**Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods**

V. F. Rodriguez-Galiano1,2, J. Luque-Espinar3, M. Chica-Olmo4 and M.P. Mendes5,\*

1,2 Physical Geography and Regional Geographic Analysis, University of Seville, Seville 41004, Spain; Geography and Environment, School of Geography, University of Southampton, Southampton, SO17 1BJ, United Kingdom; vrgaliano@us.es

3 Unidad del IGME en Granada, Urbanización Alcazar del Genil, 4, 18006 Granada, Spain; ja.luque@igme.es

4 Departamento de Geodinámica, Universidad de Granada, Avenida Fuentenueva s/n, 18071 Granada, Spain; mchica@us.es

5,\* CERIS, Civil Engineering Research and Innovation for Sustainability, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal; mpaulamendes@tecnico.ulisboa.pt

**Abstract (250 words)**

Recognising the various sources of nitrate pollution and understanding system dynamics are fundamental to tackle groundwater quality problems. A comprehensive GIS database of twenty parameters regarding hydrogeological and hydrological features and driving forces were used as inputs for predictive models of nitrate pollution. Additionally, key variables extracted from remotely sensed Normalised Difference Vegetation Index time-series (NDVI) were included in database to provide indications of agroecosystem dynamics.

Many approaches can be used to evaluate feature importance related to groundwater pollution caused by nitrates. Filters, wrappers and embedded methods are used to rank feature importance according to the probability of occurrence of nitrates above a threshold value in groundwater. Machine learning algorithms (MLA) such as Classification and Regression Trees (CART), Random Forest (RF) and Support Vector Machines (SVM) are used as wrappers considering four different sequential search approaches: the sequential backward selection (SBS), the sequential forward selection (SFS), the sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). Feature importance obtained from RF and CART was used as an embedded approach.

RF with SFFS had the best performance (mmce=0.12 and AUC=0.92) and good interpretability, where three features related to groundwater polluted areas were selected: i) industries and facilities rating according to their production capacity and total nitrogen emissions to water within a 3 km buffer, ii) livestock farms rating by manure production within a 5 km buffer and, iii) cumulated NDVI for the post-maximum month , being used as a proxy of vegetation productivity and crop yield.

1. Introduction

Nitrate in groundwater has been reported as a major problem all over the world. The Nitrates Directive ([91/271/EEC, 1991](#_ENREF_1)) is an integral part of the water policy of the European Union (EU) and it was drawn up with the specific purposes of reducing water pollution caused by nitrates from agricultural sources and preventing further pollution.

Different knowledge-driven and data-driven models can be used to recognise various sources of nitrate pollution and understand system dynamics. Knowledge-driven are models based on expert knowledge of processes that might have led to contamination in a given hydrogeological setting, but where no or very few data sample/pollution evidences are known to occur ([Aller, 1987](#_ENREF_2); [Doerfliger and Zwahlen, 1997](#_ENREF_17); [Ribeiro, 2005](#_ENREF_60)). Data-driven models use objective evidence based on the associations between predictive variables and known occurrences of nitrate pollution ([Solomatine et al., 2008](#_ENREF_66)). Within data-driven models, supervised machine learning algorithms (MLA) are normally applied from a set of training instances where each instance is described by a feature vector or attribute values (input variables) and a target feature expressed as a class label (classification) or a continuous value (regression) ([Kohavi and John, 1998](#_ENREF_39)). In this case, the primary goal of predictive modelling is to maximise the accuracy ([Motoda and Liu, 2002](#_ENREF_48)). Thus, the applicability of MLA on groundwater pollution issues is a consequence of their ability to recognise patterns of relationships among attributes and target feature, considering that there is some degree of uncertainty associated ([Dixon, 2005](#_ENREF_16)). Indeed, MLA have been gradually used to predict nitrate concentration in groundwater, e.g., Random Forest (RF) ([Rodriguez-Galiano et al., 2014](#_ENREF_62); [Tesoriero et al., 2017](#_ENREF_68); [Wheeler et al., 2015](#_ENREF_71)), Support Vector Machines (SVM) ([Dixon, 2005](#_ENREF_16); [Khalil et al., 2005](#_ENREF_38); [Mohamad and Hassan, 2017](#_ENREF_47)), Artificial Neural Networks ([Dixon, 2005](#_ENREF_16); [Khalil et al., 2005](#_ENREF_38); [Mohamad and Hassan, 2017](#_ENREF_47); [Nolan et al., 2015](#_ENREF_51)), Boosted Regression Trees and Bayesian Networks ([Nolan et al., 2015](#_ENREF_51)), and Locally Weighted Projection Regression and Relevance Vector Machines ([Khalil et al., 2005](#_ENREF_38)). Likewise, MLA have been applied to optimise subjective indexes methods for groundwater vulnerability assessment, e.g. ([Fijani et al., 2013](#_ENREF_25)) and ([Nadiri et al., 2017](#_ENREF_49)).

Common to all aforementioned studies is an undeniable fact that for the induction of a MLA, the groundwater experts can use all available features, or select a smaller number of them. Nevertheless, if there is a large number of features, different negative effects might occur, i.e.: i) irrelevant features can result in overfitting training data (i.e. poor generalisation), thus, reducing the model accuracy; ii) models with high complexity may limit their interpretability and, therefore, hamper the decision making process and; iii) models with several features can be impractical and hard to replicate to other areas. To address this issue, it is possible to precede learning with a feature selection stage that strives to eliminate some noise and redundant data, establishing the most significant attributes ([Reunanen, 2006](#_ENREF_59); [Witten and Tibshirani, 2010](#_ENREF_72)).

Feature selection (FS) is a process that selects a subset of original attributes, so that the feature space is optimally reduced according to a certain criterion ([Blum and Langley, 1997](#_ENREF_4); [Dash and Liu, 1997](#_ENREF_14); [Zhang et al., 2006](#_ENREF_74)). The goal of FS is to reduce the amount of features, focusing on the relevant data and improving their quality and hence contribute to a better understanding of the processes (i.e. nitrate pollution of groundwater) that is driven by the selected features ([Guyon and Elisseeff, 2003](#_ENREF_28); [Motoda and Liu, 2002](#_ENREF_48)). Several statistical methods can be employed in FS such as filters, wrapper and embedded methods (Figure 1). The filter approach is a preprocessing step and use criteria not involving any learning machine and, by doing that, it does not consider the effects of a selected feature subset on the performance of the algorithm ([Guyon and Elisseeff, 2006](#_ENREF_29); [Kohavi and John, 1998](#_ENREF_39); [Lal et al., 2006](#_ENREF_41)). Wrapper methods evaluate a subset of features according to accuracy of a given predictor ([Guyon and Elisseeff, 2003](#_ENREF_28); [Kohavi and John, 1998](#_ENREF_39)). Search strategies are used within wrapper methods to yield nested subsets of variables, the variable selection being based on the performance of the learned model ([Guyon and Elisseeff, 2003](#_ENREF_28); [Hilario and Kalousis, 2008](#_ENREF_33)). Embedded methods perform variable selection during the process of training and are generally specific to given learning machines ([Guyon and Elisseeff, 2003](#_ENREF_28)). In this case, the learning step and the feature selection part cannot be separated ([Lal et al., 2006](#_ENREF_41)).



Figure 1-. Conceptual chart of feature selection for predictive modelling of groundwater nitrate pollution.

FS has been used to identify which variables are more relevant to predict nitrate concentration in groundwater, such as wrapper ([Dixon, 2005](#_ENREF_16); [Khalil et al., 2005](#_ENREF_38); [Nolan et al., 2015](#_ENREF_51); [Wheeler et al., 2015](#_ENREF_71)) and embedded methods ([Rodriguez-Galiano et al., 2014](#_ENREF_62); [Tesoriero et al., 2017](#_ENREF_68)). Wrappers or embedded methods include the use of non-parametric algorithms like decision trees, neural networks and support vector machines ([Bazi and Melgani, 2006](#_ENREF_3); [Del Frate et al., 2005](#_ENREF_15); [Pal and Foody, 2010](#_ENREF_53); [Rodriguez-Galiano et al., 2012](#_ENREF_63); [Yu et al., 2002](#_ENREF_73)). Establishing features that are strongly related to nitrate pollution of groundwater can contribute to the establishment of better measures in the Action Programs ([91/271/EEC, 1991](#_ENREF_1)), ensuring an effective reduction of groundwater pollution caused by nitrates and preventing further such pollution. In this study we aim to assess the performance of different FS methods (filters, wrapper and embedded) for defining which features can predict groundwater pollution by nitrates, using the following MLA: CART, Support Vector Machine and Random Forest. Furthermore, we intend to use a comprehensive database, where, as a novelty, new features are extracted from remotely-sensed time series of vegetation indices (weekly composites on an annual basis), allowing to infer the importance of agriculture in the prediction of groundwater nitrate pollution. The objectives of this study were: i) Evaluation of the usefulness of different FS approaches; ii) Recognition of the principal sources of nitrate contamination and understanding system dynamics and, iii) mapping of classifying probabilities of nitrate occurrence in groundwater above a threshold value.

1. Methods and materials
   1. Filters

Filtering is a preprocessing step prior to classification and it is therefore independent of the choice of prediction method, i.e., no learning algorithm is performed ([Guyon and Elisseeff, 2003](#_ENREF_28)). Many different mathematical expressions have been proposed to evaluate feature importance such as correlation based algorithms, gain ratio, or information gain ([Quinlan, 1993](#_ENREF_57)), among others.

Correlation based feature selection greedyalgorithm (CFS) finds attribute subsets by considering the individual predictive ability of each feature along with the degree of redundancy between them. Good feature subsets contain features which are highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other ([Hall and Smith, 1997](#_ENREF_30)). Thus, subsets of features, which are highly correlated with the class (in our case, nitrate concentrations above 50 mg/l) but with low intercorrelation, are preferred. Given a number of k features and c classes, CFS defined the relevance of features subset by using Pearson’s correlation equation ([Ghiselli, 1964](#_ENREF_27)):

(1)

Where MeritS is the relevance of feature subset S containing k features, is the mean feature class correlation and is the average feature-feature intercorrelation ([Karthikeyan and Thangaraju, 2015](#_ENREF_37)). The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator, of how much redundancy there is among them. For estimating the feature-class correlation and feature-feature inter-correlations in equation 1, all features must be treated in a uniform manner and, discretised by using information theoretic binning ([Fayyad, 1993](#_ENREF_23)).

The information gain ranker, evaluates the worth of an attribute by measuring the information gain with respect to the class ([Fürnkranz, 2010](#_ENREF_26)):

(2)

Where is a measure of the uncertainty or unpredictability in a system, is one of the tests on feature which partitions the set into non-overlapping disjoint subsets , and Impurity can be any impurity measure.

However, information gain is biased in favour of features with more values. To counter this, one can use the gain ratio. The gain ratio ranker evaluates the worth of a feature by measuring the gain ratio with respect to the class. For that evaluation, this filter normalises the gained entropy with the entropy :

(3)

* 1. Machine learning algorithms and feature selection

### Wrappers

Wrapper algorithms select a subset of relevant features based on a performance measurement of a learning method. One can schematise the wrapper methodology in three steps: the definition of the performance measure that serves as feature selection criterion and the resampling strategy for validation; the setting of the search strategy for the establishment of the order in which the variable subsets are evaluated, and, the learning method adopted. The predictive performance measurement of a classification-learning model will establish the subset of relevant features ([Guyon and Elisseeff, 2003](#_ENREF_28)). Moreover, a bootstrap routine can be incorporated to the wrapper or embedded models, to evaluate the generalisation of the prediction model.

Different searching strategies can be used, e.g., exhaustive search, genetic algorithms, random search and deterministic forward and/or backward search, among others. This latter method was the one selected for this study due to a better trade-off between performance and computation cost ([Guyon and Elisseeff, 2003](#_ENREF_28)). The sequential search can be executed in four different ways: the sequential backward selection (SBS), the sequential forward selection (SFS), the sequential forward floating selection (SFFS) and the sequential backward floating selection (SBFS). A summarised description of these search strategies is provided below. SBS starts with all the candidate features, and the initial performance of learned model is computed. Then, progressively, the features of less importance for the prediction accuracy are excluded until the MLA results are too poor or, until a prespecified number of variables are left. The sequential forward selection (SFS) is similar to SBS. The difference lies in that, in this case, it starts with an empty set and proceeds by adding features. Gradually, the algorithm adds features to the set until no improvement of the MLA results is observed anymore or until a pre-specified number of variables is reached ([Reunanen, 2006](#_ENREF_59)). [Pudil et al. (1994](#_ENREF_56) presented the concept of floating search methods. The SFFS starts with an empty set and the first step is identical to SFS, the difference is that when a subset is defined by SFS, a SBS is performed as long as the obtained variable set is the best one of its size found so far. When this is no longer the case, the SFS begins again. The SBFS works similar to SFFS but in inverse order, and so, it starts with all possible candidates and a SBS is initially executed.

The .632+ bootstrap method ([Efron and Tibshirani, 1997](#_ENREF_20)) was used to estimate the mean misclassification error (mmce) of the wrapper methods. This method uses the test folders to assess the mmce, and hence the feature importance.

### Classification trees and Random Forest for classification

A decision tree represents a set of constraints or conditions that are organised hierarchically, and are successively applied from the root to terminal node or leaf ([Breiman, 2001](#_ENREF_6); [Quinlan, 1993](#_ENREF_57)). A classification (CART) tree grows as follows ([Hastie et al., 2009a](#_ENREF_31)): given a training set of input-output pairs for , with ( is the number of features or predictors), the algorithm needs to split the predictor space into a number of regions based on a criterion such that, the categorical response variable is constant and well characterised in each region. In a node representing a region with observations and the proportion of class observations in node (I is an indicator function returning 1 if its argument is true and 0 otherwise). We classify the observations in node to class, the majority class in node . If we adopt the Gini Index as a criterion, the splitting criterion is based on the lowest Gini impurity index:

. (4)

Random forests ([Breiman, 2001](#_ENREF_6)) is a substantial modification of bagging that builds a large collection of de-correlated trees, and combine them using majority voting. Bagging is used for training data creation by resampling randomly the original dataset with replacement, i.e., with no deletion of the data selected from the input sample for generating the next subset {h(x,Θk), k = 1, …, K}, where {Θk} are independent random vectors with the same distribution. Hence, some data may be used more than once in the training of trees, while others might never be used. When the RF makes a tree grow, it uses the best feature/split point within a subset of evidential features which has been selected randomly from the overall set of input evidential features. The random forest for classification obtains a class vote from each tree, and then classifies using majority vote ([Hastie et al., 2009b](#_ENREF_32)). In this work, we used RF as both an embedded method and a wrapper. Embedded RF uses a cross-validation process to construct a feature importance measure, to evaluate the prediction strength of each feature, based on the decrease in Gini index ([Breiman et al., 1984](#_ENREF_7)). Although the out of bag (oob) samples can be used to evaluate performance, we used the b632+ bootstrapping to compute the misclassification rate to obtain results that can be compared to those of other methods.

### Support Vector Machine (SVM)

SVM produces a model that can be applied to nonlinear problems using kernel functions. SVM aims at learning “good” separating N-dimensional hyper-planes in a high dimensional space ([Cristianini and Shawe-Taylor, 2000](#_ENREF_13)), being the optimal line based only on a training set of N input-output pairs , called support vectors, in a black box modelling approach ([Lauer and Bloch, 2008](#_ENREF_42)). Given training vectors (where represents the number of features), they are associated with vector labels such that ; let be the function that maps the input vectors into a very high dimensional feature space ([Jankowski and Grabczewski, 2006](#_ENREF_36)). The. SVM solves a quadratic optimisation problem:

(5)

with the constrains where defines a threshold and is the number of training samples, represents a weight vector, C is a regularisation constant that controls the balance between training accuracy and the margin width and, are slack variables. For any testing instance , the decision function is . We need the kernel function, to train the SVM ([Chen and Lin, 2006](#_ENREF_9)), and we used the RBF kernel function:

(6)

* 1. Induction of MLA models and accuracy assessment

Data processing for the induction of the MLA consisted in three main stages: (i) training and parameterisation of the algorithms; (ii) accuracy assessment and; (iii) post-processing requiring converting the output values to a map.

All of the MLA models were created using the R studio 1.0.136 version free software. Within this environment, “mlr” library was used for inducting the embedded and wrapper FS models. Filters were computed using the Weka 3.8 version free software. With the aim of obtaining robust and generalisable models, all possible embedded and wrapper methods were assessed for different hyper-parameter combinations. CART were built considering tree depths from 2 to 29, with a minimum number of observations per node between 1 and 50. The range of the number of trees for RF induction was set to 100, 200, 300, 400, 500, 1,000, 2,500 and 5,000, and the number of split evidential features, between 1 and 20, at 1 intervals. For the building of SVM we used a Radial Basis kernel function with the cost fixed between 0.1 and 2, at 0.1 intervals; and gamma between 0.05 and 1, at 0.05 intervals.

To assess the optimal value of the different parameters of every method, the predictions derived from all possible parameter combinations were evaluated using the Mean Square Error (MSE) using a 10-fold cross validation procedure. The “best” model was the one with the lowest MSE. The methodology followed in the selection of optimal parameters of each method was based on a manual search for them, since one of the goals of this study is to show variation in the mapping accuracy of results according to the parameter selection. Commonly, the percentage of instances that are correctly classified (respectively incorrectly classified) or a complementary measurement such as the misclassification error (mmce) has been used as a measure of the quality of classifiers ([Ferri et al., 2002](#_ENREF_24)).

The best-fit models resulting from the application of each of the methods were compared in terms of ROC curves (Receiver Operating Characteristic). The ROC is usually performed for assessing the tradeoff between true-positive rate (TPR) and false-positive rate (FPR) ([Hastie et al., 2009a](#_ENREF_31)). Generally, the FPR result is plotted on the x-axis vs. TPR on the y-axis. Each threshold result in a (TPR, FPR) pair and a series of such pairs are used to plot the ROC curve. These are also known as the “sensitivity (TPR)” and “specificity (1- FPR)” ([Rodriguez-Galiano et al., 2014](#_ENREF_62)). The sensitivity is the probability of predicting nitrate pollution given true state is polluted. The specificity is the probability of predicting non-nitrates polluted given true state is non-polluted ([Hastie et al., 2009a](#_ENREF_31)). The area under the ROC curve statistic (AUC) was used as a measure of a classifier's performance ([Bradley, 1997](#_ENREF_5)) for random forest, support vector machine and CART wrappers. An AUC value of 1 is considered perfect and AUC value equal to 0.5 is considered as random guessing ([Bradley, 1997](#_ENREF_5)).

Moreover, to identify the optimal value of the different parameters of every method, the predictions derived from all possible parameter combinations were evaluated using the mmce, since it counts the number of times that a sample is badly classified. If no substantial differences in the accuracy of the methods exist, the comparison among algorithms should be based on other factors such as operational capacity, ease of use or the interpretability of results.

* 1. The Vega de Granada aquifer

The Vega de Granada (VG) aquifer is located in the South of Spain, in the region of Andalusia (Figure 2), in the environmental region of the Mediterranean south ([Metzger et al., 2005](#_ENREF_45)). This Quaternary basin-fill aquifer has an approximate extension of 200 km2 (22 km × 8 km) with thicknesses varying between 50 and 300 m, and renewable water resources of 160 hm3/year ([Castillo, 2005](#_ENREF_8)). Towards the west the thickness of the aquifer decreases considerably leading to an important groundwater mean discharge of about 190 hm3/ year into the River Genil ([Kohfahl et al., 2008](#_ENREF_40)). The study area is considered to be semi-arid, with long dry summers (May–September) and wet winters (October–April). The groundwater levels are lower between August and November and closer to the surface between March and May ([Castillo, 2005](#_ENREF_8)). The annual mean rainfall over the aquifer amounts to 450 mm, although it can reach 1,000 mm above some points of the drainage basin (such as those on the Sierra Nevada range), giving an average of around 600 mm/year ([Luque-Espinar et al., 2008](#_ENREF_43)).

The area registered high nitrate contents in groundwater ([Castillo, 2005](#_ENREF_8)) as result of decades of fertiliser application. Consequently, the aquifer was classified as Nitrate Vulnerable Zone by the Spanish authorities by implementing the Nitrates Directive ([Comission, 2013](#_ENREF_11)). The surface limits coincide with an area of irrigated agriculture representing most of land use (49.2%) ([CLC, 2012](#_ENREF_10)), with an estimated groundwater use of 21.35 hm3 ([Confederación Hidrográfica del Guadalquivir, 2015](#_ENREF_12)). Other sources of nitrate can be related to high population density and to industrial activities ([Pardo-Igúzquiza et al., 2015](#_ENREF_54)). The livestock industry is also important in this area.

The mean groundwater flow direction is from east to west, with the steepest gradients in the northeast and eastern sectors. The main component of recharge is precipitation ([Luque-Espinar et al., 2008](#_ENREF_43)), though contributions are also received by seepage from the main rivers Genil, Dilar and Cubillas ([Kohfahl et al., 2008](#_ENREF_40)).



Figure 2- A) Geographical setting of the study area; B) Overall population of the Vega de Granada- adapted from [IECA (2015](#_ENREF_35); C) groundwater sampling points and nitrate concentrations.

* 1. Database design

A comprehensive GIS database of twenty parameters related to hydrogeological and hydrological features, driving forces (sectors of activities that may produce a series of pressures, either as point and non-point sources) and remotely sensed variables (Normalized Difference Vegetation Index data—NDVI data) were used as inputs for a predictive model of nitrate pollution (Figures 3 and 4, Table 1). These explanatory variables, measured in 110 wells, were used to build a predictive model of nitrate occurrence above 50 mg/l (as NO3−) in groundwater. Sampling campaigns took place during November 2016, in the wet season and after the harvest of the summer crops. The descriptive statistical measures of nitrates were: maximum of 547.3 mg/l, minimum of 1.3 mg/l, lower quartile of 44.9 mg/l and higher quartile of 110.8 mg/l, mean and median of 91.7 and 80.4 mg/l, respectively. Around one quarter (26%) of groundwater samples presented nitrate concentrations lower than the quality standards of 50 mg/l ([Comission, 2013](#_ENREF_11)). The nitrate content was binarised according to the cut-off value of 50 mg/l for being used as the response variable.



Figure 3 – Raster layers of intrinsic proprieties of the Vega de Granada aquifer: module of hydraulic gradient, transmissivity, vadose zone thickness, surface flow direction, drop surface and groundwater table elevation.



Figure 4 –Raster layers of the remotely sensed time series of NDVI (Normalised Difference Vegetation Index): maximum level of photosynthetic activity in the canopy (NDVImax), time of maximum photosynthesis in the canopy (NDVItime) and cumulated NDVI for the post-maximum month (NDVIpostmax) and; potential sources of nitrate pollution: overall population and population, land cover classified, distance from irrigation canals, distance from cemeteries, kernel densities of manure production rates for three search radius distances (1, 3 and 5 km) and, kernel densities of industries and facilities rating according to their production capacity and total nitrogen emissions to water for three search radius (1, 3 and 5 km).

The first step was to obtain continuous and standardised variables for the entire study area by applying different approaches to transform all data into a raster format at a resolution of 250 meters. The kernel density ([Silverman, 1986](#_ENREF_65)) or Euclidean distances were used for the rasterisation of features related to the potential point sources of nitrate pollution. In the case of kernel density, a weighted mean centre of these point sources can be used (Figures 3 and 4). For instance, industries (e.g. manufacture of fertilisers and nitrogen compounds, preparation of dairy products, brewing, processing and preservation of meat) and facilities (e.g. wastewater collection and treatment and collection of non-hazardous waste) were rating according to their production capacity and total nitrogen emissions to water in 2015 ([Ministerio de Agricultura y Pesca, 2017](#_ENREF_46)). The extent of nitrate leaching is strongly influenced by dynamic factors such as various land use and management practices ([Hooda et al., 2000](#_ENREF_34); [Rebolledo et al., 2016](#_ENREF_58)). Across the EU, there are evident positive relationships between regional livestock densities and nitrate concentrations in groundwater ([Velthof et al., 2009](#_ENREF_69)). The manure production rating was determined by the amount and type of livestock in 2016 ([Eurostat, 2013](#_ENREF_21)), considering the excretion coefficients used in Spain ([NIR, 2011](#_ENREF_50)). Three search radius distances were used - 1,000, 3,000 and 5,000 meters - being created six raster layers of these two features: industries and facilities (Ind&Fac1, Ind&Fac3 and Ind&Fac5) and manure production (LStock1, LStock3 and LStock5). Raster layers of irrigation canals and cemeteries (DCm) were calculated by Euclidian distances. Only the distance to irrigation canals with water quality problems (IrrC), as a result of discharge of effluents, was taken into consideration in the IrrC estimation ([Luque-Espinar et al., 2015](#_ENREF_44)).

Concerning the non-point sources, the land-use categories (legend level III of Corine Land Cover 2012 ([CLC, 2012](#_ENREF_10)) were reclassified according to their potential impact on nitrate pollution (LC) ([Ribeiro et al., 2017](#_ENREF_61)). For example, permanently irrigated lands were rated 90 and account for most of land use (49.2%). Other uses, such as permanent crops, were rated 70 (representing 16.6%), pastures and agro-forested areas 50 (14%) and forests 0 (1.3%). A raster of overall population (Ovpop) based on the census of January 2014 ([IECA, 2015](#_ENREF_35)) was used to evaluate the possible indirect effects of the population (e.g. possible contributions of damage septic tanks and leaky sewers; ([Nolan et al., 2002](#_ENREF_52)); [Sorichetta et al. (2012](#_ENREF_67)). Moreover, distance from cities was calculated by inverse distance weight (PopD).

Assessing hydrogeological and hydrological features related to nitrogen loss from the soil system was also considered. A raster of surface water flow direction (SWd) was created to differentiate potential zones of nitrogen runoff of agricultural fields. Eight surface water flow directions were established, where most directions are to west (28.6%), northwest (23.0%) and north (21.2%) towards the River Genil. Additionally, a drop raster (SWdrop) was created mapping the percent rise in the path of steepest descent from each cell.

The groundwater table depth (GWt) and Vadose Zone thickness (VZt) indicate if the contaminant leaching to saturated zone occurs rapidly (the deeper the water table level, the lesser the change for contamination occurrence, since, in the unsaturated zone, physical and chemical processes occur that can affect the volume and rate of movement of potential contaminants). The range of transmissivity values is between 14,505 m2/day and 63 m2/day, where the higher values are located in the eastern and western areas. The transmissivity raster (T) of this unconfined aquifer was based on 46 pumping tests provided by FAO and the “Instituto Geológico y Minero de España” ([FAO-IGME, 1972](#_ENREF_22)) for several years. The module of the hydraulic gradient (Grd) has a range of between 0.02% and 3% and defines the horizontal direction of groundwater flow. For all these hydrogeological features, a geostatistical approach was used for their interpolation ([Rodriguez-Galiano et al., 2014](#_ENREF_62)) (Figure 3).

Key variables were extracted from smoothed time-series NDVI data to provide information of agroecosystem dynamics. These NDVI features were extracted from the 2016 annual time series, formed by weekly composite images of 250 meters pixel size. These composite images were generated following the methodology proposed by [Vuolo et al. (2012](#_ENREF_70) for the global MODIS Level-3 16-day VI products available from both MODIS Terra (MOD13Q1) and Aqua (MYD13Q1) satellites. Spanning one growing season, maximum level of photosynthetic activity in the canopy (NDVImax), time of maximum photosynthesis in the canopy (NDVItime) and cumulated NDVI for the post-maximum month (NDVIpostmax) were used to indirectly contemplate nitrogen loss from crop removal, and/or nitrogen leaching to groundwater due to nitrogen fertiliser and irrigation management practices. NDVImax is associated with the type of vegetation, its vigour and density, being the highest values located in agro-forestry areas and the lowest values mainly situated in artificial areas (industrial and continuous urban areas). NDVItime is dependent on the type of vegetation and most highest values were located in the northwest and southeast borders of the VG aquifer (September and October; Figure 4). NDVIpostmax is used as a proxy of vegetation productivity and crop yield being the highest values located in agro-forested areas followed by agricultural irrigated areas. The first two NDVI features can indirectly establish amounts of fertilisers since they reflect the different nitrogen crops requirements. The NDVIpostmax can appraise the N removed from crops and potential quantity of field residues.

Table 1- Abbreviations of the features of the database and description.

|  |  |
| --- | --- |
| **Abbreviated form** | **Short Description** |
| **Hydrogeological and hydrological features** | |
| SWd  SWdrop | Surface water flow direction  Drop raster |
| GWt  VZt  T  Grd | Groundwater table depth  Vadose zone thickness  Transmissivity  Module of hydraulic gradient |
| **Remotely sensed variables** | |
| NDVImax  NDVItime  NDVIpostmax | Maximum level of photosynthetic activity in the canopy  Time of maximum photosynthesis in the canopy  Cumulated NDVI for the post-maximum month |
| **Driving forces** | |
| Ovpop  PopD | Overall population based on the census as of January 2014  Distance from cities |
| LC | Land cover reclassified according to its potential impact on nitrate pollution |
| IrrC | Distance to irrigation canals with water quality problems |
| DCm | Distance to cemeteries |
| Ind&Fac1 Ind&Fac3 Ind&Fac5 | Density of industries and facilities extended to a radius of 1 km  Density of industries and facilities extended to a radius of 3 km  Density of industries and facilities extended to a radius of 5 km |
| LStock1  LStock3  LStock5 | Livestock density within 1 km radius from the livestock farms  Livestock density within 3 km radius from the livestock farms  Livestock density within 5 km radius from the livestock farms |

1. Results and Discussion

Filters estimate the importance of features by using heuristics based on general characteristics of the data. CFS Greedy ranked the features according to average merit, where the samples related with non-urban areas (Ovpop, PopD and DCm), presence of irrigated crops (NDVImax and LC), distance from irrigation canals (IrrC) and surface water flow direction (SWd), were linear correlated with nitrate concentrations above 50 mg/l. The average merit significantly decreased in the following features. Although in a different order, Gain ratio and Information Gain rankers have selected as first five variables the same as those selected by the aforementioned filter (Figure 5). These two last rankers have attributed 13 features with low average merit when compared with the first five. However, the five ranking features were the same; these three rankers are based in different measures: the CFS greedy ranker considers the linear relationship between features and nitrate concentrations above 50 mg/l (target variable) and, the Gain Ratio and Information Gain rankers focus in class separability (i.e. nitrate contents in groundwater exceeding 50 mg/l). From a practical perspective, all these filters are easy to use with low computational cost, but do not necessarily optimise the predictive capacity of a given learner. Considering our results, the Gain Ratio ranker seems to be a possible good choice, since non-linear correlation might be found between features and target variable, and information based theory rankers merit fewer variables.

 Figure 5– Features ranked according to filters type: CFS Greedy, Correlation, Gain Ratio and Info Gain rankers. The average attribute selection of these filters is plotted. The names of predictors use the following notation: Density of industries and facilities extended to a radius of 1 km (Ind&Fac1), to a 3 km buffer (Ind&Fac3), and to a 5 km buffer (Ind&Fac5). Livestock density within 1 km radius from the livestock farms (LStock1), within a 3 km radius (LStock3) and, within a 5 km radius (LStock5). Distance to irrigation canals with water quality problems (IrrC). Distance from cemeteries (DCm). Land cover reclassified according to their potential impact on nitrate pollution (LC). Overall population based on the census as of January 2014 (Ovpop). Distance from cities (PopD). Surface water flow direction (SWd); Drop raster (SWdrop). Groundwater table depth (GWt); vadose zone thickness (VZt); transmissivity (T); Module of hydraulic gradient (Grd). Maximum level of photosynthetic activity in the canopy (NDVImax), time of maximum photosynthesis in the canopy (NDVItime) and cumulated NDVI for the post-maximum month (NDVIpostmax).

The CART are simple, easy to interpret, and can be graphically represented, as illustrated by Figure 6A. This figure shows that the wells located in unpopulated areas (PopD<10,248 inhabitants) within a radius distance lesser than 1.211m of the livestock farms (LStock1) are more likely to have groundwater polluted by nitrates. Additionally, higher values of NDVIpostmax are also indicative of polluted groundwater. On the other hand, lower values of NDVImax (>0.295) and flat populated areas (SWdrop<0.165) are more likely to be non-polluted. Nonetheless, the spatial representation of tree results revealed that most of the area was designated as having a high probability of nitrate contents (>75%), to exceed the 50 mg/l in groundwater (Figure 6B).

A)



B)

|  |
| --- |
|  |
|  |

Figure 6 – A) Embedded CART. Each feature is accompanied by the respective threshold value; B) Map output.

In the case of the embedded RF, the model with a better trade-off between number of features and mmce was chosen as the basis for estimating the likely of groundwater being polluted by nitrates (Figure 7). Only four variables (PopD, NDVImax, DCm and LStock5) were identified as the most important to determine the areas of the VG aquifer being polluted with a mmce equal to 0.138. The VG unpopulated areas (i.e. measured by PopD and DCm), covered by agro-forestry areas (NDVImax) and within the radius of 5 km from livestock farms were chosen.



Figure 7– Random Forest embedded: Relative importance of each independent variable in predicting groundwater polluted by nitrates. Different models derived from the feature selection approach are represented in each column. The figures over each column represent the coefficient determination of each model. The names of predictors use the following notation: Density of industries and facilities extended to a radius of 1 km (Ind&Fac1), to a 3 km buffer (Ind&Fac3), and to a 5 km buffer (Ind&Fac5). Livestock density within 1 km radius from the livestock farms (LStock1), within a 3 km radius (LStock3) and, within a 5 km radius (LStock5). Distance to irrigation canals with water quality problems (IrrC). Land cover reclassified according to their potential impact on nitrate pollution (LC). Overall population based on the census as of January 2014 (Ovpop). Distance from cities (PopD). Surface water flow direction (SWd); Drop raster (SWdrop). Groundwater table depth (GWt); vadose zone thickness (VZt); transmissivity (T); Module of hydraulic gradient (Grd). Maximum level of photosynthetic activity in the canopy (NDVImax), time of maximum photosynthesis in the canopy (NDVItime) and cumulated NDVI for the post-maximum month (NDVIpostmax).

The evaluation of the wrapper algorithms is based on the performance of a learned method, where the establishment of the order in which the variable subsets are evaluated depends on the search strategy. CART, RF and SVM were the learning algorithms within the wrappers, four different sequential searches performed: SBS, SFS, SFFS and SBFS (Table 2).

Table 2- – Summarised results of feature selection using wrappers. MLA: Machine Learning Algorithm; CART: Cart trees; RF: Random Forest; SVM: Support Vector Machine. Sequential Forward Selection (SFS); Sequential Forward Floating Selection (SFFS); Sequential Backward Selection (SBS); Sequential Backward Floating Selection (SBFS). The names of predictors use the following notation: Density of industries and facilities extended to a radius of 1 km (Ind&Fac1), to a 3 km buffer (Ind&Fac3), and to a 5 km buffer (Ind&Fac5). Livestock density within 1 km radius from the livestock farms (LStock1), within a 3 km radius (LStock3) and, within a 5 km radius (LStock5). Distance to irrigation canals with water quality problems (IrrC). Land cover reclassified according to their potential impact on nitrate pollution (LC). Overall population based on the census as of January 2014 (Ovpop). Distance from cities (PopD). Surface water flow direction (SWd); Drop raster (SWdrop). Groundwater table depth (GWt); vadose zone thickness (VZt); transmissivity (T); module of hydraulic gradient (Grd). Maximum level of photosynthetic activity in the canopy (NDVImax), time of maximum photosynthesis in the canopy (NDVItime) and cumulated NDVI for the post-maximum month (NDVIpostmax).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MLA | Sequential Search | mmce | N. of features  selected | Features selected |
| CART | SFS | 0.127 | 2 | Ovpop, PopD |
| SFFS | **0.120** | 2 | NDVIpostmax,T |
| SBS | 0.151 | 19 | IrrC, DCm, SWd, SWdrop, LC, Ind&Fac1, Ind&Fac3, Ind&Fac5, Lstock1, Lstock3, Lstock5,GWt, PopD, Grd, NDVImax, NDVItime, NDVIpostmax, T, VZt |
| SFBS | 0.131 | 15 | IrrC, DCm, SWd, SWdrop, LC, Ind&Fac1, LStock3, LStock5, PopD, Grd, NDVImax, NDVItime, NDVIpostmax, T, VZt |
| RF | SFS | **0.230** | 5 | LC, Ind&Fac5, LStock5, GWt, NDVImax |
| SFFS | 0.234 | 3 | Ind&Fac3, LStock5, NDVIpostmax |
| SBS | 0.233 | 14 | DCm, SWdrop, LC, Ind&Fac1, Ind&Fac5,LStock1, LStock5, Ovpop, GWt, Grd, NDVImax, NDVItime, NDVIpostmax,T |
| SFBS | 0.246 | 15 | IrrC, Dcm, SWd, LC, Ind&Fac5, LStock1, LStock3,Ovpop, GWt, PopD, Grd, NDVItime, NDVIpostmax, vadose\_zon |
| SVM | SFS | **0.239** | 3 | IrrC,Ind&Fac1,LStock5 |
| SFFS | 0.318 | 3 | IrrC, Ind&Fac1, LStock5 |
| SBS | 0.274 | 7 | IrrC, SWd, LStock5, GWt, PopD, NDVImax,NDVIpostmax |
| SFBS | 0.256 | 10 | IrrC, DCm, SWd, SWdrop, Lstock1, Ovpop, PopD, NDVImax, NDVItime, NDVIpostmax |

As regards the CART wrapper, the SFS search strategy gave the smaller mmce of 0.230, being chosen only three features: IrrC, Ind&Fac1 and LStock5. Only one feature (although for a different buffer, LStock5) is similar to those chosen by the embedded CART. Embedded CART allowed the graphical display of the decision tree, showing the synergies between the selected features and their tipping values, and therefore, providing a better interpretability of the results than that of the wrapper method (Figure 6A).

In RF with SFFS, only three features (Ind&Fac3, LStock5 and NDVIpostmax) were chosen. According to this result, groundwater polluted areas can be related with industries and facilities within a 3 km buffer and higher manure production density (within a 5 km radius from the livestock farms). NDVIpostmax is a proxy of vegetation productivity and crop yield, and may be related with higher use of fertilisers ([EEA, 2015](#_ENREF_19)). It is also interesting to note that the error obtained by RF (SFFS) (mmce= 0.120) was lower than the one obtained by embedded RF (mmce=0.138) and, in this case, only three variables were selected. However, wrapper RF had a higher computational cost when compared to embedded RF.

Regarding SVM wrappers, SFS SVM outperformed the rest (mmce=0.239), being, in this case, only two redundant features related to non-urban areas chosen (Ovpop and PopD).

Corroborating the idea of Guyon & Elisseeff (2003), wrappers built using forward sequential search were computationally more efficient, identifying a smaller feature subset at a lower error rate. The best-performing wrappers for each learner were obtained by RF using SFFS, CART with SFS, and SVM with SFS.

Figure 8 shows the results of a ROC analysis which considers both TPR and FPR according to different likelihood thresholds for being classified as above the quality standards of 50 mg/l. SVM with SFS had the worst performance (AUC=0.72), followed by CART with SFS (AUC=0.82). Relying on three driving forces, RF with SFFS had a remarkable value of AUC, showing that almost all groundwater samples with nitrate concentrations above 50 mg/l were classified well. Even with a model dependant on all features, the embedded RF had a value of mmce (0.135) larger than the one obtained by the previous wrapper (mmce= 0.12; Figure 7. Furthermore, a good agreement is reached between this method and that of Pardo-Igúzquiza et al., 2015, who pointed out the irrigated agriculture and sewage from the City of Granada as nitrate pollution sources of groundwater in the VG aquifer. Using embedded RF trained with binarised nitrates dated from 2003, Rodriguez-Galiano et al., 2014 showed that the best-performing model relied on four variables where only one of the driving forces, the distance from dairy farms, was considered to be important for nitrate prediction. In this previous study is emphasised that the distance from driving forces was Euclidian instead of being a kernel density where excretion coefficients were taken into account.



Figure 8 – ROC curves of the best-performing wrappers: CART (SFS) - CART with sequential forward search; RF (SFFS) - Random Forest with Sequential Forward Floating Selection and; SVM (SFS): Support Vector Machine with Sequential Forward Search.

Moreover, within this earlier study (Rodriguez-Galiano et al., 2014), the NDVI feature was not based on a time series reporting information on the whole crop growing season, but a snapshot of a particular date. The introduction of NDVI time series added more information than just one image, since the NDVImax and NDVItime give information on crop phenology and therefore crop type, and NDVIpostmax is a proxy of vegetation biomass and might be related to crop yield ([Duncan et al., 2015](#_ENREF_18); [Pettorelli et al., 2005](#_ENREF_55); [Sakamoto et al., 2005](#_ENREF_64)).

For the three best-performing wrappers, the likelihood of groundwater being polluted by nitrates was mapped (Figure 9). Most of the VG aquifer (around 88% of the whole area) was defined as having medium to high probabilities of being polluted by nitrates (values between 0.50 and 0.75). The SVM wrapper method defined almost every aquifer within the same range of probabilities. CART also had a high frequency (around 73%) in the upper class of probability (<0.75). RF (SFFS) was the learning model which had a more heterogeneous distribution, since it could better differentiate the upper classes of probabilities, showing 32.5% of the values between 0.5 and 0.75, and, 52% of the values above 0.75. As in 2003 (Rodriguez-Galiano et al., 2014), the area delimited as non-polluted (Figure 9), defined by the quality standard for nitrates, was mainly in the south-east. This spatial distribution obtained ensures that agriculture, livestock and, agro-industries and facilities are the principal sources of nitrates in groundwater. In the central area of the aquifer, nitrate concentration is associated with agricultural practices (NDVImax; Figure 4). The NDVImax and its importance concerning nitrate contents in groundwater found in November 2016, can express most intensively that farmed areas boosted by large amounts of nitrogen fertilisers.

The livestock is other driving force responsible for high levels of nitrates in groundwater. Considering the radius of influence of 5 km, the surface spreading of animal manure, perhaps, is not being managed properly (Figure 3). Close to the urban areas (within a 3 km buffer), the wastewater and/or waste collection of the villages and agro-industries may not be receiving the appropriate treatment, and, therefore, be contributing to groundwater pollution by nitrates.



Figure 9- Probability of nitrate concentration in groundwater ≥50 mg/l for the three best wrapper methods results: RF (SFSS): Random Forest with Sequential Forward Floating Selection, CART (SFS): CART with sequential forward search and, SVM (SFS): Support Vector Machine with Sequential Forward Search.

1. Conclusions

FS methods have been revealed as important approaches for predictive modelling of nitrate pollution. Different approaches can be used for feature selection, such as filters, embedded and wrapper methods, increasing in complexity and functionality, respectively.

Manure nitrogen production density, the density of industries and facilities and cumulated NDVI for the post-maximum month were selected by the FS methods as the most important for reaching good performances. The remotely sensed NDVI time series variables showed to be important features for nitrate pollution prediction in groundwater, especially when almost the entire area of the Vega de Granada aquifer is covered by irrigated crops. NDVImax has proven to be an important feature for establishing intensively farmed areas boosted by large amounts of nitrogen fertilisers.

Within embedded methods (CART and RF), the most important features were identified and the model prediction was optimised by minimising the prediction error; however, the reduction of the number of features to include in the model was only possible by using wrapper methods. In fact, although more computationally demanding, the wrappers could tick three important boxes: i) Selection of the most important features; ii) optimisation of the prediction model and; iii) dimensionality reduction of the feature space. A wrapper composed of a RF learner and a SFFS searching strategy outperformed the rest, showing the best accuracy, a good interpretability and a smoother spatial distribution of probabilities for above 50 mg/l nitrate occurrence (mmce=0.12 and AUC=0.92).

Acknowledgements section

Maria Paula Mendes was funded by FCT-MEC Post doctoral Grant (SFRH/BDP/110346/2015). Data are publicly available from websites referenced in the paper. We are grateful for the financial support given by the Spanish Ministerio de Economía, Industria y Competitividad (Project CGL2017-84739-R).

References

91/271/EEC D. Council Directive of 21.05.1991 concerning urban waste water treatment. Official Journal of the European Communities. 91/271/EEC 1991, pp. 8.

Aller L, Bennett, T., Lehr, J. H., Petty, R.J., and Hackett G. DRASTIC: A standardized system for evaluating ground water pollution potential using hydrogeologic settings. In: NWWA/EPA, editor, 1987.

Bazi Y, Melgani F. Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 2006; 44: 3374-3385.

Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence 1997; 97: 245-271.

Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997; 30: 1145-1159.

Breiman L. Random Forests. Machine Learning 2001; 45: 5-32.

Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Chapman and Hall/CRC, Belmont, CA 1984.

Castillo A. El acuífero de la Vega de Granada. Ayer y hoy (1966-2004). Agua, Minería y Medio Ambiente, Libro Homenaje al Profesor Rafael Fernández Rubio. López Geta et al. , 2005, pp. 161-172

Chen Y-W, Lin C-J. Combining SVMs with Various Feature Selection Strategies. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA, editors. Feature Extraction: Foundations and Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 315-324.

CLC. CORINE Land Cover. Copyright Copernicus Programme, European Environment Agency, 2012.

Comission E. on the implementation of Council Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources based on Member State reports for the period 2008–2011 Report from the Comission to the Council and the European Parliament Brussels, 2013, pp. 11.

Confederación Hidrográfica del Guadalquivir CHd. Plan Hidrológico de la demarcación hidrográfica del Guadalquivir (2015 –2021). Anejo nº 3–Descripción de usos, demandas y presiones, 2015, pp. 373.

Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press, 2000.

Dash M, Liu H. Feature Selection for Classification. Intell. Data Anal. 1997; 1: 131-156.

Del Frate F, Iapaolo M, Casadio S, Godin-Beekmann S, Petitdidier M. Neural networks for the dimensionality reduction of GOME measurement vector in the estimation of ozone profiles. Journal of Quantitative Spectroscopy and Radiative Transfer 2005; 92: 275-291.

Dixon B. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis. Journal of Hydrology 2005; 309: 17-38.

Doerfliger N, Zwahlen F. EPIK: a new method for outlining of protection areas in karstic environment. International symposium and field seminar on “karst waters and environmental impacts. Gunay G and Jonshon AI, , Antalya, Turkey, Balkema, Rotterdam, 1997, pp. 117–123.

Duncan JMA, Dash J, Atkinson PM. Elucidating the impact of temperature variability and extremes on cereal croplands through remote sensing. Global Change Biology 2015; 21: 1541-1551.

EEA. EEA Signals 2015 - Living in a changing climate. EEA, Copenhagen, 2015, pp. 37.

Efron B, Tibshirani R. Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association 1997; 92: 548-560.

Eurostat. Nutrient Budgets –Methodology and Handbook. Eurostat and OECD, Luxembourg, 2013.

FAO-IGME. Proyecto piloto de utilización de aguas subterráneas para el desarrollo agrícola de la cuenca del guadalquivir. Utilización de las aguas subterráneas para la mejora del regadío de la Vega de Granada, 1972.

Fayyad UaI, K. Multi-interval discretization of continuous-valued attributes for classification learning. Proċ10th Inṫ Conf Machine Learning, 1993, pp. 194–201.

Ferri C, Flach P, Hernandez-Orallo J. Learning decision trees using the area under the ROC curve. Proceedings of the 19th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., Sydney, Australia, 2002, pp. 139–146.

Fijani E, Nadiri AA, Asghari Moghaddam A, Tsai FTC, Dixon B. Optimization of DRASTIC method by supervised committee machine artificial intelligence to assess groundwater vulnerability for Maragheh–Bonab plain aquifer, Iran. Journal of Hydrology 2013; 503: 89-100.

Fürnkranz J. Decision Tree. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Springer US, Boston, MA, 2010, pp. 263-267.

Ghiselli EE. Theory of Psychological Measurement: McGraw-Hill Education, 1964.

Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Researc 2003; 3: 1157-1182.

Guyon I, Elisseeff A. An Introduction to Feature Extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA, editors. Feature Extraction: Foundations and Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 1-25.

Hall MA, Smith LA. Feature Subset Selection: A Correlation Based Filter Approach. 1997.

Hastie T, Tibshirani R, Friedman J. Additive Models, Trees, and Related Methods. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer New York, New York, NY, 2009a, pp. 295-336.

Hastie T, Tibshirani R, Friedman J. Random Forests. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer New York, New York, NY, 2009b, pp. 587-604.

Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. Briefings in Bioinformatics 2008; 9: 102-118.

Hooda PS, Edwards AC, Anderson HA, Miller A. A review of water quality concerns in livestock farming areas. Science of The Total Environment 2000; 250: 143-167.

IECA. Distribución espacial de la población de Andalucía. Instituto de Estadística y Cartografía de Andalucía. 2017. Instituto de Estadística y Cartografía de Andalucía (es responsabilidad exclusiva de los autores el grado de exactitud o fiabilidad de la información derivada de ese procesamiento ), 2015.

Jankowski N, Grabczewski K. Learning Machines. In: Guyon I, Gunn, S., Nikravesh, M., Zadeh, L.A. , editor. Feature Extraction: Foundations and Applications. Springer-Verlag Berlin Heidelberg, 2006, pp. 29-64.

Karthikeyan T, Thangaraju P. Best First and Greedy Search Based CFS- Naïve Bayes Classification Algorithms for Hepatitis Diagnosis. Biosci Biotech Res 2015; 12.

Khalil A, Almasri MN, McKee M, Kaluarachchi JJ. Applicability of statistical learning algorithms in groundwater quality modeling. Water Resources Research 2005; 41: n/a-n/a.

Kohavi R, John GH. The wrapper approach. In: Liu H, Motoda H, editors. Feature Extraction, Construction and Selection: A Data Mining Perspective. Springer Verlag, 1998.

Kohfahl C, Sprenger C, Herrera JB, Meyer H, Chacón FF, Pekdeger A. Recharge sources and hydrogeochemical evolution of groundwater in semiarid and karstic environments: A field study in the Granada Basin (Southern Spain). Applied Geochemistry 2008; 23: 846-862.

Lal TN, Chapelle O, Weston J, Elisseeff A. Embedded Methods. In: Guyon I, Gunn S, Nikravesh M, Zadeh LA, editors. Feature Extraction: Foundations and Applications. Springer-Verlag Berlin Heidelberg, 2006, pp. 137-165.

Lauer F, Bloch G. Incorporating prior knowledge in support vector regression. Machine Learning 2008; 70: 89-118.

Luque-Espinar JA, Chica-Olmo M, Pardo-Igúzquiza E, García-Soldado MJ. Influence of climatological cycles on hydraulic heads across a Spanish aquifer. Journal of Hydrology 2008; 354: 33-52.

Luque-Espinar JA, Navas N, Chica-Olmo M, Cantarero-Malagón S, Chica-Rivas L. Seasonal occurrence and distribution of a group of ECs in the water resources of Granada city metropolitan areas (South of Spain): Pollution of raw drinking water. Journal of Hydrology 2015; 531, Part 3: 612-625.

Metzger MJ, Bunce RGH, Jongman RHG, Mücher CA, Watkins JW. A climatic stratification of the environment of Europe. Global Ecology and Biogeography 2005; 14: 549-563.

Ministerio de Agricultura y Pesca AyMA. Registro Estatal de Emisiones y Fuentes Contaminantes. © PRTR España, 2017.

Mohamad S, Hassan R. Statistical Learning Methods for Classification and Prediction of Groundwater Quality Using a Small Data Record. International Journal of Agricultural and Environmental Information Systems (IJAEIS) 2017; 8: 37-53.

Motoda H, Liu H. Feature selection, extraction and construction. Towards the Foundation of Data Mining Workshop, Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD’02), Taipei, Taiwan 2002, pp. 67–72.

Nadiri AA, Gharekhani M, Khatibi R, Sadeghfam S, Moghaddam AA. Groundwater vulnerability indices conditioned by Supervised Intelligence Committee Machine (SICM). Science of The Total Environment 2017; 574: 691-706.

NIR. Inventario de Emisiones de Gases de efecto Invernadero de España e Información adicional años 1990-2009. Comunicación a la Secretaría del Convenio Marco sobre el Cambio Climático y Protocolo de Kioto. Ministerio de Medio Ambiente, y Medio Rural y Marino Secretaría de Estado de Cambio Climático Dirección General de Calidad y Evaluación Ambiental D.G., 2011, pp. 706 pp.

Nolan BT, Fienen MN, Lorenz DL. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. Journal of Hydrology 2015; 531, Part 3: 902-911.

Nolan BT, Hitt KJ, Ruddy BC. Probability of Nitrate Contamination of Recently Recharged Groundwaters in the Conterminous United States. Environmental Science & Technology 2002; 36: 2138-2145.

Pal M, Foody GM. Feature Selection for Classification of Hyperspectral Data by SVM. IEEE Transactions on Geoscience and Remote Sensing 2010; 48: 2297-2307.

Pardo-Igúzquiza E, Chica-Olmo M, Luque-Espinar JA, Rodríguez-Galiano V. Compositional cokriging for mapping the probability risk of groundwater contamination by nitrates. Science of The Total Environment 2015; 532: 162-175.

Pettorelli N, Vik JO, Mysterud A, Gaillard J-M, Tucker CJ, Stenseth NC. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends in Ecology & Evolution 2005; 20: 503-510.

Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters 1994; 15: 1119-1125.

Quinlan JR. C4.5 programs for machine learning. San Mateo, CA: Morgan Kaurmann, 1993.

Rebolledo B, Gil A, Flotats X, Sánchez JÁ. Assessment of groundwater vulnerability to nitrates from agricultural sources using a GIS-compatible logic multicriteria model. Journal of Environmental Management 2016; 171: 70-80.

Reunanen J. Search Strategies. In: Guyon *I, Gunn, S., Nikravesh, M., Zadeh, L.A.*, editor. Feature Extraction: Foundations and Applications. Springer-Verlag Berlin Heidelberg, 2006, pp. 119-136.

Ribeiro L. Desenvolvimento e aplicação de um novo índice de susceptibilidade dos aquíferos à contaminação de origem agrícola. In: APRH, editor. 7º Simpósio de Hidráulica e Recursos Hídricos dos Países de Língua Oficial Portuguesa,, Évora, Portugal, 2005.

Ribeiro L, Pindo JC, Dominguez-Granda L. Assessment of groundwater vulnerability in the Daule aquifer, Ecuador, using the susceptibility index method. Science of The Total Environment 2017; 574: 1674-1683.

Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of The Total Environment 2014; 476–477: 189-206.

Rodriguez-Galiano VF, Chica-Olmo M, Abarca-Hernandez F, Atkinson PM, Jeganathan C. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sensing of Environment 2012; 121: 93-107.

Sakamoto T, Yokozawa M, Toritani H, Shibayama M, Ishitsuka N, Ohno H. A crop phenology detection method using time-series MODIS data. Remote Sensing of Environment 2005; 96: 366-374.

Silverman BW. Density Estimation for Statistics and Data Analysis: Taylor & Francis, 1986.

Solomatine D, See LM, Abrahart RJ. Data-Driven Modelling: Concepts, Approaches and Experiences. In: Abrahart RJ, See LM, Solomatine DP, editors. Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 17-30.

Sorichetta A, Masetti M, Ballabio C, Sterlacchini S. Aquifer nitrate vulnerability assessment using positive and negative weights of evidence methods, Milan, Italy. Computers & Geosciences 2012; 48: 199-210.

Tesoriero AJ, Gronberg JA, Juckem PF, Miller MP, Austin BP. Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resources Research 2017; 53: 7316-7331.

Velthof GL, Oudendag D, Witzke HP, Asman WAH, Klimont Z, Oenema O. Integrated Assessment of Nitrogen Losses from Agriculture in EU-27 using MITERRA-EUROPE All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Journal of Environmental Quality 2009; 38: 402-417.

Vuolo F, Mattiuzzi M, Klisch A, Atzberger C. Data service platform for MODIS Vegetation Indices time series processing at BOKU Vienna: current status and future perspectives. 8538, 2012, pp. 85380A-85380A-10.

Wheeler DC, Nolan BT, Flory AR, DellaValle CT, Ward MH. Modeling groundwater nitrate concentrations in private wells in Iowa. Science of The Total Environment 2015; 536: 481-488.

Witten DM, Tibshirani R. A framework for feature selection in clustering. Journal of the American Statistical Association 2010; 105: 713-726.

Yu S, De Backer S, Scheunders P. Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery. Pattern Recognition Letters 2002; 23: 183-190.

Zhang H, Ho TB, Zhang Y, Lin M-S. Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform. Informatica 2006; 30 305–319.