Journal of Hydrology 603 (2021) 127092

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Predictive modelling benchmark of nitrate Vulnerable Zones at a regional scale based on Machine learning and remote sensing

Aaron Cardenas-Martinez [a,*], Victor Rodriguez-Galiano [a], Juan Antonio Luque-Espinar [b], Maria Paula Mendes [c]

[a] Departamento de Geografía Física y Análisis Geográfico Regional, Universidad de Sevilla, 41004 Seville, Spain
[b] Geological Survey of Spain (IGME), Urb. Alcázar del Genil, 4, 18006 Granada, Spain
[c] Civil Engineering Research and Innovation for Sustainability (CERIS), Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

ARTICLE INFO

Keywords:
Nitrates
Machine learning
Feature selection
Groundwater
Nitrate Vulnerable Zones

ABSTRACT

Nitrate leaching losses from arable lands into groundwater were a main driver in designating Nitrate Vulnerable Zones (NVZs) according to the Nitrates Directive, with a view to enhancing their water quality. Despite this, developing common strategies for effective water quality control in these areas remains a challenge in the European Union. This paper evaluates the performance of the Random Forest (RF) machine learning algorithm combined with Feature Selection (FS) techniques in predicting nitrate pollution in NVZs groundwater bodies in different periods and using updated environmental features in Andalusia, Spain. A set of forty-four features extrinsic to groundwater bodies were used as environmental predictors, with an aim to make this methodology exportable to other regions. Phenological features obtained through remote-sensing techniques were included to measure the dynamics of agricultural activity. In addition, other dynamic features derived from weather and livestock effluents were included to analyse seasonal and interannual changes in nitrate pollution. Three feature stacks and two nitrate databases were used in the predictive modelling: Period 1 (2009), with 321 nitrate samples for training; Period 2 (2010), with 282 nitrate samples for validation and initial spatial prediction; and Period 3 (2017), to assess the changes in the probability of groundwater nitrate content exceeding 50 mg/L. Random Forest as a wrapper with four sequential search methods was considered: sequential backward selection (SBS), sequential forward selection (SFS), sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). From among all the Feature Selection methods applied, Random Forest with SFS had the best performance (overall accuracy = 0.891 and six predictor features) and linked the highest probability of nitrate pollution with three dynamic features: the Normalized Difference Vegetation Index (NDVI) base level, NDVI value for the end of the growing season and accumulated manure production of livestock farms; and three static features: slope, sediment depositional areas and valley depth.

1. Introduction

Protecting strategic resources such as water is currently a major challenge, not only in the field of environmental research, but also at all levels of government. Changes in water resource quality significantly affect society and the economy, especially in the production of basic resources such as food (WHO, 2017). The application of nitrogen fertilisers and pesticides to crops in recent decades, coupled with the diffuse pollution that these generate, is one of the main causes behind the failure to achieve a good chemical status for groundwater in the European Union (EU) (European Environmental Agency, 2018). The surplus of chemical and organic nitrogen fertilisers from agricultural activity can leach to groundwater as nitrate (Babiker et al., 2004; Juntakut et al., 2019) and from the action of other sources, such as livestock farming (Cho et al., 2000; Tullo et al., 2019). High concentrations of nitrate in drinking water can be a serious threat to human health and can cause illnesses with long-term exposure, including methemoglobinemia and thyroid cancer (Fewtrell, 2004; Ward et al., 2010). Groundwater is a major source of drinking water in the EU, where 24.5% of freshwater extracted for drinking purposes was groundwater in 2017 (European Environmental Agency, 2020). To address the problem of water pollution, the Water Framework Directive (WFD) (2000/60/EC) aimed to

* Corresponding author.
E-mail address: acardenasm@us.es (A. Cardenas-Martinez).

https://doi.org/10.1016/j.jhydrol.2021.127092
Received 29 April 2021; Received in revised form 15 October 2021; Accepted 17 October 2021
Available online 23 October 2021

protect and improve water quality in the EU, establishing monitoring controls to evaluate impacts and control long-term trends. Together with the WFD, the Nitrates Directive (ND) (91/676/EEC) had a huge impact on controlling pollution caused by agricultural nitrates in ground and surface waters (Oenema et al., 2011). This Directive required that all Member States designate areas within their territories that drain into waters which could be affected by nitrate pollution or that could be affected if action is not taken to decrease nitrate leaching as Nitrate Vulnerable Zones (NVZs) (Goodchild, 1998; Velthof et al., 2014). The report on the implementation of the Nitrates Directive between 2012 and 2015 revealed that 13.2% of the groundwater stations in the EU recorded nitrate concentrations exceeding the legal limit of 50 mg/L (European Commission, 2018). Approximately 21.5% of measuring stations in Spain registered averages higher than 50 mg/L in the same period (European Commission, 2018). The same situation occurred in Andalusia (the most highly populated NUTS2 region in Spain), where 51 of the 176 groundwater bodies had a "poor chemical status" due to high nitrate concentrations in 2015 according to the Hydrological plans (Confederación Hidrográfica del Guadalquivir, 2015).

Determining nitrate concentration in groundwater enables the identification and delimitation of areas at spatial risk of pollution, a narrowed focus in terms of resource management and compliance with the WFD and ND. Usual methods based on time series forecasting for the evaluation of nitrate concentrations consist of trend analysis (Ducci et al., 2019; Hansen et al., 2012), theoretical gross nitrogen balance (Wick et al., 2012) and numerical modelling of fate and transport of contaminants (Akbariyeh et al., 2018; Esmaeili et al., 2014). These methods are applied at various scales (principally at the farm, regional and national agency levels) and possess inherent uncertainties due to weather variability, soil conditions, manure management, crop cultivation practices and socio-economic circumstances (Kawagoshi et al., 2019; Mendes et al., 2012; Mendes & Ribeiro, 2010; Wick et al., 2012). The selection of these methods is also based on the availability of data (i.e. input parameters), the spatial and temporal representativity of monitoring networks and underlying assumptions.

Recently, hydrology studies on nitrate pollution have benefited from the development of machine-learning algorithms (MLA) that learn from examples and thus do not require pre-established rules based on expert criteria (Buduma & Locascio, 2017). MLAs have been used both alone (Dixon, 2005; Knoll et al., 2020; Nolan et al., 2014; Rodriguez-Galiano et al., 2014; Tesoriero et al., 2017; Wagh et al., 2018) and as part of ensembles (Barzegar et al., 2018; Khosravi et al., 2018; Knoll et al., 2019; Motevalli et al., 2019; Rodriguez-Galiano et al., 2018; Sajedi-Hosseini et al., 2018; Singh et al., 2014; Wheeler et al., 2015) to predict the spatial distribution of nitrate concentrations in groundwater. Generally, these studies have been applied primarily at local and regional scales, although others have been applied at a national scale, such as Knoll et al. (2020) in Germany; and at a continental scale, such as Ouedraogo et al. (2019) in Africa. One of the most widely used MLAs for predicting nitrate concentrations in groundwater is Random Forest (RF), with examples such as Knoll et al. (2020); Nolan et al. (2014); Ouedraogo et al. (2019); Rodriguez-Galiano et al. (2014); and Wheeler et al. (2015). Random Forest has multiple advantages as a non-parametric method that allows different data sets to be handled, and is very effective for non-linear relationships between features and for outlier values (Breiman, 2001a; Biau and Scornet, 2016). In addition, Random Forest allows assessment of the relative importance of predictive features within the prediction, thus providing an understanding of how each predictive feature influences the model in order to select the best features for modelling (Ghimire et al., 2010; Rodriguez-Galiano et al., 2012). The performance of Random Forest has been evaluated with other machine-learning algorithms in several nitrate pollution studies (Band et al., 2020; Knoll et al., 2019; Messier et al., 2019) and it was the MLA with the best predictive performance. The use of Random Forest can be improved by using Feature-Selection (FS) algorithms, which allow a subset of original attributes to be selected and to optimally reduce the feature space according to a specific criterion (Blum & Langley, 1997; Dash & Liu, 1997; Zhang & Bao Ho, 2006). The aim of Feature Selection is to select a subset of relevant features to build robust learning models, thus reducing the number of features used in the prediction (Blum & Langley, 1997; Rodriguez-Galiano et al., 2018; Saeys et al., 2007) and contributing to a better understanding of the processes. Feature Selection allows for increased model accuracy, reducing the effect of the curse of dimensionality (Bellman, 2003), obtaining more generalisable models, accelerating the learning process and increasing their interpretability (Guyon and Elisseeff, 2003). Different Feature Selection statistical methods are available: filters, embedded and wrappers. Filters are usually applied in a pre-processing stage and do not need a machine-learning algorithm for feature selection (Kohavi & John, 1997). Embedded methods perform feature selection during the training process and are generally specific to given learning machines (Guyon and Elisseeff, 2003). Likewise, wrapper-based methods combine a machine-learning algorithm with a feature-search method, selecting the subset of features with the best predictive performance (Guyon and Elisseeff, 2003). The ability to select the most relevant features has allowed the use of Feature Selection in predicting nitrate concentrations in groundwater using either the wrapper method (Dixon, 2005; Khalil et al., 2005; Wheeler et al., 2015), or the embedded method (Rodriguez-Galiano et al., 2014; Tesoriero et al., 2017). With several feature selection methods available, studies such as Rodriguez-Galiano et al. (2018) and Effrosynidis & Arampatzis (2021) assessed the performance of different Feature Selection methods and concluded that wrapper-based feature selection methods had higher predictive performance than other methods.

Although the use of MLA has been a huge step forwards in the study of nitrate pollution, regularly updating these studies to monitor particularly vulnerable groundwater bodies remains challenging. Many studies include intrinsic properties of groundwater bodies, which influence the direction and migration rate of nitrates within groundwater bodies. Some commonly used features include groundwater recharge rate, hydraulic conductivity, transmissivity and depth to groundwater (Motevalli et al., 2019; Ransom et al., 2017; Wheeler et al., 2015). Nevertheless, the hydraulic properties evaluated by pumping tests are usually few and have limited spatial representation, not representing the anisotropic behaviour of the groundwater bodies. The exclusive use of features extrinsic to groundwater bodies for nitrate prediction allows these studies to be easily upgraded and does not require measurements in a specific environment or at great expense. Thus, previous studies such as Wells et al. (2021) used static and dynamic extrinsic features of groundwater bodies to predict vadose and saturated zone transport rates and lag times using Random Forest. Studies based on predicting the spatial distribution of nitrates in groundwater commonly use land cover images as a proxy to measure the impact that agriculture has on groundwater pollution. Rodriguez-Galiano et al. (2014) used images from vegetation indices obtained through remote-sensing techniques for a specific date to measure the importance of agriculture in making predictions. However, the use of a land cover image or a single image from a vegetation index (Normalized Difference Vegetation Index – NDVI) to measure agricultural activity and its impact on nitrate pollution lacks fundamental information, such as agricultural seasonality, interannual crop variability and productivity, due to the fact that it is not representative of the entire growing season. The inclusion of NDVI time series adds more information by including all stages of the growing season. In addition, they can serve as an indicator of vegetation biomass and crop type, and may be related to crop yields (Duncan et al., 2015; Sakamoto et al., 2005).

This study develops a method based on Random Forest and Feature Selection to predict the probability of nitrate concentrations above 50 mg/L in groundwater bodies located in Nitrate Vulnerable Zones in Andalusia, Spain in the present (Period 3-2017), using models trained with past environmental conditions (Period 1-2009) validated at an intermediate time point (Period 2-2010). The main novelties of this

study are based on the ability to export the applied methodological procedures to other geographical areas using only features extrinsic to groundwater bodies available from national and international public agencies. An easily updatable, free data-source and operational method for identifying potential nitrate pollution of groundwater from agricultural activities is proposed that can help to establish Nitrate Vulnerable Zones where nitrate monitoring data and hydrogeological parameters for aquifers are scarce. To determine the impact that agriculture has on nitrate pollution of Nitrate Vulnerable Zones, a set of phenological metrics were included that were derived from the functional analysis of vegetation indices obtained via remote-sensing techniques. Other features used were weather data, terrain features, soil properties and livestock effluents. The main objectives of this study were: (1) to map zones of high nitrate concentration in groundwater bodies at different time periods; (2) to identify the most important features in nitrate pollution of Nitrate Vulnerable Zones; (3) to evaluate the effectiveness of different Feature Selection approaches; and (4) to evaluate the probability of exceeding the 50 mg/L nitrate threshold in Nitrate Vulnerable Zones for different types of groundwater bodies.

2. Materials and methods

2.1. Case study

Andalusia is the southernmost Spanish administrative region (NUTS2) on the Iberian Peninsula. It is located in Southwestern Europe and has a land area of >87,000 km^2. Its terrain consists of various mountainous systems, separated by alluvial depressions. The region is characterised by a Mediterranean climate, with a dry period in the summer that separates two rainfall peaks in spring and autumn. Average annual rainfall ranges between 300 and 2000 mm and the average temperature is 17 °C, with maximum values in July–August and minimum values in January. The climate promotes agricultural use in the alluvial river basins, primarily through intensive agriculture. The main arable land areas are irrigated and dry farming, olive groves, orchards and greenhouses. Besides the Baetic Depression, which is the main region for agricultural production, other production regions of note include Vega de Granada and Campo de Dalías (Fig. 1).

Groundwater in Andalusia comprises approximately 27% of the water used for irrigation of arable crops (INE, 2021). Large agricultural areas in Andalusia are located in the river basins of its main rivers, such as the Guadalquivir and Genil rivers (Fig. 1). Most of the detrital aquifers are located in these river basins and receive the excess nitrogen pollutants from fertilisers. Andalusia is the region in Spain that saw the greatest use of nitrogen fertilisers between 2009 and 2017, with an average of 240.57 thousand tonnes per year. Fertiliser application has been generally trending upwards between 2009 (210.3 thousand tonnes) and 2017 (309.5 thousand tonnes). First, fertiliser inputs were decreasing until 2014, when fertiliser application reached its minimum level (285.6 thousand tons of nitrogen). The trend reversed the

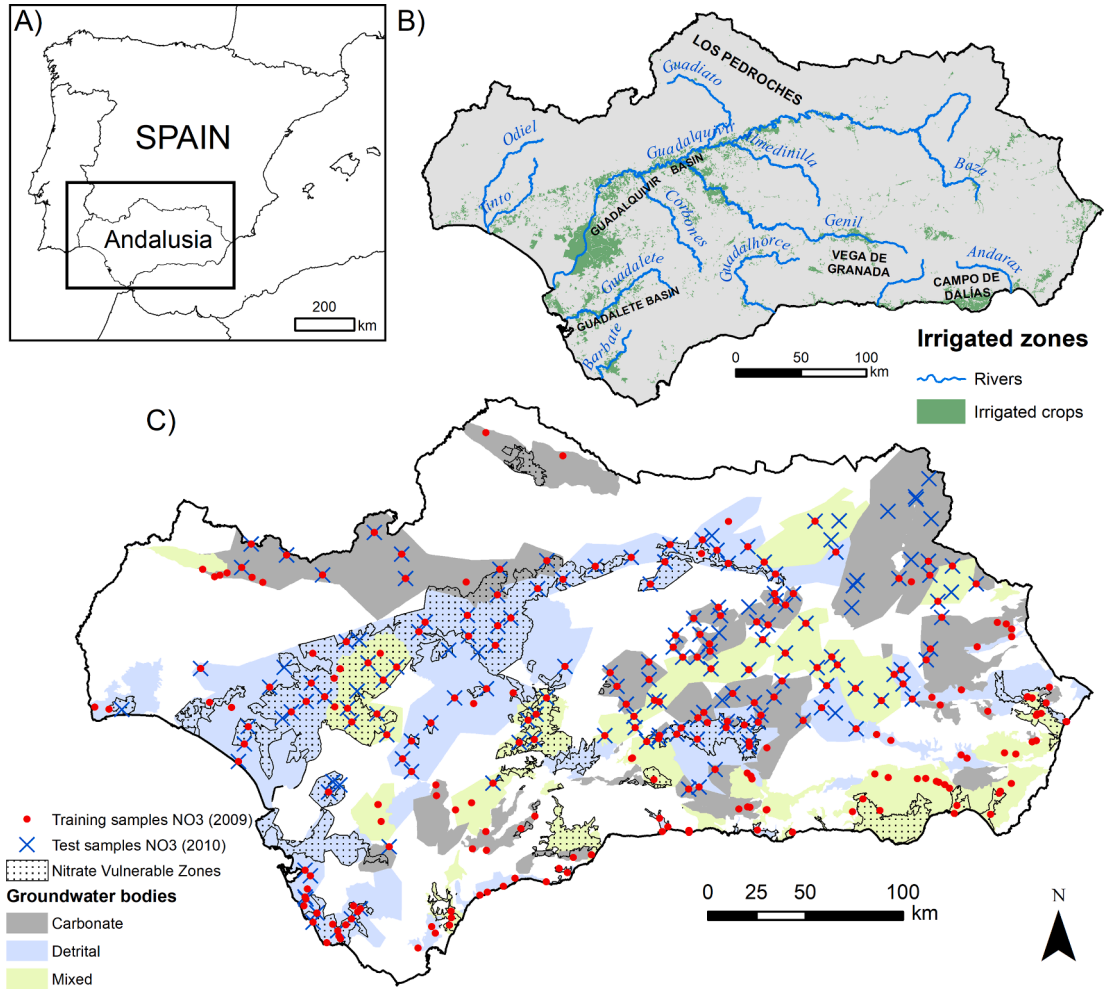

Fig. 1. A) Location of the Autonomous Community of Andalusia in Spain. B) Distribution of the main irrigated arable land areas in the river basins. C) Types of groundwater bodies and delimitation of the Nitrate Vulnerable Zones (NVZs) located above groundwater bodies. Nitrate samples used for training (321wells) and testing (282 wells); predictive models are represented here (see Section 2.2).

following years, reaching the maximum fertiliser application in 2017 (Ministerio de Agricultura Pesca y Alimentación, 2021). The Nitrates Directive (91/676/EEC) required that EU countries identify surface waters and groundwaters with nitrate concentrations above 50 mg/L or those at risk of being polluted by agricultural nitrates. The identification of polluted waters or waters at risk of being polluted was the first step towards the designation of Nitrate Vulnerable Zones. Regional governments in Spain designated Nitrate Vulnerable Zones using bodies of water that had been identified as polluted or at risk of being polluted, and the identification of intensive agricultural and livestock plots whose seepage affected bodies of water due to nitrate input (Junta de Andalucía, 2008).

The hydrogeological properties of groundwater bodies in Andalusia are diverse. Carbonate aquifers are mainly located in the east and north of Andalusia and are composed of limestones and dolomites. Detrital and mixed bodies of water are located in alluvial river basins and areas with low relief complexity, and are primarily composed of materials such as sands, gravels, silts and sandstones. The main recharge method for all groundwater bodies is infiltration by rainwater, but to a lesser extent also by infiltration of agricultural irrigation surpluses, lateral inputs from other aquifers and surface runoff, as the upper layers are usually composed of permeable materials (López Geta, 1998). This study uses the groundwater bodies identified in Fig. 1 to predict the distribution of nitrate concentrations and the area of Nitrate Vulnerable Zones within the groundwater bodies to assess their susceptibility to nitrate pollution.

2.2. Data

Groundwater nitrate data was obtained from a national scale water quality monitoring network database (Ministerio para la Transición Ecológica y el Reto Demográfico, 2021). Nitrate samples were obtained from open interval at varying depths. Thus, the final sample is a mixture of the water collected from the borehole throughout the open interval, at an average depth of 10–60 m, depending on the groundwater body. The initial database was filtered to collect groundwater samples from Andalusia and to remove erroneous data and sampling problems. The final database consisted of 1,312 samples from 451 different measurement sites with an average temporal frequency of six months (i.e. two sampling campaigns per year). Nitrate samples were obtained considering the months and years that had the highest number of samples between June and October of 2002–2010 (Fig. 2). These two periods in the year coincided with the period before and after the fertilisation process. The years 2002–2006 and 2008 had fewer samples and showed higher seasonal and interannual variability with greater dispersion. June samples showed a tendency to accumulate higher amounts of nitrates with a lower number of samples, whereas October samples showed lower amounts of nitrates when the number of samples increased. The 2009 and 2010 campaigns were selected for predictive modelling, as they had a higher number of samples and a more balanced distribution between the pre- and post-fertilisation period (June and October). The mean and median nitrate values for 2009 were lower than those for 2010 (43.8 mg/L and 14 mg/L in 2009, and 59 mg/L and 19.5 mg/L in 2010). The maximum and minimum values were 356 mg/L and 1 mg/L for 2009, and 420 mg/L and 0 mg/L (<LOD mg/L) for 2010. Two data subsets were generated: i) 321 samples from the June and October campaigns in 2009, used for model training; and ii) 282 samples from the June and October campaigns in 2010, used as an independent test for model validation (see Section 2.3 Predictive modelling). All available samples from each campaign were used for model training, applying a 10-fold cross validation to obtain an independent test. Of all samples, 248 (41.1%) were obtained in detrital groundwater bodies; 196 (32.5%) in carbonate groundwater bodies; and 159 (26.4%) in mixed groundwater bodies.

A set of 44 independent features related to driving forces (activities that may affect the environment) and nitrate infiltration was used for prediction. Only features for which data could be easily obtained on a large scale were selected. A requirement in this study was to use data available from public agencies, with a view to making this methodology exportable to other regions or a national level in a simple and straightforward way. The features that were selected were related to: i) crop phenology; ii) livestock effluents; iii) digital terrain models (DTMs); iv) weather; and v) soil textures. Table 1 shows the independent features included, data sources and extraction methodology. Features were resampled to a grid with common coordinate origin and spatial resolution (250 m × 250 m).

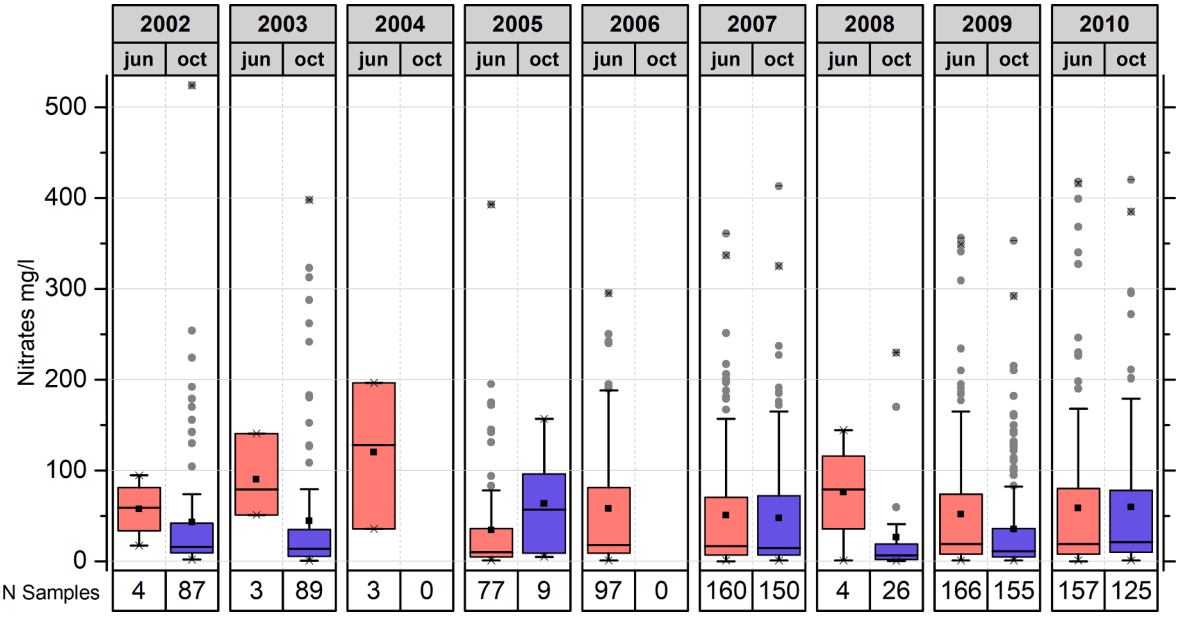

Fig. 2. Box plots of nitrate samples for the months of June and October from 2002 to 2010. The box plots for June and October are shown in red and blue respectively. The horizontal line in each boxplot is the median value and the square symbol is the mean value. The edges of each box are the 25th and 75th percentiles (i.e. the interquartile range), and the whiskers extend to 1.5 times the interquartile range. The 1st and 99th percentiles with an × and the number of samples is displayed at the bottom of the graph. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Summary of the spatial features, data sources and methodology.

Name	Abbreviation	Source	Methods	Units
Phenology		MODIS/Terra Surface Reflectance 8-Day L3 Global 250 m data (MOD09Q1) obtained from NASA's LP DAAC (https://lpdaac.usgs.gov/).	Savitzky-Golay filtering (Jönsson & Eklundh, 2004) and a threshold-based method of 10% was used to extract the phenological features.	
Time for start of the season	Start season NDVI			Julian days
Time for end of the season	End season NDVI			Julian days
Length of the season	Length season NDVI			Number of days
Base level	Base level NDVI			Unitless
Time for middle of the season	Mid-season NDVI			Unitless
Largest data value during the season	Largest value fitted NDVI			Unitless
Seasonal amplitude	Seasonal amplit. NDVI			Unitless
Rate of increase at the beginning of the season	Rate incr. NDVI			Unitless
Rate of decrease at the end of the season	Rate dec. NDVI			Unitless
Large seasonal integral	Large integral NDVI			Unitless
Small seasonal integral	Small integral NDVI			Unitless
Value for start of the season	Start season val. NDVI			Unitless
Value for end of the season	End season val. NDVI			Unitless
Livestock effluents		Livestock census for the Andalusia region (NUTS2)(https://www.juntadeandalucia.es/). Excretion coefficients UNFCCC (2021).	Calculation of livestock effluents from livestock farm data and excretion coefficients for Spain in 2010 using the methodology of Eurostat (2013) at a radius of 1, 3 and 5 km from livestock farms.	
Average livestock effluents with a search radius of 1 km	Lstock mean 1			Tonnes of nitrogen (N)
Average livestock effluents with a search radius of 3 km	Lstock mean 3			Tonnes of nitrogen (N)
Average livestock effluents with a search radius of 5 km	Lstock mean 5			Tonnes of nitrogen (N)
Sum of livestock effluents with a search radius of 1 km	Lstock sum 1			Tonnes of nitrogen (N)
Sum of livestock effluents with a search radius of 3 km	Lstock sum 3			Tonnes of nitrogen (N)
Sum of livestock effluents with a search radius of 5 km	Lstock sum 5			Tonnes of nitrogen (N)
Digital Terrain Model (DTM)		Digital Terrain Model from LiDAR.(http://centrodedescargas.cnig.es/).	Extracted using SAGA GIS (2.3.2.).	
Valley depth	Valley depth			Metres (m)
Total catchment area	Tot. Catchment area			Square metres (m^2)
Convexity	Convexity		(Iwahashi & Pike, 2007).	Percentage (%)
Slope Length and Steepness factor	LS-Factor		(Moore et al., 1991).	Unitless
Multi-resolution Ridge Top Flatness	MRRTF		(Gallant & Dowling, 2003).	Unitless
Topographic wetness index	Topographic wet. Idx.			Unitless
Convergence Index	Convergence Idx.		(Koethe & Lehmeier, 1996).	Unitless
Terrain ruggedness index	Terr. Ruggedness Idx.		(Riley et al., 1999).	Unitless
Slope	Slope			Degrees (°)
Channel network	Ch. Network			Metres (m)
Relative slope	Rel. slope			Unitless
Channel network base level	Ch. Net. Base level			Metres (m)
Profile curvature	Prof. Curvature		(Krcho, 1973).	Unitless
Aspect	Aspect			Degrees (°)
Plan curvature	Plan curvature			Unitless
Analytical hillshading	Analytical hillshading			Unitless
Multiresolution Index of Valley Bottom Flatness	MRVBF		(Gallant & Dowling, 2003).	Unitless
Altitude	Altitude			Metres (m)
Weather		Andalusian Environmental Information Network (REDIAM)(https://laboratoriorediam.cica.es/).	Spatial interpolation of monthly precipitation data using Inverse Distance Weighted (IDW) spatial interpolation method.Multiple regression from daily maximum and minimum temperature data, using elevation, distance to the coast and orientation as factors.	
Mean precipitation 3 previous months	Acc. Precip. 3 months			Millimetres (mm)
Mean precipitation 1 previous month	Acc. Precip. 1 month			Millimetres (mm)
Mean temperature 3 previous months	Mean temp. 3 months			Degrees Celsius (°C)
Mean temperature 1 previous month	Mean temp 1 month			Degrees Celsius (°C)
Soil texture		Spanish National Soil Erosion Inventory (INES). (https://www.mapa.gob.es/).	Cañero & Rodríguez Galiano (2019).	
Silt	Silt			Percentage (%)
Sand	Sand			Percentage (%)
Clay	Clay			Percentage (%)

Phenological observations are essential in many aspects of agricultural practice, such as defining the growing season length for crops in a specific region or the timing of irrigation and fertilisation (Caparros-Santiago et al., 2021; Chmielewski, 2013). In this study, phenology was considered as a proxy to identify agricultural areas and separate them from natural vegetation and urban areas. Phenological features provide more information than a static land cover image, as they not only record a measure of which agricultural areas are in production, but can also be used to predict the degree of agricultural production. Their use allows identification of the date on which the growing season begins (sprouting of the plant) and the date of harvest, enabling distinctions between different types of crops and their phenological stages. Phenology also allows distinguishing between irrigated and rainfed crops, with the former typically being more demanding in terms of nitrogen fertilisation

and with better soil conditions for nitrate leaching occur (Dzurella et al., 2015; Merchán et al., 2020). Thirteen phenological features were extracted from the analysis of NDVI time series calculated from weekly surface reflectance composites of the MODIS MOD09Q1 product. Livestock effluent features were obtained from the livestock farms census in Andalusia, which identifies each livestock farm, livestock quantity and species. The excretion coefficient was obtained for each type of livestock using those available for Spain (UNFCCC, 2021). Livestock effluents were quantified by considering their cumulative and average impact. To this end, three search radii (1, 3 and 5 km) were specified, yielding 6 features. Livestock effluents were calculated using the methodology proposed by Eurostat (2013). $N_{excretion}$ was predicted by multiplying the excretion coefficients per head and type of livestock (Nc) by the average annual population ($AAPi$) of that livestock species:

$$N_{excretion}(tonnes\ N) = \sum_i \{AAPi\ (1000\ heads) * N_c(Kg\ N\ per\ head\ p.a.)\} \quad (1)$$

A set of eighteen terrain model-related features and six weather-related features was included in the study. Terrain features were extracted from the Digital Terrain Model (DTM) for Spain, obtained from the airborne LiDAR point cloud at a resolution of 25 m on the website of the Spanish National Geographic Institute (IGN), which was then resampled to 250 m. The terrain features included in the study were: (i) basic topography metrics (i.e. elevation); (ii) morphometry, through roughness, concavity, and convexity measurements (i.e. Slope, Terrain Ruggedness Index, Multiresolution Ridge Top Flatness Index - MRRTF and Multiresolution Valley Bottom Flatness Index - MRVBF); and (iii) hydrological analysis, including measurements of channel network identification and flow catchment area (i.e. Channel Network Base Level and Total Catchment Area). Weather was included with two features: precipitation and temperature. The monthly average for the three months and the month before the nitrate samples were used, taking into account the time it might take for nitrogen pollutants to leach into the groundwater bodies under different meteorological conditions and after precipitation events, as well as the flow and transport mechanisms in the unsaturated zone of the aquifers (Al-Jaf et al., 2021; Kawagoshi et al., 2019; Menció et al., 2011; Mendes & Ribeiro, 2010). High temperatures are associated with higher plant evapotranspiration, which may have an impact on lower nitrate leaching (Wick et al., 2012). In contrast, high average temperatures may lead to higher rates of soil mineralisation, which could increase nitrate concentration in groundwater (Schweigert et al., 2004). Meanwhile, heavy precipitation is associated with increased leaching of pollutants such as nitrates that have previously been stored in the soil (Wageningen University & Research, 2011). Features with different temporal amplitudes for precipitation and temperature were included to train the model for spatio-temporal variability of pollutant infiltration into groundwater. Both precipitation and temperature features were obtained as final products from the Andalusian Environmental Information Network (REDIAM - laboratoriorediam.cica.es). Temperature and precipitation data was obtained from the meteorological stations of the Spanish Meteorological Agency (AEMET) in both cases. Likewise, the textural fractions (sand, silt and clay) were estimated using predictive models built with the Random Forest algorithm (Cañero & Rodríguez Galiano, 2019), also using predictor features derived from land surface phenology, terrain attributes and meteorology. The soil samples for the textural fractions were obtained from the Spanish National Soil Erosion Inventory (INES), which has surface soil samples for the whole of Spain, obtained at a depth of approximately 10–30 cm.

2.3. Predictive modelling

The aim of the study was to take a three-phase methodology that is commonly used for land-use change modelling and adapt it for predicting nitrate pollution in groundwater. Figure 3 shows the flowchart for methodology that was applied. Three stacks of 44 features with a common spatial resolution (250 m × 250 m) were generated for predictive modelling, including static features (terrain features and textures) and dynamic features (phenology, livestock effluents and weather). It should be noted that the DTM and texture derived features are static and therefore identical for 2009, 2010 and 2017. The models were trained using two different campaigns of 321 nitrate concentration samples, both for the Period 1 (June and October 2009), and a stack of predictor features for 2009. The predictive performance of the models trained in Period 1 was evaluated internally using a 10-fold Cross Validation (CV). The use of CV allows the algorithm to learn from the totality of the data and is an unbiased and iterative procedure, which is useful for cases where there is a low number of training samples. Once evaluated by CV, the models were applied to perform a spatial prediction of the probability of nitrate concentrations in the groundwater bodies identified in Fig. 1 exceeding 50 mg/L for the Period 2 (2010) by using a new feature stack for 2010. Two nitrate maps were obtained, corresponding to June and October, to assess the prediction differences associated with the dynamic features. The spatial predictions made for Period 2 were used to perform a statistical analysis of nitrate concentrations in the area of groundwater bodies in the Nitrate Vulnerable Zones (NVZs) (see Fig. 1) by model and groundwater body type (detrital, carbonate and mixed), in order to find out which bodies of water have a higher probability of nitrate pollution (>50 mg/L nitrate). The spatial predictions for Period 2 were also applied to perform an additional test on a different dataset to the training dataset. This test of 282 nitrate samples from June and October 2010 enabled assessment of the generalisation capacity and uncertainty associated with applying models trained in an earlier period (Period 1-2009) to the dynamic features of Period 2 (2010). In addition, a third feature stack corresponding to Period 3 (2017) was used to perform a new spatial prediction, where it was assumed that the associated accuracy or uncertainty would be similar to those of the independent test performed on the 282 nitrate samples from Period 2 (2010), as the features in both cases were different to those from Period 1 (2009). Thus, changes in the probability of nitrate concentrations above 50 mg/L in groundwater bodies were quantified for the period 2010–2017.

All models were built in R 3.5.3. software, using the 'mlr' package (Bischl et al., 2016) for training and validation. Random Forest alone and in combination with Feature Selection approaches were employed to identify the drivers of high nitrate concentrations. Random Forest for classification was used in the study because of its successful results in predicting nitrate pollution in groundwater (Knoll et al., 2019; Tesoriero et al., 2017). As this is a probability analysis, the response feature was binarised prior to model training. Samples of nitrate concentrations below 50 mg/L were reclassified as 0 (not polluted) and those above were reclassified as 1 (polluted). Random Forest only needs two parameters to be defined to generate the predictive models: the number of trees generated in the classification (k) and the number of predictor features (m) used in each tree (Breiman, 2001). For classification, each tree provides a unit vote for the most popular class in each input instance, so that the final result is determined by the majority vote of all trees (Guo et al., 2011; Rodriguez-Galiano et al., 2012). The hyperparameters were tuned to obtain more robust and generalisable models, establishing parameter combinations that optimise the predictive performance of the models (Probst et al., 2019). The k parameter used values of 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000, and the m parameter used a value of 1:30, running 300 different models.

In order to avoid creating very complex and uninterpretable models with features that provide redundant information, a Feature Selection process was applied to the models trained in Period 1 (2009). Different Feature Selection approaches based on the Random Forest wrapper method were applied in this study, due to the better performance compared to other feature selection methods (Effrosynidis & Arampatzis, 2021; Rodriguez-Galiano et al., 2018). Different search strategies can be used in feature selection such as exhaustive search, genetic

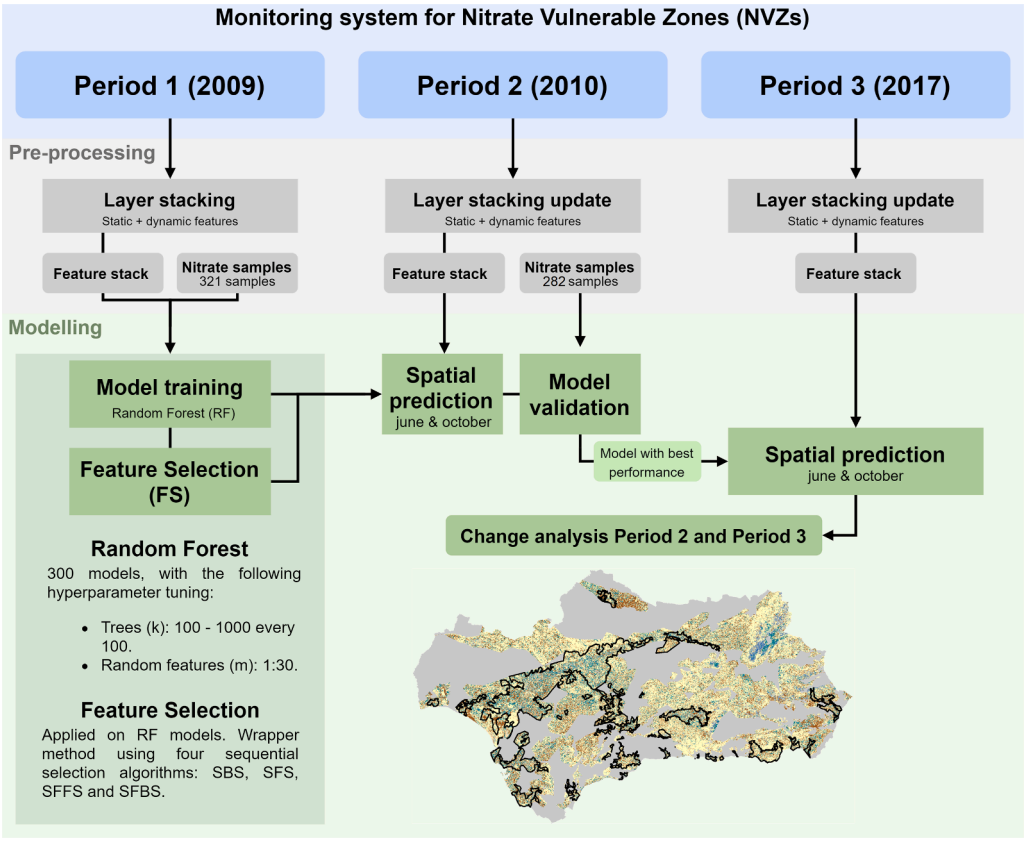

Fig. 3. Flow chart of methodology based on machine learning and driving forces for the monitoring of Nitrate Vulnerable Zones. Three periods were used: Period 1 (2009) for training of the predictive models, Period 2 (2010) for validation and initial prediction and Period 3 (2017) to analyse the updated capacity of the models.

algorithms, random search and forward and/or backward deterministic search. The last search method was selected because it has the best ratio balance between performance and computational cost (Guyon and Elisseeff, 2003). Four sequential feature search strategies were used: sequential backward selection (SBS), sequential forward selection (SFS), sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). The evaluation of the predictive performance of all models generated in Period 1 with CV and in Period 2 with the additional fixed test was performed using an objective function such as the Mean Misclassification Error (MMCE). The MMCE is the average misclassification error expressed as a proportion and is commonly used as a measure of classifier accuracy (Ferri et al., 2002). In terms of predicting nitrate pollution in groundwater, the MMCE has been used to analyse misclassified cases (Rodriguez-Galiano et al., 2018). Likewise, the importance of features in predicting Random Forest and Feature Selection models was measured with Mean Decrease in Gini, which is the average of a feature's total decrease in node impurity weighted by the proportion of samples reaching that node in each individual decision tree in the random forest (Breiman, 1998). The greater the value of the mean decrease Gini score, the greater the importance of the feature in the model.

The Kappa coefficient (K) was employed as a complementary measure to assess the suitability of applying a model trained using the driving forces of Period 1 (2009) to the Period 2 (2010) feature stack. The K coefficient is a measure of the overall agreement of an error matrix after discounting for matches that may be due to chance (Cohen, 1960). It is calculated by comparing the proportion observed along the diagonal (P_0) minus the proportion expected to be obtained by chance (P_e), divided by the maximum chance of agreement that can be expected for the marginal totals ($1–P_e$): $K = (P_0 - P_e)/(1 - P_e)$. Likewise, the evaluation of the model accuracy obtained by Feature Selection was performed by estimating the true positive rate (TPR) and the false positive rate (FPR). TPR indicates the percentage of nitrate samples correctly classified in polluted areas. Conversely, FPR indicates the percentage of misclassified samples in polluted areas. Each threshold results in a (TPR, FPR) pair and a series of such pairs are used to plot the Receiver Operating Characteristic (ROC) curve. These are also known as the "sensitivity (TPR)" and "specificity 1–(FPR)" (Rodriguez-Galiano et al., 2014). The sensitivity is the probability of predicting nitrate pollution given that the true state is polluted. The specificity is the probability of predicting no nitrate pollution given that the true state is unpolluted (Hastie et al., 2009). The area under the ROC curve statistic (AUC) was used as a measure of a classifier's performance; an AUC value of 1 is considered perfect and an AUC value of 0.5 is considered to be random guessing (Bradley, 1997).

3. Results

3.1. Feature importance

Figure 4 shows the 15 most important predictor features included in the models. In general, terrain features were the most important, followed by phenological features. The models trained in Random Forest (RF) alone and those in Random Forest with Feature Selection using sequential backward search (RF + SBS and RF + SBFS) used more features in the prediction than the Random Forest models with Feature Selection using sequential forward search (RF + SFS and RF + SFFS), resulting in large differences in importance values. In RF + SFS only 6 features were used in the prediction (Slope, MRRTF, End season val. NDVI, Valley Depth, Base level NDVI and Lstock sum 1), making it the simplest model. In comparison, RF, RF + SBS and RF + SBFS used 44, 41 and 36 predictor features, respectively, while RF + SFFS used 7 features (Slope, MRTRF, Altitude, Silt, Convexity, Start season NDVI and Mean temp. 1 month).

3.2. Modelling assessment

Table 2 shows the evaluation of the models in Cross Validation (CV) with nitrate samples from Period 1 and in the fixed test with nitrate samples from Period 2. In Cross Validation, the models trained in RF + SFS and RF + SFFS achieved better performance (MMCE = 0.101) than Random Forest (RF) alone (MMCE = 0.133) and the remaining Random Forest models with Feature Selection. The results show that the Random Forest models with Feature Selection outperformed Random Forest alone. The models based on sequential backward searches (RF + SBS and RF + SBFS) had greater accuracy (MMCE = 0.102) than those based on sequential forward search (MMCE = 0.109). However, RF + SFS and RF + SFFS prediction used fewer features, achieving models with a lower computational cost. *K* showed a good degree of agreement across all models, although certain differences existed, and RF + SBS was the model that obtained the greatest degree of agreement (K = 0.766). Random Forest alone achieved the lowest degree of agreement in this metric (K = 0.731), using the complete set of features for model building.

Figure 5 shows the results of an ROC curve analysis considering the True Positive Rate and False Positive Rate according to different thresholds for the possibility of being classified as zones with high probability of nitrate pollution. Both RF + SFS/RF + SFFS (AUC = 0.958) and RF + SBS/RF + SBFS (AUC = 0.961) achieved identical accuracies, showing that nearly all groundwater samples with high nitrate levels (>50 mg/L) were classified correctly. Therefore, the accuracy of the models trained in RF + SFS/RF + SFFS was very similar to that of the models trained in RF + SBS/RF + SBFS despite using a set with a lower number of driving forces, confirming the results obtained in the fixed test for Period 2 (see Table 2).

3.3. Spatial prediction for Period 2 and groundwater typology analysis

The model used for spatial prediction was the one obtained with all predictor features – Random Forest (RF) alone; the simplest model – Random Forest with Sequential Forward Selection (RF + SFS); and the model with the lowest MMCE – Random Forest with Sequential Backward Selection (RF + SBS). Figure 6 represents the probability of identifying nitrate concentrations above 50 mg/L in the groundwater bodies in Andalusia for Period 2 (2010). The spatial prediction analysis shows the same nitrate distribution pattern for all models, with differences existing mainly at the local level and particularly in RF + SFS. In general, the probability of identifying high nitrate concentrations increased in areas with larger arable land extensions. The highest probabilities were primarily found in the alluvial depressions of rivers and depositional zones where sedimentation is the dominant process. The results obtained for Random Forest alone and RF + SBS were spatially similar, which may be due to the use of a set of similar features in the prediction. However, the small difference between the features used for Random Forest alone and RF + SBS made this latter model the one with the highest accuracy. Figure 7 shows the box plots for predictions generated by the models. RF + SFS obtained more dispersed probability values than Random Forest alone and RF + SBS, which mainly predicted intermediate probability values (between 0.3 and 0.6). RF + SFS showed a lower probability of identifying large quantities of nitrates overall, with 34% of the area of the Nitrate Vulnerable Zones having a probability of > 0.5 for October, compared with Random Forest alone (44%) and RF + SBS (45%). The larger number of dynamic features used in the Random Forest alone and RF + SBS models favoured seasonal variability in making predictions. Thus, the difference in average probability between campaigns was not significant for both models with and without feature selection, with higher probability values in October (0.45 in RF alone and 0.46 in RF + SBS) than in June (0.43 in RF alone and 0.44 in RF + SBS). RF + SFS did not use any dynamic features in the prediction (e.g. temperature or precipitation), thus the spatial predictions for June and October were identical, with an average probability of 0.38.

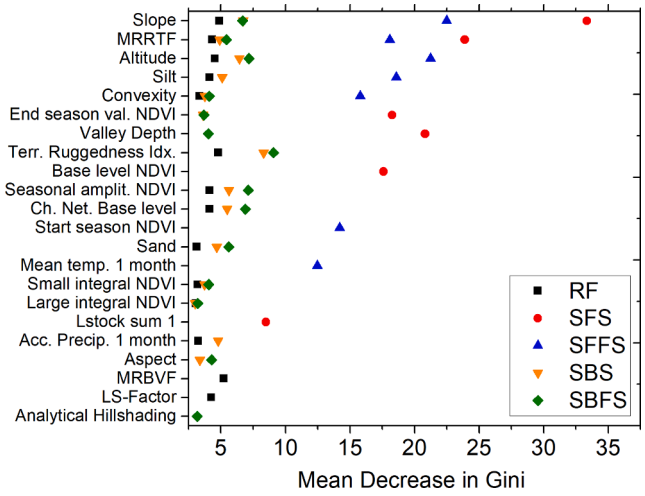

Fig. 4. Feature importance according to Mean Decrease in Gini for Random Forest and the Feature Selection algorithms. The 15 most important features for the models generated by Random Forest alone (RF), Random Forest with Sequential Backward Selection (SBS) and Random Forest with Sequential Backward Floating Selection (SBFS), the 6 features for Random Forest with Sequential Forward Selection (FS) and 7 features for Random Forest with Sequential Forward Floating Selection (SFFS) were selected.

Table 2
Predictive performance of Random Forest alone and Random Forest with Feature Selection in Cross Validation for Period 1 and in spatial prediction for Period 2.

Learner	N Features	MMCE CV	MMCE Period 2	Kappa
RF	44	0.133	0.113	0.731
RF + SFS	6	0.101	0.109	0.743
RF + SFFS	7	0.103	0.109	0.744
RF + SBS	41	0.122	0.102	0.766
RF + SBFS	36	0.106	0.102	0.758

Figure 8 shows the box plots for the probability of nitrate concentrations above 50 mg/L in the Nitrate Vulnerable Zones (NVZs) in Andalusia by learning model and groundwater body typology (detrital, carbonate and mixed) for Period 2 (2010). Box plots were created to assess the probability of nitrate pollution (>50 mg/L nitrate) in the different groundwater body types in Andalusia (detrital, carbonate and mixed) in June and October. Overall, the prediction results show that detrital and mixed groundwater bodies are more likely to have nitrate concentrations above 50 mg/L than carbonate groundwater bodies. Model analysis shows how RF + SFS predicted lower mean probability of nitrate pollution in detrital and mixed groundwater bodies (Detrital = 0.38 in June and October; Mixed = 0.40 in June and October), compared to Random Forest alone (Detrital = 0.45 in June and 0.48 in October; Mixed = 0.42 in June and October) and RF + SBS (Detrital = 0.46 in June and 0.48 in October; Mixed = 0.43 in June and October). For carbonate groundwater bodies, a similar probability was predicted by the models for Random Forest alone (Carbonate = 0.29 in June and 0.30 in October), RF + SFS (Carbonate = 0.25 in June and 0.25 in October) and RF + SBS (Carbonate = 0.31 in June and October). The analysis by models and groundwater body typologies shows that RF + SFS dispersed the values more than the other models, which is reflected in the spatial prediction as an absence of central values (yellow tones – see Fig. 6).

3.4. Spatial prediction for Period 3 and change analysis

The model built in Random Forest with Sequential Forward Selection (RF + SFS) was used to perform a new spatial prediction using the Period

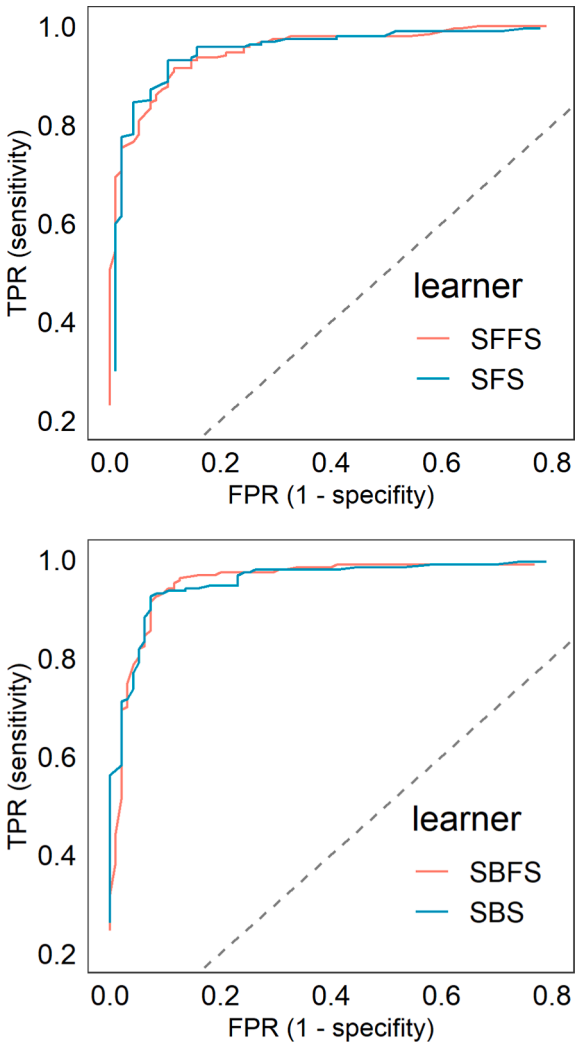

Fig. 5. Receiver Operating Characteristic (ROC) curves of the Random Forest with Feature Selection models Top: Random Forest with Sequential Forward Selection (RF + SFS) and Random Forest with Sequential Forward Floating Selection (RF + SFFS). Bottom: Random Forest with Sequential Backward Selection (RF + SBS) and Random Forest with Sequential Backward Floating Selection (RF + SBFS). The dashed line indicates a random guess.

3 (2017) feature stack. This model had the best predictive performance of all the trained models and was computationally more efficient as it had a similar predictive performance with a much smaller number of features than the other trained models. Figure 9 shows the spatial prediction of the probability of nitrate concentrations above 50 mg/L in groundwater bodies for June and October of Period 3 (2017), as well as the changes in the spatial prediction made by the learning model between 2017 and 2010. It should be noted that since RF + SFS did not use any seasonal dynamic features (i.e., temperature or precipitation), the prediction was identical for June and October. The map displays changes in probabilities of nitrate concentrations exceeding 50 mg/L as the areas where the probability of nitrates increased in 2017 (in brown) and areas where it decreased (in blue). The areas that show no change (in yellow) range from −0.05 to 0.05. Overall, the probability remained stable at 0.38, although, according to the prediction, the area affected by the highest probability of nitrate pollution (>50 mg/L nitrate) decreased in 2017 to 33%, slightly lower than the Period 2 value (34%). Feature Selection has the advantage of improving the interpretability of the models by selecting the most relevant feature set for classification. The model built in RF + SFS used three dynamic features: End season val. NDVI; Base level NDVI; and Lstock sum 1. The dynamic features enabled detection of changes at the local scale. Thus, the probability was reduced in those areas where the End season val. NDVI (NDVI value for the end of the growing season - EOS) and Base level NDVI (BL) were higher compared to Period 2. In other areas where NDVI was slightly lower, Lstock sum 1 (accumulation of livestock effluent within 1 km) was found to have increased significantly compared to Period 2, thus livestock effluent may also have influenced the increase in the probability of nitrate pollution in specific regions.

4. Discussion

4.1. Use of extrinsic features of groundwater bodies and their effect on nitrate pollution

The probability of nitrate concentrations exceeding the thresholds established in WFD (50 mg/L nitrates) can be predicted using environmental features extrinsic to groundwater bodies. To this end, features were used that are easily updatable on a spatio-temporal scale and can be adapted to the complex and dynamic interplay between the hydrological cycle and diffuse pollutant transport (Kumar et al., 2020). Existing predictive modelling of nitrate pollution in the literature uses hydrogeological features as part of subsets of aquifer-scale predictive features (i.e. Boy-Roura et al., 2013; Motevalli et al., 2019; Rahmati et al., 2019; Tesoriero et al., 2017). The complex measurement of hydrogeological parameters or the lack of large-scale measurements of these parameters in many regions hampers the ability to replicate the methodology of these studies in larger settings. Knoll et al. (2020) and Ouedraogo et al. (2019) performed predictions of nitrate concentrations at national and continental scales using MLA that overcome the limitations of obtaining hydrogeological information for several bodies of water. However, its applicability in the continuous analysis of nitrate may be limited by the need for updated hydrogeological information in areas where, in the case of countries such as Spain, said information is not provided by government administrative bodies. This study attempted to avoid the limitations associated with producing updated hydrogeological information by using dynamic environmental features to train multi-temporal predictive models sensitive to seasonal changes in nitrate concentration in groundwater.

Random Forest with Sequential Forward Selection (RF + SFS) was the model that had the best predictive performance in relation to the number of features and used both dynamic and static features in the prediction. The three dynamic features were: End season val. NDVI and Base level NDVI, which are related to phenology; and Lstock sum 1, which is related to livestock farming. The phenological features were Base Level NDVI and the NDVI value of the EOS. Base Level NDVI is the mean between the minimum values at the beginning and at the end of the growing season (Eklundh & Jönsson, 2017) and both Base Level NDVI and the NDVI value of the EOS are associated with the period with less photosynthetic activity of vegetation (i.e. the period where the NDVI value is lower). The results show that the probability of higher nitrate concentrations in the Nitrate Vulnerable Zones is linked to the zones where the Base Level NDVI and the NDVI value of EOS are lower. This coincides with Zhao et al. (2020), who identified a negative relationship between nitrate leaching and the increase in the vegetation index value in autumn and winter months, in such a way that nitrate leaching is reduced in zones with increased above-ground biomass and plant nitrogen, as was also shown in Macdonald et al. (2005). Zhao et al. (2020) also linked a decrease in soil water nitrate concentration with an increase in the NDVI value and greater plant nitrogen storage. Higher nitrate concentrations in Nitrate Vulnerable Zones also showed a positive relationship with high-density manure production areas. The accumulation of livestock effluents within a 1 km radius (Lstock sum 1) was the feature included in the model, indicating that the greatest effluent impact would be in areas closest to farms. This positive relationship has been analysed in studies such as De Notaris et al. (2018) and could be explained by the leaching of livestock effluents deposited in

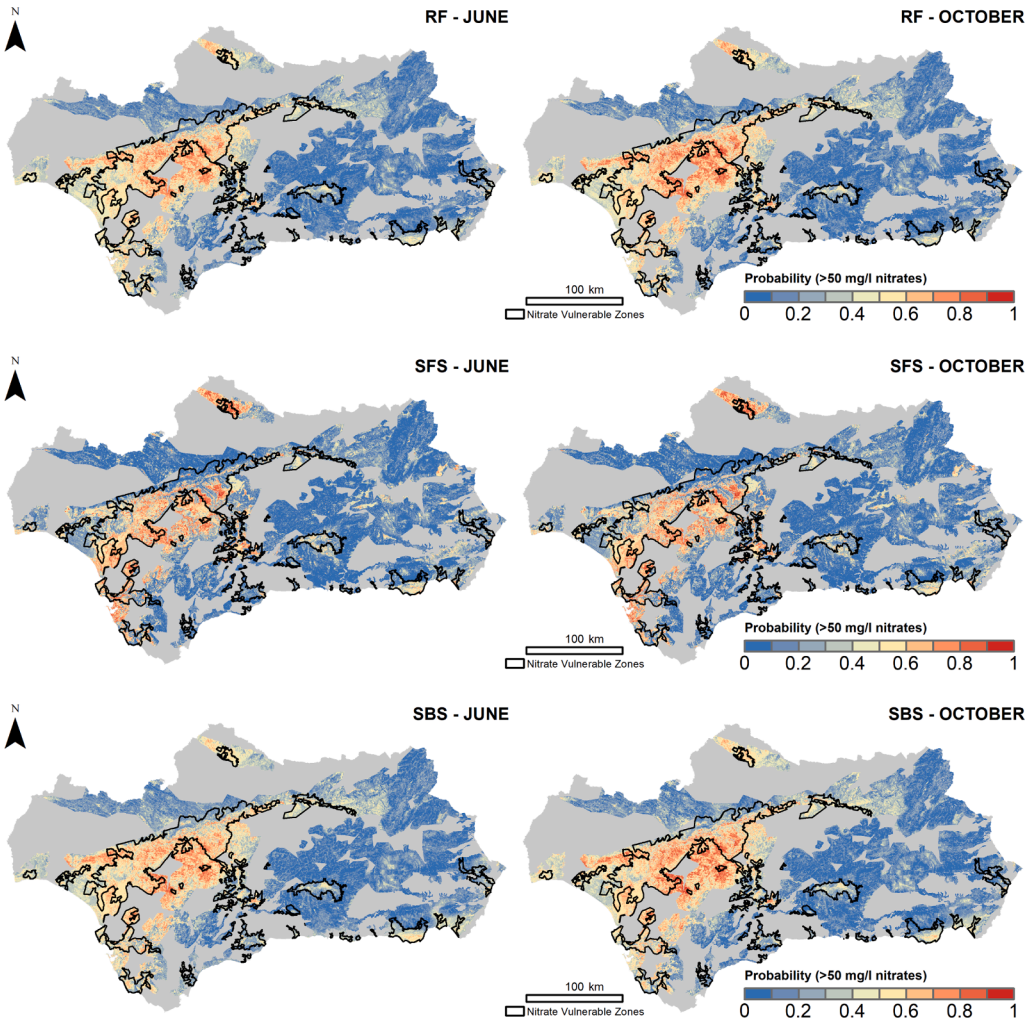

Fig. 6. Spatial prediction of the probability of nitrate concentrations above 50 mg/L in the groundwater bodies in Andalusia, Spain in Random Forest alone, Random Forest with Sequential Forward Selection (RF + SFS) and Random Forest with Sequential Backward Selection (RF + SBS) for June and October of Period 2 (2010). The prediction was performed using the 44-feature stack with a common spatial resolution (250 m × 250 m) for all groundwater bodies in the outlined Nitrate Vulnerable Zones (NVZs).

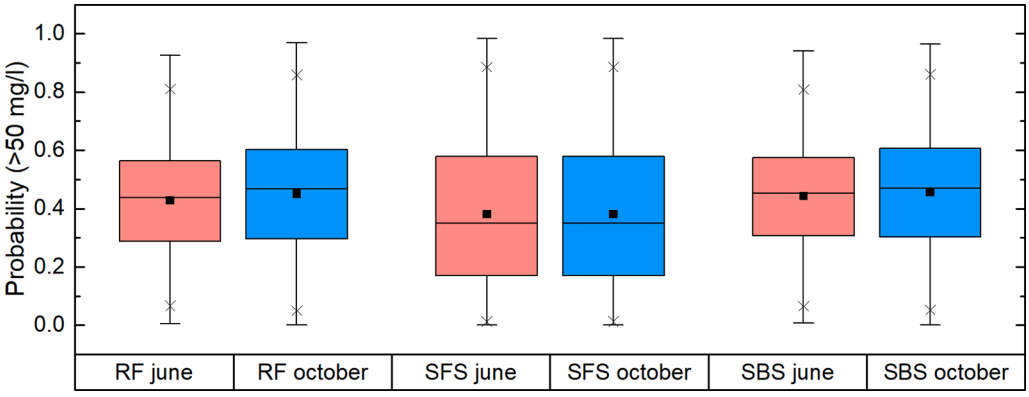

Fig. 7. Box plot of the probability of nitrate concentrations above 50 mg/L in the Nitrate Vulnerable Zones (NVZs) in Andalusia, Spain in Random Forest alone, Random Forest with Sequential Forward Selection (RF + SFS) and Random Forest with Sequential Backward Selection (RF + SBS) for June and October of Period 2 (2010). The horizontal line in each boxplot is the median value and the square symbol is the mean value. The edges of each box are the 25th and 75th percentiles (i.e. interquartile range), and the whiskers extend to 1.5 times the interquartile range. The 1st and 99th percentile as an ×, with the search methods and months represented at the bottom of the graph.

and around buildings (Oenema et al., 2007) or used as fertilisers.

The static features included in SFS and with the greatest importance in all models were terrain features (see Fig. 4). Terrain features were also the most important in SFFS, although this model additionally included soil textural features such as silt. A controversial aspect of Feature Selection is the multiplicity of good models, which is also common in statistical algorithms, such as multiple regression or logistic regression. Different feature subsets might share good and similar accuracy, thus resulting in a non-unique solution or physical model explaining a phenomenon (Rashomon effect, Breiman, 2001b). Slope was the most important in SFS and SFFS; the zones with the shallowest slope favoured infiltration processes affecting groundwater bodies (Antonakos & Lambrakis, 2007). The average surface terrain slope over groundwater bodies in Andalusia is 8.16%, while this slope decreases to 3.52% in Nitrate Vulnerable Zones. This is because Nitrate Vulnerable Zones usually occur within the alluvial zones of large rivers, which are more suitable for agricultural activities. In fact, features such as slope, altitude or ruggedness are linked in the literature to the groundwater potential

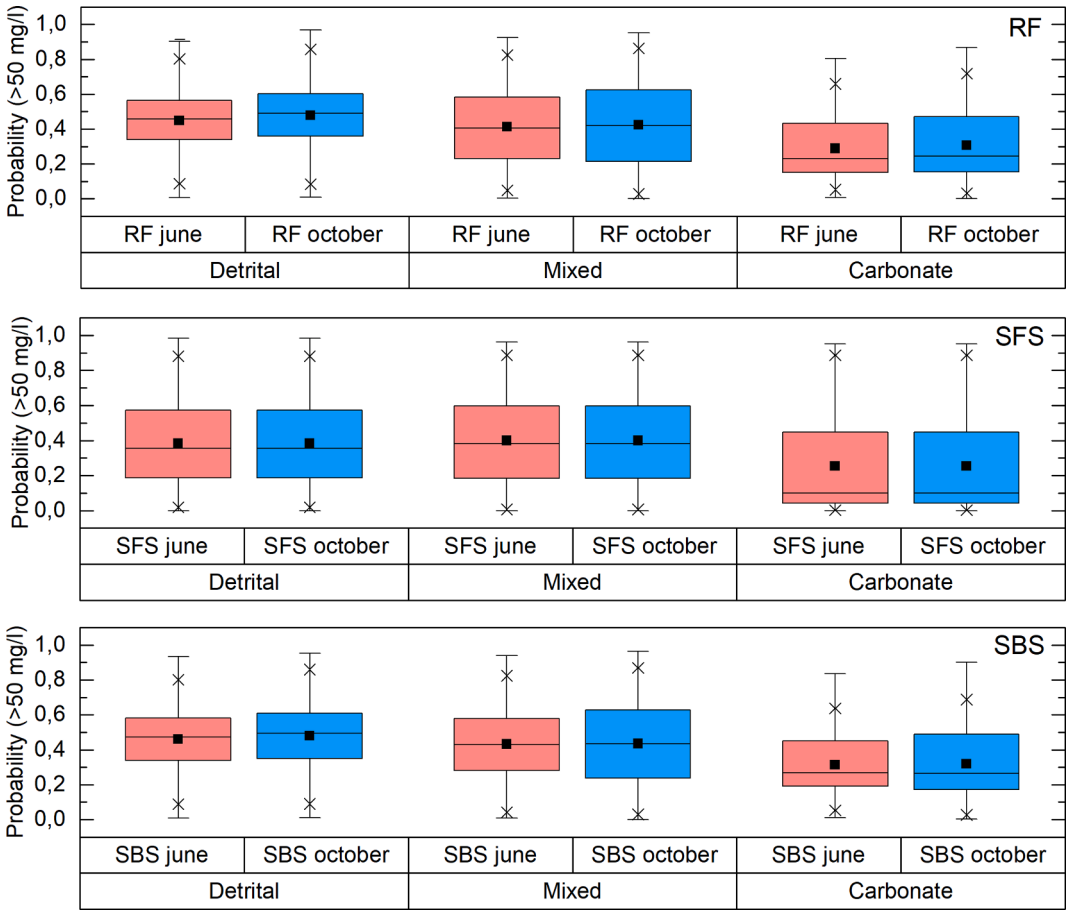

Fig. 8. Box plot of the probability of nitrate concentrations above 50 mg/L in the Nitrate Vulnerable Zones (NVZs) for each type of groundwater body (detrital, mixed and carbonate) in Random Forest (RF) alone, Random Forest with Sequential Forward Selection (SFS) and Random Forest with Sequential Backward Selection (SBS) for June and October of Period 2 (2010). The horizontal line in each boxplot is the median value and the square symbol is the mean value. The edges of each box are the 25th and 75th percentiles (i. e. interquartile range), and the whiskers extend to 1.5 times the interquartile range. The 1st and 99th percentile as an ×,with the search methods, month and type of groundwater body represented at the bottom of the graph.

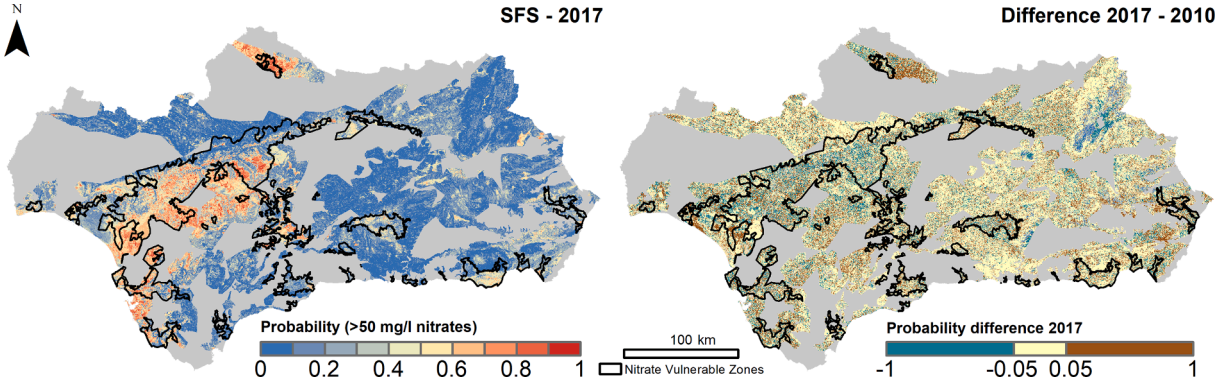

Fig. 9. Left: spatial prediction of the probability of nitrate concentrations above 50 mg/L in groundwater bodies in Andalusia, Spain in Random Forest with Sequential Forward Selection (RF + SFS) for June and October Period 3 (2017). Right: difference in probability between Period 3 (2017) and Period 2 (2010). The zones in blue indicate reduction in probability in Period 3 compared to Period 2. The zones in yellow indicate no difference in probability. The zones in brown indicate an increase in probability. The prediction was performed using the 44-feature stack with a common spatial resolution (250 m × 250 m) for all groundwater bodies in the outlined Nitrate Vulnerable Zones (NVZs). It should be noted that the prediction was identical for June and October because the model did not use seasonal dynamic features. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

zones (Naghibi et al., 2017; Rahmati et al., 2018), as well as to the higher nitrate pollution in groundwater (Mfumu Kihumba et al., 2016; Motevalli et al., 2019; Ouedraogo et al., 2016). Other features included in the models such as MRRTF and MRVBF, identified the summits of ridges and the bottoms of valleys, respectively (Gallant & Dowling, 2003). The importance of these features, which are related to the morphological conditions of the terrain, might be linked to the accumulation of run-off surpluses and the infiltration process affecting the vadose zone (Mendes & Ribeiro, 2014). Thus, Parra Suárez et al. (2019) and Zhu et al. (2009) observed that nitrate leaching processes in hillside

agroforestry environments increased in lower area of the hillsides and sediment deposition areas (which might be related to slope, MRRTF and MRVBF), due to nitrate transport by soil water (Young & Briggs, 2005). The results also showed that areas of lower elevation in Nitrate Vulnerable Zones are more susceptible to nitrate pollution. Creed & Band (1998) found a positive relationship between lower elevation and nitrate leaching, stating that lower elevation might be related to areas that receive less precipitation. Rahmati et al. (2019) reported the highest concentration of nitrates in the groundwater of valleys, which tend to have greater agricultural activity and are more susceptible to soil water retention. In the case of Andalusia, agricultural areas are spread over the catchments of the main rivers, where the average elevation does not exceed 350 m, thus the highest nitrate concentrations might be related to agricultural areas.

4.2. Comparison of feature selection approaches

The application of Feature Selection algorithms enabled the selection of subsets of explanatory environmental features, obtaining simple methodological procedures that could be exported to other Nitrate Vulnerable Zones with similar properties in the EU. The sequential backward search (SBS and SBFS) strategies obtained a lower MMCE compared to the sequential forward search (SFS and SFFS) strategies, with success rates of 89.8% and 89.1%, respectively (see Table 2 – MMCE Period 2). These results stood in contrast to those of Rodriguez-Galiano et al. (2018), who obtained better results with models trained with Sequential Forward Selection (SFS) and Sequential Forward Floating Selection (SFFS). Likewise, Random Forest alone was the model with the worst performance, correctly classifying 88.7% of cases (see Table 2 – MMCE Period 2). However, the prediction obtained from Random Forest alone showed better performance than other probability-based studies, such as Tesoriero et al. (2017), who obtained a success rate of 77% for probabilities of nitrate concentrations above 5 mg/L in Wisconsin, USA and Rodriguez-Galiano et al. (2014), who had a success rate of 80.46% for probabilities of nitrate concentrations above 50 mg/L in a region of southern Spain. This is an advance in nitrate prediction in large-scale areas, generating highly accurate spatial predictions for groundwater bodies with different hydrogeological properties without requiring nitrate sampling campaigns for each Nitrate Vulnerable Zone. To this end, different innovative aspects have been included, using dynamic (e.g. phenology) and static (e.g. soil attributes) features that are extrinsic to groundwater bodies. In addition, a novel three-step methodology has been applied to select representative drivers of nitrate pollution and their seasonal variability. Among all the models generated, RF + SFS was considered the model that performed best, obtaining an MMCE similar to the most accurate models with only six features. The difference in the number of features confirms the idea proposed by Guyon and Elisseeff (2003) and verified by Rodriguez-Galiano et al. (2018) that wrappers built using forward sequential search were computationally more efficient, identifying a smaller feature subset at a similar error rate.

Nitrate samples from Period 2 (2010) were used for model validation. In this sense, nitrate samples were not available for validation purposes at Period 3 (2017), as the idea behind this methodology is to predict the “current” status of groundwater bodies for operative and early diagnosis (see Section 2.3 Predictive modelling). It should be noted that this is a standard approach in certain areas of science, such as land-use change modelling (Camacho Olmedo et al., 2015; Msofe et al., 2020), but has never been applied to groundwater pollution. Additionally, the usefulness of this approach is not only to have spatial predictions of the probability of nitrate pollution using nitrate measurements from previous years, but also to have a prediction for groundwater bodies at a regional level. The proposed methodology allows for assessment of the probability of pollution in all groundwater bodies, including those for which the WFD quality network does not obtain measurements, exemplified in an agricultural region with complex orography such as Andalusia.

4.3. Spatial probability of nitrate pollution in nitrate Vulnerable Zones

The probability of nitrate pollution had a similar spatial pattern across all models, showing differences in specific regions. In general, the irrigated agricultural zones in the Nitrate Vulnerable Zones over detrital groundwater bodies showed a strong relationship with the highest nitrate concentrations, especially in the Guadalquivir and Guadalete basins in south-western Andalusia (see Fig. 1). This trend is shown in similar studies for other agricultural regions located on detrital groundwater bodies (e.g. Boy-Roura et al., 2013; Knoll et al., 2019). In these zones, the interaction of groundwater bodies with surface flow, the permeability of alluvial deposits and agricultural practices are associated with a greater vulnerability to diffuse nitrate pollution (Arauzo et al., 2011; Kazakis & Voudouris, 2015). Other flat areas with significant livestock farming activity also seemed to be especially linked to nitrates (e.g. Los Pedroches, in the north of Andalusia). Nitrate Vulnerable Zones over carbonate aquifers had a lower probability of nitrate pollution, which can be observed in the east of Andalusia. These zones are characterised by having few valley bottoms and are zones with greater ruggedness, favouring infiltration processes through the dissolution of carbonates. However, soil cover in these karstic regions is usually thin or absent (Leibungut, 1998), which may hinder large-scale agricultural practices and explains the reason why the majority of Nitrate Vulnerable Zones were delimited in detrital aquifers (see Fig. 1). Thus, Ducci et al. (2019) related low nitrate levels in carbonate aquifers to mountainous areas with less agricultural activity and anthropogenic pressure. The analysis of changes between 2010 and 2017 (see Fig. 9) showed a slight decrease in the probability of nitrate pollution in Nitrate Vulnerable Zones, especially in irrigated areas. The decrease in the probability was mainly in the Nitrate Vulnerable Zones, while increases occurred mostly in the remaining bodies of water. These results are in line with the monitoring report on Directive 91/676/EEC for the period 2016–2019 for Spain (Ministerio para la Transición Ecológica y el Reto Demográfico, 2020), which shows a strong decrease in most of the measurement stations in the Nitrate Vulnerable Zones in Andalusia. At the same time, the report warned of the increase in nitrate concentrations in groundwater bodies outside the Nitrate Vulnerable Zones (especially in areas such as Los Pedroches - see Section 2.1. Case study). Although the unavailability of predictions for other years made it impossible to identify clear trends, the decrease in the probability in Nitrate Vulnerable Zones might be related to the constraints and measures on agricultural practices imposed by the WFD and Nitrates Directive (Oenema et al., 2011). Thus, the decrease in probability in the modelling was related to the increase in Base Level NDVI and NDVI value at the end of the season in 2017. The impact of the Nitrates Directive on agriculture has been examined in studies such as Velthof et al. (2014), who linked the decrease in N losses in the 2000–2008 period with the reduced use of fertilisers and manure in the EU. Despite the improvement in water quality in the Nitrate Vulnerable Zones, the development of tools and control mechanisms for monitoring the Nitrate Vulnerable Zones was identified as one of the main challenges for the EU (European Commission, 2018). Considering that Spain does not have a unified water quality data system, a different approach was taken that considers the good performance of a trained model for an intermediate year (Period 2 – 2010). This is because most of the nitrate campaigns are carried out by regional and supra-regional agencies, using different sampling frequencies, methodologies and databases, hindering the possibility of training models with larger input data and assessing the interannual variability of the nitrate content of groundwater bodies in a more comprehensive way. Nevertheless, the results obtained in this study could serve as a starting point to enhance operational nitrate monitoring in Nitrate Vulnerable Zones.

5. Conclusion

This study assessed the application of machine learning and feature selection algorithms for the predicting nitrate pollution in Nitrate Vulnerable Zones using spatial features extrinsic to groundwater bodies. Predictive modelling based on extrinsic features, such as those derived from remote sensing, made it possible to identify the area of groundwater bodies identified as Nitrate Vulnerable Zones that might be most susceptible to pollution without the need to carry out specific analysis of each groundwater body. Phenology and terrain features were selected as the most important in the predictive modelling. Phenology enabled the incorporation of fundamental knowledge about the effects of productivity, area and the agricultural calendar and their possible relationship to nitrogen fertilisers application. Thus, features such as the NDVI value for the end of the season or Base Level NDVI might be an important source of information in measuring the impact that agriculture has on nitrate pollution in Nitrate Vulnerable Zones.

The creation of highly accurate spatial predictions for a period beyond the model learning period may be useful in establishing mitigation measures, and it would also help to provide complete and reliable information on the state of groundwater bodies in the Nitrate Vulnerable Zones, in accordance with the requirements of the Nitrates Directive. Feature Selection methods allowed for optimisation of Random Forest performance, generating more accurate models and achieving a reduction in the dimensionality of the feature space. Wrapper-based Random Forest with Sequential Forward Selection (RF + SFS) was the model with the best performance in relation to the number of features used (MMCE = 0.109, AUC = 0.958 and six predictor features). This model predicted that 34% of the area of the Nitrate Vulnerable Zones was susceptible to having nitrate concentrations above 50 mg/L in Period 2. The application of the model to Period 3 resulted in a 1% reduction in the area of the Nitrate Vulnerable Zones susceptible to nitrate pollution, down to 33%. The applicability of this methodology to other regions of the EU might serve as the basis for periodic prediction of the state of Nitrate Vulnerable Zones groundwater bodies and the establishment of new criteria for identifying trends that require the application of control and protection measures.

CRediT authorship contribution statement

Aaron Cardenas-Martinez: Writing – original draft preparation, Investigation, Methodology, Software, Data curation, Writing – review & editing, Visualization. **Victor Rodriguez-Galiano:** Writing – review & editing, Methodology, Software, Data curation, Supervision, Conceptualization. **Juan Antonio Luque-Espinar:** Visualization, Resources, Formal analysis. **Maria Paula Mendes:** Supervision, Formal analysis, Conceptualization, Writing – review & editing.

Declaration of Competing Interest

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

Acknowledgements

The first author is a FPU grant holder funded by the Spanish "Ministerio de Universidades" (Reference FPU19/00384). The authors would like to express their gratitude for the financial support provided by the projects RTI2018-096561-A-I00 and US-1262552, funded by the "Ministry of Science and Innovation (Ministerio de Ciencia e Innovación)", the "State Research Agency (Agencia Estatal de Investigación)" and the "European Regional Development Fund (Fondo Europeo de Desarrollo Regional (FEDER)", and the "Junta de Andalucía" and FEDER, respectively.

References

Akbariyeh, S., Bartelt-Hunt, S., Snow, D., Li, X., Tang, Z., Li, Y., 2018. Three-dimensional modeling of nitrate-N transport in vadose zone: Roles of soil heterogeneity and groundwater flux. J. Contam. Hydrol. 211, 15–25. https://doi.org/10.1016/j.jconhyd.2018.02.005.

Al-Jaf, P., Smith, M., Gunzel, F., 2021. Unsaturated zone flow processes and aquifer response time in the chalk aquifer, Brighton, South East England. *Groundwater* 59 (3), 381–395. https://doi.org/10.1111/gwat.v59.310.1111/gwat.13055.

Antonakos, A.K., Lambrakis, N.J., 2007. Development and testing of three hybrid methods for the assessment of aquifer vulnerability to nitrates, based on the drastic model, an example from NE Korinthia, Greece. *J. Hydrol.* 333 (2–4), 288–304. https://doi.org/10.1016/j.jhydrol.2006.08.014.

Arauzo, M., Valladolid, M., Martínez-Bastida, J.J., 2011. Spatio-temporal dynamics of nitrogen in river-alluvial aquifer systems affected by diffuse pollution from agricultural sources: implications for the implementation of the Nitrates Directive. J. Hydrol. 411 (1–2), 155–168. https://doi.org/10.1016/j.jhydrol.2011.10.004.

Babiker, I.S., Mohamed, A.A., Terao, H., Kato, K., Ohta, K., 2004. Assessment of groundwater contamination by nitrate leaching from intensive vegetable cultivation using geographical information system. Environ. Int. 29, 1009–1017.

Band, S.S., Janizadeh, S., Pal, S.C., Chowdhuri, I., Siabi, Z., Norouzi, A., Melesse, A.M., Shokri, M., Mosavi, A., 2020. Comparative analysis of artificial intelligence models for accurate estimation of groundwater nitrate concentration. Sensors 20 (20), 5763. https://doi.org/10.3390/s20205763.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2018. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci. Total Environ. 621, 697–712. https://doi.org/10.1016/j.scitotenv.2017.11.185.

Bellman, R. (2003). *Dynamic Programming*. Dover Publications. https://doi.org/34.

Biau, G., Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M., 2016. mlr: Machine learning in R. J. Machine Learn. Res. 17 (170), 1–5. http://jmlr.org/papers/v17/15-066.html.

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97 (1–2), 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5.

Boy-Roura, M., Nolan, B.T., Menció, A., Mas-Pla, J., 2013. Regression model for aquifer vulnerability assessment of nitrate pollution in the Osona region (NE Spain). J. Hydrol. 505, 150–162. https://doi.org/10.1016/j.jhydrol.2013.09.048.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Breiman, L., 2001a. Random forests. Machine Learn. 45, 5–32.

Breiman, L. (Ed.), 1998. Classification and regression trees. Chapman & Hall/CRC Press.

Breiman, L., 2001b. Statistical modeling: the two cultures. Statistical Sci. 16 (3), 199–231. https://doi.org/10.1214/ss/1009213726.

Buduma, N., Locascio, N. (2017). *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms* (M. Loukides & S. Cutt (Eds.)). O'Reilly Media.

Camacho Olmedo, M.T., Pontius, R.G., Paegelow, M., Mas, J.-F., 2015. Comparison of simulation models in terms of quantity and allocation of land change. Environ. Modell. Software 69, 214–221. https://doi.org/10.1016/j.envsoft.2015.03.003.

Cañero, F., Rodríguez Galiano, V. (2019). Mapping organic material and texture fractions of soils in Spain using satellite-derived vegetation phenology. *ESA Living Planet Symposium 2019*.

Caparros-Santiago, J.A., Rodriguez-Galiano, V., Dash, J., 2021. Land surface phenology as indicator of global terrestrial ecosystem dynamics: a systematic review. ISPRS J. Photogramm. Remote Sens. 171, 330–347. https://doi.org/10.1016/j.isprsjprs.2020.11.019.

Chmielewski, F.-M., 2013. Phenology in Agriculture and Horticulture. In: Schwartz, M.D. (Ed.), Phenology: An Integrative Environmental Science. Springer, Netherlands, pp. 539–561.

Cho, J.-C., Cho, H.B., Kim, S.-J., 2000. Heavy contamination of a subsurface aquifer and a stream by livestock wastewater in a stock farming area, Wonju, Korea. Environ. Pollut. 109 (1), 137–146.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Measur. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Confederación Hidrográfica del Guadalquivir. (2015). *Plan Hidrológico de la Demarcación Hidrográfica del Guadalquivir. Segundo ciclo de planificación: 2015 - 2021* (p. 161). Ministerio de Agricultura, Alimentación y Medio Ambiente. https://www.chguadalquivir.es/descargas/PlanHidrologico2015-2021/Planes_2DO_Ciclo/Guadalquivir/MEMORIA_PHD_GUADALQUIVIR.pdf.

Creed, I.F., Band, L.E., 1998. Export of nitrogen from catchments within a temperate forest: Evidence for a unifying mechanism regulated by variable source area dynamics. Water Resour. Res. 34 (11), 3105–3120. https://doi.org/10.1029/98WR01924.

Dash, M., Liu, H., 1997. Feature selection for classification. Intell. Data Anal. 1 (1–4), 131–156. https://doi.org/10.1016/S1088-467X(97)00008-5.

De Notaris, C., Rasmussen, J., Sørensen, P., Olesen, J.E., 2018. Nitrogen leaching: a crop rotation perspective on the effect of N surplus, field management and use of catch crops. Agric. Ecosyst. Environ. 255, 1–11. https://doi.org/10.1016/j.agee.2017.12.009.

Dixon, B., 2005. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: a GIS-based sensitivity analysis. J. Hydrol. 309 (1–4), 17–38. https://doi.org/10.1016/j.jhydrol.2004.11.010.

Ducci, D., Della Morte, R., Mottola, A., Onorati, G., Pugliano, G., 2019. Nitrate trends in groundwater of the Campania region (southern Italy). Environ. Sci. Pollut. Res. 26 (3), 2120–2131. https://doi.org/10.1007/s11356-017-0978-y.

Duncan, J.M.A., Dash, J., Atkinson, P.M., 2015. Elucidating the impact of temperature variability and extremes on cereal croplands through remote sensing. Glob. Change Biol. 21 (4), 1541–1551. https://doi.org/10.1111/gcb.12660.

Dzurella, K.N., Pettygrove, G.S., Fryjoff-Hung, A., Hollander, A., Harter, T., 2015. Potential to assess nitrate leaching vulnerability of irrigated cropland. J. Soil Water Conserv. 70 (1), 63–72. https://doi.org/10.2489/jswc.70.1.63.

Effrosynidis, D., Arampatzis, A., 2021. An evaluation of feature selection methods for environmental data. Ecol. Inf. 61, 101224. https://doi.org/10.1016/j.ecoinf.2021.101224.

Eklundh, L., Jönsson, P. (2017). *TIMESAT 3.3 with seasonal trend decomposition and parallel processing Software Manual*.

Esmaeili, S., Thomson, N.R., Tolson, B.A., Zebarth, B.J., Kuchta, S.H., Neilsen, D., 2014. Quantitative global sensitivity analysis of the RZWQM to warrant a robust and effective calibration. J. Hydrol. 511, 567–579. https://doi.org/10.1016/j.jhydrol.2014.01.051.

European Commission. (2018). Report from the Commission to the council and the European Parliament on the implementation of Council Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources. https://ec.europa.eu/environment/water/water-nitrates/pdf/nitrates_directive_implementation_report.pdf.

European Environmental Agency. (2018). *European waters — Assessment of status and pressures 2018*. https://doi.org/10.2800/303664.

European Environmental Agency. (2020). *Waterbase - Water Quantity*. https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quantity-12.

Eurostat. (2013). *Nutrient Budgets. Methodology and Handbook. Version 1.02*. https://ec.europa.eu/eurostat/documents/2393397/2518760/Nutrient_Budgets_Handbook_%28CPSA_AE_109%29_corrected3.pdf/4a3647de-da73-4d23-b94b-e2b23844dc31.

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

Fewtrell, L., 2004. Drinking-water nitrate, methemoglobinemia, and global burden of disease: a discussion. Environ. Health Perspect. 112 (14), 1371–1374. https://doi.org/10.1289/ehp.7216.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12) https://doi.org/10.1029/2002WR001426.

Ghimire, B., Rogan, J., Miller, J., 2010. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sens. Lett. 1 (1), 45–54. https://doi.org/10.1080/01431160903252327.

Goodchild, R.G., 1998. EU Policies for the reduction of nitrogen in water: the example of the Nitrates Directive. Environ. Pollut. 102 (1), 737–740. https://ac.els-cdn.com/S0269749198801061/1-s2.0-S0269749198801061-main.pdf?_tid=6ce70527-1e45-4a50-b5b6-2693755fcc50&acdnat=1550515618_356da4b5807798728d656687513c829b.

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS J. Photogramm. Remote Sens. 66 (1), 56–66. https://doi.org/10.1016/j.isprsjprs.2010.08.007.

Guyon, I., Elisseeff, A., 2003. In: Studies in Fuzziness and Soft ComputingFeature Extraction. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–25. https://doi.org/10.1007/978-3-540-35488-8_1.

Hansen, B., Dalgaard, T., Thorling, L., Sørensen, B., Erlandsen, M., 2012. Regional analysis of groundwater nitrate concentrations and trends in Denmark in regard to agricultural influence. Biogeosciences 9 (8), 3277–3286. https://doi.org/10.5194/bg-9-3277-2012.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Additive Models, Trees, and Related Methods. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, pp. 295–336. https://doi.org/10.1007/978-0-387-84858-7_9.

INE. (2021). *Encuesta sobre el uso del agua en el sector agrario. 2000 - 2012*. Instituto Nacional de Estadística. https://www.ine.es/jaxi/Tabla.htm?path=/t26/p067/p03/a2000-2012/l0/&file=02001.px&L=0.

Iwahashi, J., J. Pike, R. (2007). Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology 86*, 409–440. https://doi.org/10.1016/j.geomorph.2006.09.012.

Jönsson, P., Eklundh, L., 2004. TIMESAT - a program for analyzing time-series of satellite sensor data. Comput. Geosci. 30 (8), 833–845.

Decreto 36/2008, de 5 de febrero, por el que se designan las zonas vulnerables y se establecen medidas contra la contaminación por nitratos de origen agrario., (2008) (testimony of Junta de Andalucia). https://www.juntadeandalucia.es/boja/2008/36/1.

Juntakut, P., Snow, D., Haacker, E., Ray, C., 2019. The long term effect of agricultural, vadose zone and climatic factors on nitrate contamination in Nebraska's groundwater system. J. Contam. Hydrol. 220, 33–48.

Kawagoshi, Y., Suenaga, Y., Chi, N.L., Hama, T., Ito, H., Duc, L.V., 2019. Understanding nitrate contamination based on the relationship between changes in groundwater levels and changes in water quality with precipitation fluctuations. Sci. Total Environ. 657, 146–153. https://doi.org/10.1016/j.scitotenv.2018.12.041.

Kazakis, N., Voudouris, K.S., 2015. Groundwater vulnerability and pollution risk assessment of porous aquifers to nitrate: modifying the DRASTIC method using quantitative parameters. J. Hydrol. 525, 13–25. https://doi.org/10.1016/j.jhydrol.2015.03.035.

Khalil, A., Almasri, M.N., McKee, M., Kaluarachchi, J.J., 2005. Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour. Res. 41 (5) https://doi.org/10.1029/2004WR003608.

Khosravi, K., Sartaj, M., Tsai, F.-T.-C., Singh, V.P., Kazakis, N., Melesse, A.M., Prakash, I., Tien Bui, D., Pham, B.T., 2018. A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment. Sci. Total Environ. 642, 1032–1049. https://doi.org/10.1016/j.scitotenv.2018.06.130.

Knoll, L., Breuer, L., Bach, M., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci. Total Environ. 668, 1317–1327.

Knoll, L., Breuer, L., Bach, M., 2020. Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning. Environ. Res. Lett. 15 (6), 064004 https://doi.org/10.1088/1748-9326/ab7d5c.

Koethe, R., Lehmeier, F. (1996). *SARA - System zur Automatischen Relief-Analyse. User Manual, 2. Edition (No publicado)*.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97 (1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.

Krcho, J., 1973. Morphometric analysis of relief on the basis of geometric aspect of field theory. Acta Geographica Universatis Comenianae, Geographico-Physica 1, 7–233.

Kumar, R., Heße, F., Rao, P.S.C., Musolff, A., Jawitz, J.W., Sarrazin, F., Samaniego, L., Fleckenstein, J.H., Rakovec, O., Thober, S., Attinger, S., 2020. Strong hydroclimatic controls on vulnerability to subsurface nitrate contamination across Europe. Nat. Commun. 11 (1), 6302. https://doi.org/10.1038/s41467-020-19955-8.

Leibungut, J.C., 1998. Vulnerability of karst aquifers. In: Karst Hydrology. IAHS Press, pp. 45–60.

López Geta, J.A. (1998). *Atlas hidrogeológico de Andalucía* . Instituto Tecnológico Geominero de España.

Macdonald, A.J., Poulton, P.R., Howe, M.T., Goulding, K.W.T., Powlson, D.S., 2005. The use of cover crops in cereal-based cropping systems to control nitrate leaching in SE England. Plant Soil 273 (1–2), 355–373. https://doi.org/10.1007/s11104-005-0193-3.

Menció, A., Boy, M., Mas-Pla, J., 2011. Analysis of vulnerability factors that control nitrate occurrence in natural springs (Osona Region, NE Spain). Sci. Total Environ. 409 (16), 3049–3058. https://doi.org/10.1016/j.scitotenv.2011.04.048.

Mendes, M.P., Ribeiro, L., Nascimento, J., Condesso de Melo, T., Stigter, T.Y., Buxo, A., 2012. A groundwater perspective on the river basin management plan for central Portugal – developing a methodology to assess the potential impact of N fertilizers on groundwater bodies. Water Sci. Technol. 66 (10), 2162–2169. https://doi.org/10.2166/wst.2012.427.

Mendes, M.P., Ribeiro, L., 2010. Nitrate probability mapping in the northern aquifer alluvial system of the river Tagus (Portugal) using Disjunctive Kriging. Sci. Total Environ. 408 (5), 1021–1034. https://doi.org/10.1016/j.scitotenv.2009.10.069.

Mendes, M.P., Ribeiro, L., 2014. The importance of groundwater for the delimitation of Portuguese National Ecological Reserve. Environ. Earth Sci. 72 (4), 1201–1211. https://doi.org/10.1007/s12665-013-3039-y.

Merchán, D., Sanz, L., Alfaro, A., Pérez, I., Goñi, M., Solsona, F., Hernández-García, I., Pérez, C., Casalí, J., 2020. Irrigation implementation promotes increases in salinity and nitrate concentration in the lower reaches of the Cidacos River (Navarre, Spain). Sci. Total Environ. 706, 135701. https://doi.org/10.1016/j.scitotenv.2019.135701.

Messier, K.P., Wheeler, D.C., Flory, A.R., Jones, R.R., Patel, D., Nolan, B.T., Ward, M.H., 2019. Modeling groundwater nitrate exposure in private wells of North Carolina for the Agricultural Health Study. Sci. Total Environ. 655, 512–519. https://doi.org/10.1016/j.scitotenv.2018.11.022.

Mfumu Kihumba, A., Ndembo Longo, J., Vanclooster, M., 2016. Modelling nitrate pollution pressure using a multivariate statistical approach: the case of Kinshasa groundwater body, Democratic Republic of Congo. *Hydrogeol. J.* 24 (2), 425–437. https://doi.org/10.1007/s10040-015-1337-z.

Ministerio de Agricultura Pesca y Alimentación. (2021). *Estadística de consumo de fertilizantes en la agricultura*. https://www.mapa.gob.es/es/estadistica/temas/estadisticas-agrarias/agricultura/estadisticas-medios-produccion/fertilizantes.aspx.

Ministerio para la Transición Ecológica y el Reto Demográfico. (2020). *Informe de seguimiento de la Directiva 91/676/CEE de contaminación del agua por nitratos utilizados en la agricultura. Cuatrienio 2016 - 2019. ESPAÑA*. https://www.miteco.gob.es/es/agua/temas/estado-y-calidad-de-las-aguas/informe-2016-2019_tcm30-518402.pdf.

Ministerio para la Transición Ecológica y el Reto Demográfico. (2021). *Sistema de Información de Redes de seguimiento del estado e información hidrológica*. https://sig.mapama.gob.es/redes-seguimiento/.

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5 (1), 3–30. https://doi.org/10.1002/hyp.3360050103.

Motevalli, A., Naghibi, S.A., Hashemi, H., Berndtsson, R., Pradhan, B., Gholami, V., 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. J. Cleaner Prod. 228, 1248–1263. https://doi.org/10.1016/j.jclepro.2019.04.293.

Msofe, N.K., Sheng, L., Li, Z., Lyimo, J., 2020. Impact of land use/cover change on ecosystem service values in the Kilombero Valley Floodplain, Southeastern Tanzania. *Forests* 11 (1), 109. https://doi.org/10.3390/f11010109.

Naghibi, S.A., Moghaddam, D.D., Kalantar, B., Pradhan, B., Kisi, O., 2017. A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. J. Hydrol. 548, 471–483. https://doi.org/10.1016/j.jhydrol.2017.03.020.

Nolan, B.T., Gronberg, J.M., Faunt, C.C., Eberts, S.M., Belitz, K., 2014. Modeling nitrate at domestic and public-supply well depths in the central valley, California. *Environ. Sci. Technol.* 48 (10), 5643–5651. https://doi.org/10.1021/es405452q.

Oenema, O., Bleeker, A., Braathen, N.A., Budnáková, M., Keith Bull, K., Cermák, P., 2011. Nitrogen in current European policies. In: The European nitrogen assessment. Cambridge University Press, pp. 62–81.

Oenema, O., Oudendag, D., Velthof, G.L., 2007. Nutrient losses from manure management in the European Union. Livestock Sci. 112 (3), 261–272. https://doi.org/10.1016/j.livsci.2007.09.007.

Ouedraogo, I., Defourny, P., Vanclooster, M., 2016. Mapping the groundwater vulnerability for pollution at the pan African scale. Sci. Total Environ. 544, 939–953. https://doi.org/10.1016/j.scitotenv.2015.11.135.

Ouedraogo, I., Defourny, P., Vanclooster, M., 2019. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeol. J. 27 (3), 1081–1098. https://doi.org/10.1007/s10040-018-1900-5.

Parra Suárez, S., Peiffer, S., Gebauer, G., 2019. Origin and fate of nitrate runoff in an agricultural catchment: Haean, South Korea – Comparison of two extremely different monsoon seasons. Sci. Total Environ. 648, 66–79. https://doi.org/10.1016/j.scitotenv.2018.08.115.

Probst, P., Wright, M.N., Boulesteix, A., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Rev. Data Min. Knowl. Disc. 9 (3) https://doi.org/10.1002/widm.2019.9.issue-310.1002/widm.1301.

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., Mollaefar, E., Tiefenbacher, J., Cipullo, S., Ahmad, B.B., Tien Bui, D., 2019. Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. Sci. Total Environ. 688, 855–866. https://doi.org/10.1016/j.scitotenv.2019.06.320.

Rahmati, O., Kornejady, A., Samadi, M., Nobre, A.D., Melesse, A.M., 2018. Development of an automated GIS tool for reproducing the HAND terrain model. Environ. Modell. Software 102, 1–12. https://doi.org/10.1016/j.envsoft.2018.01.004.

Ransom, K.M., Nolan, B.T., Traum, J.A., Faunt, C.C., Bell, A.M., Gronberg, J.A.M., Wheeler, D.C., Rosecrans, C.Z., Jurgens, B., Schwarz, G.E., Belitz, K., Eberts, S.M., Kourakos, G., Harter, T., 2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. Sci. Total Environ. 601-602, 1160–1172. https://doi.org/10.1016/j.scitotenv.2017.05.192.

Riley, S., Degloria, S., Elliot, S.D., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Int. J. Sci. 5, 23–27.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.

Rodriguez-Galiano, V.F., Luque-Espinar, J.A., Chica-Olmo, M., Mendes, M.P., 2018. Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. Sci. Total Environ. 624, 661–672. https://doi.org/10.1016/j.scitotenv.2017.12.152.

Rodriguez-Galiano, V.F., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., Ribeiro, L., 2014. 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). Sci. Total Environ. 476–477, 189–206. https://doi.org/10.1016/j.scitotenv.2014.01.001.

Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344.

Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., Pradhan, B., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054.

Sakamoto, T., Yokozawa, M., Torinati, H., Shibayama, M., Ishitsuka, N., Ohno, H., 2005. A crop phenology detection method using time-series MODIS data. Remote Sens. Environ. 96 (3–4), 366–374. https://doi.org/10.1016/j.rse.2005.03.008.

Schweigert, P., Pinter, N., van der Ploeg, R.R., 2004. Regression analyses of weather effects on the annual concentrations of nitrate in soil and groundwater. J. Plant Nutr. Soil Sci. 167 (3), 309–318. https://doi.org/10.1002/jpln.200321291.

Singh, K.P., Gupta, S., Mohan, D., 2014. Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. J. Hydrol. 511, 254–266. https://doi.org/10.1016/j.jhydrol.2014.01.004.

Tesoriero, A.J., Gronberg, J.A., Juckem, P.F., Miller, M.P., Austin, B.P., 2017. Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour. Res. 53 (8), 7316–7331. https://doi.org/10.1002/2016WR020197.

Tullo, E., Finzi, A., Guarino, M., 2019. Review: environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy. Sci. Total Environ. 650, 2751–2760. https://doi.org/10.1016/j.scitotenv.2018.10.018.

UNFCCC. (2021). *UNFCCC National Inventory Submissions 2011*. https://unfccc.int/process/transparency-and-reporting/reporting-and-review-under-the-convention/greenhouse-gas-inventories/submissions-of-annual-greenhouse-gas-inventories-for-2017/submissions-of-annual-ghg-inventories-2011.

Velthof, G.L., Lesschen, J.P., Webb, J., Pietrzak, S., Miatkowski, Z., Pinto, M., Kros, J., Oenema, O., 2014. The impact of the Nitrates Directive on nitrogen emissions from agriculture in the EU-27 during 2000–2008. Sci. Total Environ. 468-469, 1225–1233.

Wageningen University & Research. (2011). Recommendations for establishing Action Programmes under Directive 91/676/EEC concerning the protection of waters against pollution caused by nitrates from agricultural sources Contract number N° 07 0307/2010/580551/ETU/B1. Part C: Analysis of the process. https://op.europa.eu/en/publication-detail/-/publication/4ec63804-0cc9-4133-ad73-31b65ef584f3/language-en/format-PDF/source-217942479.

Wagh, V., Panaskar, D., Muley, A., Mukate, S., Gaikwad, S., 2018. Neural network modelling for nitrate concentration in groundwater of Kadava River basin, Nashik, Maharashtra, India. Groundwater Sustainable Dev. 7, 436–445. https://doi.org/10.1016/j.gsd.2017.12.012.

Ward, M.H., Kilfoy, B.A., Weyer, P.J., Anderson, K.E., Folsom, A.R., Cerhan, J.R., 2010. Nitrate intake and the risk of thyroid cancer and thyroid disease. Epidemiology 21 (3), 389–395. https://doi.org/10.1097/EDE.0b013e3181d6201d.

Wells, M.J., Gilmore, T.E., Nelson, N., Mittelstet, A., Böhlke, J.K., 2021. Determination of vadose zone and saturated zone nitrate lag times using long-term groundwater monitoring data and statistical machine learning. Hydrol. Earth Syst. Sci. 25 (2), 811–829. https://doi.org/10.5194/hess-25-811-2021.

Wheeler, D.C., Nolan, B.T., Flory, A.R., DellaValle, C.T., Ward, M.H., 2015. Modeling groundwater nitrate concentrations in private wells in Iowa. Sci. Total Environ. 536, 481–488. https://doi.org/10.1016/j.scitotenv.2015.07.080.

WHO. (2017). *Guidelines for drinking-water quality, 4th edition, incorporating the 1st addendum*. https://www.who.int/water_sanitation_health/publications/drinking-water-quality-guidelines-4-including-1st-addendum/en/.

Wick, K., Heumesser, C., Schmid, E., 2012. Groundwater nitrate contamination: factors and indicators. J. Environ. Manage. 111, 178–186. https://doi.org/10.1016/j.jenvman.2012.06.030.

Young, E.O., Briggs, R.D., 2005. Shallow ground water nitrate-N and ammonium-N in cropland and riparian buffers. Agric. Ecosyst. Environ. 109 (3–4), 297–309. https://doi.org/10.1016/j.agee.2005.02.026.

Zhang, H., Bao Ho, T., 2006. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. Informatica 30, 305–319.

Zhao, J., De Notaris, C., Olesen, J.E., 2020. Autumn-based vegetation indices for estimating nitrate leaching during autumn and winter in arable cropping systems. Agric. Ecosyst. Environ. 290, 106786. https://doi.org/10.1016/j.agee.2019.106786.

Zhu, B., Wang, T., Kuang, F., Luo, Z., Tang, J., Xu, T., 2009. Measurements of nitrate leaching from a Hillslope cropland in the Central Sichuan Basin, China. Soil Sci. Soc. Am. J. 73 (4), 1419–1426. https://doi.org/10.2136/sssaj2008.0259.