Journal of Archaeological Science 59 (2015) 80–88

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

A supervised machine-learning approach towards geochemical predictive modelling in archaeology

Stijn Oonk [a, b, *], Job Spijker [c]

[a] Institute for Geo- and Bioarchaeology, Faculty of Earth and Life Sciences, VU University Amsterdam, De Boelelaan 1085, 1081HV Amsterdam, The Netherlands
[b] Independent Researcher at AT52, Amsterdam, The Netherlands
[c] Centre for Environmental Quality, National Institute for Public Health and the Environment (RIVM), PO Box 1, NL-3720 BA Bilthoven, The Netherlands

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
25 March 2015
Accepted 3 April 2015
Available online 15 April 2015

Keywords:
Geochemistry
Archaeology
Predictive modelling
Machine-learning
Soil

ABSTRACT

In this work, data fusion of multi-element XRF results from archaeological feature soils and regional background soils was applied to assess the complementary value of geochemistry and machine-learning on predictive modelling in archaeology.

Our principal aim was to integrate multiple data sources, train learning models for classification of archaeological soils and background soils, and compare model predictions for three validation areas with current archaeological interpretation and established predictive models. This was done using three supervised machine-learning algorithms (k-nearest neighbors, support vector machines and artificial neural networks) which were trained, cross-validated and tested. The validation areas included a high archaeological potential area (n = 247 samples), the Dutch province of Zeeland (n = 261 samples) and an excavated imprint of an ancient farmhouse (n = 38 samples). The predictive models showed good overall performance and correctly classified about 95% of all test instances. According to the learning models, the first validation site has a top soil horizon that shows limited parallels with archaeological horizons used in model training, whereas features of high archaeological probability become more apparent below this horizon. This is in good correspondence with geochemical depth profiles and current archaeological interpretation. As for the second validation site, the models highlighted several archaeological hotspots that to some extend spatially coincide with areas of high archaeological potential as indicated by established predictive modelling. Reversely, the classifiers attributed high archaeological potential status to the most southern region of Zeeland, thereby complementing established modelling results. For the third validation site none of the instances were correctly classified and these results clearly show the limitations of geochemical predictive modelling of significantly different soil types (fine-to-coarse sands) compared to the training set (clayey sands).

Present proof-of-concept study shows that modelling of multiple-source geochemical soil data using machine-learning algorithms can be successfully accomplished and that model predictions nicely complement current interpretation and/or established archeological predictive modelling of areas of archaeological interest. Limitations of our approach were found to reside in lithological differences between sites used for model training and prediction sites.

1. Introduction

Domestic and agricultural activities in the past have had a significant impact on soil processes and the present chemical soil composition. Due to the different uses of space, habitation areas may thus hold chemical anomalies that coincide spatially with archaeological features (see e.g., Griffith, 1981; Linderhold and Lundberg, 1994; Middleton and Price, 1996; Pierce et al., 1998; Fernandez et al., 2002; Knudson et al., 2004; Terry et al., 2004; Wilson et al., 2005). Detection and delineation of such areas is of value because it aids defining the focus of archaeological research and excavations. Modern agricultural activities and soil pollution are, however, likely to obscure ancient chemical soil anomalies. Hence, it is essential to attain a geochemical understanding of

* Corresponding author. Institute for Geo- and Bioarchaeology, Faculty of Earth and Life Sciences, VU University Amsterdam, De Boelelaan 1085, 1081HV Amsterdam, The Netherlands. Tel.: +31 (0)645103984.
E-mail address: stijnoonk@yahoo.com (S. Oonk).

http://dx.doi.org/10.1016/j.jas.2015.04.002
0305-4403/

archaeological sites, with reference to the relationship between chemical data, ancient and modern anthropogenic activities and soil pollution. Multi-element soil analyses yield a detailed, yet complex account of data for this purpose, as shown by e.g. Aston et al. (1998), Entwistle et al. (1998, 2000), Schlezinger and Howes (2000), Eckel et al. (2002), Parnell et al. (2002), Sullivan and Kealhofer (2004), Wells (2004), Cook et al. (2005) and Oonk et al. (2009a, b). However, controversy exists on how to assess multi-element anomalies in soils and a multitude of unsupervised modelling techniques have been proposed and tested. These include various on-site vs. off-site element ratios (i.e. enrichment factors), latent variable models (e.g. factor analysis, principal component analysis) and clustering techniques. Geochemical predictive modelling through supervised machine-learning provides another means to analyze patterns in multivariate (geochemical) data, and identify and locate soil anomalies. Although such methods have recently seen successful application in e.g. geological prospection, and geological and mineral mapping (Ingham et al., 2014; Baudron et al., 2013; Abedi and Norouzi, 2012, Abedi et al., 2012), none has found use in archaeological applications so far. Strong parallels exist between these geological applications and archaeological prospection, and verifies that a machine-learning approach based on geochemical data is feasible for archaeological predictive modelling. In contrast to unsupervised methods, supervised machine-learning relies on prior knowledge of a geochemical system. Thus, class memberships need to be known and assigned to each instance in order to constitute a training dataset. The labeled geochemical data is thence used to construct models that proxy for class characteristics. Validation steps are usually applied to optimize model parameters and select the best performing model which can be used to predict class memberships of new instances.

This study was set out to test the feasibility of archaeological predictive modelling by means of combined multi-element soil profiling and supervised machine-learning. Note that machine-learning is essentially a black box approach, which, in this respect, utilizes covariance in multidimensional data in order to characterize various patterns of variation in the training data. Spatial anomalies, as predicted by the models constructed and applied in this work can thus best be regarded as different patterns of variation rather than actual geochemical anomalies.

Soil XRF analysis data from previously published Dutch archaeological prospection studies and regional geochemical surveys were labeled with their respective classification status (archaeology vs. no-archaeology) and assessed by three machine-learning algorithms; weighted *k*-nearest neighbors analysis (*k*NN), support vector machines (SVM) and artificial neural networks (NN).

K-nearest neighbors algorithms are considered the simplest of learning algorithms; training of a *k*NN classifier is merely a process of storing features and classification labels Classification is then based on calculated distance measures and a voting scheme. The here applied weighted *k*NN algorithm additionally includes transformation of distance measures to similarity measures using a kernel function. This makes the technique more robust towards *k*-values that lead to high misclassification rates during training.

Support vector machines use non-linear decision boundaries in high dimensional variable space to train classifiers. The rationale behind this is that for non-linearly separable two-class data there are an infinite number of hyperplanes that divide the classes. To select a hyperplane that optimally separates the two classes (i.e. a decision boundary) a subset of training samples, also known as support vectors, is used. With regards to cases that are not linear separable SVM make use of kernels that transform input variables and allows to separate non-linear separable support vectors using a linear hyperplane. An optimal decision boundary is then represented by the maximal margin (M) between support vectors, which in turn is determined by penalizing misclassifications using a cost parameter (C).

Lastly, NN use a network of primitive functions arranged in layers that receive multiple inputs that are weighted according to their ability to discriminate classes. Hereby, different function types and network configurations are created, and hence this results in different models. During training, network connection weights are adjusted so as to minimize errors due to separation of inputs and classes, while convergence proceeds until between-iteration errors reach a decay-threshold.

In this study, a 75% portion of the total dataset was used to train the models, whereas the remaining data was used for model testing. Best performing models were then applied to multi-element data from three distinct validation areas (VA1, VA2 and VA3). Here, VA1 concerns an area of high archaeological potential and was chosen to i) assess the spatial probability of ancient anthropogenic impact at different depths (ca. 20, 40 and 60 cm below surface level) and ii) compare these results with field survey and excavation findings. Additional geochemical data from a much larger area (VA2) was interrogated by the models in order to test similarities between established predictive modelling and geochemistry based modelling. Here, established modelling is regarded as modelling the archaeological impact of an area by means of e.g. coring campaigns, surface finds, physical- and historical geography. Lastly, samples from an archaeological house plan embedded in fine-to-coarse sands were classified so as to test the effects of alternative site lithology and possible limitations of the here presented approach.

2. Materials and methods

2.1. Study areas

2.1.1. Training and testing areas

For model training, geochemical data from two archaeological sites was used. These sites (at Tiel and Zijderveld) are situated in the Rhine/Meuse delta in the center of the Netherlands and consist mainly of fluvial clayey sands. All sites were previously excavated and could be dated to 240–270 AD (Tiel) and 1500–1100 BC (Zijderveld). Samples were all taken inside house plans, whilst a small sample set (n = 10) from the Tiel site was taken off-site. Further details on the lithology and archaeology of these sites can be found in Oonk et al., 2009a. In addition, Dutch geochemical background data was used for non-archaeological training instances (see below).

2.1.2. Validation area 1

Validation area 1 is located in the southeastern part of The Netherlands along the Meuse river (see Fig. 1) near to the village of Borgharen. The area mainly consists of alkaline (pH_{water} = 7.2–8.4) clayey river sediments deposited during floods. Primary sedimentary sequences, consisting of yellow to brown sandy clays, were deposited on top of fluvial sands and gravels during the Weichselien period. These deposits also constitute the immediate parent soil material. Also present in these sequences are distinct grayish layers that consist of sands, gravels and charcoal fragments. Superimposed on these layers are brownish calcareous and silty clays deposited during the early Holocene. Being part of a gravel rich point bar, the study site is manifested as a slightly elevated area at 43.9 ± 0.37 m above sea level. The study area itself has little topography except for a minor elevation at the far south of the area (44.40–44.60 m above sea level).

Pollution of the area is likely to be extensive, given the long-term agricultural activities adding (artificial) manure and