Groundwater

Mapping Groundwater Potential Through an Ensemble of Big Data Methods

by P. Martínez-Santos[1] and P. Renard[2]

Abstract

Groundwater resources are crucial to safe drinking supplies in sub-Saharan Africa, and will be increasingly relied upon in a context of climate change. The need to better understand groundwater calls for innovative approaches to make the best out of the existing information. A methodology to map groundwater potential based on an ensemble of machine learning classifiers is presented. A large borehole database ($n = 1848$) was integrated into a Geographic Information Systems (GIS) environment and used to train, validate and test 12 machine learning algorithms. Each classifier predicts a binary target (positive or negative borehole) based on the minimum flow rate required for communal domestic supplies. Classification is based on a number of explanatory variables, including landforms, lineaments, soil, vegetation, geology and slope, among others. Correlations between the target and explanatory variables were then generalized to develop groundwater potential maps. Most algorithms attained success rates between 80% and 96% in terms of test score, which suggests that the outcomes provide an accurate picture of field conditions. Statistical learners were observed to perform better than most other algorithms, excepting random forests and support vector machines. Furthermore, it is concluded that the ensemble approach provides added value by incorporating a measure of uncertainty to the results. This technique may be used to rapidly map groundwater potential for rural supply or humanitarian emergencies in areas where there is sufficient historical data but where comprehensive field work is unfeasible.

Introduction

Groundwater is crucial in sub-Saharan Africa. This largely due to its abundance and reliability during droughts, which make it the resource of choice both in urban and rural areas across the continent. While millions own a domestic well, access to safer groundwater supplies is constrained by a number of factors, including cost, the absence of adequate operation and maintenance strategies and the high rate of negative boreholes (Harvey 2004; Foster et al. 2006; Foster 2013; Danert 2015). Causes should be found in the fact that groundwater development is often demand-driven, generally with little regard for hydrogeological considerations. The absence of competition in the drilling sector and certain unethical practices also hamper water access (Foster et al. 2006).

[1]Corresponding author: UNESCO/UNITWIN Chair Appropriate Technologies for Human Development. Departamento de Geodinámica. Estratigrafía y Paleontología. Facultad de Ciencias Geológicas. Universidad Complutense de Madrid. C/José Antonio Novais 12, 28040 Madrid, Spain; pemartin@ucm.es

[2]Centre for Hydrogeology and Geothermics, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland.

Article impact statement: This paper presents a novel method based on machine learning classifiers to map groundwater potential in remote regions.

Received March 2019, accepted August 2019.

doi: 10.1111/gwat.12939

Although some of these considerations are clearly beyond the technical scope, there is an impending need to develop approaches to make groundwater exploration as cost and time effective as possible. Groundwater potential mapping may underpin field surveys in areas where hydrogeological knowledge is limited. This technique allows to narrow down the choice of drilling locations, as well as to gain insight as to where the most favorable hydrogeological formations are. When coupled with adequate on-site exploration, groundwater potential maps contribute to improve siting of water supply boreholes, thus maximizing the chances of obtaining suitable yields for local communities. Furthermore, groundwater potential maps may be used to better understand groundwater flow patterns and ecosystem dependencies, as well as to convey information to planners and users.

Groundwater potential mapping typically relies on different factors. A common assumption is that groundwater occurrence can sometimes be predicted from surface features. These usually include soil, lineaments, slope, geology, landforms, lithology, and drainage density (Díaz-Alcaide and Martínez-Santos 2019a). Groundwater potential mapping is an application of predictive mapping, a forecasting technique that consists in developing spatially distributed estimates for a target variable based on a series of indirect indicators (explanatory variables). Predictive mapping involves the compilation of data derived from existing maps, aerial photographs,

satellite imagery, and airborne geophysical information (Schetselaar et al. 2008). In the past, predictive mapping has been successfully applied in fields as diverse as the delineation of forest structures (Ohmann and Gregory 2002), geological exploration (Schetselaar et al. 2008), studying the spatial distribution of vegetation species (Von Wehrden et al. 2009), soil-landscape relationships (Regmi and Rasmussen 2018), the potential spread of mosquito-borne disease (Jain and Kumar 2018), or fecal contamination in domestic wells (Díaz-Alcaide and Martínez-Santos 2019b).

This research deals with the application of machine learning in predictive groundwater mapping. Machine learning algorithms are frequently classified in two broad categories, namely, supervised and unsupervised methods. The main difference is the availability of ground-truth to calibrate the outcomes. Supervised algorithms attempt to unravel complex patterns in cases where the target and explanatory variables are known. Supervised learning focuses on finding the function or set of functions that link them. Once these are found, the target variable may be predicted in instances where it is no longer known. In contrast, unsupervised learning is used when the value of target variable is seldom or never available. Because there is no way to validate how precise an unsupervised algorithm is, these typically focus on discovering associations among variables.

The aim of this paper is threefold. From a methodological standpoint, we aim at demonstrating how an ensemble of machine learning classification algorithms can provide a self-validated approach to map groundwater potential, as well as to provide a measurement of uncertainty in the predictions. Furthermore, this research tests the ability of a specifically developed piece of software (MLMapper v 1.0) whose purpose is to elaborate predictive maps. The third goal is site-specific, and consists in illustrating this approach by developing a groundwater potential map for the Baoulé basin, southern Mali.

Research Method

Study Site

The study site is the Baoule subcatchment, southern Mali, which spans 60,000 km^2 within the upper Senegal basin. The Baoule River presents a length of approximately 550 km from the Mandingue plateau to its confluence with the Bakoye river (Figure 1). Temperatures are hot and relatively uniform across the basin (yearly average 28 °C). The coolest weather takes place in January in the southern area (25 °C yearly average in Kita), whereas the warmest occurs in the northern part of the basin in May (33 °C in Diema). Rainfall varies significantly, with a clear north–south gradient. In the northern part, which borders the Sahara desert, conditions are arid to semiarid (<500 mm/year), whereas rainfall exceeds 1000 mm/year in the south. Precipitation is subject to the West African monsoon throughout. The wet season lasts from June to October, and accounts for over 90% of the total annual precipitation. There is virtually no rain across the entire catchment between December and March.

The Baoule River is the sole permanent surface water course in the basin. Due to the absence of rain for most of the year, the population relies almost exclusively on groundwater. The area features three hydrogeological regions (Traore et al. 2018). The northern part consists of sedimentary Paleozoic rocks (sandstones, limestones, and shales) of the Cambrian-Carboniferous periods. These may exceed 1000 m in thickness, and present moderate to high productivity by local standards (transmissivities up to 450 m^2/d and average borehole yields of about 6 m^3/h). Groundwater flow occurs predominantly in fractures within sandstone and limestone, and is constrained by the presence of low permeability shale and regional-scale dolerite intrusions. A large part of the basin presents Permian–Triassic volcanic outcrops (basalts and gabbros). These are poorly fissured aquifers of moderate to low productivity, boreholes yielding about 1 m^3/h on average. The third region consists of Holocene sediments, which make up alluvial aquifers of local importance. Sand dunes behave as local-scale reservoirs toward the northern end of the basin.

Conceptual Model

The target outcome of groundwater potential maps is the feasibility of drilling successful boreholes in different parts of a given region. These are typically defined as "high" or "low" groundwater potential areas. In this case, the groundwater potential for each point in the map is expressed as "positive" or "negative" based on the likelihood of obtaining enough water to underpin communal domestic supplies. Thus, a positive borehole is that whose yield justifies the installation of a hand pump (at least 0.5 m^3/h), whereas a negative one is that whose flow rate falls below 0.5 m^3/h. From the hydrogeological viewpoint, the positive/negative dichotomy is a simplification. It is, however, appropriate here because yield is the decision criterion that determines whether a newly drilled borehole will be equipped or abandoned.

A groundwater potential map is necessarily affected by the conceptual model of the study area, that is, by those variables which are considered to have an influence on the presence or absence of groundwater. Because machine learning approaches operate on "brute force" (i.e., the computer needs to examine many individual cases in order to learn to generalize), a suitable large number of independent variables can be expected to render better outcomes. However, an excess of meaningless variables may incorporate undesired noise, potentially leading to spurious outcomes.

Explanatory variables must be operational (correlated to the target), complete (sufficiently represented across the study area), nonuniform (must vary spatially), measurable (should be susceptible of being expressed in some kind of scale) and nonredundant (its effect should not "double-count" toward the final result) (Ayalew and Yamagishi 2005). While most groundwater potential studies consider a more or less standard number of explanatory variables,

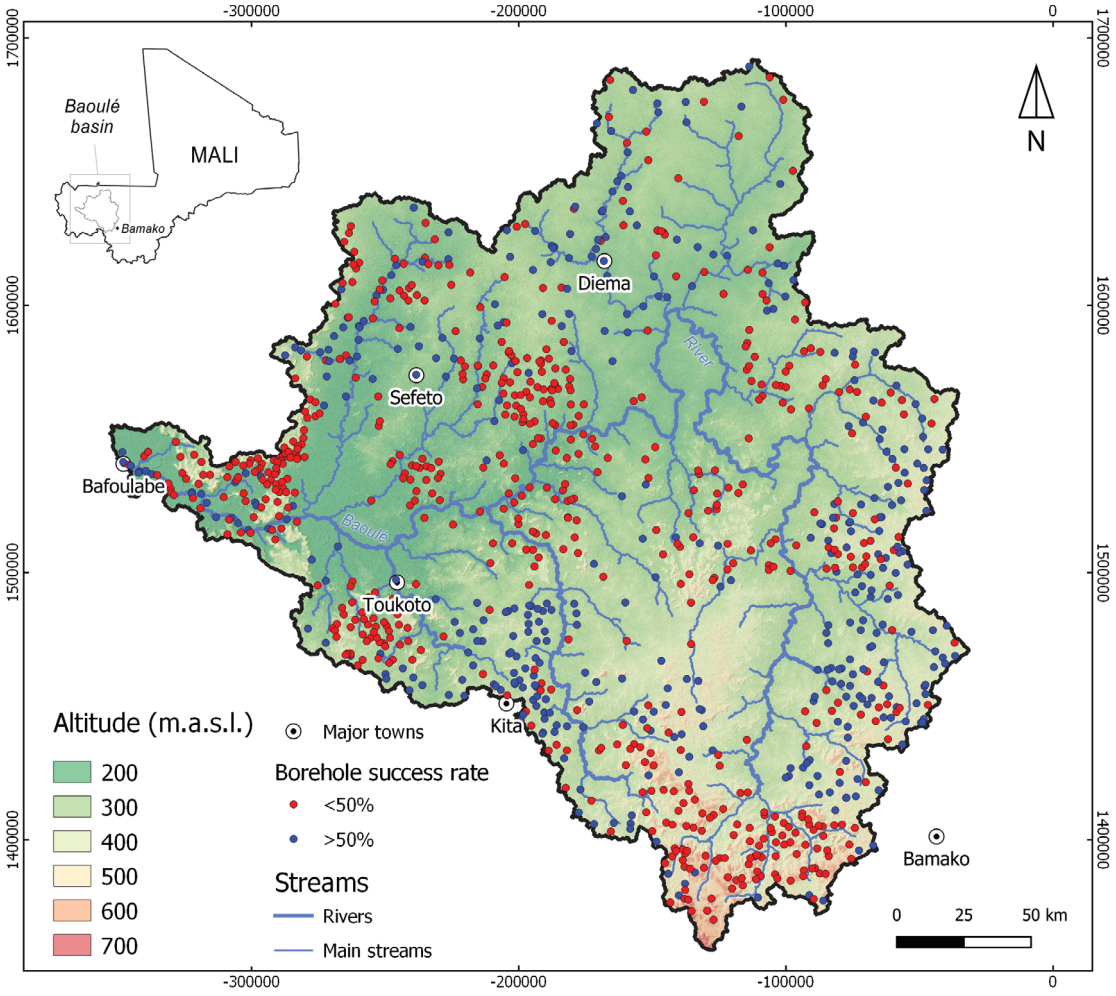

Figure 1. Study area. The Baoulé basin is located in the upper Senegal basin, southwestern Mali.

the importance of each factor may change based on specific geological, topographical and climatic conditions. This means, for example, that slope may not be as important in flat areas of a given catchment than in the mountains.

In the case at hand, the conceptual model uses 13 different explanatory variables, including geology and related features (lineaments), topography and related features (drainage density, drainage network, slope, topographic wetness index [TWI]), soils, land use and land cover, rainfall, and vegetation vigor (normalized difference vegetation index). These were selected based on an extensive survey of the groundwater mapping literature (Díaz-Alcaide and Martínez-Santos 2019a), as well as on the available information for the study site. Each variable is classified as "primary" or "secondary" (Table 1). Primary layer types are those elaborated from scratch or obtained directly from a given source, whereas secondary layer types are those which are elaborated from the primary ones.

Groundwater occurrence is first constrained by geology. Unconsolidated sediments and weathering mantles typically accumulate groundwater throughout, whereas groundwater only really occurs in fractures in the case of fresh crystalline rock. Lineament mapping is thus important when dealing with geological domains where fracture flow is expected to predominate over diffuse flow. Lineament maps rely on a variety of sources, including aerial photographs, digitally-processed color composites, radar data and digital elevation models. Both the methods and the outcomes present shortcomings. Whenever lineament mapping relies on operator intuition, nonrepeatable results can be expected. In such cases, comparing and integrating the results obtained by several operators may offset the effects of subjectivity (Sander et al. 1997; Sander 2007). Similarly, the combination of automated methods with an expert eye can render more reliable results than those obtained separately by each method (Meijerink 2007). On the other hand, outcome wise it is important to note that a fracture does not necessarily hold groundwater. Hydrogeologically, compressive fractures are less favorable than tensional or shear fractures. Besides, fractures filled with low-permeability sediments may be close to impervious. Hence, a knowledge of tectonic history and a map of fracture orientations may yield valuable insights. The potential of lineament maps can be enhanced by working with related variables such as lineament density or distance to major lineaments.

Table 1
Physical Model Layers (Explanatory Variables for Groundwater Potential)

Layer	Layer Type (Primary or Secondary)	Layer Origin	Classification
Geology, lineaments	Primary	Existing cartography, own elaboration	1. Miocene igneous 2. Precambrian 3. Holocene alluvial 4. Holocene undifferentiated
Lineament density	Secondary	Geology, lineaments	1. <175 m/km^2 2. 175 to 350 m/km^2 3. 350 to 500 m/km^2 4. >500 m/km^2
Lineament distance	Secondary	Geology, lineaments	1. <100 m 2. 100 to 250 m 3. 250 to 500 m 4. >500 m
Proximity to surface water (permanent or intermittent)	Secondary	Topography, satellite photo, surface water	1. <100 m 2. 100 to 250 m 3. 250 to 500 m 4. >500 m
Landforms	Secondary	Topography	1. Narrow valley bottoms, alluvial fans, incised valleys 2. Plains, flat areas 3. Plateaus, local valleys within plateaus 4. Cliffs, ridges, mountain tops
Drainage density	Secondary	Topography	1. <0.0004 m/km^2 2. 0.0004 to 0.0005 m/km^2 3. 0.0005 to 0.0006 m/km^2 4. >0.0006 m/km^2
Slope	Secondary	Topography	1 $<2\%$ 2 2% to 5% 3 5% to 15% 4. $>15\%$
Topographic wetness index	Secondary	Topography	1. Low flow accumulation 2. Moderate flow accumulation 3. High flow accumulation 4. Very high flow accumulation
NDVI (end of dry season)	Primary	Satellite image (Landsat 8)	1. <0.00 2. 0.00 to 0.25 3. 0.25 to 0.50 4. >0.50
Soil	Primary	ESDAC soil database	1. Leached, slightly acid soil, clay-enriched subsoil 2. Moderately developed soil (not acid) 3. Moderately developed soil (swelling clays) 4. Sandy soil with distinct clay accumulation 5. Shallow soil over continuous hard rock 6. Indurated soil with accumulation of hardened iron 7. Strongly cemented soil with iron nodules 8. Weakly developed soil in unconsolidated material
Land use/land cover	Primary	ESA Climate Change Initiative	1. Rainfed crops 2. Mosaic crops—Shrubland 3. Deciduous forest 4. Mosaic forest—Shrubland 5. Shrubland 6. Barren 7. Water
Rainfall	Primary	Own elaboration	1. <650 2. 650 to 800 3. 800 to 950 4. >950

Note: "Primary" layer types are those elaborated from scratch or obtained directly from a given source, whereas "secondary" layer types are those which are elaborated from the primary ones.

Landform cartography also provides hints as to the presence of shallow groundwater, as geomorphological features tend to define areas of preferential infiltration and storage. Alluvial fans, sand dunes, weathering mantles, and, in general, accumulations of unconsolidated materials, can be expected to store groundwater. In contrast, inselbergs, scarps, and ridges can be assumed unlikely to contain groundwater and impractical for drilling (Martín-Loeches et al. 2018). Geomorphological maps can be developed from field surveys, aerial photos and satellite images.

Soil data is also useful because soil permeability is related to effective porosity, grain shape, and size and void ratio, which suggests that soil type also plays a role in infiltration. A higher infiltration potential is expected in sandy and gravelly soils, while clayey and silty soils are less favorable. In the case at hand, the presence of indurated soils (laterites) across large sectors of the study area may prevent infiltration altogether, thus resulting in a low recharge potential. Soil descriptions were obtained from the European Soil Data Centre online database (Dewitte et al. 2013).

Human activities impact hydrological dynamics. This is the reason why land use and land cover cartography is frequently used in groundwater potential mapping. Cropland and forests are associated with high groundwater potential because plowing, root development, and biological activity favor infiltration. Permanent water bodies may also act as recharge or discharge mechanisms for the underlying aquifer (Naghibi et al. 2017). Human settlements and wastelands are generally assumed to be of low groundwater potential due to the widespread presence of impervious surfaces and the absence of moisture, respectively (Magesh et al. 2012).

Digital elevation models (DEM) contribute to groundwater potential mapping in a variety of ways. Reliance of DEM information assumes that infiltration and groundwater flow are partially driven by surface features. For instance, gentle slopes can be correlated with slow runoff and longer residence times at the surface, which favors recharge. Steeper gradients imply greater erosion and short residence time, while the presence of unconsolidated sediments in steep slopes is also less likely. Hence, so is the potential for groundwater accumulation (Fashae et al. 2014). Topographic control on variables such as infiltration or soil moisture can be inferred from the TWI (Sorensen et al. 2006). Groundwater occurrence can be expected to correlate well with a high TWI because this index computes the relation between the water that accumulates at any point of a given catchment and the gravitational force that drives water down slope (Nampak et al. 2014).

Drainage density is obtained from DEM data, and can be used to complement TWI. Drainage density is the total length of the streams per subcatchment area. Thus, it depicts how close together runoff channels are. If the drainage density is high, erosion potential is high and runoff can be evacuated quickly and infiltration potential is low (Magesh et al. 2012; Fashae et al. 2014).

Satellite images provide valuable information on shallow groundwater. Perhaps the most important variables are the occurrence and vigor of vegetation. The normalized vegetation index (NDVI) distinguishes the response of vegetation to visible red and infrared wavelengths, thus providing and indicator of vegetation vigor (Xie et al. 2008; Xue and Su 2017). Because NDVI is sensitive to seasonal changes, assessing it over time is interesting in dry climates or in regions subject to very clear seasonality. Satellite images may also lead to identify rock types based on mineral spectral response (Gupta 2018).

Mapping Methodology

Overview of the Machine Learning Classifiers

Both the way data are integrated and the validation procedure are of crucial importance to ensure representative results. MLMapper v 1.0 was developed and used for this purpose. MLMapper is a QGIS 3 plugin created by the authors to produce predictive maps based on point-source data. It provides a set of tools to process efficiently and integrate the explanatory variables in order to predict a Boolean outcome (positive or negative borehole).

Within the machine learning terminology, an algorithm predicting a categorical outcome such as a Boolean one is called a classifier. MLMapper uses 12 supervised machine learning classifiers from the SciKit-Learn 0.19.2 toolbox (Pedregosa et al. 2011). These include support vector machines (SVCs), logistic regression (LRG), decision tree classifier (CRT), random forest classifier (RFC), K-neighbor classification (KNN), linear discriminant analysis (LDA), Gaussian naïve Bayes classification (NBA), multilayer perceptron neural network (MLP), Ada-boost classifier (ABC), quadratic discriminant analysis (QDA), gradient boosting classification (GBC), and Gaussian process classifier (GPC).

These classifiers can be grouped in six major families: statistical learners (LDA, QDA, NBA, GPC, LRG), decision trees (CRT, RFC), instance learners (KNN), support vector machines (SVCs), ensemble methods (ABC, GBC, RFC), and neural network models (MLP). The theory behind each of these algorithms has been discussed in depth by different authors (Kotsiantis 2007; Hastie et al. 2009; Pedregosa et al. 2011).

Data Processing

The main assumption behind this research is that supervised classifiers can identify meaningful associations between the target and explanatory variables for those points in space where both are known. Once found, these may be generalized to develop a predictive map because the value of each explanatory variable is known for every pixel in the spatial database.

Thus, MLMapper requires two inputs. The first one is a borehole dataset. A geographic database with information from 1848 boreholes distributed across 550 villages was used as input (DNH 2010). Most records include location (village), number of positive and negative

boreholes, success rate, average yield, average depth of the boreholes, average depth of the static groundwater level, and average electric conductivity of the water.

The borehole database is a point vector shapefile that includes all attributes from the original database, as well as the pixel value of each borehole for each layer of the GIS (explanatory variables). Each point also includes a Boolean attribute representing "success" (0 = negative; 1 = positive), which is the target variable for classification. Success is defined as the likelihood of siting a positive borehole. In turn, each borehole is defined as "positive" and "negative" based whether it exceeds a flow rate of 0.5 m^3/h.

The second input file is also a point vector shapefile. Every pixel in the database corresponds to a point whose attributes are its values for each layer. Explanatory variables include geology, soil type, landforms, proximity to lineaments, lineament density, proximity to permanent surface water courses, proximity to ephemeral surface water courses, land use and cover, rainfall, drainage density, TWI, slope, and normalized difference vegetation index (Figure 2 and Table 1). If the borehole database presents a suitably large number of records, machine learning algorithms may be able to identify those patterns among the explanatory variables that lead to a positive or a negative borehole. Then the outcomes can be extrapolated to every pixel in the database.

The standard machine learning protocol involves splitting the input database into training and testing datasets (Figure 3). Each classifier loops through the training dataset in an attempt to detect associations between the target and explanatory variables. At this stage the computer uses the target, so it can determine which patterns of explanatory variables lead to each outcome.

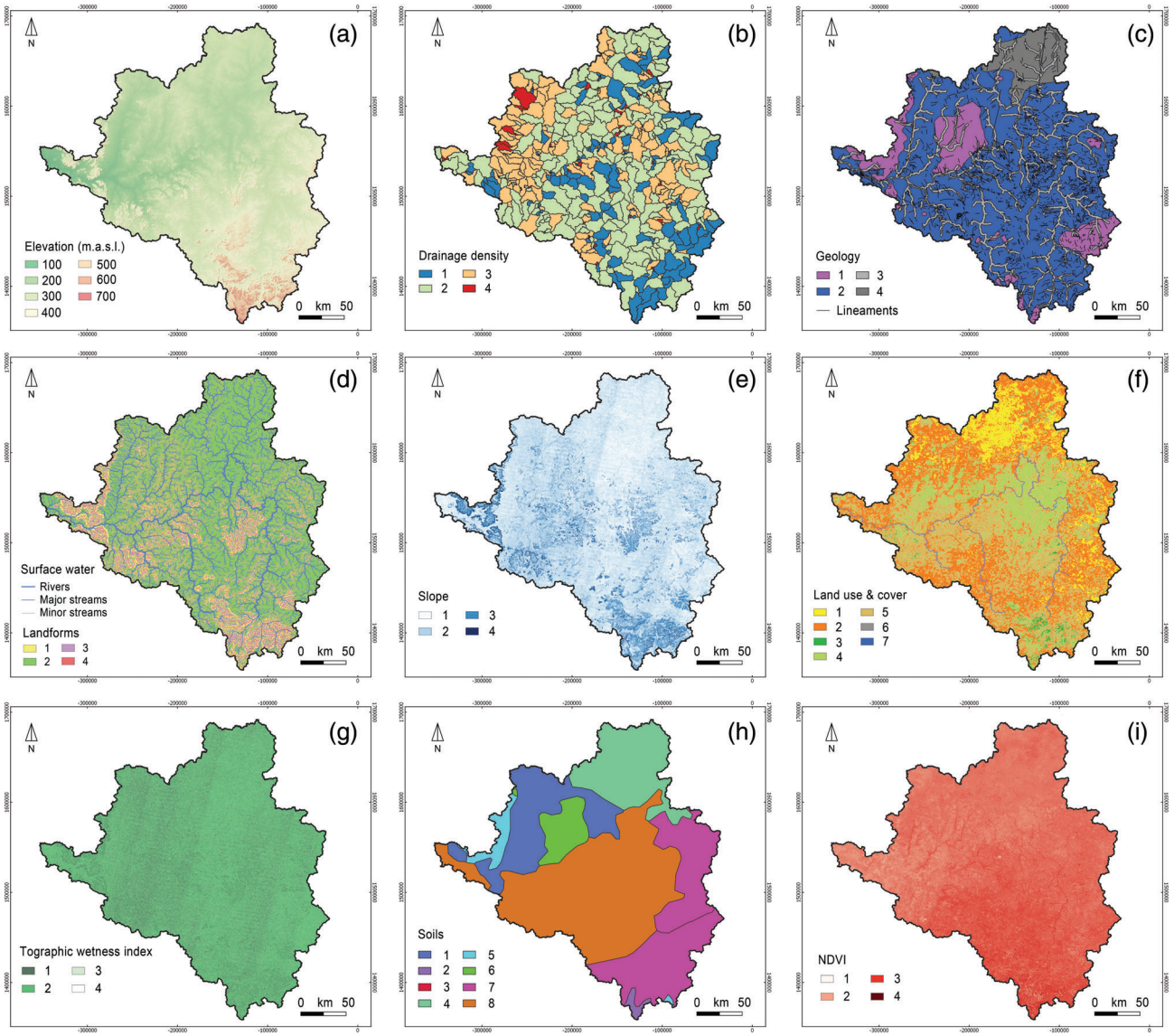

Figure 2. Explanatory variables. (a) Topography; (b) Drainage density; (c) Geology and lineaments; (d) Surface water and landforms; (e) Slope; (f) Land use and land cover; (g) Topographic wetness index; (h) Soil map; (i) Normalized difference vegetation index. Numeric values refer to the classification indices in Table 1.

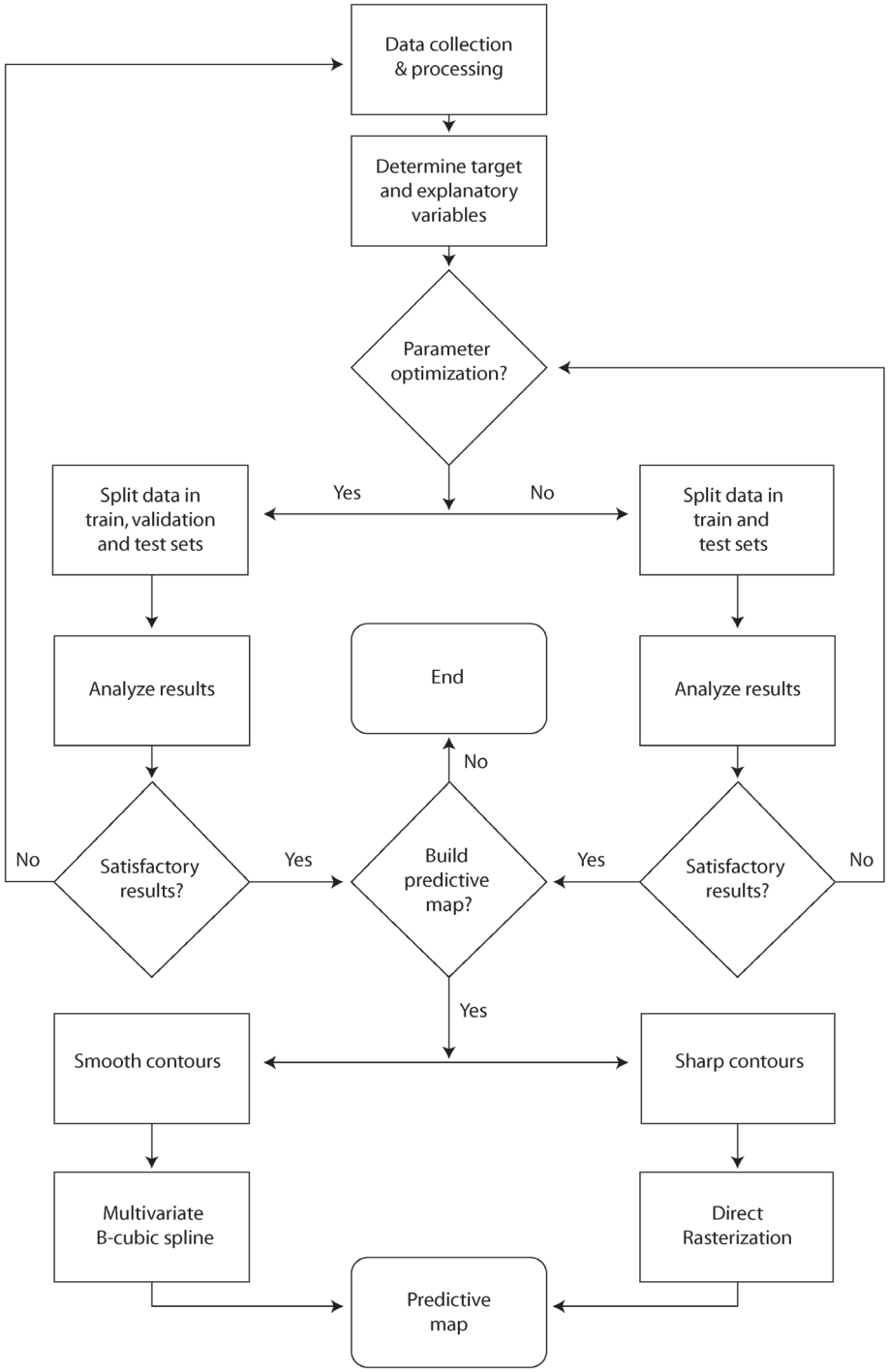

Figure 3. MLMapper v1.0 conceptual design. MLMapper is an open-source QGIS 3 plugin that can be integrated seamlessly within the QGIS environment.

Testing takes place immediately after. In this second stage, the computer attempts to predict the target values based on the associations it developed during the training stage. The purpose of the test is to check whether the associations found by the computer can be generalized into reliable predictions.

If there are no clear correlations between the input and explanatory variables, classifiers have a tendency to overfit the training set. In plain words, this means that the computer memorizes the training set in order to maximize its own success rate. This results in a high training score coupled with a low ability to generalize into meaningful predictions. Overfitting may also stem from an imbalanced input dataset (i.e., if one of the target values is considerably more frequent than the others). Under such conditions, machine learning algorithms often focus on explaining the most frequent outcome, while simply ignoring the rest.

Improving the explanatory variables and finding ways to enrich the input dataset with additional data are

common approaches to prevent overfitting. Furthermore, overfitting may be reduced by automated parameter tuning. For this purpose, a validation subset needs to be derived from the input dataset. Validation is an intermediate stage that takes place between training and testing. It is best described as a process whereby the user is allowed to specify a range of values for the internal parameters of each algorithm, so that the computer will pick whichever combination renders a better test score.

In this case, we used two techniques to reduce overfitting. Since the original dataset was perceptibly skewed toward the positive-type outcome, the first runs with MLMapper were observed to render very low test scores for the negative-type outcome. Thus, a series of negative boreholes were added to the dataset at locations where drilling a borehole would be likely to fail (inselbergs, cliffs). The second approach consisted in using a validation set with automated parameter optimization. MLMapper uses a parameter tuning routine known as grid search. Grid-search allows for the optimization of those parameters shown in the supporting information section (Table S1). Some of them are relatively straightforward to understand. Take for instance the depth of a decision tree, the number of trees used to develop a forest, the number of iterations to be performed or the mathematical solver used to reach the results. Others are significantly more complex, and govern the bias of algorithms toward rendering certain outcomes by including more or less severe penalties to test scores each time the algorithm makes the wrong prediction.

Parameter optimization is set by default to maximize accuracy (i.e., the number of "correct guesses"), but advanced users may gear it toward the optimization of any of the other metrics by modifying the source code. Train/test/validation splits of 60/40, 70/30 and 80/20 were considered for this purpose. A 70/30 split means that the 70% of the original dataset was used for training; it also implies that 70% of the remainder was used for validation and 30% for testing (this means 70/21/9 for training, validation and testing, respectively).

Since data is aggregated at the village level, all boreholes within the same village have the same coordinates (i.e., all are located at the exact same point in the GIS database). This poses no problem in those villages where all boreholes in the database are either positive (100% success rate) or negative (0% success rate). However, in cases where there are both successful and unsuccessful boreholes there is a need to determine whether the point in question should be labeled "positive" or "negative." There are two ways to approach this. The most obvious one would be to establish the actual success rate (or a multilevel classification) as the target variable. However, this was observed to result in an imbalanced dataset, ultimately leading to overfitting. Besides, this choice restricted the number of classifiers that could be used, as some are limited by design to Boolean outcomes. Hence adopted an approach based on confidence thresholds (or "thresholds," for short) was adopted. For the purpose of the ensuing discussions, a confidence threshold of 0.2 means that those villages with a success rate in excess of or equal to 20% were all taken as positive, whereas the rest were considered negative. Confidence thresholds of 0.2, 0.4, 0.6, and 0.8 were used for analysis.

Classifier Ensemble

Since each algorithm relies on different principles, some will inevitably perform better than others. By establishing adequate metrics and examining the individual performance of each one, it is possible to determine which ones are better suited to each case. Ensemble methods can then be used to combine the better ones into a single predictive model (e.g., Martre et al. 2015). This contributes to decrease variance and bias, as well as to improve the reliability of predictions. Thus, mapping is carried out in two steps. All algorithms are run separately at first. Then they are ranked based on how well they perform on the test set for each split and threshold. Each selected algorithm is run separately to come up with algorithm-specific groundwater potential maps. Then these are averaged out to produce the ensemble map.

For this purpose, MLMapper incorporates two rasterization methods from QGIS' SAGA toolbox: direct rasterization and multilevel b-cubic spline interpolation. Direct rasterization simply copies the score of each point to a raster grid, often resulting in sharp contrasts from one pixel to the next. Conversely, the spline approach uses geostatistical interpolation between points, thus providing a smoother output.

Results

Individual Classifiers

Figure 4 presents the outcomes of individual classifiers for each split and threshold, comparing the training, validation, optimized training, and test scores. As shown, all algorithms tend to yield similar accuracies (i.e., number of correct predictions relative to the total number of predictions), with small standard deviations for each train/test/validation split. This suggests that the effect of the split is relatively small, which in turn implies that the dataset appears to be large enough for practical purposes. Confidence thresholds are however important. The best results are obtained for a threshold of 0.2, accuracies gradually dropping until the 0.8 threshold. In other words, all models get worse at predicting the likelihood of siting a positive borehole as the confidence threshold gets higher (Tables S2 and S3 in "Supporting Information").

Tree-based algorithms (RFC, CRT), were found to have a greater tendency to overfit the data at first, as shown by the near 1.0 training accuracy and a relatively large difference between this and the test score in most cases (Figure 4). The fact that the optimized training score is far closer to the test score suggests the need to carry out optimization routines whenever these algorithms are used, even though this implies a 10-fold increase in computational time. In contrast, optimization procedures were observed to turn out counterproductive results

Figure 4. Comparative performance of all 12 classifiers in terms of training, validation, optimized training, and testing scores for each split and threshold. The training and optimized training scores coincide in the case of those algorithms which do not allow for parameter tuning.

on occasion, like in the case of the KNN and MLP algorithms. Automated optimization rendered marginal improvements in the case of discriminant analyses (LDA, QDA) and Bayesian models (NBA), which could be attributed to the limited choice of parameter tuning options.

Since test score determines generalization potential, it provides a valid metric to determine which classifiers to include in the ensemble. Those classifiers that are better ranked on average also yield the lower standard deviations for ranking, which points at a consistently high performance. Furthermore, there is a noticeable gap between the fifth- and sixth-ranked classifiers (LRG [4.3] and GBC [6.2], respectively). Consequently, only those algorithms ranked above sixth (LDA, LRG, NBA, RFC, and SVC) were used for ensemble purposes.

Figure 5a to 5e present the map outcomes of each of the best performing classifiers for a threshold of 0.2 and a split of 80/20. Some common features are observed. For instance, all five classifiers tend to respect the major fluvial courses, as well as the northern plains and southern alluvial valleys, as areas of high groundwater potential. All but RFC render largely similar maps, the difference being that RFC tends to identify larger high groundwater potential areas. The more restrictive algorithms in terms of identifying areas of high groundwater potential were LDA and LRG.

Ensemble Mapping

Ensemble averages of models have been shown to provide consistently results of higher reliability than individual models in various fields (Tebaldi and Knutti 2007; Martre et al. 2015). The approach in this case is limited

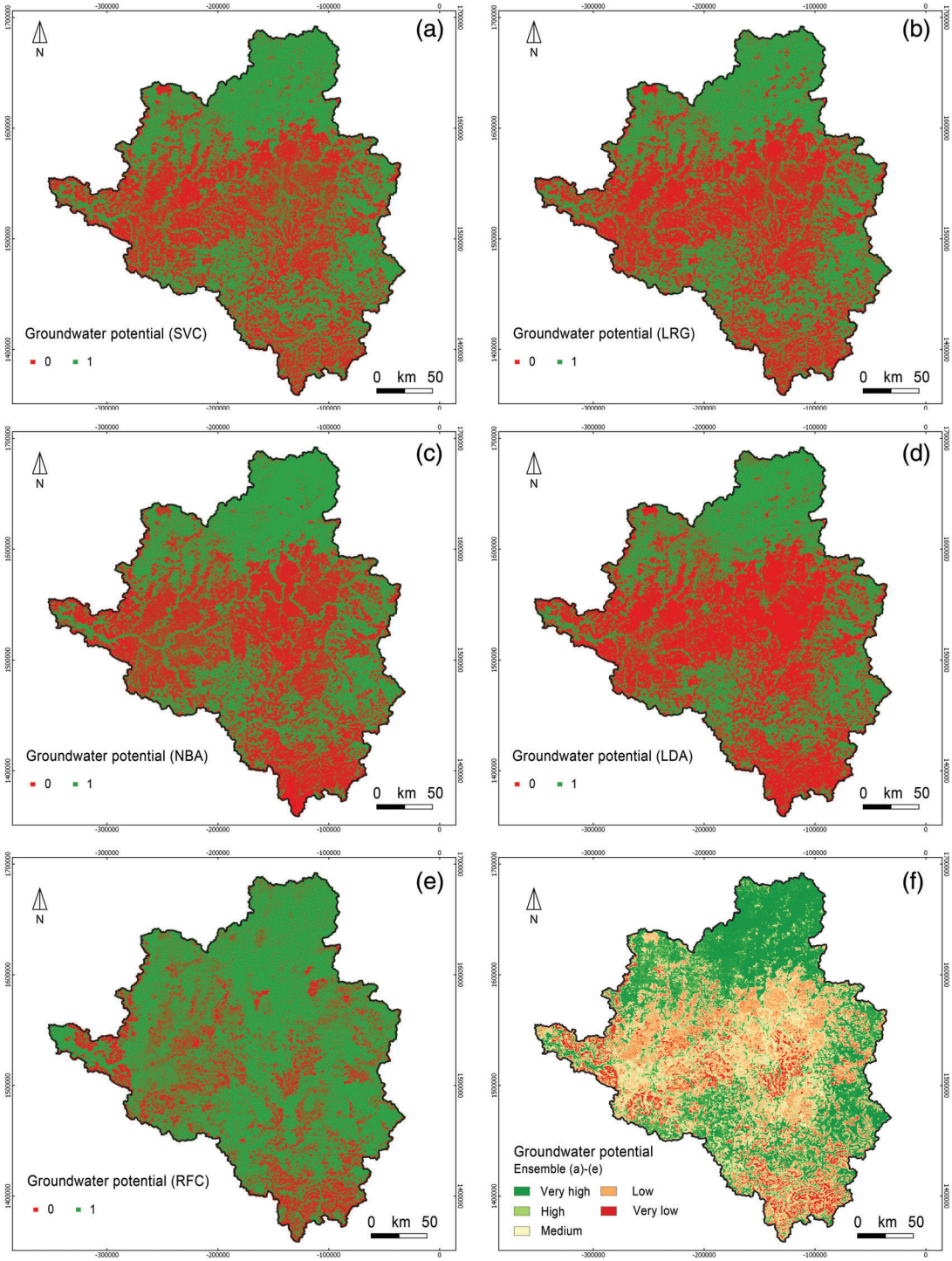

Figure 5. Groundwater potential maps calculated individually for each of the five best classifiers (LDA, LRG, NBA, RFC, and SVC), (a) through (e), compared to the ensemble of the five (f). The outcome in figures (a) through (e) is binary. A "0" value means low potential and a "1" score represents high potential. The threshold is 0.2 and the split is 80/20 in all six cases.

to computing the arithmetic mean of the best classifiers. In practice, the ensemble map represents the degree of agreement among classifiers for each possible outcome (Figure 5f). Each pixel score is computed as the arithmetic mean of the values obtained by each classifier, the underlying assumption being that the best performing classifiers will render the better predictions. Thus, a dark green pixel implies that all five classifiers "agreed" to define that location as "positive" whereas a dark red one implies all classifiers agreed on a "negative" outcome. Intermediate values (orange) indicate disagreement between classifiers, and thus, greater uncertainty. Thus, the combined weight of NBA, LDA, LRG, and SVC explains why the central area in the ensemble map presents a medium groundwater potential despite the influence of RFC outcomes.

Discussion

According to the better performing algorithms, the key features constraining groundwater potential in the area are proximity to fluvial courses, slope and landforms, whereas rainfall, TWI, and lineaments present a comparatively lesser bearing on the results. This can be readily observed in Figure 5, by examining the resemblance of the outcomes with the explanatory variables of Figure 2. A key aspect of groundwater potential mapping, however, is the evaluation of uncertainties. These may stem either from the physical model or from the method used to integrate the data (i.e., the algorithm), and will be discussed separately under the two following headings. A third one will deal with the uses of MLMapper.

Explanatory Variables and Input Dataset

An important aspect in evaluating the results is that machine learning relies by design on "big data," whereas groundwater is often "small data." Indeed, machine learning algorithms require huge datasets (typically thousands or millions of points) to smooth out inconsistencies and improve generalization potential. Such datasets may be easily obtained in several domains of environmental science, but not so often in groundwater. Thus, some degree of uncertainty in the results can be attributed to the input dataset being relatively small by machine learning standards. Furthermore, input data may suffer from inaccuracies, as well as from measurement errors, sampling limitations and geographical bias.

Another explanation for ambiguity is the physical model itself. Díaz-Alcaide and Martínez-Santos (2019a) found over 20 variables that are frequently involved in groundwater potential studies, out of which eight are almost always present (geology, lineaments, landforms, soil, land use/land cover, rainfall, drainage density, and slope). An important detail is that these are all surface variables, while by definition groundwater takes place underground. This means that groundwater potential mapping methods are typically suitable for areas where the presence of shallow groundwater may be inferred from surface features, but much more difficult to apply in cases where groundwater is found at great depth. Furthermore, there is often no means to determine a priori which ones among the choice of explanatory variables are actual predictors for groundwater occurrence and which ones just incorporate noise to the model. Because the algorithms rely on mere association between physical variables and borehole outcome, these aspects can become a major source of uncertainty.

On the other hand, each explanatory variable may present a different bearing in each region of the study area. In this sense, machine learning algorithms provide a welcome addition to groundwater potential mapping by incorporating automated nonlinear decision mechanisms. These can be expected to improve classic expert criteria approaches in situations where the interaction among variables may be too complex for the human eye to interpret. In certain cases, the outcomes of automated mapping may be counter-intuitive. For instance, in the case at hand there is a relatively low association between positive boreholes and the presence of lineaments. This could be attributed to the fact that not all of the study region is made up of fissured aquifers. In fact, a large part of the system is actually made up of loose sediments, where there are many positive boreholes despite the absence of lineaments. Hence, other variables were found to be more relevant.

Machine Learning Classifiers

All the top-five classifiers performed well above 90% in terms of the test score for the 0.2 confidence threshold, which means that, in theory, all of them could reliably predict favorable areas of groundwater potential. Two statistical learners (linear discriminant analysis, LDA, and Gaussian naïve Bayes classification, NBA) ranked first and second among all classifiers, whereas a third one (logistic regression, LRG) ranked fifth. This suggests that this family of supervised learners is generally useful for groundwater mapping. The other two top performers in terms of test score are random forest and SVCs. The advantages and disadvantages of these methods will be discussed in the ensuing paragraphs.

An important characteristic of statistical learning algorithms is that these build probabilistic models to estimate the chance that a set of explanatory variables will render a given target. Hence, unless the explanatory variables incorporate spatial consideration explicitly (proximity to lineaments, proximity to fluvial courses), the models will be based on statistical association alone. This should be handled adequately in mapping studies such as the one at hand so as to prevent spurious results.

NBA classifiers rely on Bayes conditional probability theorem, and present several practical advantages. NBA are easy to build and interpret, and can be expected to be robust in most circumstances. Moreover, Bayesian approaches have a very narrow scope for parameter estimation, which makes NBA computationally effective. These reasons imply that NBA is an interesting choice in comparative studies, although it is precisely due to some of these advantages that it cannot be expected to consistently outperform other classifiers (Wu et al. 2008). On the other hand, since NBA assumes the explanatory variables to be independent, its application can be problematic when these are heavily correlated (Hastie et al. 2009). This could justify the good performance of NBA in this case study. However, the literature shows that Bayesian approaches should be handled with care whenever used in the characterization of groundwater potential, precisely because groundwater occurrence is sometimes constrained by explanatory variables that resemble each other to a large extent (Naghibi et al. 2017).

In some ways, LDA can be interpreted as an evolution of LRG. LRG is a generally simple—but powerful—algorithm that has been tested extensively (Ayalew and Yamagishi 2005; Chen et al. 2018). However, it suffers from stability issues in cases where targets are clearly separated or when the number of training examples is small. In this context, LDA is generally better

suited to multiclass problems. Much like NBA, LDA relies on Bayes principles. LDA assumes that data is Gaussian and that each explanatory variable has the same variance. In the case at hand, both methods present largely similar results. A better performance on the part of LDA can possibly be attributed to greater affinity with NBA.

Random forest's (RFC) solid scores across the board suggests that this approach is largely insensitive to the noise derived from explanatory variables, even when these are many (Breiman 2001). This is because random forests are calculated by averaging out the results of a large number of unbiased decision trees. Cracknell and Reading (2014) contend that RFC algorithms are easy to train, largely stable, computationally efficient and more accurate than others when dealing with spatially dispersed training data, which makes them an appropriate choice in the case of lithological maps. The findings of the case at hand suggest that this may well be extrapolated to groundwater. However, these also suggest that raw RFC applications tend to overfit training data, which points at the need to use optimization routines whenever this algorithm is involved.

RFC can be seen as an upgraded version of the worst two performers (K-neighbor classification, KNN, and decision tree classifier, CRT). The fact that it renders better results is therefore coherent. However, RFCs suffers from interpretability issues, as it is seldom possible to find meaning to a large number of trees.

SVCs proved computationally expensive, particularly during the optimization stage. Nevertheless, its natural resistance to overfitting makes up for this limitation. This is exemplified by the fact that parameter tuning did not enhance performance significantly. SVC methods have proven efficient in dealing with relatively sparse training datasets and large numbers of explanatory variables (Trustorff et al. 2011), which are both welcome features in groundwater mapping studies.

A key aspect to consider is the meaning of "confidence threshold." A 0.2 threshold tends to overestimate groundwater potential, because this means a lower cutoff value for what is considered a "favorable" setting for groundwater occurrence. In other words, a 0.2 threshold means that an area where at least 20% of the boreholes have historically been positive will be considered favorable for drilling. Similarly, a 0.8 threshold implies that the user defined as high potential areas only those where 80% or more of the boreholes were successful. Thus, a tradeoff can be expected between the threshold and the degree of accuracy of predictions: the higher the threshold, the more difficult to predict becomes the outcome, which in turn results in a lower test score (Figure 6). For the same reason, when picking a lower threshold, the discrepancies between classifiers can be expected to become greater. A different way of looking at this is that lower thresholds naturally present higher test values because the model is trying to predict an "easier" outcome (Figure 4). The threshold approach implies that the results can be presented as a set of groundwater potential maps, rather than as a single one. The choice of a threshold values is important for practical purposes. Thus, the rationale behind using 0.2, 0.4, 0.6, and 0.8 thresholds is to provide a hypothetical decision maker with a choice of options, out of which the most appropriate one could be picked based on management considerations.

A second important factor is the classification process intrinsic to each algorithm. For example, LRG is known to present greater difficulties in cases where the relation between explanatory variables and the outcomes are nonlinear, or wherever explanatory variables interact heavily with each other. Decision tree methods such as RFC and CRT excel in such instances, but are weaker at handling linear relations. In this context, the approach presented in this paper demonstrates how ensemble mapping can provide a means to express discrepancies among classifiers as a measure of uncertainty. This in itself provides an added value to a mere binary classification.

Software Advantages and Limitations

Despite the growing importance of machine learning, the number of tools available to environmental scientists remains limited. QGIS 3.2 can be considered exceptional in this regard, as it incorporates several supervised classification algorithms from the SAGA toolbox (K-neighbors, support vector machine, or boosting classifiers, among others). In this context, MLMapper complements existing functionalities, but can be considered a unique tool that delivers considerable added value to the user. For one, MLMapper incorporates a series of new algorithms based on the SciKit-Learn 0.19.2 toolbox into a single application with a deliberately uncomplicated interface. This provides a wide array of computational alternatives, as well as the possibility of performing ensemble mapping automatically. Furthermore, grid search optimization allows for customized parameter tuning, which enables users to deal with overfitting. Finally, MLMapper brings along abundant graphical output in the form of confusion matrices, receiver operator characteristic curves and standard machine learning metrics, all of which may be used to compare algorithms.

On the other hand, the current version of MLMapper presents two functional limitations. The first one is that MLMapper only accepts point-source information as input. This can be an advantage when training algorithms with items whose spatial location is clearly defined, but is less efficient than polygon-based training in those cases where the information may be diffusely spread in space, or where the item signal may present a wide range of values. A second limitation is that MLMapper only supports binary classification as per its current design. This means that the results of each classifier can be expressed ether in Boolean terms or as an ensemble mean (Figure 5). Multiclass mapping is planned for future release.

Conclusions

Big data approaches are gaining recognition in environmental science and, by extension, in the field of water

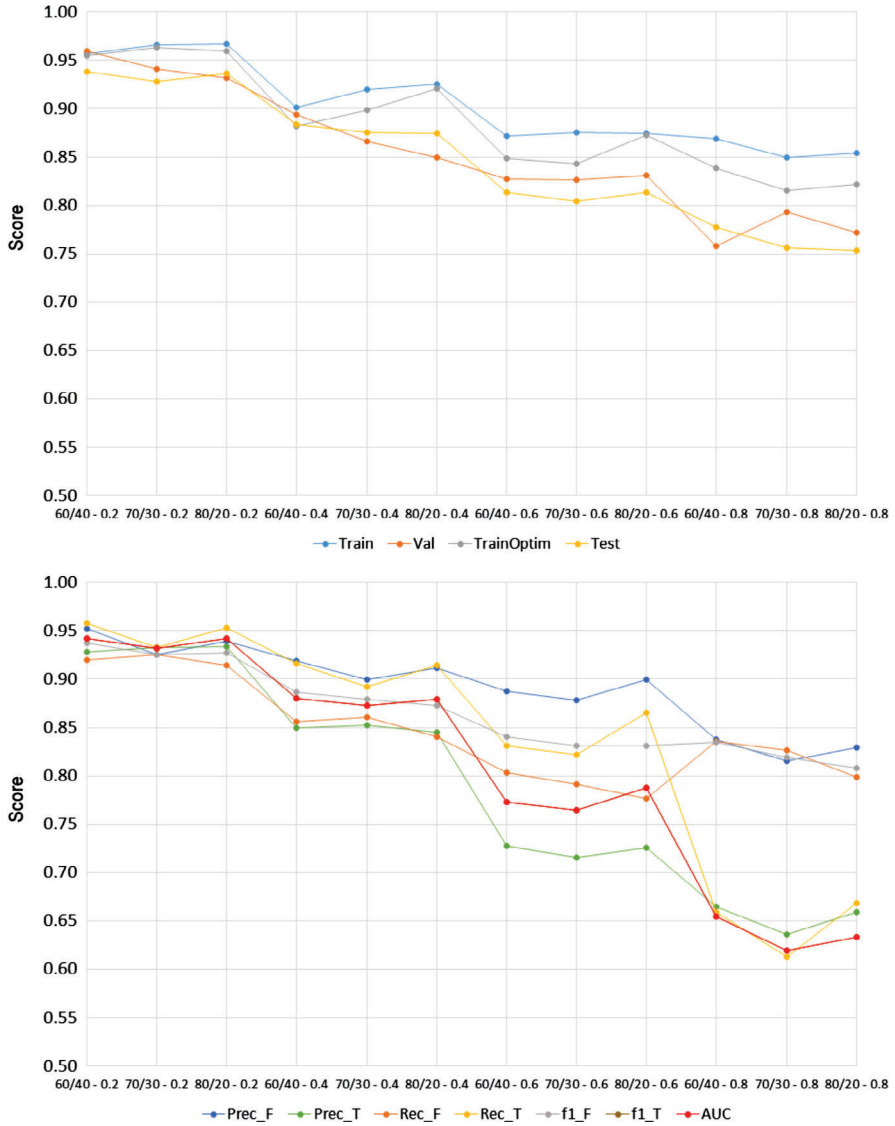

Figure 6. Graphical overview of scoring metrics for the ensemble maps under each split and threshold (Train = training score; Val = validation score; TrainOpt = optimized training score; Test = test score; Prec_F = precision false; Prec_T = precision true; Rec_F = recall false; Rec_T = recall true; f1_F = f−1 score false; f1_T = f−1 score true; AUC = area under curve; TN = true negatives; TP = true positives; FP = false positives; FN = false negatives).

resources. This attests to the potential of machine learning methods to find complex associations in large datasets, as well as to their ability to develop meaningful predictions. Machine learning provides a welcome addition to groundwater mapping in remote regions, particularly in instances where historical databases exist and where carrying out extensive field work is impractical. In particular, machine learning techniques may contribute to narrow down the choice of locations for field investigation, thus underpinning borehole siting efforts.

The ensemble method presented in this paper is versatile enough to be extrapolated to any setting where (a) there is enough ground-truth data and (b) a sufficiently meaningful set of explanatory variables is available. In the case at hand, statistical learners (LRG, NBA, and LDA algorithms) were observed to perform just as well as RFCs and SVCs. In contrast, simple decision trees, multilayer perceptrons, and k-nearest neighbor algorithms rendered the worst results. In any case, anticipating which algorithm will render the most accurate outcomes is typically unfeasible, due to the complexity of big data approaches. This is because each machine learning algorithm presents specific biases based on the way it develops associations between explanatory and target datasets. Hence, ensemble approaches based on the selection of best performing classifiers out of a suitable large sample of algorithms is advocated as the approach of choice to maximize reliability and depict uncertainty.

Acknowledgment

This paper was prepared under research grants 2016/ACDE/1953 and 2018/ACDE/0799 of the Agencia Española de Cooperación Internacional al Desarrollo and

grant number RTI2018-099394-B-I00 of the Ministerio de Ciencia, Innovación y Universidades. The first author received a Salvador de Madariaga grant from Spain's Ministerio de Educación, Cultura y Deporte (PRX18/00235) to carry out a 3-month research stay at the Université de Neuchâtel, Switzerland, where the underlying work was carried out. The authors thank the Direction Générale de l'Hydraulique, Mali, for making borehole data available for this research, and declare no conflicts of interests. A sample of the data used for the development of this research and the open source Python code can be downloaded from https://www.ucm.es/hidrogeologia/programas-software.

Authors' Note

The author(s) does not have any conflicts of interest

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article. Supporting Information is generally *not* peer reviewed.

Table S1 MLMapper's preset grid search value for automated parameter optimization. Descriptions and role of optimization parameters after Pedregosa et al. (2011).
Table S2. Test score and relative ranking of machine learning classifiers for each split and threshold (ranking in brackets). Computational cost is ranked qualitatively, between one (high) and three (low) stars, for unoptimized and optimized predictions.
Table S3. Summary of scoring metrics of the ensemble of the five best performing algorithms for each split and threshold (Train = training score; Val = validation score; TrainOpt = optimized training score; Test = test score; Prec_F = precision false; Prec_T = precision true; Rec_F = recall false; Rec_T = recall true; f1_F = f-1 score false; f1_T = f-1 score true; AUC = area under curve; TN = true negatives; TP = true positives; FP = false positives; FN = false negatives).

References

Ayalew, L., and H. Yamagishi. 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology* 65: 15–31.

Breiman, L. 2001. Random forests. *Machine Learning* 45, no. 1: 5–32.

Chen, W., H. Li, E. Houa, S. Wang, G. Wang, M. Panahi, T. Li, T. Peng, C. Guo, C. Niua, L. Xiao, J. Wang, X. Xie, and B.B. Ahmad. 2018. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Science of the Total Environment* 634: 853–867.

Cracknell, M.J., and A.M. Reading. 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences* 63: 22–33.

D.N.H. 2010. In *Données Hydrogeologiques et des Forages. Direction Nationale de l'Hydraulique*, ed. Ministère de l'Environnement, de l'Eau et de l'Assainissement, 2010. Bamako, Mali: Direction Generale de l'Hydraulique.

Danert, K. 2015. Manual drilling compendium 2015. *Rural Water Supply Network* 2015-2. 1-40.

Dewitte, O., A. Jones, O. Spaargaren, H. Breuning-Madsen, M. Brossard, A. Dampha, J. Deckers, T. Gallali, S. Hallet, R. Jones, M. Kilasara, P. Le Roux, E. Michéli, L. Montanarella, L. Thiombiano, E. Van Ranst, M. Yemefack, and R. Zougmore. 2013. Harmonisation of the soil map of Africa at the continental scale. *Geoderma* 211-212: 138–153.

Díaz-Alcaide, S., and P. Martínez-Santos. 2019a. Review: Advances in groundwater potential mapping. *Hydrogeology Journal*. https://doi.org/10.1007/s10040-019-02001-3

Díaz-Alcaide, S., and P. Martínez-Santos. 2019b. Mapping fecal pollution in rural groundwater supplies by means of artificial intelligence classifiers. *Journal of Hydrology* 577: 124006. https://doi.org/10.1016/j.jhydrol.2019.124006

Fashae, O.A., M.N. Tijani, O.A. Talabi, and O.I. Adedeji. 2014. Delineation of groundwater potential zones in the crystalline basement terrain of SW-Nigeria: An integrated GIS and remote sensing approach. *Applied Water Science* 4: 19–38. https://doi.org/10.1007/s13201-013-0127-9

Foster, T. 2013. Predictors of sustainability for community-managed hand pumps in sub-Saharan Africa: Evidence from Liberia, Sierra Leone, and Uganda. *Environmental Science and Technology* 47: 12037–12046.

Foster, S., A. Tuinhof, and H. Garduño. 2006. Groundwater development in sub-Saharan Africa. A strategic overview of key issues and major needs. *Case Profile Collection* 15: 1–12.

Gupta, R.P. 2018. *Remote Sensing Geology*, 3rd ed. Berlin Heidelberg: Springer.

Harvey, P. 2004. Borehole sustainability in rural Africa: An analysis of routine field data. 30th WEDC International Conference, Vientiane, Laos PDR, 339–346.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, 745. New York: Springer.

Jain, V.K., and S. Kumar. 2018. Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *Journal of Computational Science* 25: 406–415.

Kotsiantis, S.B. 2007. Supervised machine learning: A review of classification techniques. *Informatica* 31, no. 2007: 249–268.

Magesh, N.S., N. Chandrasekar, and J.P. Soundranayagam. 2012. Delineation of groundwater potential zones in Theni district, Tamil Nadu, using remote sensing, GIS and MIF techniques. *Geoscience Frontiers* 3, no. 2: 198–196.

Martín-Loeches, M., J. Reyes-López, J. Ramírez-Hernández, J. Temiño-Vela, and P. Martínez-Santos. 2018. Comparison of RS/GIS analysis with classic mapping approaches for siting low-yield boreholes for hand pumps in crystalline terrains. An application to rural communities of the Caimbambo province, Angola. *Journal of African Earth Sciences* 138, no. 2018: 22–31.

Martre, P., D. Wallach, S. Asseng, F. Ewert, J.W. Jones, R.P. Rötter, K.J. Boote, A.C. Ruane, P.J. Thorburn, D. Cammarano, J.L. Hatfield, C. Rosenzweig, P.K. Aggarwal, C. Angulo, B. Basso, P. Bertuzzi, C. Biernath, N. Brisson, A.J. Challinor, J. Doltra, S. Gayler, R. Goldberg, R.F. Grant, L. Heng, J. Hooker, L.A. Hunt, J. Ingwersen, R.C. Izaurralde, K.C. Kersebaum, C. Müller, S.N. Kumar, C. Nendel, G. O'Leary, J.E. Olesen, T.M. Osborne, T. Palosuo, E. Priesack, D. Ripoche, M.A. Semenov, I. Shcherbak, P. Steduto, C.O. Stöckle, P. Stratonovitch, T. Streck, I. Supit, F. Tao, M. Travasso, K. Waha, J.W. White, and J. Wolf. 2015. Multimodel ensembles of wheat growth:

Many models are better than one. *Global Change Biology* 21: 911–925. https://doi.org/10.1111/gcb.12768

Meijerink, A.M.J. 2007. Remote sensing applications to groundwater, IHP-VI, Series on Groundwater No.16. UNESCO. 311p.

Naghibi, S.A., D.D. Moghaddam, B. Kalantar, B. Pradhan, and O. Kisi. 2017. A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. *Journal of Hydrology* 548: 471–483.

Nampak, H., B. Pradhan, and M.A. Manap. 2014. Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *Journal of Hydrology* 513: 283–300.

Ohmann, J., and M.J. Gregory. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, U.S.A. *Canadian Journal of Forest Research* 32: 725–741.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12: 2825–2830.

Regmi, N.R., and C. Rasmussen. 2018. Predictive mapping of soil-landscape relationships in the arid Southwest United States. *Catena* 165: 473–486.

Sander, P. 2007. Lineaments in groundwater exploration: A review of applications and limitations. *Hydrogeology Journal* 15: 71–74.

Sander, P., T.B. Minor, and M.M. Chesley. 1997. Groundwater exploration based on lineament analysis and reproducibility tests. *Groundwater* 35, no. 5: 888–894.

Schetselaar, E.M., J.R. Harris, T. Lynds, and E.A. De Kemp. 2008. Remote predictive mapping 1. Remote predictive mapping (RPM): A strategy for geological mapping of Canada's north. *Geoscience Canada* 34, no. 3–4: 93–111.

Sorensen, R., U. Zinko, and J. Seibert. 2006. On the calculation of the topographic wetness index: Evaluation of different methods based on field observations. *Hydrology and Earth System Sciences* 10: 101–112.

Tebaldi, C., and R. Knutti. 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences* 365: 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Traore, A.Z., H. Bokar, A. Sidibe, K.Ó. Upton, B.É. Dochartaigh, and I. Bellwood-Howard. 2018. *Africa Groundwater Atlas: Hydrogeology of Mali*. British Geological Survey. http://earthwise.bgs.ac.uk/index.php/Hydrogeology_of_Mali (accessed January 2019).

Trustorff, J.H., P. Konrad, and J. Leker. 2011. Credit risk prediction using support vector machines. *Review of Quantitative Finance and Accounting* 36, no. 4: 565–581.

Von Wehrden, H., H. Zimmermann, J. Hanspach, K. Ronnenberg, and K. Wesche. 2009. Predictive mapping of plant species and communities using GIS and Landsat data in a southern Mongolian Mountain range. *Folia Geobotanica* 44: 211. https://doi.org/10.1007/s12224-009-9042-0

Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14, no. 1: 1–37.

Xie, Y., Z. Sha, and M. Yu. 2008. Remote sensing imagery in vegetation mapping: A review. *Journal of Plant Ecology* 1, no. 1: 9–23.

Xue, J., and B. Su. 2017. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors* 2017: 1–17. https://doi.org/10.1155/2017/1353691