Science of the Total Environment 857 (2023) 159655

Contents lists available at ScienceDirect

Science of the Total Environment

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

Prediction of ecological status of surface water bodies with supervised machine learning classifiers

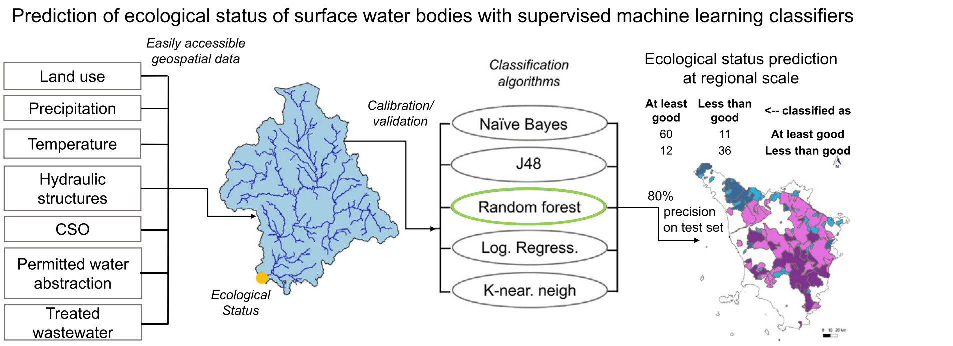
Chiara Arrighi *, Fabio Castelli

Department of Civil and Environmental Engineering, Università degli Studi di Firenze, via di S. Marta 3, 50139 Florence, Italy

HIGHLIGHTS

- Prediction of ecological status at large scales with few spatial data.
- Performances of 5 Machine Learning classification algorithms are compared.
- Ecological status is correlated to land use, summer climate and water exploitation.
- Random Forest predicts ecological status with 80 % precision.

GRAPHICAL ABSTRACT


Prediction of ecological status of surface water bodies with supervised machine learning classifiers
Easily accessible geospatial data
Land use
Precipitation
Temperature
Hydraulic structures
CSO
Permitted water abstraction
Treated wastewater
Ecological Status
Calibration/ validation
Classification algorithms
Naïve Bayes
J48
Random forest
Log. Regress.
K-near. neigh
Ecological status prediction at regional scale
At least good
Less than good
<-- classified as
60
11
At least good
12
36
Less than good
80% precision on test set

ARTICLE INFO

Editor: Ashantha Goonetilleke

Keywords:
Water resources management
River catchment
Hydrology
Ecological status

ABSTRACT

Ensuring a good ecological status of water bodies is one of the key challenges of communities and one of the objectives of the European Water Framework Directive. Although recent works identified the most significant stressors affecting the ecological quality of rivers, the ability to predict the overall ecological status of rivers based on a limited amount of easily accessible geospatial data has not been investigated so far. Most of the analyses focus on detailed local modelling and measurements which cannot be systematically applied at regional scales for the purposes of water resources management. The aim of this work is to understand the capabilities of five supervised machine learning classifiers of predicting the ecological status of rivers based on land use, climate, morphology, and water management parameters extracted over the river catchments corresponding to the ecological monitoring stations. Moreover, the performances of machine learning classifiers are compared to the results of the canonical correlation analysis. The method is applied to 360 catchments in Tuscany (central Italy) with a median size of 33.6 km^2 and a Mediterranean climate. The results show (i) a significant correlation of ecological status with summer climate (i.e., maximum temperatures and minimum precipitation), land use and water exploitation, (ii) an 80 % precision of Random Forest algorithm to predict ecological status and (iii) higher capability of all classifiers to predict at least good ecological status. In perspective, such predictive capabilities can support decision making in the land and water resources management and highlight strategies for river eco-hydrological conservation.

1. Introduction

River water quality is significantly affected by numerous stressors (Reid et al., 2019; Herrero et al., 2018; EEA, 2018). The European Water framework Directive WFD (60/2000/EC) (EU Parliament, 2000) prescribes an intense monitoring program of ecological status of river water bodies with the final objective of achieving at least a “good” condition. The interaction among different stressors in determining the ecological status is still poorly understood due to the complex non-linear feedbacks of living ecosystems and does not facilitate the adoption of mitigation measures (Carvalho et al., 2019). Moreover, the WFD recognizes the crucial

* Corresponding author.
E-mail address: chiara.arrighi@unifi.it (C. Arrighi).

http://dx.doi.org/10.1016/j.scitotenv.2022.159655
Received 10 August 2022; Received in revised form 19 October 2022; Accepted 19 October 2022
Available online 21 October 2022

importance of water quantity and dynamics in maintaining the quality of aquatic ecosystems and requires River Basin District Authorities to set out ecological flows. Ecological flows are considered within the context of the WFD as “a hydrological regime consistent with the achievement of the environmental objectives of the WFD in natural surface water bodies” (European Commission, 2016). Thus, understanding the key ecological stressors allows to set out more appropriate ecological flows.

Ecological status is usually determined by measuring and combining specific biological indicators and recognizing that different water types and supporting quality elements demand different threshold levels, thus a good ecological status cannot be defined across Europe using absolute standards (Voulvoulis et al., 2017). The ecological status classification is based on the worst among the selected biological indicators. Common indicators for rivers are macroinvertebrate-based indices, which are sensitive to pollution and habitat degradation (Azzellino et al., 2013). Nitrogen, phosphorus, and dissolved oxygen are accounted for in the LIMeco index which is widely recognized as a proxy for ecological quality (Larsen et al., 2019; ARPAT, 2021). Macrophytes indices (e.g., Macrophyte Biological Index for Rivers, IBMR) and diatom (Trophic Diatom Index, TDI) (Lu et al., 2020; Bytyqi et al., 2020) are also widely used. The river quality classification sensu WFD also includes the integrity and continuity of river morphology and riparian zones (Belletti et al., 2020).

These indicators require periodic on-site surveys by environment agencies that collect samples for analysis. Obviously, such a detailed on-site monitoring can be carried out only at predetermined locations and water bodies of main interest, with limited capabilities of extrapolating ecological quality indicators at unmonitored locations. In recent years, the role of different stressors on water quality has been investigated. Grizzetti et al. (2017) considered the catchments of the whole European Union, with an average size of 180 km^2 to study the most significant indicators in explaining ecological quality. They identified twelve indicators in terms of (i) pollution pressures (e.g., phosphorus concentrations), (ii) hydrological alterations (e.g., low flow alteration), (iii) hydromorphological alterations (e.g., artificial land cover in floodplains, density of infrastructures) and (iv) integrated pressures (e.g., agricultural land cover in the catchment area). The correlation among indicators was explored and three types of classification techniques were adopted to test the combined effects of multiple pressures. Their results showed that a good ecological status of rivers is mainly driven by the presence of natural areas in floodplains, nutrient concentration (especially nitrogen), infrastructures in floodplains and urbanisation and agriculture in the drained catchment. It should be noted that the indicators related to pollution and hydrological alterations were obtained by numerical simulations, while others were elaborated from spatial data.

Lemm et al. (2021) analysed ca. 50,000 catchments of a median size of 60 km^2 and seven stressors, i.e., urban and agricultural land use in the riparian zone, alteration of mean annual flow and base flow, phosphorus and nitrogen load and mixture toxic pressure. They distinguished the river type in terms of geological substrate, catchment size, altitude, and climate. Also in this case, hydrological alterations and nutrient load were simulated to build the dataset. The dataset was analysed with Spearman's correlation to check how individual stressors are correlated and through nonlinear Boosted Regression Tree models. They found counter-intuitive results such as a significant role of nutrient enrichment in mountain rivers rather than in lowland rivers and similar effects of hydrological alteration in Mediterranean rivers and overall catchment population.

Visser et al. (2022) adopted 10 Machine Learning (ML) models and a multiple regression model on a number of samples of about 200 records to calculate the effect of restoration and mitigation measures on the ecological status of surface waters and to support decision makers in the Netherlands. They select 15 stressors which include, besides pollution ones, water transparency, bank design, hydraulic connectivity, meandering, impoundments etc. to fit the common specific water body conditions in the country. They found that Random Forest algorithm provides the best prediction of ecological quality ratios but with a limited transparency of model structure.

A ranking of stressors by using Random Forest algorithm was also performed by Herrero et al. (2018) for the Ebro catchment. They identified agricultural surface, population, and altitude as good predictors for biological quality elements, i.e., diatoms, macrophytes and invertebrates. They also considered future climate and socio-economic scenarios and predicted a decrease in diatom and invertebrate indices.

Valerio et al. (2021) used ML (Random Forest and gradient boosted regression trees) to model biological response to multiple stressors, such as land use and nutrient concentrations. They obtained an accuracy between 70 % and 90 % in predicting macroinvertebrates, diatoms and macrophytes indices in the Tagus River basin.

Nevertheless, the prediction of the overall ecological status at large spatial scales, though crucial for planning effective policies, is still a complex task, especially when the required information, e.g., nutrient concentration/load needs to be simulated in many catchments with limited ability to validate the estimates with on-site data. Some authors also highlight a low attention on the role of chemical pollution and other major pressures besides eutrophication (Posthuma et al., 2020; Poikane et al., 2020).

The mentioned studies considered nutrients load, simulated at large scales, land use, simulated discharges and morphological alterations to understand the response of biological quality in surface water highlighting the importance of phosphorus (and nitrogen) and land use (Visser et al., 2022; Lemm et al., 2021; Valerio et al., 2021; Grizzetti et al., 2017). ML algorithms are key tools (Chen et al., 2020) to rank the significance of single and combined stressors or predict biological indices, which constitute the basis for the determination of ecological status. In the hypothesis that land use and nutrients load in rivers are correlated and that also biological indices are correlated to the overall ecological status, the aim of this work is to investigate the ability of ML classification algorithms to predict the overall ecological status of rivers by using a limited number of geospatial data related to climate, land use, water management and morphological alterations, without any model to simulate physico-chemistry stressors (e.g. dissolved oxygen, ammonium, etc.) and hydrological alterations. The ML classification algorithms are trained and tested on two independent datasets of the Tuscany Region (Central Italy) starting from the ecological monitoring stations of the regional environment agency. To the authors' knowledge this is the first example of prediction of the overall ecological status for a diverse set of catchments sizes from 1 km^2 to c. 8000 km^2 without explicitly accounting for physico-chemistry stressors, but only for a limited number of easily available geospatial data. The proposed approach has the potential for a systematic application on large scales, e.g., region or country, based on readily available geospatial datasets and might help water resources management.

2. Materials and method

The methodological workflow is summarized in Fig. 1. First, the river catchments corresponding to the location of the ecological monitoring sites are identified based on a Digital Terrain Model in a Geographic Information System environment. Second, the vector layer of the river catchments is enriched with several attributes that are combined into synthetic parameters to build the working dataset (Section 2.1). In the third phase the dataset is split into calibration and validation set to test the predictive performances of 5 ML classification algorithms (Section 2.2) in comparison with a canonical correlation analysis. Finally, the prediction errors on the validation set are interpreted on a geographic basis through the spatial mapping of the components of the confusion matrix.

2.1. Dataset creation

The catchment dataset is created in Quantum Geographic Information System (QGIS) starting from the text file of coordinates of the ecological monitoring sites converted into a vector shapefile. In order to resolve minor issues related to the position of monitoring sites, i.e., few meters of distance between river centerline and monitoring stations, the QGIS plugin ClosestPoint (Baudin, 2020) is used to associate the points to the river

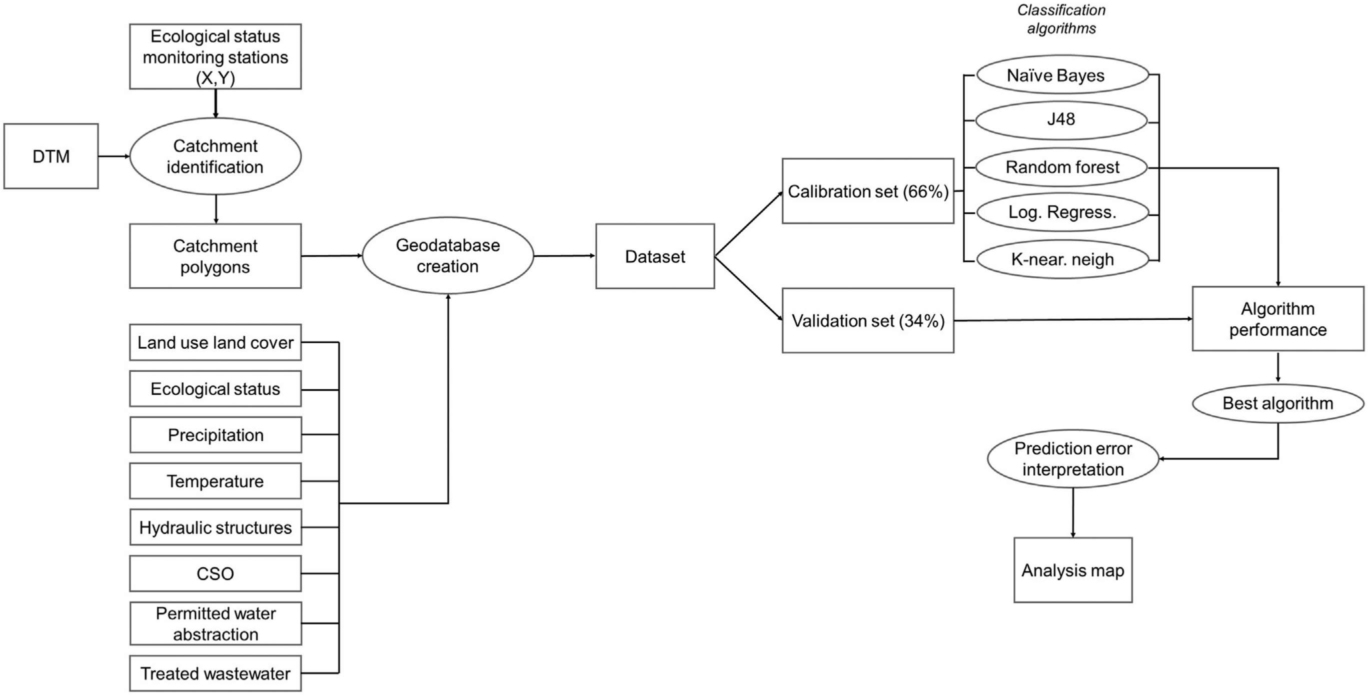


Fig. 1. Methodological workflow, ellipses stand for numerical operations, rectangles for data flows.

network. The river catchments upstream each ecological monitoring site are obtained by a filled 10 m resolution Digital Terrain Model (DTM). The hydrology SAGA tools *channels* and *upslope area* allow creating the catchments and associate to them the monitoring station ID and the ecological quality class ECO_St (from 1-High to 5-Bad). The procedure univocally associates river ID, corresponding catchment, and ECO_St.

Once the catchments are created, a series of geospatial data are extracted over the catchment area A (km^2). Table 1 lists the geospatial data included in the attribute table. The data include climate conditions (temperature and precipitation), land use, presence of hydraulic structures, water abstraction and treatment.

From the GIS data extracted on the catchment areas a series of normalized parameters have been calculated to summarize key stressors in terms of climate, water management, morphological alterations, and land use (Table 2). With respect to land use, the percentage of agricultural, artificial and forest areas have been calculated (parameters 3–5). Climate parameters on the catchments are mean annual precipitation, minimum summer precipitation, and maximum summer temperature (parameters 6–8). They have been selected to account both for annual average conditions and for summer conditions, that in Mediterranean climate are the driest months. Maximum summer temperature, although lasting for a short time, show high correlation with mean annual temperature (Pearson's $r = 0.85$) and can highlight the most critical period for ecological status when higher surface water temperatures have consequences on dissolved oxygen concentration at saturation.

Alterations in river continuity and morphology are accounted for in two parameters measuring the density of linear and point hydraulic structures over the length of the water body in the catchment (parameters 9 and 10).

The ratio between permitted water abstraction and precipitation volumes in the catchment, both on annual and summer basis are the stressors on renewable water quantity (parameters 12–13). Proxies for water quality are the number of Combined Sewer Overflow (CSO) per unit length of water body expressed in km, and volume fraction of treated wastewater with respect to permitted water abstraction (parameters 11 and 14). Nutrients load from CSO and treated wastewaters is not here explicitly accounted for. In the case of CSO the information is not available due to the occasional activation of these systems, for treated wastewaters the limits for concentration of nutrients at the outlet are prescribed by law, thus we consider the treated volume as driving variable.

2.2. ML algorithm and predictive performances

The final dataset is composed by the 14 parameters calculated in the catchments (Table 2) and the class attribute of the ecological status assigned to rivers at the monitoring stations (ECO_St). As a preliminary analysis, the Spearman's rank correlation coefficients between all pairs of

Table 1
GIS data extracted over the catchments vector layer.

GIS original data	Description
DTM	Digital Terrain Model 10 × 10
Ecological status (ECO_St)	Classification 1-high, 2-good, 3-moderate, 4-poor, 5-bad (2019–2020)
Land use land cover (LULC)	Level 1 (artificial surfaces, agricultural areas, forest and semi-natural areas)
Annual and monthly precipitation (P_y, P_m)	Raster map based on a twenty years' time series (1999–2019)
Monthly temperatures (Tm)	Raster map based on a twenty years' time series (1999–2019)
Linear hydraulic structures (LHS)	Polyline data including dikes, levees etc.
Point hydraulic structure (PHS)	Point data including weirs, abstraction structures, sluice gates, spillway, groynes etc.
Combined Sewer Overflow (CSO)	Point data with occasional untreated discharges
Permitted Water Abstraction (PWA)	Point data with annual volumes withdrawn
Treated Wastewater Volume (WWV)	Point data with annual volumes discharged from wastewater treatment plants

Table 2
Elaborated and normalized parameters subdivided per stressor typology for the creation of the dataset.

Type of stressor	Parameters		Elaboration	Unit of measurement
Catchment characteristics	1	Catchment area *A*	–	km^2
	2	Elevation of catchment outlet *E*	–	m a.s.l.
	3	Agricultural surface *Agr%*	A_{agr}/A	%
	4	Artificial surface *Urb%*	A_{urb}/A	%
	5	Forest and semi-natural areas *For%*	A_{for}/A	%
Climate	6	Mean annual precipitation in the catchment P_{y_mean}	$\frac{1}{n\ cells}\cdot\sum_{j=1}^{n\ cells} P_{y,j}$	mm
	7	Min. summer precipitation P_{s_min} in the catchment (July *j*, August *a*, September *s*)	$min(P_{j,\,k} + P_{a,\,k} + P_{s,\,k})$ $k = 1, \ldots, n\ cells$	mm
	8	Maximum summer temperature T_{s_max} in the catchment (July *j*, August *a*, September *s*)	$max(T_{i,\,k}), i = j, a, s; k = 1, \ldots, n\ cell$	°C
Morphology	9	Density of linear hydraulic structures *DLHS*	L_{LHS}/L_{river}	km/km
	10	Density of point hydraulic structures *DPHS*	PHS/L_{river}	Number/km
Water resources management	11	Density of Combined Sewer Overflow *DCSO*	CSO/L_{river}	Number/km
	12	Permitted Water Abstraction/Precipitation (annual) *PWAP*	$PWA/(P_{y\_mean} \cdot A)$	%
	13	Permitted Water Abstraction/Precipitation (summer) *PWAP_S*	$(0.25 \cdot PWA)/(P_{s\_min} \cdot A)$	%
	14	Treated Water Fraction *TWF*	WWV/PWA	%

variables are calculated to detect statistically significant relationship between ecological status (5 classes) and stressors, and between stressors. Moreover, the canonical correlation analysis, a widely adopted technique to find linear combinations of vectors which show maximum correlation with each other (Hardoon et al., 2003) is performed to act as benchmark for ML classification algorithms.

Predicting classes is one of the most common supervised learning tasks (Géron, 2019). In order to simplify the classification problem into a binary classification, the ecological status of river is transformed into two classes: (i) at least good, for those rivers classified as high or good, and (ii) less than good, for those rivers classified as moderate, poor or bad, sensu WFD. The dataset is randomly split into a training set (66 % of the instances) and a test set (34 % of the instances). The sensitivity with respect to different randomly selected training and test set has been verified by repeating the calibration-validation exercise over different training and test set to avoid the risk of overfitting.

Five classification algorithms are considered to compare their predictive ability in the WEKA (Waikato Environment for Knowledge Analysis) environment (Frank et al., 2016): Naïve Bayes (John and Langley, 1995), Random Forest (Breiman, 2001), J48 classification tree (Quinlan, 1993), Logistic regression (Strickland, 2017) and K-nearest neighbour (Aha and Kibler, 1991). The performances of the classifiers are evaluated by means of precision, recall and F-measure (Géron, 2019; Chen et al., 2020) obtained by the evaluation of the confusion matrix. Precision is the ratio between true positive TP and the sum of true positive and false positive FP.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall is the ratio of positive instances that are correctly detected by the classifier

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where FN is the number of false negative.

F-measure is the harmonic mean of precision and recall

$$\text{F} = \frac{\text{TP}}{\text{TP} + \frac{\text{FN}+\text{FP}}{2}} \tag{3}$$

Finally an analysis of the most significant parameters for ecological status classification is carried out through algorithms that iteratively create all possible subsets from the feature vector and then use a classification algorithm to assess which subset performs the best (Hall, 1998). Predictive performances with a reduced number of attributes can be then compared to the classification with all 14 parameters. This final analysis provides key indications for the application of the proposed methodology in regions where not all the 14 parameters considered here may be available, or for designing data collection/retrieval experiments aimed at ecological status prediction.

2.3. Case study

The study area is the Tuscany Region located in central Italy (Fig. 2, panel a). The surface area is c. 23,000 km^2 and the population is 3.7 million approximately. The northern boundary is characterized by mountains, with an altitude of the order of 1000–1500 m a.s.l., the western part is bounded by the Tyrrenian Sea. Climatic conditions are semiarid in southern coastal areas and perhumid in the northern mountainous regions. Annual average temperatures are irregularly distributed within the study area, ranging from 8 °C in the northern mountain peaks to 17 °C in southern coastal area. The average annual precipitation is 1190 mm (min. 618 mm, max. 2748 mm at point rainfall gauges), with a high variability with respect to season and elevation. The mountains and the southern portion of the region have a limited population density. The population is mostly concentrated within the Arno River catchment in the metropolitan area of Florence, including the municipalities of Prato and Pistoia (ca. 1.2 million inhabitants) and the province of Pisa (ca. 450,000 inhabitants).

The Regional environment agency (ARPAT, 2021) monitors 360 points within rivers in the area (violet points in Fig. 2, panel b). The five indicators used for ecological status classification are benthic macroinvertebrates, macrophytes, benthic diatoms, LIMeco (dissolved oxygen, phosphorus, ammonium, nitrate) and the concentration of selected hazardous substances (according to Italian Law *d.lgs. 172/2015*). According to the monitoring results, WFD objectives are achieved for 215 river catchments whose ecological status is classified at least good (green polygons in Fig. 2, panel b), while 145 rivers fail to achieve the WFD objectives (orange polygons in Fig. 2, panel b). The corresponding 360 catchments identified upstream of the ecological monitoring sites have a surface area ranging from 1 to 8122 km^2, with an average size of 205 km^2 and median size of 33.6 km^2 (Fig. 2, panel c). The Arno River has the largest river basin in the region with a surface of 8200 km^2. The Tuscany Region cartographic data portal and the regional hydrologic service provide the data listed in Table 1.

3. Results and discussion

3.1. Correlations between ecological status and stressors

Table 3 shows Spearman's correlation coefficients between all pairs of variables considering all the 5 classes of ecological status (ECO_St). Non-statistically significant correlations between pairs are marked as *nss* using a p threshold of 0.001. ECO_st appears correlated to all land use classes, i.e., Urb%, Agr%, For%, and is negatively correlated to For% that means that ecological quality decreases with the decrease of forest and natural

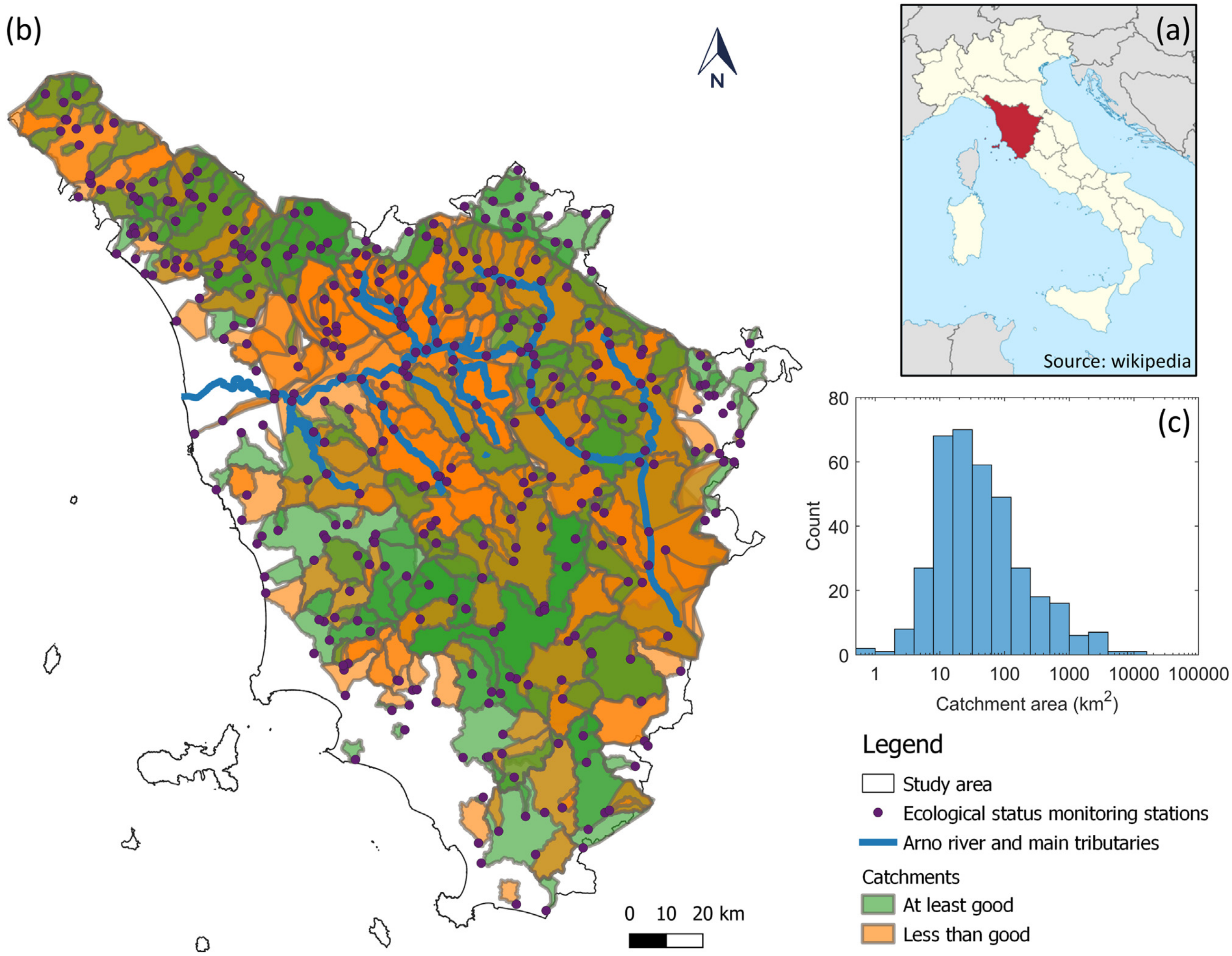


Fig. 2. Setting of the study area in central Italy (a). Catchments identified upstream each ecological monitoring station and their compliance with respect to WFD objectives at least good (green) and less than good (orange) (b), frequency histogram of catchment surface areas (c).

areas cover. Agricultural and artificial surfaces are instead positively correlated showing a detrimental effect of both Urb% and Agr% on ecological quality. ECO_St also shows statistically significant correlations with climate parameters especially with P_{s_min} and T_{s_max} highlighting the presence of worse ecological status of rivers in catchments with drier and warmer summer conditions. Elevation is correlated to ECO_St, For% and all climate parameters (especially temperatures). Some works considered elevation as a proxy of anthropization (Larsen et al., 2019) and also for our study area it is possible to notice a Spearman's correlation coefficient of −0.45 between E and Urb% (see also Fig. 3 panel c). In fact, (i) the elevation is also negatively correlated with DLHS, i.e., increased morphological alterations with linear infrastructures (e.g., levees) in floodplains, and (ii) positively correlated with DPHS, i.e., higher density of point infrastructures (e.g., weirs) in the mountains. The density of point hydraulic structures is slightly correlated to ECO_St, no correlation exists with the density of CSO, however, a non-negligible correlation is found with the density of linear hydraulic structures DLHS. On the side of water management, the ratio between permitted water abstraction and precipitation, both in annual and summer conditions are positively correlated to ECO_St showing, as largely expected, a decreasing ecological status with increasing water abstraction. Nevertheless, it is worth mentioning that some authors found some biological indicators, used in ecological status classification, unsuitable for detecting specific hydrologic pressures highlighting gaps that should be addressed in future (Larsen et al., 2019; Arrighi et al., 2021). The positive correlation of ECO_St with the catchment area A highlights a decrease in ecological status with increasing area, that can be interpreted in terms of combination of multiple pressures occurring upstream creating impacts in downstream areas due to advection and bio-chemical reactions.

Fig. 3 shows some significant catchment parameters for the ecological status. Panel (a) represents the five classes of ecological status (1-High to 5-Bad) versus minimum summer precipitation and maximum summer temperature. It shows a quite evident trend of decreasing ecological status towards drier and warmer climatic conditions, i.e., moving towards the bottom right part of the plot, which are concerning for future climate scenarios. Particularly, poor and bad ecological statuses appear clustered in the area with $P_{S\_min} < 200$ mm and $T_{S\_max} > 27$ °C. Panels (b–c) represent the five classes of ecological status vs elevation and land use, forest and artificial cover, respectively. Higher elevations and less artificial land cover in favour of forest land cover appear also quite important to determine a good ecological status.

The canonical correlation analysis returns a canonical variable *V* as a linear combination of the 14 attributes. The coefficients of the linear combination are shown in the Supplementary Table 1. The higher the linear combination coefficients obtained, the greater the influence that the single variable has in explaining ECO_St. Again, the canonical correlation analysis confirms the importance of land use, summer maximum temperature and permitted water abstraction, which show the higher order of magnitude of the coefficients. The correlation coefficient between the variables ECO_St and the canonical variable V is equal to 0.62, it constitutes the benchmark for ML algorithms.

Table 3
Rho Spearman's correlation coefficients between all pairs of variables. Positive and negative correlations are highlighted in yellow and blue shades respectively with a linear scale of four intervals with colors darkening with higher values.

	ECO_St	A	Urb%	Agr%	For%	P_{y_mean}	P_{s_min}	T_{s_max}	DPHS	E	DLHS	DCSO	TWF	PWAP	PWAP_S
ECO_St	1.00	0.44	0.44	0.44	-0.53	-0.25	-0.34	0.47	0.28	-0.45	0.42	nss	0.32	0.31	0.31
A		1.00	0.20	0.30	-0.40	-0.20	-0.45	0.54	0.35	-0.48	0.49	0.42	0.55	0.34	0.35
Urb%			1.00	0.56	-0.61	-0.30	-0.30	0.45	0.27	-0.45	0.39	0.22	0.30	0.32	0.30
Agr%				1.00	-0.88	-0.74	-0.67	0.59	nss	-0.47	0.19	nss	0.29	nss	nss
For%					1.00	0.68	0.69	-0.66	-0.20	0.61	-0.34	nss	-0.35	nss	nss
P_{y_mean}						1.00	0.84	-0.54	nss	0.38	nss	nss	nss	0.22	0.26
P_{s_min}							1.00	-0.68	nss	0.56	-0.18	nss	-0.23	nss	nss
T_{s_max}								1.00	0.26	-0.71	0.37	0.20	0.35	0.18	nss
DPHS									1.00	-0.25	0.59	0.21	0.22	0.33	0.34
E										1.00	-0.44	nss	-0.36	-0.33	-0.33
DLHS											1.00	0.29	0.36	0.46	0.46
DCSO												1.00	0.23	0.24	0.25
TWF													1.00	0.22	0.22
PWAP														1.00	1.00
PWAP_S															1.00

The significance of land use parameters in the indices determining the river ecological status confirms previous studies (Valerio et al., 2021; Lemm et al., 2021; Grizzetti et al., 2017; Molina-Navarro et al., 2020). However, Spearman's rho of land use parameters is slightly higher than in Grizzetti et al. (2017) (0.44 for Agr% here with respect to 0.29) and higher than in Lemm et al., 2021 (0.44 for Agr% and URB% here with respect to 0.23 and 0.18 respectively). Higher Spearman's rho in our study can be due to the smaller number of observations and on the limited regional variability of catchment characteristics.

The scarce correlation of ECO_St and the presence of point hydraulic structures can be a consequence of the scarce pressures from river barriers with respect to other EU countries (Belletti et al., 2020; European Environment Agency, 2020), while linear morphological alterations, such as levees or river bed protections appear quite significant. However, summer maximum temperatures and summer precipitations haven't emerged as significant drivers in previous studies, which focused on different geographical areas. Only Molina-Navarro et al. (2020) identified mean annual temperature as a stressor for ecological status. The study by Grizzetti et al. (2017) is the only one investigating the role of water demand in ecological quality, although measured in absolute terms (mm per day) rather than in comparison with precipitation. In that analysis the significance of water demand on ecological status was less clear with respect to land use and nutrients concentration. Here instead, the correlation of ECO_St with the stressors related to permitted water exploitation and treated wastewater (TWF, PWAP and PWAP_S) appears significant.

With respect to the canonical correlation analysis, which to the authors' knowledge has been rarely applied to predict ecological status or biological indicators, it is possible to say that the interpretability of the method, as defined by Visser et al. (2022) is high and can be preferred by decision makers and stakeholders, with respect to some ML methods whose structure resembles to a *black box*. Nevertheless, a compromise between interpretability of model structure and predictive capabilities should be carefully analysed.

3.2. ML classification performances

With ECO_St transformed into two classes (*at least good* and *less than good*), the ML classification algorithms are calibrated on the training set and provide an estimated predictive ability on the validation set, as shown in Table 4.

All the ML classification algorithms have a better performance with respect to the correlation coefficient obtained by the canonical correlation analysis, that we consider here as a benchmark. All the applied classification algorithms perform better in terms of precision in predicting a status of at least good. This can be interpreted in terms of difficulties in predicting the effects of multiple climate and anthropic pressures. Overall, the best classification performance is obtained by Random Forest approach with an 80 % precision, recall and F-measure. An accuracy of the order of 70–90 % was also found in the case study of Ebro river catchment in predicting macroinvertebrates, diatoms and macrophytes indices (Valerio et al., 2021). Random forest algorithm confirms to be one of the most accurate predictors also in similar studies (Chen et al., 2020; Visser et al., 2022; Grizzetti et al., 2017).

In order to better understand what the reasons for a bad prediction are, the catchments used as validation set are mapped and compared with the actual ecological status assessed by ARPAT. The results are shown in Fig. 4. The catchments in light blue and pink are those correctly predicted by the Random Forest classifier, the light blue ones and the pink ones correspond to at least good and less than good rivers respectively. The dark blue polygons are predicted at least good by the model but assessed less than good. They are mostly located in the north western part of the region with mid altitudes, high summer precipitation and forest land cover, however the analysis of chemical substances, sensu WFD, reveals in these river the presence of mercury, lead and polybrominated diphenylethers (pBDEs), which are recognized in EU as the main responsible of failure in achieving good chemical status (Posthuma et al., 2020; European Environment Agency, 2018). Mercury compounds contaminate

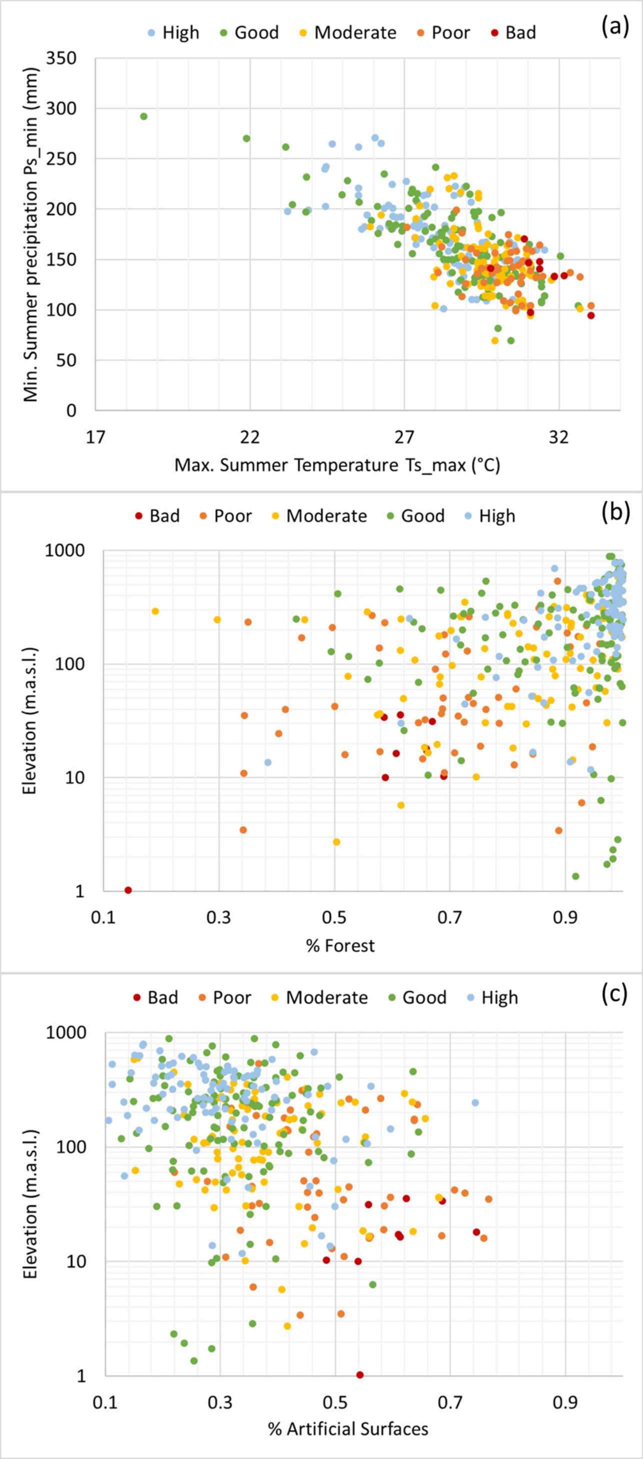


Fig. 3. Ecological status of rivers with respect to climate and land use conditions of the upstream catchment.

waters through atmospheric deposition (from energy sector, industrial processes, cement manufacturing etc.), thus their presence is not easily guessed by land use in the catchment. Urban settlements are the main sources of pBDEs which are persistent in sediments and soils and have harmful effect on endocrine system of aquatic organisms, adversely affecting reproduction and growth (European Environment Agency, 2018). The presence of pollution sources which are not directly linked to the characteristics of the catchments, such as those from atmospheric deposition, make weaker the assumption that is possible to predict ecological status with a certain degree of accuracy with a few geospatial information, because our selected parameters are not able to capture this phenomenon. This aspect should be addressed in future works to possibly identify different parameters or more appropriate combinations of parameters for the ML models and by better understanding overfitting issues and most sensitive parameters for model training.

The purple polygons, predicted less than good but actually at least good, are located in the southern part of the region in the two provinces of Siena and Grosseto. Here the reasons for a wrong classification could reside in the large agricultural land use, which however counts a significant amount of organic farming (ca. 30 % of the agricultural area) (Regione Toscana, 2006, 2012). However, the exact geographic location of organic agricultural areas is currently not available as geospatial data. The availability of this type of geospatial data could further improve the predictive capabilities of ML classifiers. Overall, ML learning classifiers perform well when the available data for training and testing are enough and have a good quality. The predictive abilities of the ML algorithms might be enhanced by (i) an improved quality of the geospatial data used for training and testing, e.g., organic agricultural surfaces, and (ii) the adoption of different parameters more capable to capture pressures on river ecology, (iii) an enlargement of the sample size.

The analysis of the most significant parameters for ecological status classification, yields as results the following 7 attributes: Agr%, For%, P_{y_mean}, P_{s_min}, T_{s_max}, DCSO, and PWAP. The application of the ML algorithm to this attribute subset yields the results shown in the Supplementary Table 2.

With a reduced number of attributes, the best predictive precision and F-measure reduces to 0.72 and 0.71, with respect to 0.8 of the whole set of attributes. The Naïve Bayes classifiers performs slightly better than Random Forest in this case. The precision is again better in predicting at least good instances rather than less than good. The performance of the ML classifiers, although reduced, are still better than the correlation coefficient obtained by the canonical correlation analysis ($r = 0.62$).

Complex biological, hydrological, morphological, and climatological interactions occur in rivers and determine ecological quality. As a result of this complexity, modelling ecological status and ecological parameters such as diatoms, macrophytes or invertebrates is extremely difficult at large scales, i.e., river district scale. Nevertheless, strategic measures are identified and prioritized at such large scales. ML techniques, which demonstrated in this study good predictive capabilities might be a useful tool for large scale planning of water resources.

4. Conclusions

This work aimed at understanding if the overall ecological status of river can be predicted with a limited, easily accessible amount of

Table 4
Ability to predict river ecological status of the ML classification algorithms with all 14 attributes.

Algorithm	Precision (at least good)	Recall (at least good)	F measure (at least good)	Precision (*less than good*)	Recall (*less than good*)	F measure (*less than good*)	Precision (weighted avg.)	Recall (weighted avg.)	F measure (weighted avg.)
Naïve Bayes	0.86	0.63	0.73	0.60	0.83	0.70	0.75	0.71	0.72
J48	0.82	0.54	0.65	0.53	0.81	0.65	0.71	0.65	0.65
Random forest	0.83	0.84	0.84	0.76	0.75	0.75	0.80	0.80	0.80
Logistic regression	0.73	0.70	0.72	0.57	0.60	0.59	0.67	0.66	0.67
K-Nearest Neighbours	0.79	0.66	0.72	0.58	0.73	0.65	0.71	0.69	0.69

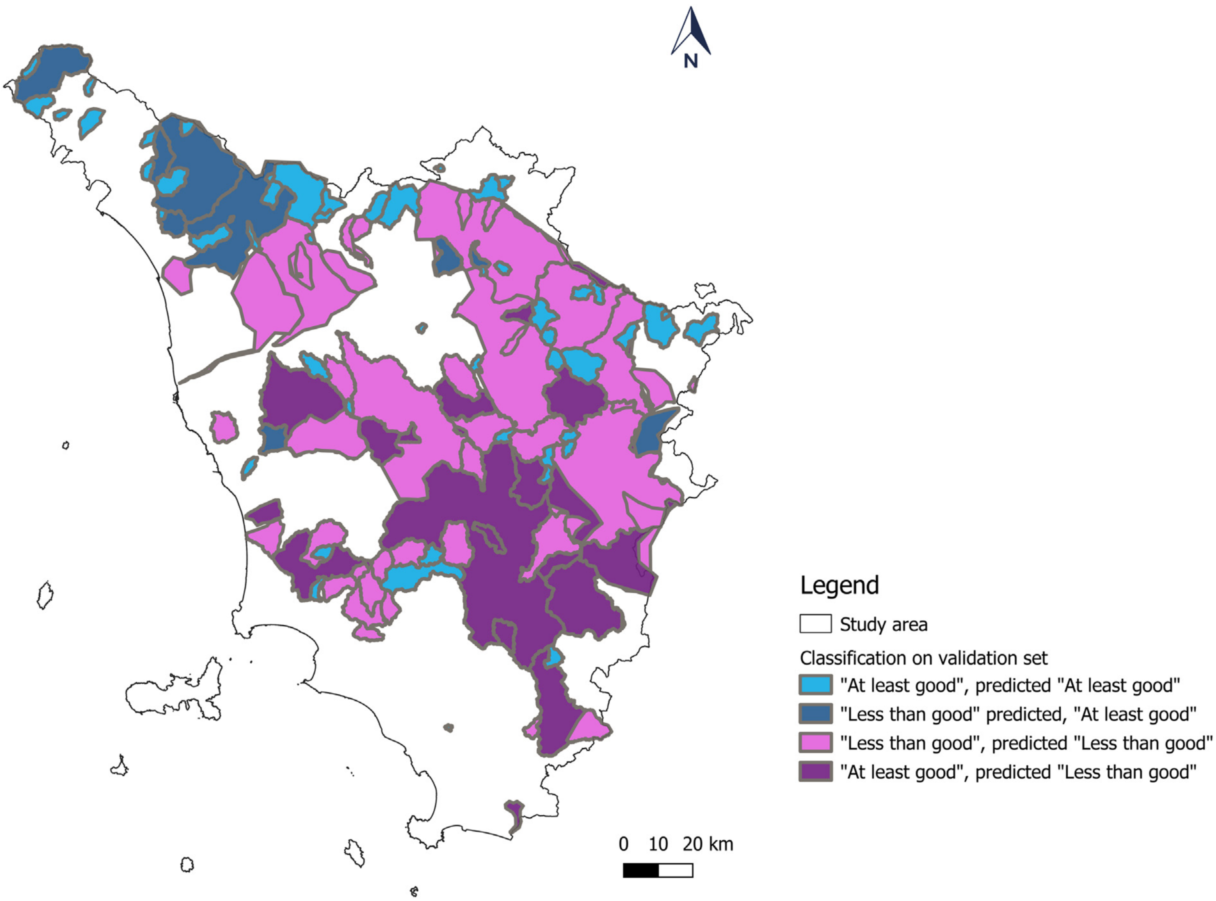

N
Legend
Study area
Classification on validation set
"At least good", predicted "At least good"
"Less than good" predicted, "At least good"
"Less than good", predicted "Less than good"
"At least good", predicted "Less than good"
0 10 20 km

Fig. 4. Spatial map of the confusion matrix obtained by the Random Forest algorithm in predicting ecological status of rivers on the validation set.

geospatial data related to land use, climate, water management and morphological alterations.

In the study area, ecological status has been found correlated to land use, but also to minimum summer precipitation and maximum summer temperature bringing to the fore an important aspect for Mediterranean climate, that was not highlighted before and raises some concerns related to future climate scenarios. Moreover, the ratio between permitted water abstraction and precipitation has been found correlated to ecological status highlighting the role of water exploitation on river ecology.

Random Forest algorithm was the best classifier with 80 % precision and F-measure, followed by Naïve Bayes classifier (F-measure = 0.72). All classifiers performed better in predicting at least good status, highlighting the difficulties in understanding multiple stressors interactions. The performances of ML classification algorithms are better than those obtained by a canonical correlation analysis, here used as a benchmark.

The spatial analysis of prediction error highlighted a potentially high significance of organic farming in reducing nutrients-related pressures on rivers. Unfortunately, a distinction of agricultural land use in "conventional" and "organic" was not possible with the currently available spatial data. Moreover, the classifiers failed in predicting some less than good ecological statuses where some chemical substances, i.e., mercury and pBDEs were detected by the environmental authority. This raise some concerns especially related to the atmospheric deposition of mercury compounds that cannot be easily predicted with the spatial information used in this study. Overall, this work contributed to the understanding of pressures determining ecological status of rivers and demonstrated a pretty good capability of ML classifiers in predicting ecological status that can help in identifying appropriate ecological flows, mitigation measures and water management practices at large scales.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.159655.

Funding

This work was supported by the Hydrographic District of the Northern Appennines, Florence.

CRediT authorship contribution statement

Chiara Arrighi was responsible of conceptualization, data curation, methodology, validation and visualization of results. She also wrote the original draft. Fabio Castelli supervised the work and acquired the funding. Both authors worked on manuscript review and editing.

Data availability

Data will be made available on request.

Declaration of competing interest

Fabio Castelli reports financial support was provided by Autorità di Bacino Distrettuale dell'Appennino Settentrionale.

References

Aha, D., Kibler, D., 1991. Instance-based learning algorithms. Mach. Learn. 6, 37–66.

ARPAT, 2021. Monitoraggio Ambientale Dei Corpi Idrici Superficiali: Fiumi, Laghi, Acque Di Transizione. Risultati Parziali Secondo Anno Monitoraggio Triennio 2019-2021. last access 9/8/22. http://www.arpat.toscana.it/documentazione/catalogo-pubblicazioni-arpat/monitoraggio-ufficiale-delle-acque-superficiali/monitoraggio-ambientale-dei-corpi-idrici-superficiali-fiumi-laghi-acque-di-transizione-risultati-2020.

Arrighi, Chiara, Bonamini, Isabella, Simoncini, Cristina, Bartalesi, Stefano, Castelli, Fabio, 2021. WFD ecological quality indicators are poorly correlated with water levels in river catchments in Tuscany (Italy). Hydrol. 8 (4), 1–9. https://doi.org/10.3390/hydrology8040185.

Azzellino, A., Antonelli, M., Canobbio, S., Çevirgen, S., Mezzanotte, V., Piana, A., Salvetti, R., 2013. Searching for a compromise between ecological quality targets, and social and ecosystem costs for heavily modified water bodies (HMWBs): the lambro-seveso-olona system case study. Water Sci. Technol. 68 (3), 681–688. https://doi.org/10.2166/wst.2013.277.

Baudin, Jean-Christophe, 2020. "ClosestPoint 4.0.1." Code Repository. last access 9/20/22 https://github.com/tomflyjc/CLOSEST-POINT-PLUGIN.

Belletti, Barbara, Garcia, Carlos, de Leaniz, Joshua, Jones, Simone Bizzi, Börger, Luca, Segura, Gilles, Castelletti, Andrea, et al., 2020. More than one million barriers fragment Europe's Rivers. Nat. 588 (7838), 436–441. https://doi.org/10.1038/s41586-020-3005-2.

Breiman, Leo, 2001. Random Forest. Mach. Learn. 45 (1), 5–32.

Bytyqi, Pajtim, Czikkely, Marton, Shala-Abazi, Albona, Fetoshi, Osman, Ismaili, Murtezan, Hyseni-Spahiu, Mimoza, Ymeri, Prespa, Kabashi-Kastrati, Edona, Millaku, Fadil, 2020. Macrophytes as biological indicators of organic pollution in the Lepenci River basin in Kosovo. J. Freshw. Ecol. 35 (1), 105–121. https://doi.org/10.1080/02705060.2020.1745913.

Carvalho, Laurence, Mackay, Eleanor B., Cardoso, Ana Cristina, Baattrup-Pedersen, Annette, Birk, Sebastian, Blackstock, Kirsty L., Borics, Gábor, et al., 2019. Protecting and restoring Europe's waters: an analysis of the future development needs of the water framework directive. Sci. Total Environ. 658, 1228–1238. https://doi.org/10.1016/j.scitotenv.2018.12.255.

Chen, Kangyang, Chen, Hexia, Zhou, Chuanlong, Huang, Yichao, Qi, Xiangyang, Shen, Ruqin, Liu, Fengrui, et al., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res. 171, 115454. https://doi.org/10.1016/j.watres.2019.115454.

EEA, 2018. European Waters - Assessment of Status and Pressures 2018. last access 8/9/2022. https://www.eea.europa.eu/publications/state-of-water/download.

EU Parliament, 2000. Directive 2000/60/EC. Off. J. Eur. Commun. 21. https://doi.org/10.1039/AP9842100196.

European Commission, 2016. Ecological Flows in the Implementation of the Water Framework Directive: Guidance Document N°31. last access 9/8/22. https://circabc.europa.eu/sd/a/4063d635-957b-4b6f-bfd4-b51b0acb2570/Guidance%20No%2031%20-%20Ecological%20flows%20%28final%20version%29.pdf.

European Environment Agency, 2018. Chemicals in European Waters. Knowledge Developments. last acces 9/8/22. https://www.eea.europa.eu/publications/chemicals-in-european-waters/download.

European Environment Agency, 2020. Tracking barriers and their impacts on European river ecosystems. Briefing 30/2020 - Water and Marine Environment, pp. 1–9.

Frank, E., Witten, I., Hall, M.A., 2016. Data mining: practical machine learning tools and techniques. Morgan Kaufmann. https://doi.org/10.1016/C2009-0-19715-5.

Géron, Aurelian, 2019. Hands-on machine learning with scikit-learn, keras and TensorFlow. 2nd Editio. O'Reilly Media, Sebastopol.

Grizzetti, B., Pistocchi, A., Liquete, C., Udias, A., Bouraoui, F., Van De Bund, W., 2017. Human pressures and ecological status of european Rivers. Sci. Rep. 7 (1), 1–11. https://doi.org/10.1038/s41598-017-00324-3.

Hall, M.A., 1998. Correlation-based feature subset selection for machine learning. Hamilton, New Zealand.

Hardoon, David R., Szedmak, Sandor, Shawe-taylor, John, 2003. Canonical correlation analysis; an overview with application to learning methods. Sci. 16.

Herrero, Albert, Gutiérrez-Cánovas, Cayetano, Vigiak, Olga, Lutz, Stefanie, Kumar, Rohini, Gampe, David, Huber-García, Verena, Ludwig, Ralf, Batalla, Ramon, Sabater, Sergi, 2018. Multiple stressor effects on biological quality elements in the Ebro River: present diagnosis and predicted responses. Sci. Total Environ. 630, 1608–1618. https://doi.org/10.1016/j.scitotenv.2018.02.032.

John, G.H., Langley, P., 1995. Estimating continuous distributions in bayesian classifiers. Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345.

Larsen, Stefano, Bruno, Maria Cristina, Zolezzi, Guido, 2019. WFD ecological status indicator shows poor correlation with flow parameters in a large Alpine catchment. Ecol. Indic.. 98, pp. 704–711. https://doi.org/10.1016/j.ecolind.2018.11.047 (September 2018)

Lemm, Jan U., Venohr, Markus, Globevnik, Lidija, Stefanidis, Kostas, Panagopoulos, Yiannis, van Gils, Jos, Posthuma, Leo, et al., 2021. Multiple stressors Determine River ecological status at the european scale: towards an integrated understanding of river status deterioration. Glob. Chang. Biol. 27 (9), 1962–1975. https://doi.org/10.1111/gcb.15504.

Lu, Xinxin, Liu, Yan, Fan, Yawen, 2020. Diatom taxonomic composition as a biological indicator of the ecological health and status of a River Basin under agricultural influence. Water (Switzerland) 12 (7). https://doi.org/10.3390/w12072067.

Molina-Navarro, Eugenio, Segurado, Pedro, Branco, Paulo, Almeida, Carina, Andersen, Hans E., 2020. Predicting the ecological status of rivers and streams under different climatic and socioeconomic scenarios using Bayesian belief networks. Limnologica 80, 125742. https://doi.org/10.1016/j.limno.2019.125742 (May 2019).

Poikane, Sandra, Herrero, Fuensanta Salas, Kelly, Martyn G., Borja, Angel, Birk, Sebastian, van de Bund, Wouter, 2020. European aquatic ecological assessment methods: a critical review of their sensitivity to key pressures. Sci. Total Environ. 740, 140075. https://doi.org/10.1016/j.scitotenv.2020.140075.

Posthuma, Leo, Zijp, Michiel C., De Zwart, Dick, Van de Meent, Dik, Globevnik, Lidija, Koprivsek, Maja, Focks, Andreas, Van Gils, Jos, Birk, Sebastian, 2020. Chemical pollution imposes limitations to the ecological status of european surface waters. Sci. Rep. 10 (1), 1–12. https://doi.org/10.1038/s41598-020-71537-2.

Quinlan, Ross, 1993. C4.5: Programs for machine learning. Morgan Kaufman, San Mateo, California. ISBN 1-55860-238-0.

Regione Toscana, 2006. Il Biologico in Toscana. last access 10/8/22. https://www.regione.toscana.it/documents/10180/23930/Guida%2C + il + biologico + in + Toscana.pdf/ea0b5c53-e18d-4e3c-9087-08d411646da7?version = 1.0&t = 1352200403397&download = true.

Regione Toscana, 2012. La Toscana AL 6 ° Censimento Generale Dell'agricoltura. last access 10/8/22. https://www.regione.toscana.it/documents/10180/320308/La + Toscana + al + 6. + Censimento + generale + dell%27agricoltura/937e306c-2408-41f0-ac69-fb44617867fb?version = 1.0.

Reid, Andrea J., Carlson, Andrew K., Creed, Irena F., Eliason, Erika J., Gell, Peter A., Johnson, Pieter T.J., Kidd, Karen A., et al., 2019. Emerging threats and persistent conservation challenges for freshwater biodiversity. Biol. Rev. 94 (3), 849–873. https://doi.org/10.1111/brv.12480.

Strickland, Jeffrey, 2017. Logistic regression inside-out. Lulu Press Inc., Morrisville, USA 978-1-365-81915-5.

Valerio, Carlotta, De Stefano, Lucia, Martínez-Muñoz, Gonzalo, Garrido, Alberto, 2021. A machine learning model to assess the ecosystem response to water policy measures in the Tagus River basin (Spain). Sci. Total Environ. 750, 141252. https://doi.org/10.1016/j.scitotenv.2020.141252.

Visser, Hans, Evers, Niels, Bontsema, Arjan, Rost, Jasmijn, de Niet, Arie, Vethman, Paul, Mylius, Sido, et al., 2022. What drives the ecological quality of surface Waters? A review of 11 predictive modeling tools. Water Res. 208, 117851. https://doi.org/10.1016/j.watres.2021.117851.

Voulvoulis, Nikolaos, Arpon, Karl Dominic, Giakoumis, Theodoros, 2017. The EU water framework directive: from great expectations to problems with implementation. Sci. Total Environ. 575, 358–366. https://doi.org/10.1016/j.scitotenv.2016.09.228.