journal of computer applications in archaeology

An Explorative Application of Random Forest Algorithm for Archaeological Predictive Modeling. A Swiss Case Study

RESEARCH ARTICLE

MARIA ELENA CASTIELLO
MARJ TONINI

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The present work proposes an innovative approach to surveys and demonstrates the effectiveness of bringing together traditional archaeological questions, such as the exploration and the analysis of settlement patterns, with the most innovative technologies related to Machine Learning. Namely, we applied Random Forest, an ensemble learning method based on decision trees, to perform archaeological predictive modeling (APM) for the Canton of Zurich, in Switzerland. This was done based on a dataset of known archaeological sites dating back to the Roman Age. The APM represents an automated decision-making and probabilistic reasoning tool that is relevant for archaeological risk assessment and cultural heritage management. Machine learning-based approaches can learn from data and make predictions, starting from the acquired knowledge, through the modeling of the hidden relationships between a set of observations, representing the dependent variable (i.e. the archeological sites), and the independent variables (i.e. the geo-environmental features prone to influence the site locations). The main objective of the present study is to assess the spatial probability of presence for Roman settlements within the study area. As results, we produced: 1) a probability map, expressing the likelihood of finding a Roman site at different locations; 2) the importance ranking of the geo-environmental features influencing the presence of the archeological sites. These outputs in our results are of paramount importance, not only in verifying the reliability of the data, but also in stimulating experts in different ways. Also, these results help evaluate the benefits and constraints of using such innovative techniques and, ultimately, help explore the performance of machine learning-based models in processing archaeological information.

CORRESPONDING AUTHOR:
Maria Elena Castiello
Institute of Archaeological Sciences, University of Bern, CH-3012 Bern, Switzerland
maria.castiello@iaw.unibe.ch

KEYWORDS:
Roman Settlements; Locational Patterns; Machine Learning; Cultural Heritage Management; Canton of Zurich

TO CITE THIS ARTICLE:
Castiello, ME and Tonini, M. 2021. An Explorative Application of Random Forest Algorithm for Archaeological Predictive Modeling. A Swiss Case Study. *Journal of Computer Applications in Archaeology*, 4(1), 110–125. DOI: *https://doi.org/10.5334/jcaa.71*

1. INTRODUCTION

The massive expansion of urban settlement areas and transport infrastructures, increasingly threatens our cultural heritage all over the world. In particular, in the few last decades, Switzerland has been experiencing constant growth in its infrastructures and modern agglomerations (Hafner 2013). Several areas, which were so far unexploited are nowadays critically threatened by modern development, which often results in permanent destruction to any possible archaeological remains not yet unearthed. Within the Swiss Confederation, archaeology *per se* is a prerogative of each single region or member state (i.e. canton). Due to the country's decentralized political organization, each canton has its own specific procedures for archaeological heritage management (Kaeser 2013; 2012; Kaenel 2002). Each canton is also responsible for the protection of nature, landscape, and cultural heritage within its territorial boundaries. Hence, a multiplicity of approaches exists. This decentralized situation further strengthens the need of exploring new solutions and to develop objectified and quantitative tools that can help detect archaeological sites, identify them, and protect them.

Prediction and modeling have always played a relevant role in this regard (Nebbia et al. 2016; Arnoldus-Huyzendveld, Citter & Pizziolo 2015; Rogers, Fischer & Huss 2014; McEwan 2012; Danese et al. 2014; van Leusen & Kamermans 2005; Lock 2000). Based on the assumption that human behavior can be patterned, the possibility to map this pattern can result in a helpful tool for the assessment of where we have the highest likelihood of (re)-discover archaeological remains not yet unearthed. Archaeological Predictive Models (APMs) have been studied and implemented both in academia, as "locational preference maps" or "distribution maps", both in cultural resource management offices, making numerous steps forward (to cite few examples of APM: Cecamore & Castiello 2014; Nicu 2019; Visentin & Carrer 2017; Anichini et al. 2011; Verhagen & Whitley 2011; Ford, Clarke & Reisen 2009; Rua 2009; Oštir et al. 2007; de Vries 2007; Verhagen 2007; Kamermans et al. 2005; Ducke & Münch 2005; Ejstrud 2003; Kvamme 1990). The APMs produced so far share a number of common aspects, such as the use of archaeological data and environmental variables and a methodological approach based on multivariate statistical techniques such as Logistic Regression (on this topic see for example: Wachtel et al. 2018; Carrer 2013; Vaughn & Crawford 2009; Espa 2006; Kvamme 1999). Verhagen and Withley (2020) have recently published a comprehensive list of the most popular predictive models developed worldwide to predict the spatial location of archaeological sites. Authors highlight strengths and weaknesses of their applications stressing how, over the last decades, the academic research had to deal with the raising availability and complexity of archaeological datasets, as well as the complexity of the questions they raise. Questions that, in time, led archaeologists to establish new collaborations and exchange with experts in different research domains (Hintz, Laabs & Castiello 2019; Carlson 2017; Barcelo & Bogdanovic 2015; Dubbini & Lodoen 2014; Djindjian 2009; Giligny et al. 2010). It is only recently that archaeological researchers have started to explore more complex models, relying on innovative applications of spatial and statistical computing, as well as on Machine Learning (ML) techniques. So far, these studies are increasingly numerous when dealing with archeological site detection relying on high-resolution satellite or drone imagery, as well as in pottery classification (see for example the works of: Garcia-Molsosa et al. 2021; Orengo et al. 2020; Orengo & Garcia-Molsosa 2019; Caspari & Crespo 2019; Gattiglia 2018; Chen et al 2013). At the best of our knowledge, applications of ML in the field of archaeological site distribution dealing with settlement patterns and location probability assessment are very rare in literature. Märker and Heydari-Guran (2009) used Random Forest (RF) to predict the location of Paleolithic sites in the Zagros Mountains of Iran, which represents a first attempt of application of data-mining approaches in this domain. In a more recent study, Roalkvam (2020) compares logistic regression with RF to formalize and quantitatively evaluate environmental factors for coastal site location in Mesolithic Norway and to determine the evolution in time of the relative importance of these variables. Our study differs from the abovementioned ones not only in that it results in both the assessment of the environmental factors and the elaboration of a predictive map for archeological site location, but also in the implementation of an accurate procedure for the model evaluation and validation. In addition, despite the several examples produced at international level, only few studies concerning application in the domain of APM and mapping were realized for Switzerland. In this regard, the work carried out by Ebersbach (2015) provides a first and almost unique example in Switzerland where the author applies exploratory spatial-statistical analysis for evaluating locational criteria of Roman villas and Neolithic sites in the Canton of Bern.

In the present study, we explore the potential of RF, a learning algorithm based on a multitude of regression trees, to elaborate predictive maps for archaeological Roman settlements in the Canton of Zurich. The predisposing factors suggesting the presence of Roman sites in the area are the environmental features. They are described by: (i) topographic indices derived from the digital elevation model (DEM); (ii) different characteristics related to the soil and its aptitude to agricultural activities; (iii) strategic and water-related criteria on which past populations may have based their site location choice. Although predictive models based on ML have been successfully applied in different environmental studies – such as geological prospection, geological and mineral

mapping (see: Oonk & Spijker 2015; Baudron et al. 2013; Abedi & Nouruzi 2012; Abedi, Nouruzi & Bahroudi 2012) and natural hazard susceptibility mapping (see: Tonini et al. 2020; Tehrany et al. 2019; Reichenbach et al. 2018; Deluigi 2018; Leuenberger et al. 2017; Zêzere et al. 2017; Pham et al. 2016; Goetz et al. 2015) – the present research represents a first example of ML application to carry out an exhaustive analysis for APM and mapping in Switzerland.

2. MATERIALS AND METHODS

2.1 STUDY AREA

The study area lies in the current administrative limits of the Canton of Zurich, located in the northeastern part of Switzerland (***Figure 1***). The territory covers 1729 km² and nowadays is considered productive for about 80% of its area. Forests cover 505 km² and lakes 73 km². Most of the canton consists of narrow river valleys that go towards the Rhine River in the north. Together with the Lakes Zurich, Greifensee, Pfäffikersee, the rivers crossing the region have played important roles for commercial and communication purposes since the antiquity. Turicum (the ancient settlement of the City of Zurich) arose in the Limmat valley as a small artisan settlement (*vicus*) occupying both sides of the valley and becoming the first military post in the area (Horisberger 2017).

The Roman epoch in Switzerland, lasting from 30 BC to 450 AD, is a well-known period of the history thanks to the numerous literature sources and archaeological discoveries. According to the Archaeological Department of the Canton of Zurich ("Amt fur Raumentwicklung Kantonsarchäologie Zurich"), the region in antiquity was particularly dense with roads and settlements (especially military camps). Numerous *vici* (small towns) were probably embedded in a wider and dense networking context connecting the heart of the Roman Empire and the Mediterranean coasts to the south, with its northern provinces. The settlements were mainly located on headlands or on the shores of main lakes and surrounded by trenches and fortifications, in an easy to defend position and suitable for the trades, (Flutsch et al. 2002; Furger et al. 2001; Frei-Stolba & Benedetti Martig 1991). At the time, they were probably provided for by a *forum*, *tabernae* or *thermae,* and one or more religious temples (Cramatte 2012; Bögli 1962). According to the historical sources, and to more recent archaeological investigations and analyses, the settlements were often located at a distance of 30 km from each other, corresponding to about one walking day. From the 1st to 3rd century AD, *vici,* as well as a certain number of urban *domus,* and hundreds of rural *villae* of varying sizes, intensively occupied the countryside, e.g. the villas of Dietikon, Neftenbach, Buchs, etc. (Ebnöther & Monnier 2002; Ebnöter 1995).

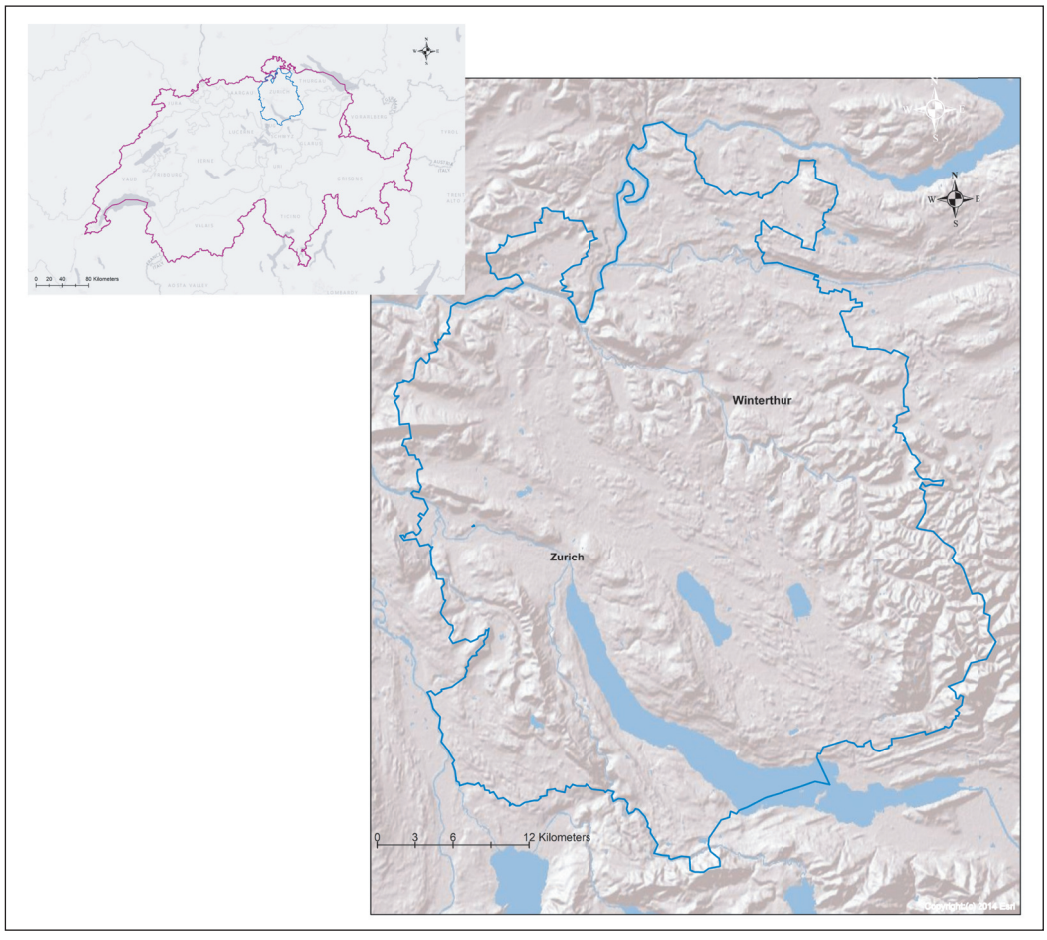

Figure 1 The case study area – Canton of Zurich.

Agriculture assumed a very important role as the main subsistence activity. Thus, environmental factors such as the suitability of the terrain for agriculture, the proximity to water resources, and the topographic indices are all highly relevant for the analysis developed in this study.

2.2 DATA ACQUISITION AND PRE-PROCESSING

2.2.1 Roman archaeological sites

The Roman archaeological sites discovered in the study area represent the dependent variable of our model. The original dataset was provided by the cantonal archaeological service of Zurich. It was in the form of a digital table containing a list of surveys carried out in previous decades (5812 entries catalogued until October 2015) covering different epochs (from Mesolithic to Middle Age). But, our current analysis is limited to the Roman period. The table was structured into different columns (i.e. identifier, municipality, type of site, assigned epoch, X and Y geographic coordinates). However, a certain underlying degree of uncertainty exists at this stage, as not all entries had exact coordinates. The dataset was reworked and standardized in order to be processed into a GIS (Geographical Information System). The pre-processing was thus necessary to check information and correct it, and to elaborate it in the form of a well-structured geo-localized punctual dataset. This included the transformation of the coordinates into the official Swiss projected coordinate system and the precision of the name of the modern municipalities, where the archaeological evidences were found. Additionally, the description of each previous survey was verified, and a short interpretation correctly defined and embedded. These interpretations provided detailed information about the nature of the findings or the nature of the site itself, such as a specific socio-economic function, whether it was a permanent or non-permanent settlement (e.g. housing, *villae urbanae, rusticae, vici,* etc.), a place of worship and of religious identity (e.g. *funum*, *temple*, etc.), a burial grave, a necropolis, an artisanal production center, etc. Further screenings allowed us to divide the dataset into two main categories: "settlements" *sensu stricto,* and "single finds", like pottery shards, coins, etc. ***Figure 2*** shows the amount of discoveries attributed to each class. The final geo-dataset, containing only the information related to the presence of Roman settlements, consists of 227 occurrences. In addition, a random set of pseudo-absences was generated, which are located across the landscape and assumed to be non-sites. This process resulted in a balanced binary geo-dataset of presences and absences for Roman settlements, which is essential for ML modeling purposes.

2.2.2 Predisposing factors

The following geo-environmental features prone to influence the Roman sites location, acting as independent variables of the model, were taken into account: Digital Elevation Model (DEM; altitude) and derivates (slope, northness and eastness); Distance to water (lakes and rivers); Agricultural suitability; Depth of vegetal soil; Soil skeleton; Water saturation and Water storage capacity; Permeability and Nutrient storage capacity (***Figure 3***). We used a DEM with a cell resolution of 100 m (pixel size = 100 m x 100 m), as for the digital layers on soil properties. Indeed, the present-day topography can be a useful factor to detect relations between site locations and their environmental surroundings (Märker & Bolus 2018). Northness and eastness, corresponding respectively to the cosine and to the minus sine of the aspect angle, were considered instead of the terrain orientation in order to avoid the use of a circular variable that can introduce bias, as very distinct values (0° and 360°) represent the same situation in reality (i.e. north orientation). We also considered the proximity of each archaeological site to lakes and rivers, intended as a source of water supply: the distance values were computed for each pixel, considering the Euclidean distance to the closest water element (i.e. lake or river).

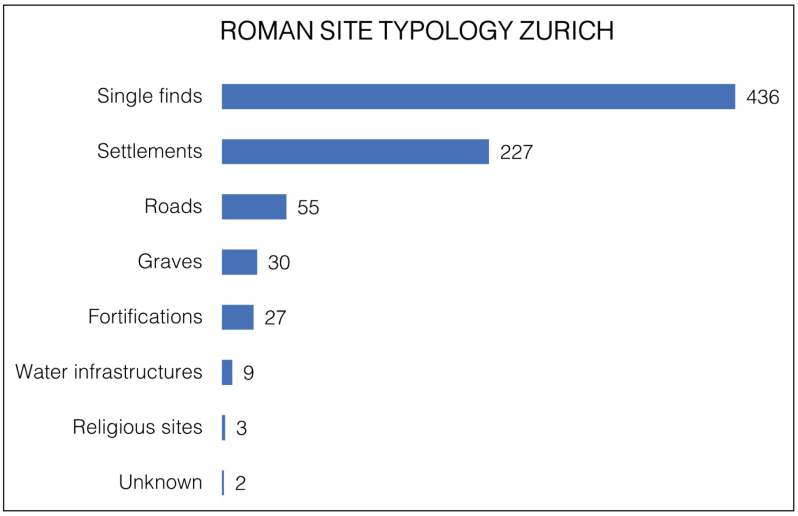

Figure 2 Typological classes of the Roman sites in the Canton of Zurich.

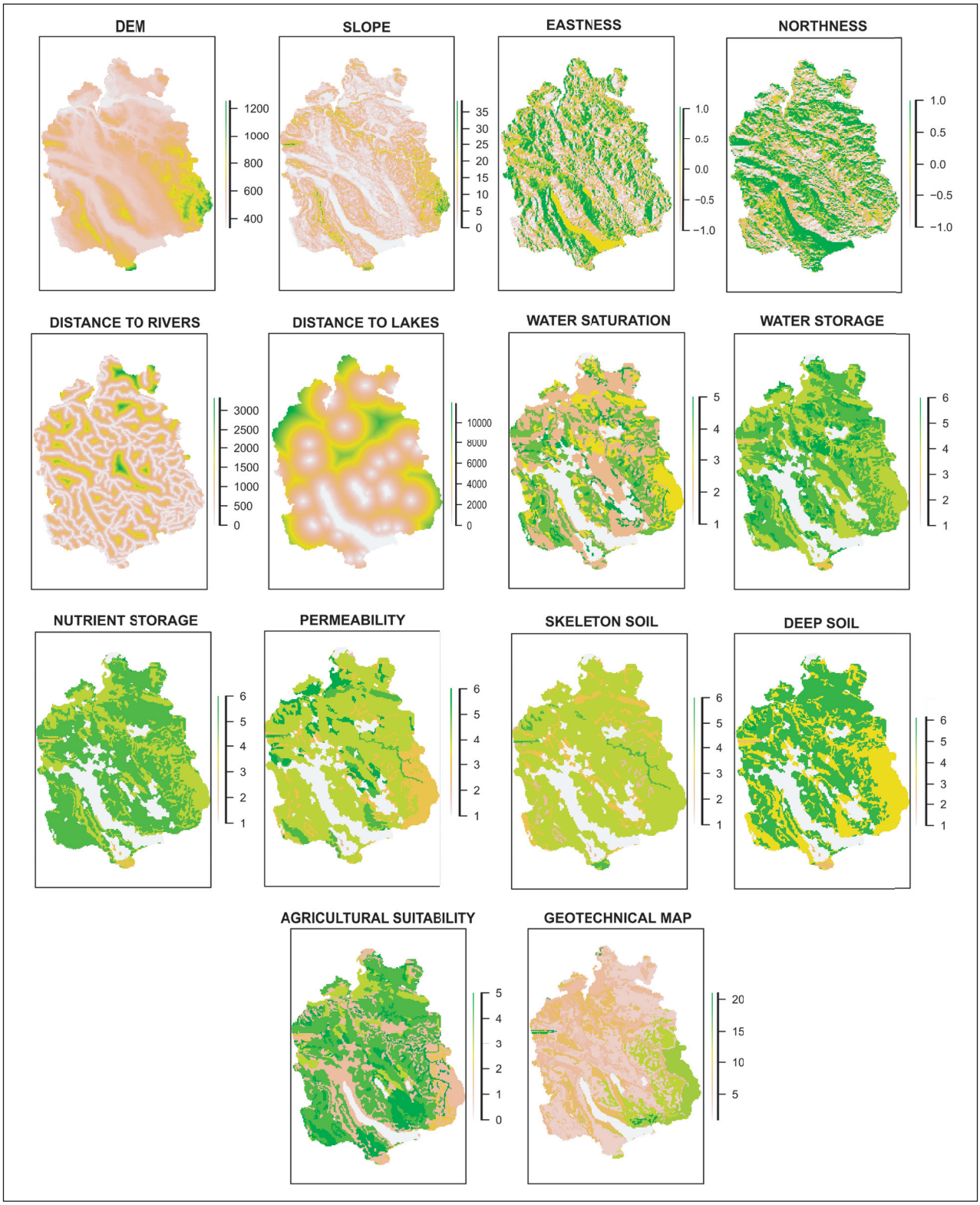

Figure 3 The geo-environmental features prone to influence the location of roman sites. The classes of each feature are visually expressed in a graduated scale of colors. DEM from 400 to 1.200 m.a.s.l.; SLOPE from 0° to 35°; EASTNESS and NORTHNESS from −1.0 to 1.0; DISTANCE TO RIVERS from 0 to 3.000 m; DISTANCE TO LAKES from 0 to 10.000 m. See Table 1 for the maps representing the soil properties (i.e. PERMEABILITY, DEEP SOIL, SKELETON SOIL, WATER SATURATION, WATER STORAGE, NUTRIENT STORAGE, AGRICULTURAL SUITABILITY); See Table 2 for the description of different categories of the geotechnical map.

Furthermore, we used the soil map[1] (*Digitale Bodeneignungskarte der Schweiz*, 2012) providing us with the most valuable information about a terrain's suitability for agricultural. Such information provided specific soil properties, each stored as single digital raster layer (agricultural suitability, depth of vegetal soil, soil skeleton, water saturation capacity, water storage capacity, soil permeability capacity, nutrient storage capacity) ranked each into five or six classes based on the different aptitude of the soil for the specific characteristic (***Table 1***). Soil properties are considered to be significant factors in determining agricultural productivity, which, in turn,

CLASS	1	2	3	4	5	6
SOIL PROPERTIES						
Agricultural Suitability	Unknown	Hindered	Good	Very good	Sufficient	__
Deep Soil	Unknown	Very shallow	Shallow	Medium	Deep	Very deep
Skeleton	Unknown	Not stony	Slightly stony	Stony	Very stony	Extremely stony
Water Saturation	Unknown	Absent	Humid	Slightly wet	Wet	__
Water Storage	Unknown	Very poor	Poor	Medium	Good	Very good
Permeability	Unknown	Very slow	Slower	Slightly slower	Normal	Extreme
Nutrients Storage	Unknown	Very poor	Poor	Medium	Good	Very good

Table 1 List of the different soil properties and their classes used as independent variables in RF.

shapes the Roman site distribution patterns (Simpson et al. 2002; Westcott & Brandon 1999). Intensive land use changes and deforestation occurred during the Roman period, probably related to the introduction of agriculture and to the mass movements of human population. These changes are discernible from the soil properties, and confirmed also by recent studies on pollen-based land-cover reconstructions, focused on northern and central Europe (Roberts et al. 2018; Wickham 2011). Hence, we decided to use these maps representing soil properties, such as agricultural suitability and nutrient storage capability, texture and specific soil depth, permeability, water saturation and water storage capacity (Nussbaum et al. 2018; Ebersbach 2015). Finally, the geotechnical map of Switzerland provided information about the distribution of the uppermost rock strata (***Table 2***).[2]

The ensemble of these spatial layers was pre-processed in order to correct and eliminate construction errors (e.g. no-data, and resampled to match the same spatial resolution of 100 meters). The pre- and post-processing of the geographical layers was performed in a GIS environment using ArcGIS Desktop, release 10.7 (ESRI).

GEOTECHNICAL MAP	
CATEGORY	**TYPE**
1	Lakes
3	Silt, Clay Pan, Ground Moraine, Frontal Moraine
4	Loam argillaceous, Clay
5	Gravel Pits and Sand (Glacial Deposit)
6	Gravel Pits and Sand (Modern Deposit)
7	Gravel Grit, blocs, Landslide
8	Marlstone with Sandstone inclusions weakly solidified
13	Conglomerates, from weakly to mildly solidified
14	Conglomerates, from weakly to mildly solidified
19	Solid Limestone
21	Schists Deposit

Table 2 List of the geotechnical categories derived from the Geotechnical map of the Canton of Zurich (The category numbers are not continuous, as the geotechnical map describes the entire Swiss territory and not all soil types are present within the Canton of Zurich).

2.3 METHOD

Random Forest (RF) (Breiman et al. 2018), a machine learning based approach, was used to estimate the probability of discovering archaeological remains, namely Roman sites, in a given area. RF is an ensemble learning algorithm based on decision trees, capable of learning from data and making predictions, starting from the acquired knowledge (i.e., observations) through the modeling of the hidden relationships between a set of input variables (i.e, geo-environmental features prone to influence the location of Roman sites) and output variables (i.e., the archaeological sites). In detail, a decision tree is a decision support system using a tree-like model. The paths from root to leaf represent the rules of the model which, for a binary classification problem (e.g., presence or absence of Roman sites) works as follows: internal nodes allow splitting the observations based on the value of a specific attribute (e.g., elevation below or above a certain value); each branch represents the outcome of the previous step, where data are split in two main groups by maximizing the difference among them in terms of presences and absences; each leaf node represents a class label, after computing all attributes (if a roman site is present or absent at pixel level). RF constructs a huge number of independent trees and the prediction of new data is finally computed taking the majority or the soft voting. The latter consists in converting the results of a binary classification, such as the prediction of presence ("yes") or absence ("no") of a Roman site, by counting how many times each observation is classified as "yes" or "no", and by normalizing the result over the total number of predictions. This provides probabilistic outputs, which can be used to elaborate maps identifying areas susceptible to experience the presence of a Roman site, over a rank from very low to very high.

Operationally, a subset of the training dataset is generated by bootstrapping (i.e. random sampling with replacement), while about one-third of the cases, called

'out-of-bag', are left out at each iteration. Then the algorithm creates a decision tree for each training subset, and a reduced number of independent variables are randomly sampled as candidates at each split, which is done by measuring the node impurity. This lets the trees grow and eventually stops when each terminal node contains less than a pre-fixed amount of observations. The out-of-bag are used to optimize the parameters of RF (basically, the number of trees and the reduced number of variables), while a third dataset, named the *testing dataset*, is used to evaluate the error rate of the final optimized model and to assess its generalization performance. Indeed, at the beginning of the computation, a fraction of the original dataset is normally held out from the construction of the learning model, and used in the final step to assess the ability of the algorithm, trained on independent data (the *training dataset*), to make good predictions on unused observations (the *testing dataset*).

Our model involves the generation of pseudo-absences representing the non-sites. To assure a good generalization of the model and to avoid the overestimation of the lower classes, a balanced number of pseudo-absences (i.e. in a number equal to the observed presences) need to be specified. Moreover, RF allows us to measure the relative importance of each variable on the prediction. This is assessed by evaluating the mean decrease accuracy, computed by looking at how much the tree nodes, which use that variable, reduce the mean square errors estimated with the out-of-bag, across all the trees in the forest. Additionally, the partial dependence plots give us a graphical depiction of the marginal effect that each variable has on the class probability.

Two models were compared in this study: the first, including all the geo-environmental features, and a second one considering only the first six most important features resulting from the previous model. The final probability map was elaborated based on the results of the second model. Analyses were performed with the R language and environment for statistical computing (R Core Team, 2018). Specifically, for probability mapping we used the package *randomForest* (Liaw and Wiener 2002).

2.3.1. Model validation

When dealing with spatial data, as in the present study, observations belonging to the testing dataset are most likely located close to the training ones. This leads to overestimating the predictive performance of the model. This circumstance is known as "spatial autocorrelation", which implies that observations close to each other hold similar characteristics. One way to solve this issue is to select the training and testing data far enough apart in the geographic space by adopting, for example, a statistical technique called spatial *k*-fold cross validation. This technique consists in splitting the original dataset into a number *k* of non-overlapping groups, training the model on *k*-1 sets, and then testing it on the hold out set. The process is repeated *k*-times and the *k*-error estimates are finally averaged to yield the overall error rate. In this study, we adopted a 5-folds spatial cross validation (***Figure 4***). The *blockCV* package (Valavi et al. 2018) was

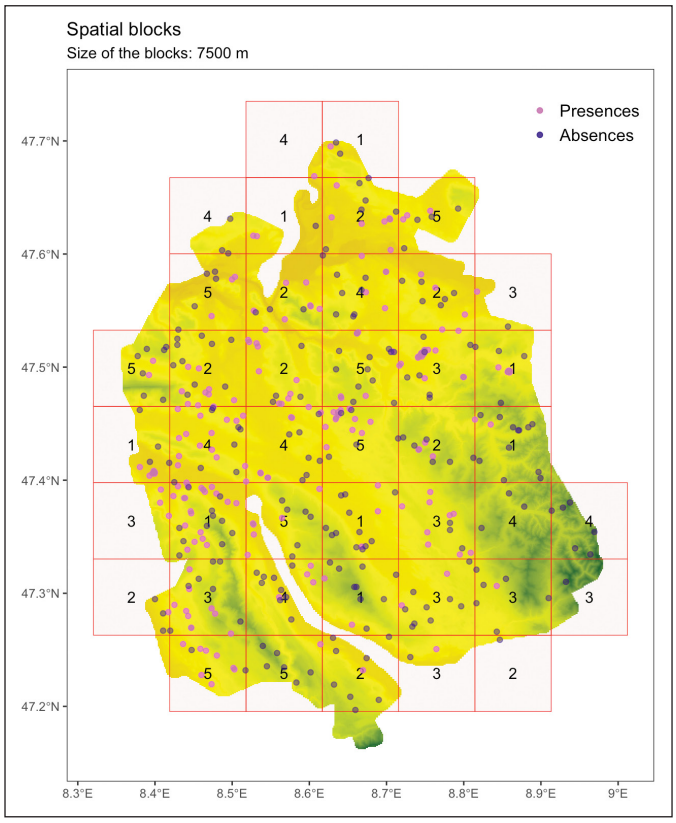

Figure 4 Spatial *k* fold cross validation. The figure shows the Study Area divided in 5-folds or spatial blocks with the real presences and pseudo absences overlaid.

used for performing spatial cross validation. The model was finally evaluated by using the ROC curve (Receiver Operating Characteristic). This is a graphical technique based on the plotting of the true positives rate (TPR) against the false positives rate (FPR), both expressed as a percentage of the total number, where true positives (TP) (or true negatives, TN) are the correct classifications, while false positives (FP) occurs when outcomes are incorrectly predicted as "yes" when it is actually "no" (and *vice-versa* for false negatives, or FN). The value of the "Area Under the ROC Curve" (AUC) lies between 0.5, denoting a bad classifier, and 1, denoting an excellent classifier. Both the ROC curve and the corresponding AUC were estimated to evaluate the performance of the model.

3. RESULTS AND DISCUSSION

The spatial probability of Roman settlements presence in the Canton of Zurich was assessed by using the RF model we implemented. As by-product, the model returns the variable importance ranking. The six most important variables (i.e., DEM, slope, water saturation capacity, geotechnical map, distance to lakes, and soil skeleton) are highlighted in ***Figure 5a*** (red rectangle): these were retained as input independent variables in the second model. Their relative importance, resulting from this last, is shown in ***Figure 5b***. Concerning the model validation and the estimation of its predictive performance (***Figure 6***), the AUC using all the variables is equal to 0.71 (blue line), while for the second model (i.e. considering only the six most important variables) it is slightly higher and equal to 0.72 (red line), which is considered as an acceptable discrimination according to the criteria of Hosmer and Lemeshow (2000). This result attests that variables low in the ranking do not add any supplementary predictive power to the model. They can thus be removed to estimate the final probabilistic output.

Figure 7 shows the output probability map obtained by the second model. The probability of finding Roman settlements in a certain area is expressed as percentage, and displayed in ten classes of equal intervals. The highest probability (red areas) are located around the modern urban agglomerates of the cities of Zurich and Winterthur, and in the middle area between these two municipalities, as well as around the main lake, Lake Zurich. This phenomenon can be interpreted in two ways. (i) Modern urban centers in the region were built upon the remaining of ancient settlements, in *continuum* with the main old *vici* (like Zurich or Winterthur). This cultural factor may have affected the preference for location choices of past populations, and may have led to new

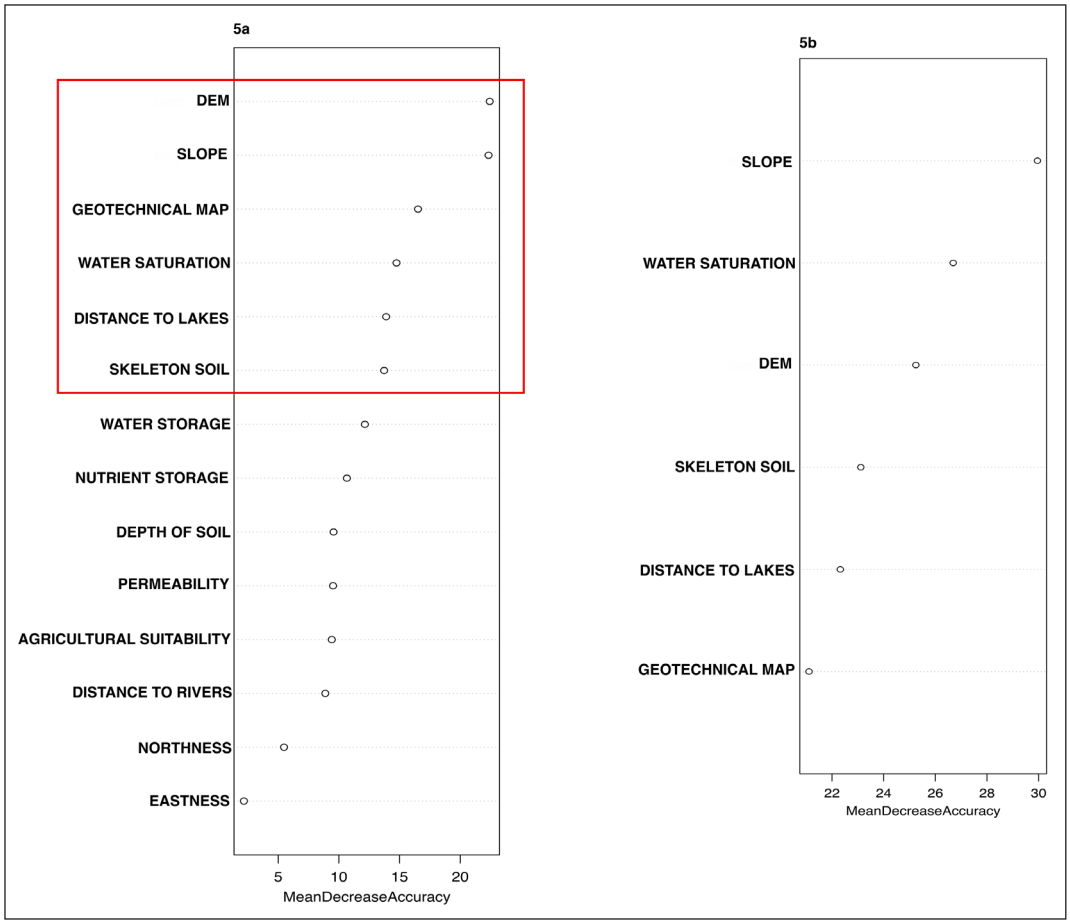

Figure 5 **Ranking of variables importance.** Mean decrease accuracy, allowing to measure the relative importance of each variable in the prediction, evaluated for all the features **(a)** and considering only the first six most important variables (red rectangle in 5a) **(b)**.

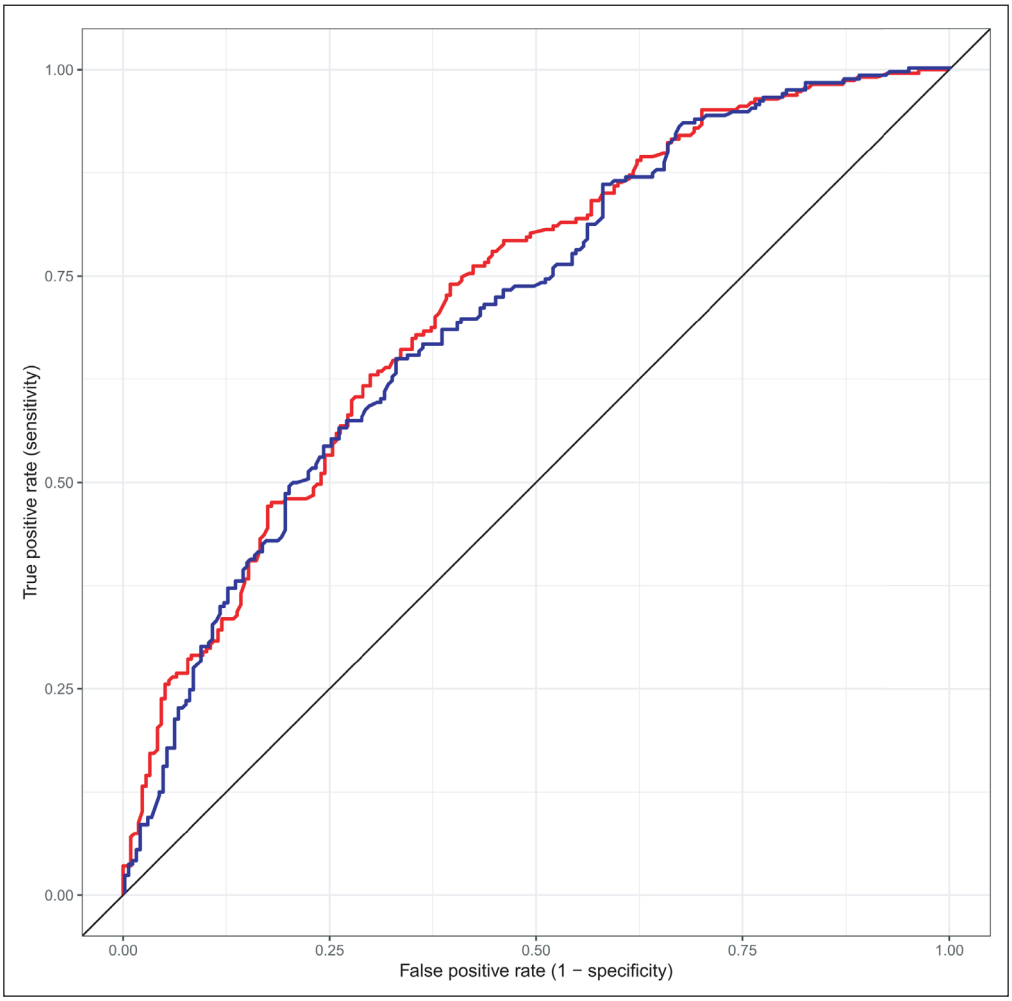

Figure 6 AUC-ROC. Receiver Operating Characteristic (ROC curves) considering all the features (blue line; AUC = 0.7077) and using only the six most important variables (red line; AUC = 0.7232).

settlements in their neighborhood. (ii) Urbanized areas are more intensively excavated as most archaeological surveys are rescue operations, which happen prior to any construction activity. By consequence, the majority of the archaeological sites discovered today are sites where such excavations take place.

Meanwhile, the lowest probability (blue areas) mainly occurs at the highest altitudes (above 700 m a.s.l. in the south-east of the region), and where unproductive zones (for agriculture) are located. Moreover, different partial dependence plots were evaluated, encapsulating how much each specific class for every single variable has positively or negatively influenced the location of Roman settlements. The partial dependence plots referring to the six most important variables are shown in ***Figure 8***. When looking closely at the partial dependence plots of the DEM and the Slope variables, we can see that the high probability to re-discover sites is mainly located between 400 and 600 m a.s.l. (within a slope of less than eight degrees). With regard to the distance to lakes, the highest probability lies between 5.5 km and 8.5 km away from the lakes. This is not surprising, as a high proportion of the study area is located in a certain distance to the nearest lake. Nevertheless, a small peak of high probability can be observed in close vicinity to the lakes, but not in their immediate vicinity. This points, in fact, to the certain importance of the lakeshores as preferred settlement areas, while avoiding the marshy or humid areas that existed along the lake coasts (also attested by the historical Dufour Map of Switzerland, published between 1845 and 1865). The partial dependence plot of the geotechnical classes, showing clays, gravels and glacial moraine deposits, as wells as marls, reveals recurrent types, corresponding to soils that are well-suited for agricultural uses. The partial dependence plots of the water saturation capacity and soil skeleton variables show that the class #1 (unknown, see ***Table 1***) occurs as the most influencing factor for the location of Roman settlements. Class #1 corresponds to the urbanized areas where no soil analyses were performed. It is essential to remember here that these variables, along with the depth of vegetal soil, water storage, permeability and nutrient storage capacity, derive directly from the soil map, and were compiled first for agricultural purposes by the Swiss Federal Office for Agriculture, in 2012. It therefore should come as no surprise that the digitization process may have produced less accurate information, or no data at all, with respect to those areas falling within modern

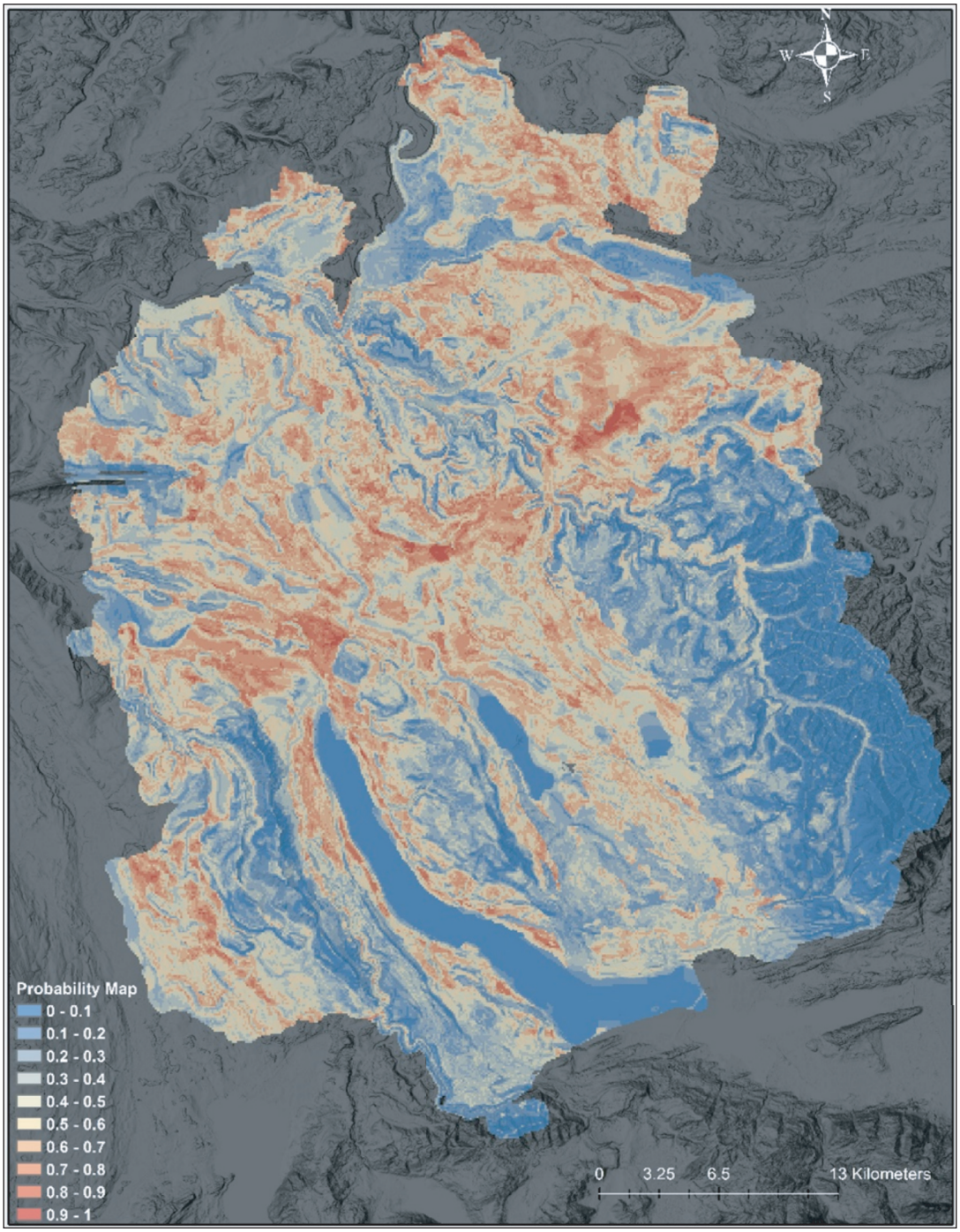

Figure 7 Probability map of the Canton of Zurich. The probability of finding Roman settlements in a certain area is expressed as a percentage and ordered in 10 classes with equal intervals.

urbanized agglomerations. This observation corroborates the correspondence we saw between high probability classes in urbanized areas, as discussed above.

Dealing with archeological remains, which are by nature under-sampled spatio-temporal data, is a fundamental issue in archeological predictive modeling. The final dataset of Roman settlements in the Canton of Zurich examined in the present study includes only 227 cases over an area of 1729 km^2. The scarcity of observations can explain the model's fair performance, as evaluated with an AUC of 0.72. Nevertheless, our model allows us to discover an interesting pattern in the distribution of the areas falling into the probability classes above the 50% threshold. These areas are generally located at: (i) about 7 km walking distance from the observed sites; (ii) an elevation of about 500 m.a.s.l. and a slope of less than 10°; (iii) more than 8 km distance from a lake; (iv) they belong to the areas defined as already urbanized in the modern soil map.

Furthermore, the machine learning modeling procedure revealed significant advantages with regard to the state of the art in archaeological predictive modeling studies. It represents an alternative to more classical statistical approaches. The reason why and the advantage of this innovative approach are exposed in the following. (i) It is not affected by any kind of subjective assumption;

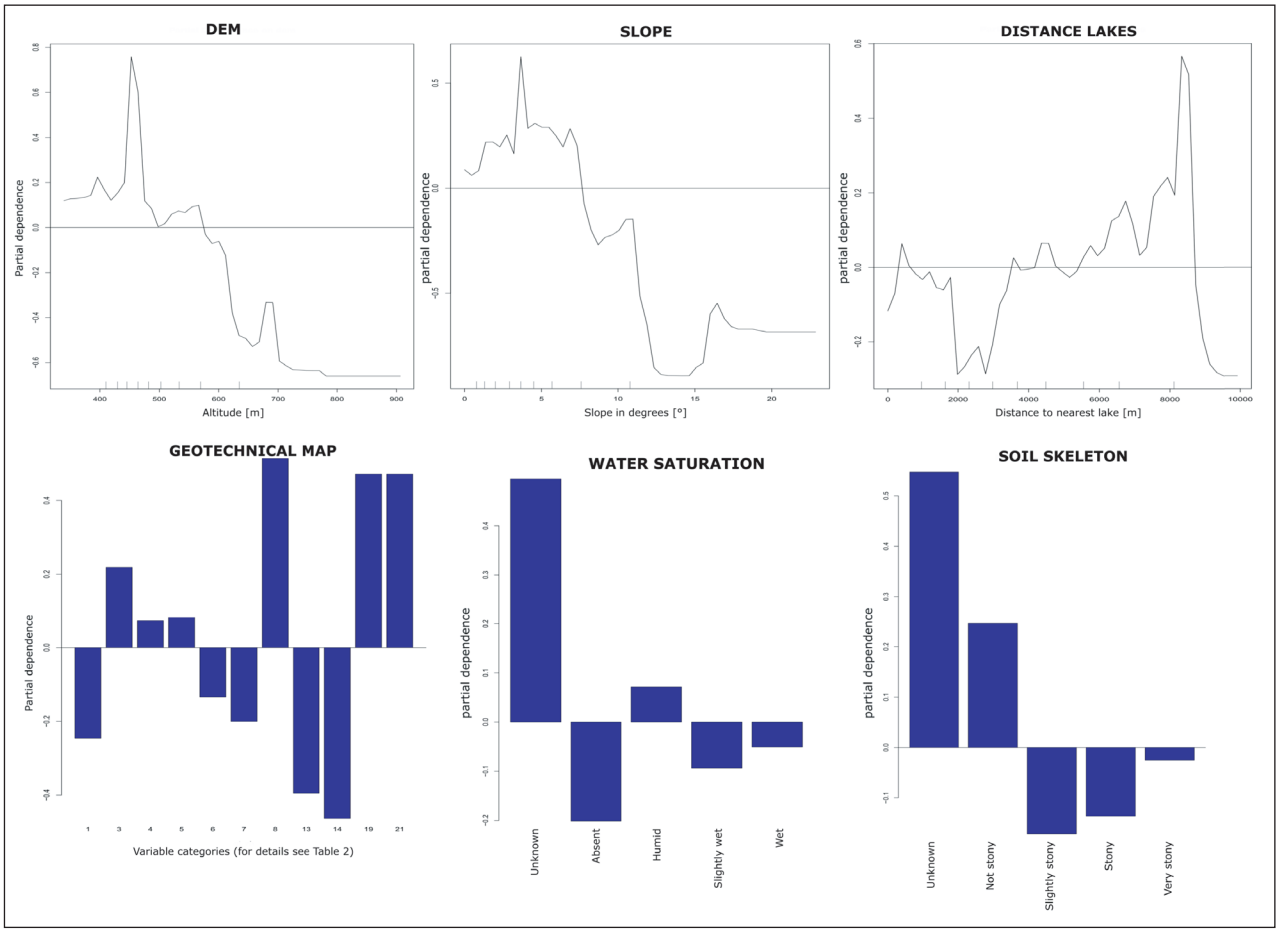

Figure 8 Partial dependence plots of the six most important variables resulting from the second model. The graphics indicate the influence (or marginal effect) of the specific class/range of values on the computed probability of Roman site location. The *Y* axis of each plot indicates the 'Partial Dependence', thus positive values (above 0.0) on this axis show that Roman sites are likely to be found for that value of the independent variable (x-axis), while negative values (below 0.0) indicate that Roman sites are less likely to be found. Zero implies no average impact on Roman site prediction according to the model.

that is, the selection of variables as well as their weight in the prediction procedure is performed independently, and without any previous assignment of reclassification or threshold values. (ii) No supplementary testing data is needed because the model quality assessment is part of the modeling procedure, as the input data are split in training, testing, and validation (performed by using the 'out of bag'). (iii) The spatial *k*-cross validation strategy for assessing the AUC using the testing dataset avoided inferring a model from new data that could be very close to, and hold the same characteristics of, the testing set. That would thus result in meaningless modeling performances. In other words, when assuming that two neighboring pixels are (almost) identical, the main goal of the spatial sampling is to avoid using archaeological site evidence for training, and identical evidence (in terms of geo-spatial characteristics) for testing the model. The quality of our classification was thus more honest since it was computed on different data. (iv) The variable ranking allows us to assess which environmental factor is a stronger player, while the partial dependence plot to determine the influence of each class belonging to the considered factors. (v) The model can manage thousands of variables and classes, both categorical and numerical at once, without internal conflict. (vi) The model we developed can be applied to investigate any kind of archaeological evidences and epochs, and (vii) it provides graphical outputs that non-expert readers can easily interpret, including a predictive map allowing them to identify areas where the likelihood of finding an archaeological site is very high.

4. CONCLUSIONS

This study aimed to outline and test a new predictive modeling technique based on Machine Learning approach, namely Random Forest, in order to identify Roman sites in the canton of Zurich, north-eastern Switzerland, with the help of institutional/legacy data. The predictive model we built here can be easily implemented and updated with the data collected during the most recent surveys (as from October 2015 to date). The model prediction can also be tested on the

field through planned surveys. However the quality and quantity of the data used are an important constraint for such such kind of applications, as it is often the case when dealing with archaeological information and modern environmental variables. Thus, the results obtained in this study inevitably raise interesting questions for archaeological managers and researchers, and also shed further light on possible future research avenues. Can Archaeological Predictive Models (APMs) really be used to comply with inventory requirements? To implement and improve data collection and storage strategies? To forecast effects, and make proactive streamlined management decisions regarding where to focus Cultural Heritage Management efforts? As this study shows, the answer to these questions is affirmative. The Random Forest based approach has demonstrated that it is a helpful instrument in overcoming issues related to data size, structure, and reliability. This study showed the importance of quantitative analysis for assessing the reliability of data on Roman settlement patterns in the Canton of Zurich and it has provided important insights for the interpretation and quantification of the variables that were only empirically considered to be important factors for locational strategies. Machine learning-based approaches are indeed able to extract insights and knowledge directly from data, and the algorithm can successfully highlight the relationships among the observed events (i.e., the archeological sites) and the prone environmental features, thus identifying trends and patterns hardly discernible by the human eye.

Finally, it is worth pointing out that the research developed in the present study is very promising in terms of technological innovation. Given the lack of previous quantitative investigations in this region, our study raises awareness on the necessity of employing quantitative methods to tackling more urgent questions, such as the protection and preservation of endangered archaeological sites. It also helps assess the importance of research biases and locational choices.

NOTES

1 *https://www.blw.admin.ch/blw/de/home/politik/datenmanagement/geografisches-informationssystem-gis/download-geodaten.html*.

2 *https://shop.swisstopo.admin.ch/en/products/maps/geology/GK500/GK500_DIGITAL*.

ACKNOWLEDGEMENTS

This study was carried out as part of my PhD in Archaeological Sciences, under the supervision of Prof. Dr. A. Hafner, at the University of Bern. Special thanks goes to the Amt für Raumentwicklung, Kantonsarchäologie Zurich for kindly providing access to the archaeological information.

COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR AFFILIATIONS

Maria Elena Castiello *orcid.org/0000-0002-0446-1301*
Institute of Archaeological Sciences, University of Bern, CH-3012 Bern, Switzerland

Marj Tonini *orcid.org/0000-0002-3592-8920*
Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, CH-1015 Lausanne, Switzerland

REFERENCES

Abedi, M and **Norouzi, G-H.** 2012. Integration of various geophysical data with geological and geochemical data to determine additional drilling for copper exploration. *Journal of Applied Geophysics*, 83: 35–45. DOI: *https://doi.org/10.1016/j.jappgeo.2012.05.003*

Abedi, M, Norouzi, G-H and **Bahroudi, A.** 2012. Support vector machine for multi-classification of mineral prospectivity areas. *Journal of Computers and Geosciences*, 46: 272–283. DOI: *https://doi.org/10.1016/j.cageo.2011.12.014*

Anichini, F, Bini, M, Fabiani, F, Gattiglia, G, Giacomelli, S, Gualandi, ML, Pappalardo, M and **Sarti, G.** 2011. MAPPA Project. Methodologies Applied to Archaeological Potential Predictivity. In: *MapPapers 1en-I*. pp. 23–43. Available at *https://www.mappalab.eu/wp-content/uploads/2019/10/MappaProject.pdf* [Last accessed 8 April 2021]. DOI: *https://doi.org/10.4456/MAPPA.2011.02*

Arnoldus-Huyzendveld, A, Citter, C and **Pizziolo G.** 2015. Predictivity –Postdictivity: A Theoretical Framework. In: Campana, S and Scopigno, R (eds.), *Keep the Revolution Going, proceedings of 43rd Computer Applications and Quantitative Methods in Archaeology. Atti del convegno internazionale*. Siena: Oxford. 2015. pp. 593–598.

Barcelo, JA and **Bogdanovic, I.** 2015. *Mathematics and Archaeology*. CRC Press. DOI: *https://doi.org/10.1201/b18530*

Baudron, P, Alono-Sarría, F, García-Aróstegui, JL, Cánovas-García, F, Martínez-Vicente, D and **Moreno-Brotóns, J.** 2013. Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification. *Journal of Hydrology*, 499: 303–315: DOI: *https://doi.org/10.1016/j.jhydrol.2013.07.009*

Bögli, H. 1962. *La Suisse à l'époque romaine*. Ed. Société Anonyme Chocolate Tobler.

Breiman, L, Cutler, A, Liaw, A and **Wiener, M.** 2018. *Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-14. DOI: *https://doi.org/10.1023/A:1010933404324*

Carlson, D. 2017. *Quantitative Methods in Archaeology Using R (Cambridge Manuals in Archaeology)*. Cambridge: Cambridge University Press. DOI: *https://doi.org/10.1017/9781139628730*

Carrer, F. 2013. An ethnoarchaeological inductive model for predicting archaeological site location. A case-study of pastoral settlement patterns in the Val di Fiemme and Val di Sole (Trentino, Italian Alps). *Journal of Anthropological Archaeology*, 32(1): 54–62. DOI: *https://doi.org/10.1016/j.jaa.2012.10.001*

Caspari, G and **Crespo, P.** 2019. Convolutional neural networks for archaeological site detection–Finding "princely" tombs. *Journal of Archaeological Science*, 110. DOI: *https://doi.org/10.1016/j.jas.2019.104998*

Cecamore, C and **Castiello, ME.** 2014. Un modello speditivo per la carta del Rischio Relativo nei Beni Culturali. In *Atti della 15a Conferenza Italiana Utenti Esri 9–10 Aprile 2014*. GEOmedia, [S.l.] 18, n. 2, giugno 2014. ISSN 2283-5687. Available at *https://www.mediageo.it/ojs/index.php/GEOmedia/article/view/873/801* [Last accessed 10 January 2021].

Chen, L, Priebe, CE, Sussman, DL, Comer, DC, Megarry, WP, et al. 2013. Enhanced Archaeological Predictive Modelling in Space Archaeology. *arXiv:1301.2738* [stat.AP]. Cornell University.

Cramatte, C. 2012. Turicum (Zürich). In: Bagnall, RS, Brodersen, K, Champion, CB, Erskine, A and Huebner, SR (eds.), *The Encyclopedia of Ancient History*. Blackwell Publishing Ltd. 2013. DOI: *https://doi.org/10.1002/9781444338386.wbeah16202*

Danese, M, Masini, N, Biscione, M and **Lasaponara, R.** 2014. Predictive modeling for preventive Archaeology: overview and case study. *Central European Journal of Geosciences*, 6(1): 42–55. DOI: *https://doi.org/10.2478/s13533-012-0160-5*

Deluigi, N. 2018. *Data-driven mapping of the potential mountain permafrost distribution*. (PhD thesis). Available at *http://nbn-resolving.org/urn:nbn:ch:serval-BIB_F417FD0D44072?siteLang=fr* [Last accessed 10 September 2020].

De Vries, P. 2007. Archaeological Predictive Models for the Elbe Valley around Dresden, Saxony, Germany. In: *Layers of Perception. Proceedings of the 35th Computer Applications and Quantitative Methods in Archaeology Conference*, Berlin, Germany, April 2–6, 2007. Bonn, pp. 1–9.

Djindjian, F. 2009. The Golden Years for Mathematics and Computers in Archaeology (1965–1985). *Archeologia e Calcolatori*, 20: 61–73.

Dubbini, N and **Lodoen, A.** 2014. Statistical and Mathematical Models for Archaeological Data Mining: A Comparison. In: *Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*. pp. 509–516.

Ducke, B and **Münch, U.** 2005. Predictive modelling and the Archaeological Heritage of Brandenburg (Germany). In: van Leusen & Kamermans (eds.), *Predictive modelling for archaeological heritage management: a research agenda*. Amersfoort. pp. 93–107.

Ebersbach, R. 2015. Eine Potentialkarte Archäologie für den Kanton Bern. In *Archäologie Bern/Archéologie Bernoise, Jahrbuch des Archäologischen Dienstes des Kantons Bern*. pp. 212–233.

Ebnöther, C. 1995. *Der römische Gutshof in Dietikon, Zürich/Egg*. Monographien der Kantonsarchäologie Zürich, Band 25.

Ebnöther, C and **Monnier, J.** 2002. Ländliche Besiedlung und Landwirtschaft. In: Laurent Flutsch, Urs Niffeler und Frédéric Rossi (eds.), *Die Schweiz vom Paläolithikum bis zum Mittelalter (SPM) V: Römische Zeit*. Basel. pp. 135–178.

Ejstrud, B. 2003. Indicative Models in Landscape Management: Testing the methods. In: Kunow & Müller (eds.), *Symposium on the archaeology of landscapes and geographic information systems. Predictive maps, settlement dynamics and space and territory in prehistory*. Wünsdorf, Germany: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum. pp. 119–134.

Espa, G, Benedetti, R, Meo, AD, Ricci, U and **Espa, S.** 2006. GIS based models and estimation methods for the probability of archaeological site location. *Journal of Cultural Heritage* 7: 147–155.

Flutsch, L, Niffeler, U and **Rossi, F.** 2002. Quand la Suisse n'existait pas. Les temps des Romains. In: *La Suisse du Paléolithique à l'âube du Moyen-Age Vol. 5 Epoque romaine*. Société suisse de préhistoire et d'archéologie, Bale.

Ford, A, Clarke, KC and **Raines, G.** 2009. Modeling settlement patterns of the late classic Maya civilization with Bayesian methods and geographic information systems. *Annals of the Association of American Geographers*, 99(3): 496–520. DOI: *https://doi.org/10.1080/00045600902931785*

Frei-Stolba, R and **Benedetti Martig, I.** 1991. *La Svizzera in epoca romana*. Schweizerische Zeitschrift für Geschichte. Revue suisse d'histoire – Rivista Storica Svizzera, 41: 111–125. Available at *https://www.e-periodica.ch/cntmng?pid=szg-006:1991:41::667* [Last accessed 15 November 2019].

Furger, A, Isler-Kerenyi, C, Jacomet, S, Russenberger, C., Schibler, J. (eds.) 2001. *Die Schweiz zur Zeit der Römer*. Archäologie und Kulturgeschichte der Schweiz, 3. Zürich: Verlag Neue Zürcher Zeitung.

Garcia-Molsosa, A, Orengo, HA, Lawrence, D, Philip, G, Hopper, K and **Petrie, CA.** 2020. Potential of deep learning segmentation for the extraction of archaeological features from historical map series. *Archaeological Prospection*, 1–13. DOI: *https://doi.org/10.1002/arp.1807*

Gattiglia, G. 2018. Classificare le ceramiche: dai metodi tradizionali all'intelligenza artificiale. L'esperienza del progetto europeo ArchAIDE. In: Malfitana, D (ed.), *Archeologia Quo Vadis? Riflessioni Metodologiche sul future di una disciplina. Atti del Workshop Internazionale Catania,* 18–19 Gennaio 2018.

Giligny, F, Djindjian, F, Costa, L, Moscati, P and **Robert, S.** 2010. CAA2014 Proceedings. In: *Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, 28(1): 1–6.

Goetz, JN, Brenning, A, Petschko, H and **Leopold, P.** 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences*, 81: 1–11. DOI: *https://doi.org/10.1016/j.cageo.2015.04.007*

Hafner, A. 2013. Archäologische Kulturgüter in der Schweiz – eine Ressource im Spannungsfeld von Zersiedlung und Verdichtung. *NIKE-Bulletin*, 28(4): 20–23. DOI: *https://doi.org/10.7892/boris.48607*

Hintz, M, Laabs, J and **Castiello, ME.** 2019. Archaeology that counts. International colloquium on digital archaeology. *Pages Magazine*, 27(1): 37. DOI: *https://doi.org/10.22498/pages.27.1.37*

Horisberger, B. 2017. *Zurigo, Cantone – Epoca romana*. Available at *https://hls-dhs-dss.ch/it/articles/007381/2017-08-24/* [Last accessed 15 November 2019].

Hosmer, DW and **Lemeshow, S.** 2000. *Applied Logistic Regression*. New York: Wiley. DOI: *https://doi.org/10.1002/0471722146*

Kaenel, G. 2002. Autoroutes et archéologie en Suisse. *Revue du Nord: archéologie de la Picardie et du Nord de la France*, 8(348): 33–41.

Kaeser, MA. 2012. *L'archéologie des grands travaux*. Laténium, Hauterive. pp. 77.

Kaeser, MA. 2013. L'archéologie, une affaire publique: les enjeux de la réglementation et du financement. *Les Nouvelles de l'archéologie. Financement et Réglementation étatique de la pratique de l'archéologie (fin XIXe-début XXe siècle)*, 133: 6–9. DOI: *https://doi.org/10.4000/nda.2081*

Kamermans, H, Deeben, J, Hallewas, D, Zoetbrood, P, van Leusen, M and **Verhagen, P.** 2005. 'Project Proposal', In: van Leusen, M and Kamermans, H (eds.), *Predictive Modelling for Archaeological Heritage Management: A research agenda.* Nederlandse Archeologische Rapporten 29. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek. pp. 13–23.

Kvamme, KL. 1990. The fundamental principles and practice of predictive archaeological modeling. In: Voorrips, A (ed.), *Mathematics and Information Science in Archaeology: A Flexible Framework*. Bonn, Germany: HOLOSVerlag. pp. 275–295.

Leuenberger, M, Parente, J, Tonini, M, Pereira, MG and **Kanevski, M.** 2017. Wildfire susceptibility mapping: Deterministic vs. stochastic approaches. *Environmental Modelling & Software*, 101: 194–203. DOI: *https://doi.org/10.1016/j.envsoft.2017.12.019*

Liaw, A and **Wiener, M.** 2002. Classification and Regression by randomForest. *R News*, 2(3): 18–22.

Lock, GR. 2000. *Beyond the Map. Archaeology and Spatial Technologies*. Amsterdam: IOS Press.

Märker, M and **Bolus, M.** 2018. Explorative Spatial Analysis of Neandertal Sites using Terrain Analysis and Stochastic Environmental Modelling. *GI_Forum*, 2: 21–38. DOI: *https://doi.org/10.1553/giscience2018_02_s21*

Märker, M and **Heydari-Guran, S.** 2009. Application of datamining technologies to predict Paleolithic site locations in the Zagros Mountains of Iran, In: *Proceedings of Computer Applications in Archaeology*, March 22–26, 2009, Williamsburg, Virginia, USA. pp. 1–7.

McEwan, DG. 2012. Qualitative Landscape Theories and Archaeological Predictive Modelling—A Journey Through No Man's Land? *Journal of Archaeological Method and Theory*, 19: 526–547. DOI: *https://doi.org/10.1007/s10816-012-9143-6*

Nebbia, M, Leone, A, Bockmann, R, Hddad, M, Abdouli, H, Masoud, A, et al. 2016. Developing a Collaborative Strategy to Manage and Preserve Cultural Heritage During the Libyan Conflict. The Case of the Gebel Nâusa. *Journal of Archaeological Method and Theory*, 23: 971–988. DOI: *https://doi.org/10.1007/s10816-016-9299-6*

Nicu, IC. 2019. Natural Risk Assessment and Mitigation of Cultural Heritage Sites in North-eastern Romania (Valea Oii river basin). *Area*, 51(1): 142–154. DOI: *https://doi.org/10.1111/area.12433*

Nussbaum, M, Spiess, K, Baltensweiler, A, Grob, U, Keller, A, Greiner, L, Schaepman, ME and **Papritz, A.** 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, 4: 1–22. DOI: *https://doi.org/10.5194/soil-4-1-2018*

Oonk, S and **Spijker, J.** 2015. A supervised machine-learning approach towards geochemical predictive modelling in archaeology. *Journal of Archaeological Science*, 59: 80–88. DOI: *https://doi.org/10.1016/j.jas.2015.04.002*

Orengo, HA, Conesa, FC, Garcia-Molsosa, A, Lobo, A, Green, AS, Madella, M and **Petrie, CA.** 2020. Automated detection of archaeological mounds using machine learning classification of multi-sensor and multi-temporal satellite data. *Proceedings of the National Academy of Sciences*, 117: 18240–18250. DOI: *https://doi.org/10.1073/pnas.2005583117*

Orengo, HA and **Garcia Molsosa, A.** 2019. A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery. *Journal of Archaeological Science*, 112. DOI: *https://doi.org/10.1016/j.jas.2019.105013*

Oštir, K, Kokalj, Ž, Saligny, L, Tolle, F, Nunninger, L, avec la collaboration de F. Pennors et K. Zaksek. 2007. Confidence Maps: A Tool to Evaluate Archaeological Data's Relevance in Spatial Analysis. In: *Layers of Perception. Proceedings of the 35th Computer Applications and Quantitative Methods in Archaeology Conference*, Berlin, Germany, April 2–6, 2007, Bonn, pp. 272–277.

Pham, BT, Pradhan, B, Tien Bui, D, Prakash, I, Dholakia, MB. 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environmental Modelling & Software*, 84: 240–250. DOI: *https://doi.org/10.5555/3006045.3006074*

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org/* [Last accessed April 2021].

Reichenbach, P, Rossi, M, Malamud, BD, Mihir, M and **Guzzetti, F.** 2018. A review of statistically based landslide

susceptibility models. *Earth-Science Reviews*, 180: 60–91. DOI: *https://doi.org/10.1016/j.earscirev.2018.03.001*

Roalkvam, I. 2020. Algorithmic Classification and Statistical Modelling of Coastal Settlement Patterns in Mesolithic South-Eastern Norway. *Journal of Computer Applications in Archaeology*, 3(1): 288–307. DOI: *https://doi.org/10.5334/jcaa.60*

Roberts, N, Fyfe, RM, Woodbridge, J, Gaillard, MJ, Davis, BAS, Kaplan, JO and **Leydet, M.** 2018. Europe's lost forests: A pollen-based synthesis for the last 11,000 years. *Scientific Reports*, 8(1). DOI: *https://doi.org/10.1038/s41598-017-18646-7*

Rogers, SR, Fischer, M and **Huss, M.** 2014. Combining glaciological and archaeological methods for gauging glacial archaeological potential. *Journal of Archaeological Science*, 52: 410–420. DOI: *https://doi.org/10.1016/j.jas.2014.09.010*

Rua, H. 2009. Geographic information systems in archaeological analysis: a predictive model in the detection of rural Roman villae. *Journal of Archaeological Science*, 36(2): 224–235. DOI: *https://doi.org/10.1016/j.jas.2008.09.003*

Simpson, I, Adderley, WP, Guðmundsson, G, Hallsdóttir, M, Sigurgeirsson, M and **Snæsdóttir, M.** 2002. Soil Limitations to Agrarian Land Production in Premodern Iceland, *Human Ecology*,30: 423–443. DOI: *https://doi.org/10.1023/a:1021161006022*

Tehrany, MS, Jones, S, Shabani, F, Martínez-Álvarez, F, Tien Bui, D. 2019. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. *Theoretical and Applied Climatology*, 137: 637–653. DOI: *https://doi.org/10.1007/s00704-018-2628-9*

Tonini, M, D'Andrea, M, Biondi, G, Degli Esposti, S, Trucchia, A, Fiorucci, P. 2020. A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. *Geosciences*, 10(3): 105. DOI: *https://doi.org/10.3390/geosciences10030105*

Valavi, R, Elith, J, Lahoz-Monfort, J, Guillera-Arroita, G. 2018. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2): 225–232. DOI: *https://doi.org/10.1111/2041-210X.13107*

Van Leusen, PM and **Kamermans, H.** (eds.) 2005. *Predictive Modelling for Archaeological Heritage Management: a research agenda*. Amersfoort: ROB, PlantijnCasparie Almere.

Vaughn, S and **Crawford, T.** 2009. A predictive model of archaeological potential: An example from northwestern Belize. *Applied Geography*, 29(4): 542–555. DOI: *https://doi.org/10.1016/j.apgeog.2009.01.001*

Verhagen, P. 2007. *Case studies in archaeological predictive modelling*. Leiden: Leiden University Press.

Verhagen, P and **Whitley, TG.** 2011. Integrating archaeological theory and predictive modeling: a live report from the scene. *Journal of Archaeological Theory and Method*, 19(1): 49–100. DOI: *https://doi.org/10.1007/s10816-011-9102-7*

Verhagen, P and **Whitley, TG.** 2020. Predictive spatial modelling. In: Gillings, M, Hacıgüzeller, P and Lock, G (eds.), *Archaeological Spatial Analysis: A Methodological Guide*. London & New York: Routledge. pp. 231–246. DOI: *https://doi.org/10.4324/9781351243858*

Visentin, D and **Carrer, F.** 2017. Evaluating Mesolithic settlement patterns in mountain environments (dolomites, eastern italian alps): The role of research biases and locational strategies. *Archaeologia e Calcolatori*, 28: 129–145. DOI: *https://doi.org/10.19282/AC.28.1.2017.08*

Wachtel, I, Zidon, R, Garti, S and **Shelch-Lavi, G.** 2018. Predictive modeling for archaeological site locations: Comparing logistic regression and maximal entropy in north Israel and north-east China. *Journal of Archaeological Science*, 92: 22–36. DOI: *https://doi.org/10.1016/j.jas.2018.02.001*

Westcott, KL and **Brandon, RJ.** (eds.) 1999. *Practical applications of GIS for archaeologists: A predictive modeling kit*. London: Taylor & Francis.

Wickham, C. 2011. *Framing the Early Middle Ages: Europe and the Mediterranean, 400–800*. Oxford University Press. DOI: *https://doi.org/10.1093/acprof:oso/9780199264490.001.0001*

Zêzere, JL, Pereira, S, Melo, R, Oliveira, SC and **Garcia, RAC,** 2017. Mapping landslide susceptibility using data-driven methods. *Science of The Total Environment*, 589: 250–267. DOI: *https://doi.org/10.1016/j.scitotenv.2017.02.188*

TO CITE THIS ARTICLE:
Castiello, ME and Tonini, M. 2021. An Explorative Application of Random Forest Algorithm for Archaeological Predictive Modeling. A Swiss Case Study. *Journal of Computer Applications in Archaeology*, 4(1), 110–125. DOI: *https://doi.org/10.5334/jcaa.71*

Submitted: 13 January 2021 Accepted: 07 April 2021 Published: 21 May 2021