spatial bias

In the first set of experiments, we compare a number of CV approaches with a k-NN classifier, in order to demonstrate both pooling and spatial biases. We used a large number of neighbors ( $k = 250$ ) as we noticed the method gave very poor results for



**Fig. 4** Comparison of leave-pair-out (LPO–SCV) and leave-one-out (LOO–SCV) spatial cross-validation results on the version of the data with  $200\text{ m} \times 200\text{ m}$  resolution and single pixel representing each deposit. Both CV algorithms were applied to k-NN ( $k = 250$ ) classifiers, in one case using the regular feature set and the other case using only x and y coordinates as features

small values of  $k$ . The experiments are performed on the data set with  $200\text{ m} \times 200\text{ m}$  resolution and a single pixel per deposit.

The first classifier is trained normally on the evidence features. The second classifier is trained only on the x and y coordinates of the instances. The second classifier is used to demonstrate the spatial bias, as clearly it cannot learn to generalize to new areas. Based on the coordinates, one can merely predict “gold deposits are found near other gold deposits”.

We compare both LOO–SCV and LPO–SCV on dead zone radii ranging from 0 to 30,000 m. When  $r = 0$ , the methods are equivalent to ordinary LOO and LPO with no correction for spatial bias. In Fig. 4, we can see a clear demonstration of both the pooling and spatial biases.

**Pooling bias:** The LOO results are much worse than the LPO results for both types of training data due to the pessimistic bias of LOO. The pooling bias increases as the dead zone grows larger; with a 30 km dead zone radius, the LPO–SCV result with a model trained on features is 0.84 AUC, whereas with LOO–SCV the result is only 0.70 AUC. Most noticeably, for the model trained on only the coordinates, the results even drop substantially below the 0.5 random level of AUC. These results are in line with the pessimistic bias of LOO for AUC estimation shown in earlier works of Airola et al. (2009), Airola et al. (2011), Parker et al. (2007) and Smith et al. (2014).

**Spatial bias:** For ordinary LPO and LOO with no dead zone ( $r = 0$ ), x and y coordinates are enough to predict well (Fig. 4, AUC = 0.81 for LPO–SCV with x and y coordinates). The predictions, however, drop to random level by  $r = 30\text{ km}$ , showing that based on only the coordinates the model cannot predict at all at 30 km distance and further from the training instances. In contrast, the models trained on the evidence features can generalize outside the training area.

LPO–SCV eliminates both sources of bias. On one hand, it eliminates the substantial pessimistic pooling bias that can be seen in the LOO–SCV results. At the same time, it shows that, whereas the model trained on the features can generalize outside the immediate surrounding area of training data, the coordinate-based models cannot.

## 4.2 Classifier comparison

We tested five different classification methods on the data set, using LPO–SCV. For SVM, logistic regression and ridge regression we present results for regularization parameter 1 ( $C = 1$  in scikit-learn,  $\text{regparam} = 1$  in RLScore), as the results for a large range of parameter values were very similar. For random forest, the results are presented for 100 trees, as little improvement was observed after increasing the number of trees beyond this point. For k-NN, we present the results both for  $k = 10$  and  $k = 250$ , as the method behaved very differently depending on whether the number of neighbors was small or large. For SVM and logistic regression we used balancing to weight both classes equally. For the ridge regression and k-NN implementations such an option was not available. For random forests balancing proved harmful and was not used.

The results are presented in Fig. 5 for the data with  $200\text{ m} \times 200\text{ m}$  resolution and single pixel per deposit, and in Fig. 6 for the data with  $50\text{ m} \times 50\text{ m}$  resolution and sixteen pixels per deposit.

The major difference between these two experiments is how k-NN with  $k = 10$  and the random forest classifier behave on  $r = 0$ , where no dead zone correction is done (compare Figs. 5 and 6). On the data set with a single pixel per deposit, the AUC for k-NN is 0.66, and that of random forest 0.79. On the data with 16 pixels, k-NN AUC is 0.98, and random forest AUC 1.00. Thus on one of the data sets, the two methods appear to work poorly, while on the other it would seem that they can classify the data perfectly. The second result is a clear example of spatial bias. On each round of CV, the

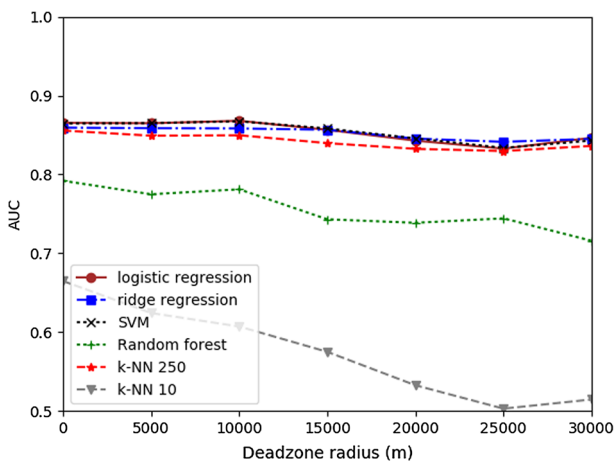
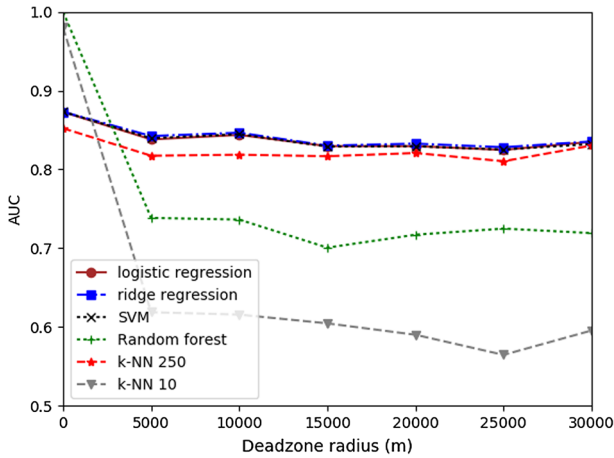


Fig. 5 Comparison of different classifiers on data with  $200\text{ m} \times 200\text{ m}$  resolution and one pixel per deposit



**Fig. 6** Comparison of different classifiers on data with 50 m  $\times$  50 m resolution and sixteen pixels per deposit. Results for k-NN ( $k = 10$ ) and a random forest classifier are highly overoptimistic when a dead zone is not used ( $r = 0$ )

methods overfit to the 15 deposit pixels left in the training set, and can thus predict the left-out pixel. When dead zone radius is increased, the effect disappears and the poor ability of the classifiers to predict beyond their immediate neighborhood is revealed. This effect is not nearly as strong for the linear methods (logistic regression, ridge regression and SVM), as they are not expressive enough to overfit to the data as much as the non-linear k-NN and random forest models.

Otherwise, the behavior of the methods is similar for the two versions of the data set. It can be seen that the linear methods outperform the non-linear ones. Their AUC starts around 0.87, and decreases to 0.85 AUC as dead zone radius grows. There are no substantial differences between the performances of these three methods. k-NN 250 results are also very close to those of the linear classifiers with AUCs ranging from 0.86 to 0.84 on the single pixel data. The random forest works poorly, with AUC always below 0.8.

Surprised by the poor performance of random forest, we also performed limited experiments to see whether by further parameter tuning, or by using other types of tree-based ensemble methods such as the extremely randomized trees (Geurts et al. 2006), results would improve. We did not find this to be the case. We also tested nonlinear kernel ridge regression (Evgeniou et al. 2000) using the RBF kernels of various widths. This did not lead to improvements over the linear ridge regression but instead resulted in a substantial increase in running time.

## 5 Discussion

The results demonstrate the clear need for spatial CV of spatial prediction models, such as MPM classifiers. Due to small number of positive instances available in many applications, CV is crucial for validating the models. We show that if the spatial depen-



dencies are not taken into account, one can obtain high AUCs even with classifiers that completely fail in generalizing outside the training area.

The data resolution and using multiple pixels versus using a single pixel to represent the deposits did not affect significantly the results for the best performing methods, when dead zone correction was properly done. However, when using several pixels to represent a deposit together with non-linear classifiers, we obtained very biased results if dead zone correction is not used.

The method comparison showed that simple linear models worked well on the MPM prediction problem. Whether the model was fitted by minimizing the logistic, least-squares (ridge), or hinge (SVM) loss did not affect the results much. The result is likely due to the small sample size, as there are only 27 positive instances of gold mineralization available in the data set. More complex models are likely to overfit to the noise in the data, rather than discover patterns that would improve the predictions beyond what the linear model already captures. This could also be seen in the k-NN results, where averaging over a very large number of neighbors ( $k = 250$ ) provided the best results, whereas more complex local models based on a smaller number of neighbors ( $k = 10$ ) did not yield a high AUC when properly validated.

In earlier work, Nykänen (2008) has shown 0.99 AUC results for both logistic regression and radial basis functional link nets on orogenic gold MPM data from the same study area. Our results are lower, though not directly comparable due to differences in data processing and experimental setup in model validation. Still, the different outcomes demonstrate the high degree to which the results depend on the chosen model validation strategy. These choices can often have a much larger effect on results than the chosen classifiers. Thus we encourage researchers dealing with spatial data to provide comprehensive spatial CV evaluations of their models in order to establish how well they can predict at different distances from training data. This approach provides additional insights about the characteristics of the data that classical model validation methods are not able to provide.

In addition to data-driven models where a classifier is trained using known occurrences or deposits, knowledge-driven approaches where this function is defined based on expert knowledge have also been popular in prospectivity modeling (see e.g. Porwal et al. 2003). Pure knowledge-driven approaches, where prior knowledge is used to define the model, and the data only to validate it, are free from the spatial bias considered in this work. Even then care should be taken to ensure that the expert will not overfit the model by studying the data used to test it in advance, as this can lead to overoptimistic results.

## 6 Conclusion

In this work, we considered the problem of evaluating the AUC of classifiers on spatial data. Standard CV methods that have been developed for i.i.d. data suffer from two sources of bias: the pooling and spatial biases. In our experiments on MPM data, we demonstrated the dangers of ignoring these biases, as one can obtain incorrect AUC values ranging from much worse than random to perfect with existing CV methods. We introduced the novel LPO-SCV method, that allows one to correct for both the pooling



and spatial biases inherent in classical CV methods. We demonstrate experimentally how the method allows one to reduce these biases and benchmarked a number of MPM classifiers showing the advantages of simple linear models. While we have considered only one MPM classification problem, the introduced evaluation approach is general and could be applied to a wide range of different types of spatial classification or ranking problems.

**Acknowledgements** Open access funding provided by University of Turku (UTU) including Turku University Central Hospital.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References


- Abedi M, Norouzi GH, Bahroudi A (2012) Support vector machine for multi-classification of mineral prospectivity areas. *Comput Geosci* 46(Supplement C):272–283
- Airo ML (2005) Aerogeophysics in Finland 1972–2004. Geological survey of Finland, Special Paper 39
- Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T (2009) A comparison of AUC estimators in small-sample studies. In: Džeroski S, Geurts P, Rousu J (eds) Proceedings of the third international workshop on machine learning in systems biology, PMLR, Ljubljana, Slovenia, Proceedings of machine learning research, vol 8, pp 3–13
- Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T (2011) An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput Stat Data Anal* 55(4):1828–1844. <https://doi.org/10.1016/j.csda.2010.11.018>
- Bamber D (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 12(4):387–415
- Bonham-Garter G (1994) Geographic information systems for geoscientists—modelling with GIS. Computer methods in geosciences, vol 13. Pergamon Press, Oxford
- Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 30(7):1145–1159
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Brown WM, Gedeon TD, Groves DI (2003) Use of noise to augment training data: a neural network method of mineral-potential mapping in regions of limited known deposit examples. *Nat Resour Res* 12(2):141–152
- Carranza E (2008) Geochemical anomaly and mineral prospectivity mapping in GIS. Handbook of exploration and environmental geochemistry, vol 11. Elsevier, Amsterdam
- Carranza EJM, Sadeghi M, Billay A (2015) Predictive mapping of prospectivity for orogenic gold, ginyan greenstone belt (South Africa). *Ore Geol Rev* 71:703–718
- Chung CJF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. *Nat Hazards* 30(3):451–472
- Cressie N (2015) Geostatistics. Wiley, Hoboken
- Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '08, pp 213–220
- Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. *Adv Comput Math* 13:1–50
- Fabbri AG, Chung CJ (2008) On blind tests and spatial prediction models. *Nat Resour Res* 17(2):107–118
- Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 9:1871–1874
- Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27(8):861–874

- Forman G, Scholz M (2010) Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explor* 12(1):49–57
- Frattini P, Crosta G, Carrara A (2010) Techniques for evaluating the performance of landslide susceptibility models. *Eng Geol* 111(1–4):62–72
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63(1):3–42
- Groves D, Goldfarb R, Gebre-Mariam M, Hagemann S, Robert F (1998) Orogenic gold deposits: a proposed classification in the context of their crustal distribution and relationship to other gold deposit types. *Ore Geol Rev* 13(1):7–27
- Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36
- Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer series in statistics. Springer, New York
- Hautaniemi H, Kurimo M, Multala J, Leväniemi H, Vironmäki J (2005) The “three in one” aerogeophysical concept of GTK in 2004. Geological Survey of Finland, Special Paper 39, 21–74
- Holyland PW, Ojala J (1997) Computer aided structural targeting: two and three dimensional stress mapping. *Aust J Earth Sci* 44:421–432
- Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng* 17(3):299–310
- Jain S, White M, Radivojac P (2017) Recovering true classifier performance in positive-unlabeled learning. In: Proceedings of the thirty-first AAAI conference on artificial intelligence, pp 2066–2072
- Kääriäinen J, Mäkinen J (1997) Airborne magnetic method: special features and review on applications. Geological survey of Finland, Special Paper 39, 77–102
- Koistinen T, Stephens M, Bogatchev V, Nordgulen O, Wennerström M, Korhonen J (2001) Geological map of the Fennoscandian shield: scale 1:2 000 000. Geological Survey of Finland, Geological Survey of Norway, Geological Survey of Sweden, Ministry of Natural Resources of Russia, Espoo, Trondheim, Uppsala, Moscow
- Korhonen JV (2005) Airborne magnetic method: special features and review on applications. Geological survey of Finland, Special Paper 39, 77–102
- Le Rest K, Pinaud D, Monestiez P, Chadoeuf J, Bretagnolle V (2014) Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Glob Ecol Biogeogr* 23(7):811–820. <https://doi.org/10.1111/geb.12161>
- Lehtonen M, Airo M, Eilu P, Hanski E, Kortelainen V, Lanne E (1998) The stratigraphy, petrology and geochemistry of the Kittilä greenstone area, northern Finland. In: Report of investigation, vol 140. Geological Survey of Finland, pp 140–144
- Longley P, Goodchild M, Maguire D, Rhind D (2005) Geographic information systems and science, 2nd edn. Wiley, Hoboken
- Nykänen V (2008) Radial basis functional link nets used as a prospectivity mapping tool for orogenic gold deposits within the Central Lapland greenstone belt, northern Fennoscandian shield. *Nat Resour Res* 17(1):29–48. <https://doi.org/10.1007/s11053-008-9062-0>
- Nykänen V, Lahti I, Niiranen T, Korhonen K (2015) Receiver operating characteristics (ROC) as validation tool for prospectivity models: a magmatic Ni–Cu case study from the Central Lapland greenstone belt, northern Finland. *Ore Geol Rev* 71(Supplement C):853–860
- Pahikkala T, Airola A (2016) RLScore: regularized least-squares learners. *J Mach Learn Res* 17(221):1–5
- Pahikkala T, Suominen H, Boberg J (2012) Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Mach Learn* 87(3):381–407. <https://doi.org/10.1007/s10994-012-5287-6>
- Parker BJ, Gunter S, Bedo J (2007) Stratification bias in low signal microarray studies. *BMC Bioinform* 8:326
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
- Pohjankukka J, Nevalainen P, Pahikkala T, Hyvönen E, Hänninen P, Sutinen R, Ala-Ilomäki J, Heikkinen J (2014) Predicting water permeability of the soil based on open data. In: Iliadis L, Maglogiannis I, Papadopoulos H (eds) Proceedings of the 10th international conference on artificial intelligence applications and innovations (AIAI 2014), IFIP advances in information and communication technology, vol 436. Springer, pp 436–446

- Pohjankukka J, Pahikkala T, Nevalainen P, Heikkonen J (2017) Estimating the prediction performance of spatial models via spatial k-fold cross validation. *Int J Geogr Inf Sci* 31(10):2001–2019. <https://doi.org/10.1080/13658816.2017.1346255>
- Porwal A, Carranza E, Hale M (2003) Knowledge-driven and data-driven fuzzy models for predictive mineral potential mapping. *Nat Resour Res* 12(1):1–25
- Rigol-Sanchez JP, Chica-Olmo M, Abarca-Hernandez F (2003) Artificial neural networks as a tool for mineral potential mapping with GIS. *Int J Remote Sens* 24(5):1151–1156
- Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol Rev* 71(Supplement C):804–818
- Salminen R, Tarvainen T (1995) Geochemical mapping and databases in Finland. *J Geochem Explor* 55:321–327
- Smith GC, Seaman SR, Wood AM, Royston P, White IR (2014) Correcting for optimistic prediction in small data sets. *Am J Epidemiol* 180(3):318–324
- Suppala I, Oksama M, Hongisto H (2005) GTK airborne EM system: characteristics and interpretation guidelines. Geological survey of Finland, Special Paper 39, 103–118
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geogr* 46(sup1):234–240
- Weihed P, Arndt N, Billström K, Duchesne JC, Eilu P, Martinsson O, Papunen H, Lahtinen R (2005) Precambrian geodynamics and ore formation: the Fennoscandian shield. *Ore Geol Rev* 27(1):273–322 (**special issue on geodynamics and ore deposit evolution in Europe**)
- Yousefi M, Carranza EJM (2015) Prediction-area (P–A) plot and C–A fractal analysis to classify and evaluate evidential maps for mineral prospectivity modeling. *Comput Geosci* 79:69–81

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Antti Airola<sup>1</sup>  · Jonne Pohjankukka<sup>1</sup> · Johanna Torppa<sup>2</sup> ·  
Maarit Middleton<sup>3</sup> · Vesa Nykänen<sup>3</sup> · Jukka Heikkonen<sup>1</sup> · Tapio Pahikkala<sup>1</sup>

Jonne Pohjankukka  
jonne.pohjankukka@utu.fi

Johanna Torppa  
johanna.torppa@gtk.fi

Maarit Middleton  
maarit.middleton@gtk.fi

Vesa Nykänen  
vesa.nykanen@gtk.fi

Jukka Heikkonen  
jukka.heikkonen@utu.fi

Tapio Pahikkala  
tapio.pahikkala@utu.fi

<sup>1</sup> Department of Future Technologies, University of Turku, 20014 Turku, Finland

<sup>2</sup> Geological Survey of Finland, Neulaniementie 5, P.O. Box 1237, FIN-70211 Kuopio, Finland

<sup>3</sup> Geological Survey of Finland, Lähteentie 2, P.O. Box 77, FIN-96101 Rovaniemi, Finland