Agrociencia Uruguay 2023 | Volume 27 | Number NE1 | Article e1192
DOI: 10.31285/AGRO.27.1192
ISSN 2730-5066

Advances in Water in Agroscience

Assessing dependence between land use/land cover and water quality: A comparison at a small and a large watershed in Uruguay

Evaluación de la dependencia entre el uso/cobertura del suelo y la calidad del agua: comparación entre una cuenca pequeña y una grande en Uruguay

Avaliação da dependência entre uso/cobertura do solo e qualidade da água: comparação entre uma pequena e uma grande bacia no Uruguai

Cal, A.[1]; Pastorini, M.[2]; Tiscornia, G.[1]; Rivas-Rivera, N.[3]; Gorgoglione, A.[4]

[1]*Instituto Nacional de Investigación Agropecuaria (INIA), Sistemas de información y Transformación Digital (GRAS), Las Brujas, Canelones, Uruguay*
[2]*Universidad de la República, Facultad de Ingeniería, Instituto de Computación (InCo), Montevideo, Uruguay*
[3]*Universidad de la República, Facultad de Ciencias, Instituto de Ecología y Ciencias Ambientales (IECA), Montevideo, Uruguay*
[4]*Universidad de la República, Facultad de Ingeniería, Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA), Montevideo, Uruguay*

Editor
Álvaro Otero
Instituto Nacional de Investigación Agropecuaria (INIA), Salto, Uruguay

Received *09 May 2023*
Accepted *04 Oct 2023*
Published *06 Feb 2024*

Correspondence
Angela Gorgoglione
agorgoglione@fing.edu.uy

Abstract

Changes in land use/land cover (LULC) directly or indirectly affect water quality in watercourses and impoundments. Sustainable management strategies aimed to enhance ecosystem health and community well-being require an accurate water-quality evaluation. This study looks into the correlation between temporal changes in LULC, represented by selected landscape variables (land cover area and proportion, patch density, Euclidean nearest-neighbor distance, mean shape index, and Shannon index), and water quality variables (nitrate, total phosphorus, and total suspended solids) at catchment scale. To compare the watershed-size influence, this analysis was performed at two different spatial scales represented by two Uruguayan basins of different sizes, San Salvador (3,118 km^2) and Del Tala (160 km^2). Partial Least Squares and Random Forest unsupervised machine-learning models were employed for this analysis. By exploiting a non-model-biased method based on game theory (SHAP), the LULC characteristics were quantified and ranked based on their level of importance in the water-quality evaluation. The main outcomes of this study proved that patch density is one of the most influencing metrics in both watersheds and for both models. Agricultural land use is the most critical one at both catchments and agricultural with a forage crop land uses are the most important ones for both algorithms. Furthermore, it is possible to state that the adopted techniques are valuable tools that can provide an adequate overview of the water-quality behavior in space and time and the correlations between water-quality variables and LULC.

Keywords: water quality, land use/land cover, unsupervised learning, feature importance

Resumen

Los cambios en el uso del suelo y la cobertura del suelo (LULC) afectan directa o indirectamente la calidad del agua en cursos de agua y embalses. Las estrategias de gestión sostenible destinadas a mejorar la salud del ecosistema y el bienestar de la comunidad requieren una evaluación precisa de la calidad del agua. Este estudio analiza la correlación entre los cambios temporales en LULC, representados por variables de paisaje seleccionadas (área y proporción de

Cal A, Pastorini M, Tiscornia G, Rivas-Rivera N, Gorgoglione Á. Assessing dependence between land use/land cover and water quality: a comparison at a small and a large watershed in Uruguay. Agrociencia Uruguay [Internet]. 2023 [cited dd mmm yyyy];27(NE1):e1192. Doi: 10.31285/AGRO.27.1192.

cobertura del suelo, densidad de parches, distancia euclidiana al vecino más cercano, índice de forma promedio e índice de Shannon), y las variables de calidad del agua (nitrato, fósforo total y sólidos suspendidos totales) a nivel de cuenca. Para comparar la influencia del tamaño de la cuenca, este análisis se realizó a dos escalas espaciales diferentes representadas por dos cuencas uruguayas de diferentes tamaños, San Salvador (3118 km^2) y Del Tala (160 km^2). Se emplearon modelos de aprendizaje automático no supervisados de Mínimos Cuadrados Parciales y Bosque Aleatorio para este análisis. Al aprovechar un método no sesgado basado en teoría de juegos (SHAP), las características de LULC se cuantificaron y clasificaron según su nivel de importancia en la evaluación de la calidad del agua. Los principales resultados de este estudio demostraron que la densidad de parches es una de las métricas más influyentes en ambas cuencas y para ambos modelos. El uso agrícola del suelo es crítico en ambas cuencas, y los usos agrícolas con cultivos forrajeros son los más importantes para ambos algoritmos. Además, es posible afirmar que las técnicas adoptadas son herramientas valiosas que pueden proporcionar una visión adecuada del comportamiento de la calidad del agua en el espacio y el tiempo, así como las correlaciones entre las variables de calidad del agua y LULC.

Palabras clave: calidad del agua, uso/cobertura del suelo, aprendizaje no supervisado, características relevantes

Resumo

Alterações no uso do solo/cobertura do solo (LULC) afetam diretamente ou indiretamente a qualidade da água em cursos d'água e reservatórios. Estratégias de gestão sustentável voltadas para melhorar a saúde do ecossistema e o bem-estar da comunidade requerem uma avaliação precisa da qualidade da água. Este estudo examina a correlação entre mudanças temporais no LULC, representadas por variáveis de paisagem selecionadas (área e proporção de cobertura do solo, densidade de manchas, distância euclidiana até o vizinho mais próximo, índice de forma média e índice de Shannon), e variáveis de qualidade da água (nitrato, fósforo total e sólidos suspensos totais) em escala de bacia hidrográfica. Para comparar a influência do tamanho da bacia hidrográfica, essa análise foi realizada em duas escalas espaciais diferentes, representadas por duas bacias uruguaias de tamanhos diferentes, San Salvador (3118 km^2) e Del Tala (160 km^2). Modelos de aprendizado de máquina não supervisionados de Mínimos Quadrados Parciais e Floresta Aleatória foram empregados para essa análise. Ao explorar um método não enviesado pelo modelo baseado na teoria dos jogos (SHAP), as características de LULC foram quantificadas e classificadas com base em seu nível de importância na avaliação da qualidade da água. Os principais resultados deste estudo mostraram que a densidade de manchas é uma das métricas mais influentes em ambas as bacias hidrográficas e para ambos os modelos. O uso agrícola da terra é crítico em ambas as bacias hidrográficas, e o uso agrícola com cultivo forrageiro é o mais importante para ambos os algoritmos. Além disso, é possível afirmar que as técnicas adotadas são ferramentas valiosas que podem fornecer uma visão adequada do comportamento da qualidade da água no espaço e no tempo e das correlações entre as variáveis de qualidade da água e LULC.

Palavras-chave: qualidade da água, uso/cobertura do solo, aprendizado não supervisionado, características relevantes

1. Introduction

In the last decades, agricultural production in the region has experienced rapid expansion and intensification[1-4] mainly for soybean production[5-6]. The predominant production systems depend on external inputs, particularly the addition of fertilizers and pesticides, as well as complementary irrigation to achieve an increase in productivity. These management strategies may lead to significant amounts of nitrogen and phosphorus (among other nutrients) being carried into rivers through runoff from agricultural lands. Approximately 20 % of nitrogen and 3 % to 20 % of phosphorus applied in agriculture are exported to surface waters, primarily due to erosion and leaching[7]. Hydrological factors, including land cover changes, storm water management, seasonal variations, river flow dynamics, sediment transport, and the influence of climate change, all play vital roles in the transport of these pollutants[8]. These factors affect the timing, magnitude, and fate of contaminants in aquatic ecosystems, making it crucial to consider them when addressing agricultural pollution. This condition can trigger eutrophication processes and affect ecosystem services such as clean water provision[9-10]. In Uruguay, since the beginning of 2000, most of the common cropping systems have included soybeans in sequences (winter crops/soybean) or intercrops (soybean/other summer crops) under no-tillage practices.

Based on these considerations, it is critical to evaluate the relationships between land use/land cover (LULC) patterns and water quality at a watershed scale to explain variations in river water quality and to assure watershed ecosystem management and water resource conservation.

Advanced spatial tools (like geographical information systems) along with machine-learning techniques

and water quality assessment methods make these types of analysis suitable. Bu and others[11] found that, during the dry season, built-up land use critically affected the concentrations of chloride, sulfate, and nutrients, which catted to the presence of possible point source pollution. During the wet season, built-up and agricultural land use had a significant influence on most water-quality variables. Moreover, all year long, the deterioration of water quality was positively associated with landscape metrics of Shannon's diversity index (SHDI), patch density (PD), and edge density (ED), and negatively related to mean shape index (SHMN), contagion (CONTAG), largest patch index (LPI), mean Euclidean nearest neighbor index (MENN), and cohesion index (COHE). Lee and others[12] demonstrated that, at the landscape level, a higher interspersion and number of land use types may worsen water quality. At the class level, the metrics ED, LPI, and PD played an important role in the relationship between land use and water quality. Xu and others[13] stated that the variation of water physicochemical variables during the wet and dry periods is better described by landscape pattern metrics at the watershed scale. Uuemaa and others[14] showed that for nutrients and biochemical oxygen demand (BOD), urban land use was the most important predictor. Particularly, for total nitrogen (TN), urban land use was accompanied by agricultural land use and ED. For BOD, PD was also a critical predictor.

Although many studies have correlated landscape patterns to water quality for different catchments worldwide, only a few studies have explored this relationship for sub-tropical streams in Southwestern America[15-16]. It is important to highlight that none of these studies evaluated the applicability of machine-learning techniques to this type of analysis.

With the aim of identifying the main landscape metrics that most affect each different water-quality variable at different watershed scales, we analyzed the relationships between seasonal water quality and landscape indexes at two Uruguayan watersheds characterized by different areas, using machine-learning techniques. The specific objective of our study is four-fold: i) assess the spatio-temporal variation of physicochemical variables of river waters; ii) evaluate the landscape composition; iii) analyze the impacts of landscape metrics on water physicochemical variables, and iv) evaluate the applicability of machine-learning techniques to this analysis.

2. Materials and methods

2.1 Study areas

To compare the watershed-size influence in the correlation between LULC changes and the water-quality variables, this study was carried out in two different sizes watersheds: Del Tala stream, and San Salvador river.

Del Tala stream watershed is in the northwest of Uruguay (between 31°14'21.9"S 57°22'48.7"W and 31°07'38.6"S 57°18'42.2"W) and is part of the Arapey Grande river watershed. The drainage area is 159 km^2, and its main course has a length of 23 km (Figure 1). It is characterized by its exclusively agricultural and livestock use, where agricultural activity occupies around 5,000 ha in the lower zone, extensive livestock farming of approximately 11,000 ha and some 25 ha are dedicated to intensive livestock (feedlot). In the 1990s, the main agricultural item was the rice crop, which was carried out in rotation with pastures. In the last 15 years, there has been a gradual diversification of agricultural rotation, including other summer crops (sorghum, corn and soybean). Given the characteristics of the basin, various research projects have been studying it for more than 10 years, which have provided historical information regarding land use, agronomic management, and water quality. The crop management practices include no-tillage and supplementary irrigation.

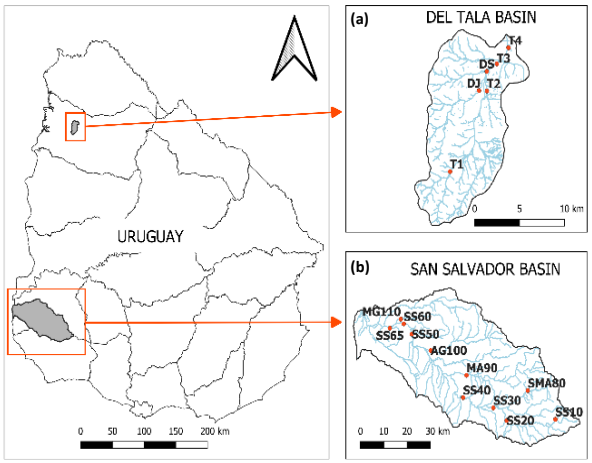


Figure 1. Study areas and water-quality sampling sites: a) Del Tala and b) San Salvador watersheds

San Salvador river watershed is in Soriano (between 33°51'27.1"S 57°30'05.8"W and 33°28'11.0"S 58°23'59.3"W). It has an area of 3,118 km^2, and the main course is 100 km long until it flows into the Uruguay river (Figure 1). It is a multi-use basin and can

be divided into three zones according to soil aptitude. The upper area is mainly used for livestock, the middle part is mainly an agricultural area, and the lower region is for urban use. Agricultural activity occupies 76 % of the watershed area, followed by natural grassland with 16 %, and forestry and native forest occupy 3 % and 4 %, respectively[17]. The main crops are soybeans, wheat, barley, corn, and sorghum. The area under irrigation is 1.3 % of the total surface[18]. Extensive and intensive cattle raising (feedlots) and an important activity of dairy farms are frequent in this catchment.

2.2 Water quality sampling collection and laboratory analysis

In the Del Tala watershed, thirteen sampling campaigns were carried out between February 2019 and December 2021 in six monitoring sites, four on the main course (T1, T2, T3, and T4) and two on tributaries (DJ and DS) (Figure 1). With the aim of having an adequate amount of data for the analysis, only DJ, DS, T3, and T4 stations were considered for this study.

In the same period, sixteen sampling campaigns were carried out in the San Salvador watershed. Five of them were executed by the research team and the remaining eleven correspond to data collected from the web database of the Monitoring Plan executed by the Ministry of Environment[19]. The monitoring sites were eleven, seven on the main course (SS10, SS20, SS30, SS40, SS50, SS60, and SS65) and four on tributaries (SMA80, MA90, AG100, MG110) (Figure 1).

In each sampling site, *in situ* pH, dissolved oxygen (DO) (mg/L), and temperature (T) (°C) were measured with a multiparametric probe. In addition, water samples were collected for analysis of fecal coliforms (FC) (UFC/100 mL), total suspended solids (TSS) (mg/L), biochemical oxygen demand (BOD) (mg/L), total phosphorus (TP), and nitrates (NO_3^-) (mg/L)[20-21]. Due to the main activities of both watersheds, NO_3^-, TP, and TSS were the selected variables to conduct the correlation analysis between water quality and land use.

In the Del Tala watershed, the data available are the result of sample analysis taken in triplicate at each site. However, in the San Salvador watershed, a single sample per site was collected, maintaining the protocol used by the Ministry of Environment[22].

The statistical analysis was performed using the Past 4.03 software[23]. Parametric analysis was executed using ANOVA with Tukey's post hoc test. Non-parametric statistics such as Kruskal Wallis or Mann-Whitney U were used when the homogeneity of the variances was not verified. All tests were done with a confidence level of 95 %.

2.3 Landscape metrics

The landscape was characterized within each basin, classifying its LULC and quantifying the spatial structure of some of the identified patterns.

2.3.1 Land use and land cover classification

The process of land cover mapping implies land cover classification using satellite imagery processed with different methods and algorithms[24]. For this study, satellite imagery corresponds to Sentinel 2 satellites from the Copernicus Programme from the European Space Agency[25]. The Sentinel 2 imagery was obtained and processed in the cloud using the Google Earth Engine[26], a planetary-scale cloud-computing platform for Earth science data and analysis. Sentinel 2 imagery was selected because it has good spatial (10/20 m), spectral (13 bands), and temporal resolutions (5 days), and is of free access.

For both watersheds, crop seasons from 2019/20 until 2021/22 were mapped to determine LULCs for winter and summer. Winter crops are sowed during the end of fall, flowering during spring, and are harvested at the end of spring. Therefore, for LULCs determination in the winter and especially winter crops detection like wheat and barley, the analyzed period comprises between June and December. In summer, crops like corn and soybean are sowed during spring, flowering during summer, and are harvested in fall. So, to detect these, the analyzed period comprised between October and May.

It was possible to identify different LULCs in an area when vegetation indices time series were analyzed. Each LULC has its characteristic photosynthetic activity dynamic through time, called temporal profile[27]; when analyzing the different vegetation time series for an area, it was possible to discriminate the different LULCs in it.

For this study, it was decided to work with the EVI2 index time series[28]:

$$EVI2 = 2.5 \cdot \frac{NEAR\ INFRARED - RED}{NEAR\ INFRARED + 2.4 \cdot RED + 1}$$

These were obtained from all available images for each period. A stack of EVI2 images was generated for each watershed and season, which was classified into ten clusters using K-means[29] so that each pixel had an assigned membership cluster (1, 2, 3,

etc.). The EVI2 time series were extracted for each image stack in a grid of regular points within each watershed. For all points belonging to the same cluster, a chart of the median signature of that cluster was generated, resulting in 10 median signatures. A hierarchical cluster analysis (HCA) was carried out to determine which median signatures were most similar and may belong to the same type of LULC[30].

Furthermore, the t-distributed Stochastic Neighbor Embedding (t-SNE) technique[31] was used to determine the classification quality with K-means. This non-linear dimensionality reduction technique allowed the visualization of high-dimensional data (in this case, time series) in a 2- or 3-dimensional space. t-SNE was used to generate graphs to visualize how well K-means clustering performed. Once the median signatures for each cluster were obtained, the shape of each signature was visualized and compared against a reference signature of a LULC class (agricultural crop, water, native forest, afforestation, forage resources, etc.) to determine which LULC the median signature most resembles and assign it accordingly.

The reference signatures were obtained from LULC maps (2018, 2020/21) generated by the National Directorate of Natural Resources (DRNR) of the Ministry of Livestock, Agriculture and Fisheries (MGAP)[32-33].

To evaluate the map accuracy, *F1 score* was calculated[34]:

$$F1\ score = 2 \cdot \frac{P \cdot R}{P + R}$$

where *P* is the precision and *R* is the recall, and are respectively equal to:

$$P = \frac{true\ positives}{true\ positives + false\ positives}$$

$$R = \frac{true\ positives}{true\ positives + false\ negatives}$$

F1 score ranges between 0 and 1; the closer to 1, the higher the map accuracy.

The classified maps were also visually validated using Google Earth to see where these classes fall to determine, based on knowledge of the study area, if it was expected to have that type of LULC in that place. Lastly, HCA determined which clusters were most similar and could belong to the same LULC, thus merging them.

The LULC categories detected in the watersheds were agriculture (*A*), forage resources in agriculture paddocks (*FRAP*), forage resources in riparian zones (*FRRZ*), forestry (*F*), forage resources (FR) and water (*W*). *A* category included areas where crops like corn, soybean, barley, wheat, and others are present. *F* included areas with native forests and forests implanted with eucalyptus and pines. *FRAP* considered agricultural areas with a forage crop in that season. *FRRZ* were riparian zones with forage resources. FR were other forage resources detected. *W* comprises areas with water bodies like rivers and water dams.

2.3.2 Landscape metrics calculation

Landscapes can be characterized quantitatively through different metrics at different scales. Within a geographic area, landscape metrics quantify the spatial structure of patterns and, historically, are indices for categorical land cover maps[35]. Composition and configuration are the two fundamentals that are quantified with landscape metrics. There are hundreds of metrics to quantify them both, but the majority measure five basic components: adjacency, amount, area, distance, and perimeter. There are three levels on which metrics can be calculated: at patch level, for each patch individually; at the class level, aggregated metrics are computed for all patches of the same class, and at a landscape level, aggregated metrics for all patches in that landscape are calculated[35]. Previous studies have adopted several metrics to represent the landscape[11][14][36-39].

To avoid the use of metrics that may provide redundant information, we selected the following metrics that in previous studies were correlated to nutrients and/or sediments: land cover (*LC*), landscape proportion (*LP*), mean Euclidean nearest neighbor distance (*MENN*), mean patch shape ratio (*MPSR*), patch density (*PD*), and Shannon index (*SI*). *LC* [ha] is a class-level metric that measures the area of all patches belonging to a class. *LP* [%] is a class-level metric that measures the percentage of the landscape belonging to a specific class. *MENN* [m] is a class-level metric that calculates the mean Euclidean nearest neighbor (*ENN*) distance of all patches belonging to a class. *ENN* measures the distance to the nearest neighboring patch of the same class. *MPSR* [none] is a class-level metric that calculates the mean SHAPE of all patches belonging to a class. SHAPE is the ratio between the patch's actual perimeter and the patch's hypothetical minimum perimeter. *PD* [n/100 ha] is a class-level and landscape-level metric that calculates the number of patches per unit area by class and landscape. *SI* [none] is a

landscape-level metric that considers the number of classes and the abundance of each class. It is a diversity metric.

Landscape metrics R package[40] was used to calculate the selected metrics. This package is an R version of the FRAGSTATS software[41]. The calculations were made in an R notebook in Google Colab[42].

2.3.3 Sub-watershed delimitation

The landscape metrics previously described were calculated within the sub-basin area upstream of each water-quality sampling site at both catchments. Hence, the water quality measured at those sites is characteristic of its own contributor sub-watershed. Combining the digital terrain model (DTM), contour curves, and the hydrographic network, it was possible to identify the basin ridgeline that delimitates the sub-basin per monitoring site.

The DTM used has a pixel of 2.5 × 2.5 meters and was obtained from *Infraestructura de Datos Espaciales del Uruguay* (IDEuy)[43]. The hydrographic network was obtained from IDEuy and the Ministry of Environment. Using the DTM, contour curves every 5 meters were extracted for each watershed with QGIS 3.16.11-Hannover software.

2.4 Machine-learning techniques

To evaluate the correlations that exist between LULC and water quality in both watersheds, partial least squares regression (PLSR) and random forest (RF) models were adopted as linear and non-linear techniques, respectively. Furthermore, the Shapely Additive Explanations (SHAP) method was used to rank the variables in terms of their degree of influence in predicting water quality variables (TP, NO_3^-, TSS).

The monthly time series of the three pollutants under study were seasonally aggregated (summer: from October till April, winter: from May till December) to be related to the LULC information. The years 2019, 2020, and 2021 were considered. The minimum, maximum, and mean concentrations for each year were calculated and used for the analysis.

For each experiment, the input variables (X) are generated from the land use indexes, neglecting the land use classes that cover less than 1 % of the watershed area. In particular, the two sets of input variables are related to the indexes calculated per class (X_{class}) and the ones computed for the entire watershed ($X_{landscape}$). The output variables (Y), three per experiment, are the minimum, mean, and maximum of each pollutant.

Sixty-six experiments were run for the San Salvador watershed (3 pollutants × 2 seasons × 11 monitoring sites). Twelve experiments were run for the Del Tala watershed (3 pollutants × 1 season (summer) × 4 monitoring sites).

2.4.1 Partial Least Squares Regression (PLSR)

The PLSR model computes the relationship between two sets of variables, the matrix $X_{m\times n}$, which is made of m variables (columns) and n objects (rows), and a response vector $y_{n\times 1}$. As well as principal component analysis (PCA), PLSR can detect a few linear combinations (components) of the original x-values that describe most of the information carried by y. In contrast to PCA, only the most significant linear combinations are considered in the PLSR-regression equation. This is mathematically calculated by maximizing the covariance between y and all possible linear functions of x[44].

In this study, the Python *hoggorm* package was used for the PLSR implementation and run[45].

2.4.2 Random Forest (RF)

RF is a supervised machine-learning technique able to depict non-linear links[46]. In this work, we used this model as a regressor: it is made of many *decision trees* (called *weak learners*) that are used to predict a value (ensemble model), and its response is the mean of the predicted values[47]. This approach overcomes the limitations of a single decision tree in terms of noise and accuracy[48]. RF is considered a *white box* model since its resulting internal structure can be easily visualized and interpreted[49].

In this study, the Python library *scikit-learn* was used for the model implementation and run[50].

2.4.3 Shapely Additive Explanations (SHAP)

The SHAP method was carried out to evaluate the contribution (importance) of each input variable to the prediction (model output)[51]. It is based on the game's theoretically optimal Shapley Values[52]. This technique was chosen over the given feature importance computed by RF for three main reasons: *i)* it is *model agnostic*, i.e., it can be calculated for any model allowing an unbiased comparison; *ii)* it not only provides the importance of each variable, but it can also compute the sign (positive or negative) of such contribution; *iii)* several studies on environmental matters have recently corroborated its success[53-55]. In this study, such analysis was implemented using the SHAP Python package.

3. Results

3.1 Characteristics and variation of water quality

The results of the spatio-temporal variations of the selected variables are shown in Figures 2 and 3. For both watersheds, the TP concentration was the only variable that did not meet the national regulations (0.025 mg/L) in any sampling campaigns, while the NO_3^- concentration values met the reference value in all cases (10 mg/L) (Decree 253/79). For TSS, only at the San Salvador watershed, 13 % of data deviated from the reference value (30 mg/L)[56].

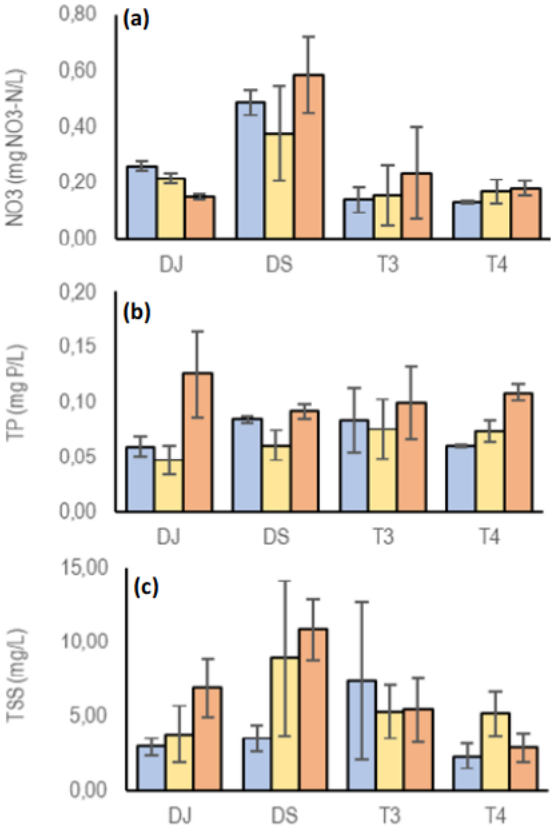


Figure 2. Water quality variables: a) nitrates (NO_3), b) total phosphorus (TP), and c) total suspended solid (TSS) in the Del Tala watershed. Blue column: 2019; yellow column: 2020; orange column: 2021

In the Del Tala watershed, the NO_3^- and TSS values at the DS site present the highest values (p=1.6E-14 and p=1.5E-04, respectively) being the year 2021 where the highest records were observed (p=0.03). It is important to note that a feedlot is located upstream of this sampling site. In this way, the obtained results agree with previous research in which increases in these variables are established in water courses associated with these undertakings[57-58]. In the case of TP, significant differences were observed among the sites. However, the values recorded in 2021 were the highest at each site (p=1.5E-08).

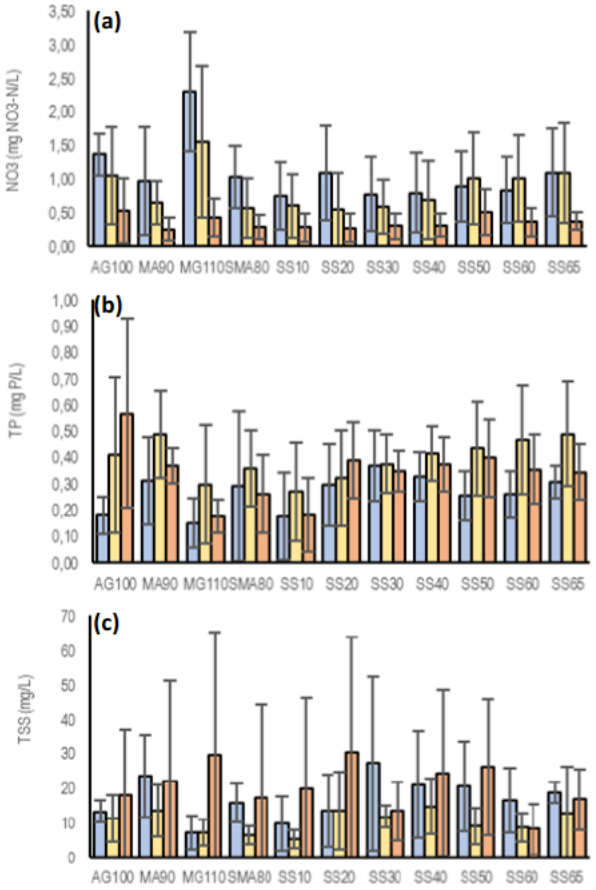


Figure 3. Water quality variables: a) nitrates (NO_3^-), b) total phosphorus (TP), and c) total suspended solid (TSS) in the San Salvador watershed. Blue column: 2019; yellow column: 2020; orange column: 2021

For the San Salvador river, an increase of nutrients from upstream to downstream was detectable. However, for TP, this trend does not present any statistical significance as it is instead observed for NO_3^- (p=0.0018). This pattern was supported by the results of the tributaries, where the highest values for both variables were observed. The TSS values do not show differences between the sampling sites, but on a temporal scale, a statistically significant increase is observed between 2020 and 2021 (p=0.002).

3.2 Land use and land cover composition

The results of the map classification process are presented in Table 1.

In Figure 4, we present the LULCs maps for season 2021/22 for the San Salvador and Del Tala watersheds. In Tables 1-2, we report the area and the proportion occupied by each LULC in both cases.

Table 1. Map classification performance for (a) San Salvador and (b) Del Tala watersheds

(a) SAN SALVADOR			
SUMMER			
LULC	**PRECISION**	**RECALL**	**F1-SCORE**
AGRICULTURE	0.75	0.89	0.82
FORAGE RESOURCES	0.68	0.46	0.55
FORESTRY	0.83	0.76	0.79
WEIGHTED AVERAGE	0.79	0.79	0.78
WINTER			
LULC	**PRECISION**	**RECALL**	**F1-SCORE**
AGRICULTURE	0.77	0.48	0.59
FORAGE RESOURCES	0.61	0.64	0.62
FORESTRY	0.6	0.8	0.69
WEIGHTED AVERAGE	0.68	0.65	0.64

(b) DEL TALA			
SUMMER			
LULC	**PRECISION**	**RECALL**	**F1-SCORE**
WATER	1	0.76	0.86
AGRICULTURE	0.54	0.74	0.63
FORAGE RESOURCES	0.92	0.85	0.89
WEIGHTED AVERAGE	0.86	0.83	0.84

Table 2. San Salvador river watershed land use and land cover (LULC) categories area for winter 2021 and summer 2021/22

LULC	SUMMER		WINTER	
	AREA (ha)	**PROPORTION (%)**	**AREA (ha)**	**PROPORTION (%)**
WATER	977	0.3	947	0.3
AGRICULTURE	155540	49.9	102683	33.0
FORESTRY	52585	16.9	28799	9.2
FRAP	57544	18.5	110787	35.6
FRRZ	44814	14.4	68245	21.9
TOTAL	**311460**			

In the San Salvador watershed, during winter, the main LULCs in area proportion in a decrease order are *FRAP*, *A*, *FRRZ*, *F*, and *W*. *FRAP* and *FRRZ* are forage resources, and they occupy almost 60 % of the area (Table 1). In summer, the agricultural area increased to 50 %, while the forage resources (*FRAP* + *FRRZ*) decreased to 33 % of the area (Table 2). This is explained by the fact that the area for summer crops is greater than for winter crops. Therefore, the areas not sown with crops in winter are occupied with forage resources like pasture. It is important to highlight that *W* occupies less than 1 % of the area, indicating the scarcity of irrigated crops.

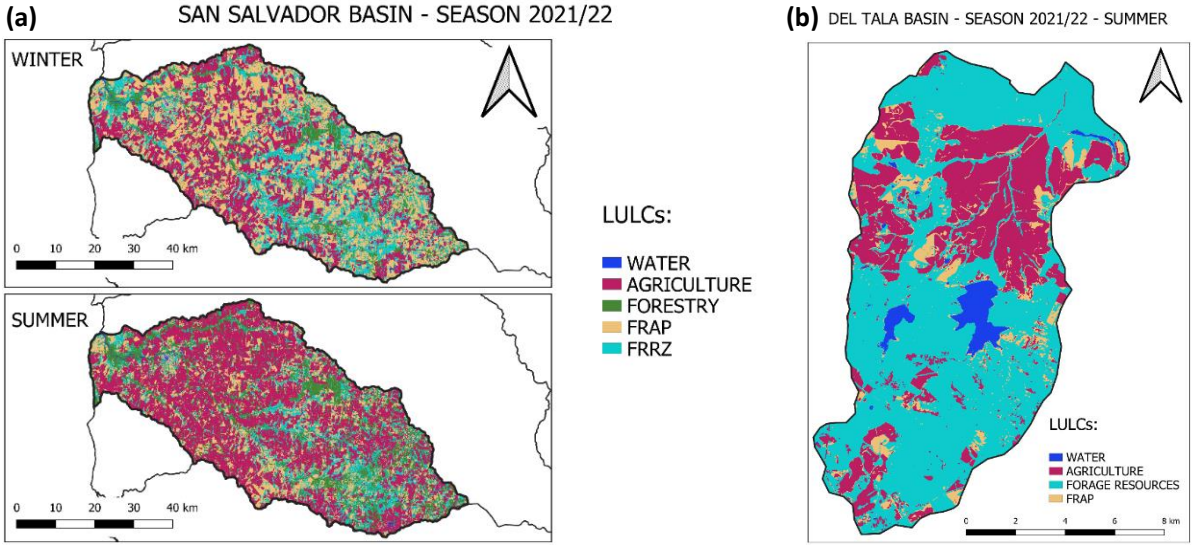


Figure 4. Land use/land cover map for a) winter 2021 and summer 2021/22 for San Salvador, and for b) summer 2021/22 for Del Tala watersheds

For the Del Tala watershed, only summer areas were detected. As this watershed is smaller than San Salvador, it is characterized by a smaller agricultural area (Table 3). The summer crop area is almost 30 %, and the forage area (forage resources + *FRAP*) is almost 70 %, contrasting with 50 % and 33 % of crops and forage resources in San Salvador. The *W* area is greater than in San Salvador, occupying 3 % of the area, and as shown in Figure 4, mainly this higher percentage is due to irrigation.

Table 3. Del Tala watershed land use and land cover (LULC) categories area for summer 2021/22.

LULC	AREA (ha)	PROPORTION (%)
WATER	480	3.2
AGRICULTURE	4231	28.3
FR	9046	60.5
FRAP	1207	8.1
TOTAL	**14964**	

A sub-watershed delineation was carried out upstream of each water-quality sampling site to quantify the proportion of landscape metrics contributing to the water quality measured at that point. In the Supplementary Materials, Tables S1 and S2 present the sub-watershed area for each site. As an example, Figure 5 shows the sub-watershed delimitations for the Del Tala T4 site and San Salvador SS65 station. We selected Del Tala T4 and San Salvador river watershed SS65 sites as representative sub-basins since they are the biggest ones, and are located close to the basin closure point. Landscape metrics of the different LULCs (class scale) are shown in Tables 4-6, and *PD* and *SI* for both sub-basins are presented in Table 7 (landscape scale).

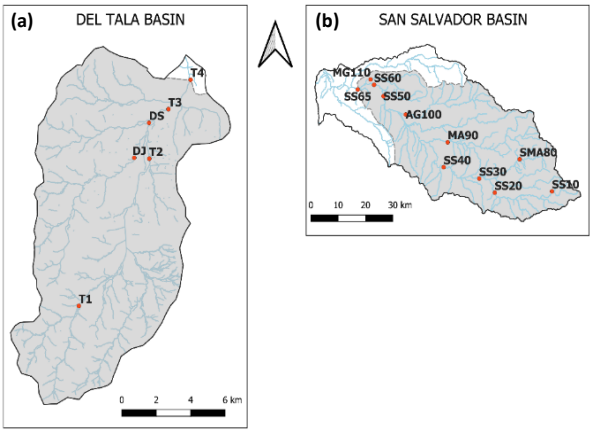


Figure 5. Example of sub-catchment delineation: a) Del Tala-T4 and b) San Salvador-SS65 sites sub-basins

PD is a metric that represents the level of fragmentation of each LULC category: the higher the value, the greater the fragmentation. Those LULCs with higher values have their total area distributed in a greater number of patches than others with a lower *PD*.

Table 4. San Salvador SS65 site winter 2021 landscape metrics (land cover: LC, landscape proportion: LP, mean Euclidean nearest neighbor distance: MENN, mean patch shape ratio: MPSR, patch density: PD) for each LULC category

LULC	LC (ha)	LP (%)	PD (n/100ha)	MENN (m)	MPSR
WATER	531	0.2	0.152	582.5	1.364
AGRICUL-TURE	75509	31.4	1.450	82.7	1.524
FORESTRY	21873	9.1	4.413	99.3	1.405
FRAP	86708	36.0	6.577	63.3	1.436
FRRZ	55990	23.3	6.239	63.8	1.511

Table 5. San Salvador SS65 site summer 2021/22 landscape metrics (land cover: LC, landscape proportion: LP, mean Euclidean nearest neighbor distance: MENN, mean patch shape ratio: MPSR, patch density: PD) for each LULC category

LULC	LC (ha)	LP (%)	PD (n/100ha)	MEN N (m)	MPSR
WATER	566	0.2	0.164	599.0	1.355
AGRICUL-TURE	115244	47.9	1.891	69.8	1.527
FORESTRY	40702	16.9	5.835	75.1	1.440
FRAP	46413	19.3	8.955	67.1	1.431
FRRZ	37685	15.7	8.870	63.9	1.493

Table 6. Del Tala T4 site summer 2021/22 landscape metrics (land cover: LC, landscape proportion: LP, mean Euclidean nearest neighbor distance: MENN, mean patch shape ratio: MPSR, patch density: PD) for each LULC category

LULC	LC (ha)	LP (%)	PD (n/100ha)	MEN N (m)	MPSR
WATER	111	4.3	0.620	284.9	1.415
AGRICUL-TURE	795	30.8	5.079	69.4	1.493
FR	1453	56.3	5.312	48.3	1.525
FRAP	220	8.5	27.837	41.1	1.428

Table 7. Patch density average and Shannon index for Del Tala-T4 and San Salvador-SS65 sites (summer season 2021/22)

SUB-WATERSHED	PD (N/100ha)	SHANNON INDEX
SAN SALVADOR - SS65	25.71	1.22
DEL TALA - T4	28.55	0.98

In the San Salvador river watershed, these LULCs were forage resources (*FRAP* and *FRRZ*), and the level of fragmentation is higher in summer when the agricultural area increases considerably com-

pared to winter. Considering the LULC of greater economic importance, such as agriculture, the value of *PD* is much lower than those of forage resources, 1.45 patches/100 hectares in winter and 1.89 in summer, against values around 6 patches/100 hectares and 9 in winter and summer for forage resources. This indicates that agriculture is much less fragmented compared to forage resources. Analyzing the image of Figure 6, where these LULCs are shown, it is observed that for agriculture, the patches are much larger, more continuous, and compact, while for forage resources, these are smaller, more isolated, and dispersed.

Regarding the Del Tala watershed, *PD* had a different pattern from San Salvador since agriculture in summer is more fragmented. It is essential to mention that Del Tala is a less agricultural basin than San Salvador, with 30 % vs. 48 % of the area occupied by agriculture, respectively. Analyzing the forage resources class, the *PD* value is like agriculture in FRRZ, while *FRAP* has a much higher *PD* value.

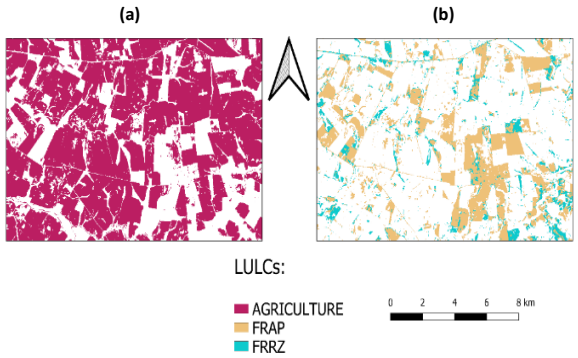


Figure 6. Patches comparison of a) agricultural and b) forage resources of San Salvador - summer 2021/22

MPSR has similar values between LULC and basins; therefore, no relevant conclusions can be inferred. For *MENN*, which indicates the average distance in meters at which the patches of a class are separated, the *W* class has the highest value in both basins, with values around 600m and 300m for San Salvador and Del Tala, respectively. This is because water occupies a small area at both basins, with few and distant patches.

3.3 Linkage between water quality and land use classes

The SHAP values were computed for the input variables of each experiment for both models, PLSR and RF. Therefore, the most influencing variables for predicting TP, NO_3^-, and TSS were detected and ranked. As an example, in Figure 7, the SHAP values obtained for PLSR and RF models are reported. In Figure 7(a), the SHAP values computed from PLSR to identify the most critical land use variables for predicting NO_3^- max at the DJ site in the Del Tala catchment are represented. In this plot, red dots depict high values of the input variables that positively impact the model outcomes. In contrast, blue dots represent low values that may not have an influence or have a negative impact on model results. In Figure 7(b), the SHAP values obtained for RF used to predict TP minimum, mean, and maximum at the site AG100 in the San Salvador watershed are reported as an example. The variables are ranked in terms of their degree of influence in predicting the pollutant under study. The different colors represent the three statistics of TP (min., mean, and max.). It is important to underline that the range of the SHAP values is only bounded by the output magnitude range of the model explained.

For the sake of brevity, in this manuscript we report a summary of the results obtained from all the experiments. In Figure 8, we report the number of times that each landscape metric resulted the most important one per pollutant, per season, at each sub-basin.

As for SHAP values based on PLSR, no landscape metric can be considered the main pollutant concentration predictor. This is particularly true in the Del Tala watershed. Here, the most frequent metrics are *PD*, *LC*, and *MPSR* associated with different land uses: *A*, *FR*, *FRAP*, and *W*.

At the San Salvador watershed, depending on the sub-watershed considered, the most influencing variables are *PD* and *MENN* associated above all with *A*, but also with *FRAP*, *F*, and *FRRZ*. Furthermore, at the San Salvador catchment, there is not a dominant predicted index for AG100, MG100, SS10, and SS30 sub-watersheds. It is also interesting that *[A] PD* is the most important predictor for the three pollutants under study. NO_3^- concentration is also well represented by *[A] MENN*. In winter, *[FRAP] MENN* and *[A] PD* are the most influencing variables of NO_3^- and TSS, respectively. While *[A] PD, [FRRZ] MPSR* and *[FRRZ] LC* are equally significant in TP concentration prediction. The *PD* measures the number of land use patches within the basins; thus, fragmented uses could negatively impact water quality by generating higher pollutant loads. Particularly in agricultural land uses, the negative effects on water quality are associated with the intensive use of fertilizers during the agricultural season and its mobilization to water courses by runoff. It

is known that approximately 20 % of the nitrogen and between 3 and 20 % of the phosphorus applied in agricultural systems are exported to surface waters, mainly by erosion and lixiviation[9][11][37].

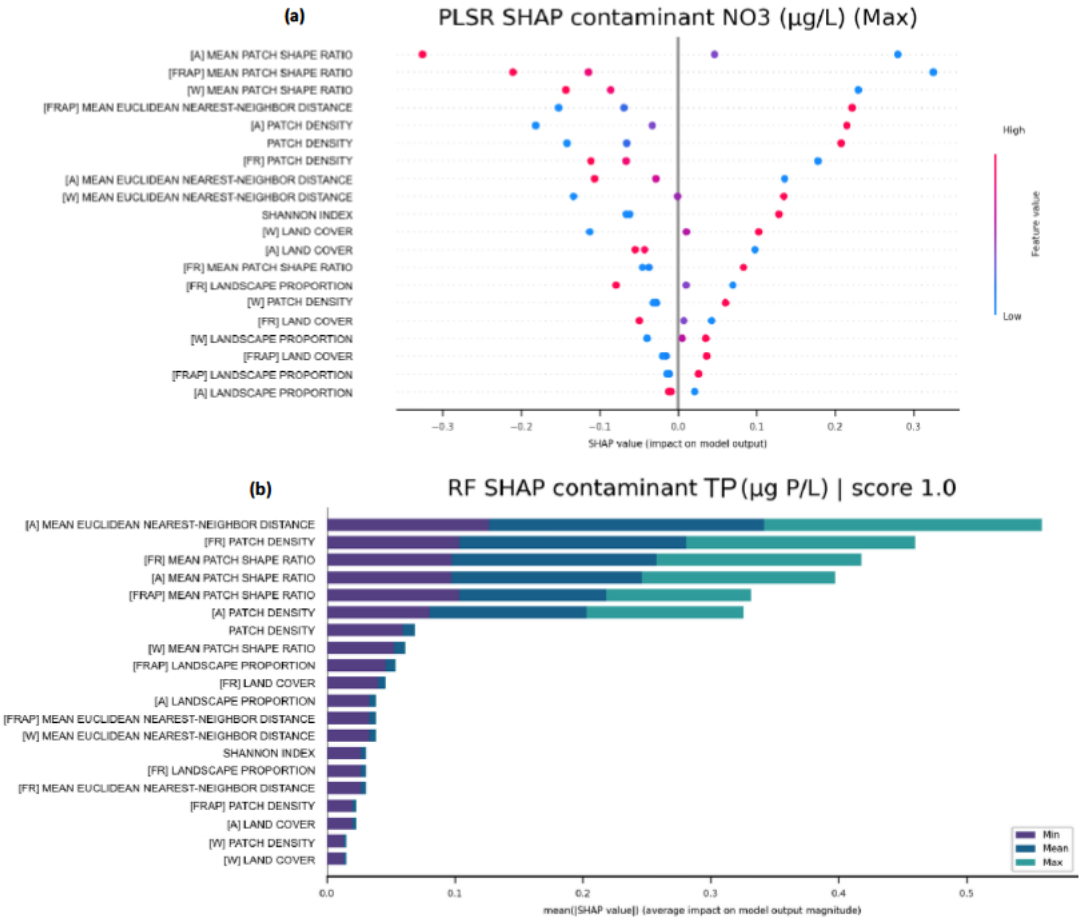
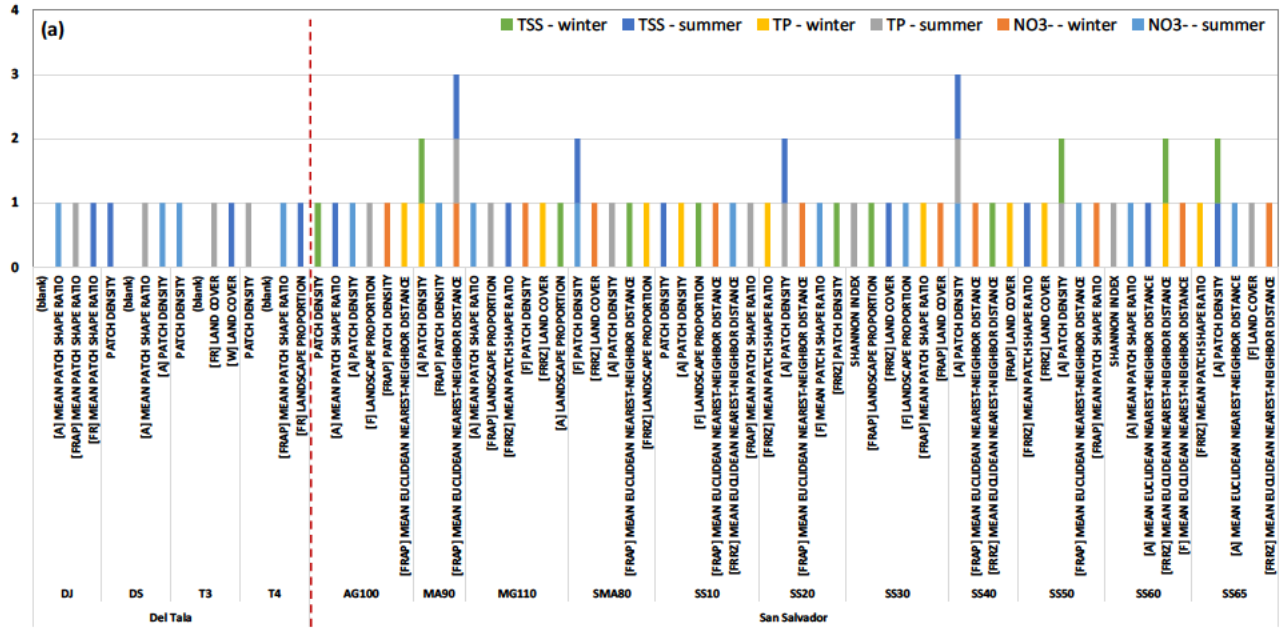


Figure 7. (a) SHAP values for nitrates (NO_3^-) max prediction at DJ tributary site in Del Tala watershed, based on partial least squares regression model (PLSR); **(b)** SHAP values for total phosphorus (TP) prediction at AG100 tributary site in San Salvador watershed, based on random forest model (RF)



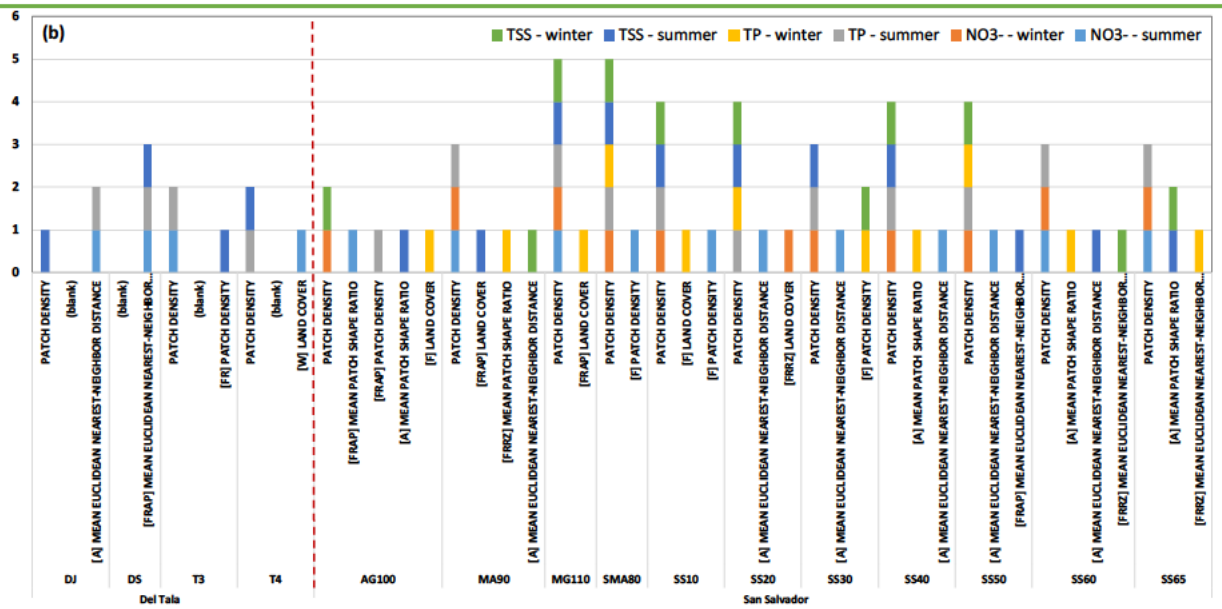


Figure 8. SHAP results: summary of the most important land use index obtained from **(a)** the partial least squares regression model (PLSR) and **(b)** the random forest model (RF). The dotted red line separates the two watersheds

The SHAP values based on RF are more homogeneous than the previous ones. This is maybe due to the ability of RF to identify non-linear relationships. *PD* strongly influences model performance for all sub-watersheds at the San Salvador basin and for T3 and T4 at the Del Tala catchment. *PD* is also frequently associated with different land uses: *FRAP* and *F* for San Salvador, and *FR* for Del Tala. Considering each variable and season analyzed, *PD* is the most influencing variable for the following: i) NO_3^- concentration, above all in winter, while in summer it is accompanied by *MENN* for *A* land use; ii) TP concentration during both seasons (above all in summer); and iii) TSS concentration during both seasons. These results are partially in accordance with Xu and others[13], who demonstrated that the *PD* metric of cropland is the key predictor of NO_3^- during the dry period. This might be related to the intensive agricultural activities that characterize the study areas. However, studies carried out by Bu and others[11] also showed that *PD* is correlated with more physicochemical variables but during the rainy season than during the dry season. They also report that during the rainy season, agricultural land use predicts more water quality variables than during the dry season probably due to agricultural runoff from soil erosion.

It is noteworthy to highlight that the statistical significance of an experiment is highly affected by any change in the selection of variables due to the lack of a suitable number of observations adopted for this study.

4. Discussion

As for the spatio-temporal variation of the water quality variables considered in this study, at the Del Tala watershed, the NO_3^- values at the DS site presented the highest values in 2021. Particularly at this site, the highest records were observed. This increase in the contribution of nitrates can be associated with the presence of a feedlot upstream. The same reason may explain the behavior of TSS, which showed a significant increase in the same site. In the case of TP, a spatial pattern was not identified. However, the values recorded in 2021 were the highest at each monitoring site. At the San Salvador basin, the TSS values did not show differences between the sampling sites, but on a temporal scale, a statistically significant increase was observed between 2020 and 2021. It is worth noting that during 2020 and 2021, La Niña occurred, which led to a significant drought in Uruguay. During dry or low precipitation periods, the lack of water can limit the leaching of nutrients from the soil into groundwater and reduce surface runoff. As a result, nutrients tend to build up in the soil, especially in the surface layers. These nutrients are partially taken up by plants but can also remain in the soil as organic or soluble mineral residues[8]. When heavy rainfall occurs after a dry period, it triggers a release of nutrients and sediments stored in the soil. The initial precipitation, often referred to as a "flushing event," has a particularly pronounced effect[46]. During this event, rainwater flows over the dry soil and begins to collect the nutrients that have accumulated on the surface and in the shallow soil layers. This surface

runoff, rich in nutrients and sediments, flows into nearby water bodies and results in a sudden increase in nutrient concentration, which can have negative effects on water quality[47].

Another important aspect to highlight is that TP values were always higher than the limit imposed by the national regulations (0.025 mg/L) at both watersheds. This may be justified by the fact that most of the fertilizers used in Uruguay are phosphate-based[17].

As for the correlation evaluated between LULC and water quality, we found that *PD* is one of the most influencing metrics in both watersheds and for both models. This outcome confirms the results obtained by several studies[11-12][37-38]. Kearns (2005) stated that there were two factors: i) *PD* and distribution, and ii) patch shape and landscape subdivision, which explained 85 % of the variation in their data set[38]. The *A* land use is the most critical one at both catchments, and *A* and *FRAP* land uses are the most important ones for both algorithms. Previous studies found that cropland land use strongly correlates with water physicochemical variables at larger scales. This was also confirmed by our results at the San Salvador watershed, where *A* land use, mainly represented by *PD*, had an important impact on nutrients and TSS. This may be due to the extensive distribution of farmland across the watershed and it recommends that the control of agricultural non-point source pollution should focus on the agricultural management at a basin scale, such as the fertilizer use control and the conservation tillage extension[11][13][44].

5. Conclusions

In this study, with the aid of unsupervised machine learning techniques, we identified and ranked the main landscape metrics that most influence different water-quality variables at different watershed scales. This aim was achieved by considering two Uruguayan basins of different sizes, San Salvador (3,118 km^2) and Del Tala (160 km^2). The main findings of this work can be summarized as follows:

- In the spatio-temporal analysis of the physicochemical variables in both watersheds, Del Tala and San Salvador, TP was the only variable that did not comply with the national regulations (0.025 mg/L), while the NO_3^- concentration values met the reference value in all cases (10 mg/L).
- During 2020 and 2021, when La Niña occurred, TP and TSS runoff was exceptionally high.
- LULC was estimated for the San Salvador river watershed in winter 2021 and summer 2021/22, and for the Del Tala basin in summer 2021/22. In San Salvador, during the winter, 30 % of the area was identified as agriculture; while during the summer, the percentage of crops was approximately 50 % for San Salvador, and 30 % for Del Tala.
- Landscape metrics were calculated for the most representative sub-basins. As expected, in the San Salvador river basin, forage resources were more fragmented (smaller, more isolated, and dispersed patches) than agriculture with much larger patches, more continuous and compact. Del Tala had the same pattern but only with *FRAP*, since *FRRZ* had a similar value to agriculture. It would be interesting to look at the patch area metric besides the *PD* metric to analyze this difference more deeply. For other landscape metrics, no relevant conclusions can be made.
- *PD* is one of the most influencing metrics for both watersheds and models. Agricultural was the most critical land use at both catchments, and agricultural with a forage crop land uses are the most important ones for both algorithms used.
- Agriculture land use showed a stronger correlation with nutrients and TSS at the San Salvador watershed (at larger scales).

It is essential to mention that the adopted techniques are valuable tools that can provide an adequate overview of the water-quality behavior in space and time, and the correlations between water-quality variables and LULC.

Acknowledgements

This work was supported by the National Research and Innovation Agency (ANII) [grant numbers: FSA_PI_2018_1_147713, SA_PI_2018_1_148628, FSA_PP_2018_1_147701].

Transparency of data

The data set supporting the results of this study is partially publicly available. The water quality data for the San Salvador river basin can be found at *https://www.ambiente.gub.uy/oan/*.

Author contribution statement

All authors contributed equally to the content.

References

1. Melgar R, Vitti G, De Melo V. Soja en Latinoamérica: fertilizando para altos rendimientos. IIP Boletín. 2011;20:81p.

2. Van Opstal NV, Caviglia OP, Melchiori RJM. Water and solar radiation productivity of double-crops in a humid temperate area. Aust J Crop Sci. 2011;5(13):1760-6.

3. Arbeletche P, Ernst O, Hoffman E. La agricultura en Uruguay y su evolución. In: García Préchac F, Ernst O, Arbeletche P, Bidegain MP, Pritsch C, Ferenczi A, Rivas M, editores. Intensificación agrícola: oportunidades y amenazas para un país productivo y natural. Montevideo: Universidad de la República; 2010. pp. 13-27.

4. Mon R, Irurtia C, Botta G, Pozzolo O, Bellora F, Rivero D, Bomben M. Effects of supplementary irrigation on chemical and physical soil properties in the rolling pampa region of Argentina. Cienc Investig Agrar. 2007;34(3):187-94.

5. Huang C, Kim S, Song K, Townshed JRG, Davis P, Altstatt A, Rodas O, Yanosky A, Clay R, Tucker CJ, Musinsky J. Assessment of Paraguay's forest cover change using landsat observations. Glob Planet Change. 2009;67:1-12. Doi: 10.1016/j.gloplacha.2008.12.009.

6. Paruelo JM, Guerschman JP, Piñeiro G, Jobbágy EG, Verón SR, Baldi G, Baeza S. Cambios en el uso de la tierra en Argentina y Uruguay: marcos conceptuales para su análisis. Agrociencia. 2006;10(2):47-61. Doi: 10.31285/AGRO.10.929.

7. De la Fuente E, Suárez SA. Problemas ambientales asociados a la actividad humana: la agricultura. Ecol Austral. 2008;18:239-52.

8. Lintern A, Webb JA, Ryu D, Liu S, Bende-Michl U, Waters D, Leahy P, Wilson P, Western AW. Key factors influencing differences in stream water quality across space. WIREs Water. 2018;5:e1260. Doi: 10.1002/wat2.1260.

9. Withers PJ, Neal C, Jarvie HP, Doody DG. Agriculture and eutrophication: where do we go from here? Sustainability. 2014;6:5853-75. Doi: 10.3390/su6095853.

10. Fisher B, Turner RK, Morling P. Defining and classifying ecosystem services for decision making. Ecol Econ. 2009;68:643-53.

11. Bu H, Meng W, Zhang Y, Wan J. Relationships between land use patterns and water quality in the Taizi River basin, China. Ecol Indic. 2014;41:187-97.

12. Lee SW, Hwang SJ, Lee SB, Hwang HS, Sung HC. Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. Landsc Urban Plan. 2009;92(2):80-9.

13. Xu S, Li SL, Zhong J, Li C. Spatial scale effects of the variable relationships between landscape pattern and water quality: example from an agricultural karst river basin, Southwestern China. Water Resour Manag. 2020;300:106999. Doi: 10.1016/j.agee.2020.106999.

14. Uuemaa E, Roosaare J, Mander Ü. Landscape metrics as indicators of river water quality at catchment scale. Hydrol Res. 2007;38(2):125-38.

15. Gorgoglione A, Gregorio J, Ríos A, Alonso J, Chreties C, Fossati M. Influence of land use/land cover on surface-water quality of Santa Lucía River, Uruguay. Sustainability. 2020;12(11):4692. Doi: 10.3390/su12114692.

16. Aubriot L, Delbene L, Haakonson S, Somma A, Hirsch F, Bonilla S. Evolución de la eutrofización en el Río Santa Lucía: influencia de la intensificación productiva y perspectivas. Innotec. 2017;14:7-17. Doi: 10.26461/14.04.

17. Centurion V, Fabre A, Kok P, Badano L, Neighbur N, Gelos M, Rodo E, Hoffmeister M, De Leon L. Evolución de la calidad de agua en la cuenca del río San Salvador: periodo 2014–2019. Montevideo: MVOTMV; 2020. 76p.

18. Caracterización de las cuencas del río San Salvador, río Yí Y río Arapey para fines de riego. Montevideo: MGAP; 2017. 198p.

19. Ministerio de Ambiente, OAN (UY). Observatorio Ambiental Nacional [Internet]. Montevideo: MA; [date unknown; cited 2023 Oct 05]. Available from: https://www.ambiente.gub.uy/oan/

20. Standard methods: for examination of water and wastewater. 15th ed. Washington: APHA; 1995. 1134p.

21. Monteiro MIC, Ferreira FN, De Oliveira NMM, Avila AK. Simplified version of the sodium salicylate method for analysis of nitrate in drinking waters. Anal Chim Acta. 2003;477(1):125-9.

22. Plan de monitoreo río San Salvador:informe de actividades y presentación de resultados: 2015. Montevideo: MVOTMA; 2016. 67p.

23. Hammer Ø, Harper DA, Ryan PD. PAST: Paleontological statistics software package for education and data analysis. Palaeontol Electron [Internet]. 2001 [cited 2023 Oct 05];4(1):9p. Available from: https://palaeo-electronica.org/2001_1/past/past.pdf

24. Horning N. Land cover classification methods [Internet]. Version 1.0. New York: American Museum of Natural History;2004 [cited 2023 Oct 05]. Available from: https://www.amnh.org/content/download/74344/1391366/file/land-cover-classification-methods.pdf

25. ESA. Sentinel-2 Mission Guide [Internet]. [place unknown]: ESA; [date unknown; cited 2023 Oct 05]. Available from: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2

26. Google. Google Earth Engine [Internet]. Mountain View: Google; [date unknown; cited 2023 Oct 05]. Available from: https://earthengine.google.com/

27. Ghazaryan G, Dubovyk O, Löw F, Lavreniuk M, Kolotii A, Schellberg J, Kussul N. A rule-based approach for crop identification using multi-temporal and multi-sensor phenological metrics. Eur J Remote Sens. 2018;51(1):511-24.

28. Jiang Z, Huete A, Didan K, Miura T. Development of a two-band enhanced vegetation index without a blue band. Remote Sens Environ. 2008;112(10):3833-45.

29. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31(8):651-66.

30. Bridges CC. Hierarchical Cluster Analysis. Psychol Rep. 1966;18(3):851-4. Doi: 10.2466/pr0.1966.18.3.851.

31. Maaten L van der, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-605.

32. Ministerio de Ganadería, Agricultura y Pesca (UY). Mapa integrado de cobertura/uso del suelo del Uruguay año 2018 [Internet]. Montevideo: MGAP; 2019 [cited 2023 Oct 05]. Available from: https://bit.ly/3rDrVN0

33. Ministerio de Ganadería, Agricultura y Pesca, DGRN (UY). Actualización de cobertura y uso del suelo del Uruguay al año 2020/2021 [Internet].Montevideo: MGAP; 2021[cited 2023 Oct 05]. Available from: https://bit.ly/45z8kM9

34. Sharma A, Mishra PK. State-of-the-art in performance metrics and future directions for data science algorithms. J Sci Res. 2020;64(2):221-38. Doi: 10.37398/JSR.2020.640232.

35. Frazier A. Landscape Metrics. In: Wilson JP, editor.The geographic information science & technology: body of knowledge. 2nd ed. [place unknown]: University Consortium for Geographic Information Science; 2019. Doi: 10.22224/gistbok/2019.2.3.

36. Amiri BJ, Nakane K. Modeling the linkage between river water quality and landscape metrics in the Chugoku district of Japan. Water Resour Manag. 2009;23(5):931-56.

37. Liu A, Duodu GO, Goonetilleke A, Ayoko GA. Influence of land use configurations on river sediment pollution. Environ Pollut. 2017;229:639-46.

38. Kearns FR, Kelly NM, Carter JL, Resh VH. A method for the use of landscape metrics in freshwater research and management. Landsc Ecol. 2005;20:113-25.

39. Uuemaa E, Roosaare J, Mander Ü. Scale dependence of landscape metrics and their indicatory value for nutrient and organic matter losses from catchments. Ecol Indic. 2005;5(4):350-69.

40. Hesselbarth MHK, Sciaini M, With KA, Wiegand K, Nowosad J. Landscape metrics: an open-source R tool to calculate landscape metrics. Ecography. 2019;42(10):1648-57.

41. McGarigal K, Marks BJ. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. Portland: USDA; 1995. 122p.

42. Google Colab [Internet]. Mountain View: Google; 2023 [cited 2023 Oct 05]. Available from: https://colab.research.google.com/

43. IDEUY: Infraestructura de Datos Espaciales [Internet]. Montevideo: Uruguay Presidencia; [date unknown; cited 2023 Oct 05]. Available from: https://visualizador.ide.uy/ideuy/core/load_public_project/ideuy/

44. Shi ZH, Ai L, Li X, Huang XD, Wu GL, Liao W. Partial least-squares regression for linking land-cover patterns to soil erosion and sediment yield in watersheds. J Hydrol. 2013;498:165-76.

45. Tomic O, Graff T, Liland KH, Naes T. Hoggorm: a python library for explorative multivariate statistics. J Open Source Softw. 2019;4(39):980. Doi: 10.21105/joss.00980.

46. Russo C, Castro A, Gioia A, Iacobellis V, Gorgoglione A. Improving the sediment and nutrient first-flush prediction and ranking its influencing factors: an integrated machine-learning framework. J Hydrol. 2023;616:128842. Doi: 10.1016/j.jhydrol.2022.128842.

47. Breiman L. Random forests. Machine learning. 2001;45:5-32.

48. Vilaseca F, Castro A, Chreties C, Gorgoglione A. Daily rainfall-runoff modeling at watershed scale: a comparison between physically-based and data-driven models. In: Gervasi O, Murgante B, Misra S, Garau C, Blečić I, Taniar D, Apduhan BO, Rocha AMAC, Tarantino E, Torre CM, editors. Computational Science and Its Applications: ICCSA 2021. Cham: Springer; 2021. pp. 18-33.

49. Wang J, Bao W, Gao Q, Si W, Sun Y. Coupling the Xinanjiang model and wavelet-based random forests method for improved daily streamflow simulation. J Hydroinformatics. 2021;23:589-604.

50. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

51. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems 30. New York: NIPS; 2017. pp. 4765-74.

52. Shapley LS. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the Theory of Games. Vol 2. Princeton: Princeton University Press; 1953. pp. 307-17.

53. Russo C, Castro A, Gioia A, Iacobellis V, Gorgoglione A. A Stormwater management framework for predicting first flush intensity and quantifying its influential factors. Water Resour Manag. 2023;37:1437-59.

54. Zhong S, Zhang K, Wang D, Zhang H. Shedding light on ‘‘Black Box’’ machinelearning models for predicting the reactivity of HO radicals toward organic compounds. Chem Eng J. 2021;405:126627. Doi: 10.1016/j.cej.2020.126627.

55. Cross T, Sathaye K, Darnell K, Niederhut D, Crifasi K. Predicting water production in the Williston basin using a machine learning model. In: Unconventional Resources Technology Conference, Virtual, 20–22 July 2020. [place unknown: publisher unknown]; 2020. pp. 3492–503. Doi: 10.15530/urtec-2020-2756.

56. CARU. Digesto sobre el uso y aprovechamiento del Río Uruguay [Internet]. Paysandú: CARU; 2019 [cited 2023 Oct 05]. 140p. Available from: https://bit.ly/3RRliBl

57. Ciganda V, Lizarralde C, Eguren G. Establecimiento de engorde a corral bovino o feedlots: cuantificación de su impacto sobre los recursos suelo y agua. Revista INIA. 2015;(41):39-44.

58. García AR, Fleite SN, Ciapparelli I, Vázquez Pugliese D, Weigandt C, Fabrizio de Iorio A. Observaciones, desafíos y oportunidades en el manejo de efluentes de feedlot en la provinicia de Buenos Aires, Argentina. Ecol Austral. 2015;25(3):255-62.

Supplementary material

Table S1. Sub-watershed delineation based on water-quality monitoring sites at Del Tala

MONITORING SITE	SITE NAME	SUB-WATERSHED AREA (ha)
T1	Del Tala Naciente	1,649
T2	Del Tala aguas abajo represa Tres Marías Gran-de	7,279
T3	Del Tala aguas abajo del área agrícola	11,813
T4	Del Tala Desembocadura	14,710
DJ	Juncal Desembocadura	2,581
DS	Sarandí Desembocadura	1,326

Table S2. Sub-watershed delineation based on water-quality monitoring sites at San Salvador

MONITORING SITE	SITE NAME	SUB-WATERSHED AREA (ha)
SS10	Colonia Larrañaga	5,835
SS20	Ruta 55, Km 39	26,388
SS30	Perseverano	62,993
SS40	Paso Zabala	83,245
SS50	Paso Ramos	214,824
SS60	Toma vieja OSE	225,690
SS65	Península Timoteo Ra-mospe	240,766
SMA80	A.° San Martin	11,196
MA90	A.° Maciel	34,600
AG100	A.° del Águila	31,579
MG110	A.° Magallanes	10,027