Informe Técnico

ANII Fondo sectorial de investigación a partir de datos - 2018

Evaluación temporal y espacial del impacto del cambio de cobertura del suelo sobre la calidad del agua: cuenca del río Santa Lucía como cuenca piloto

Número propuesta:	FSDA_1_2018_1_153967	
Instituciones:	Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA), Facultad de Ingeniería (FIng), Universidad de la Republica (UdelaR). Instituto de Computación (InCo), Facultad de Ingeniería (FIng), Universidad de la Republica (UdelaR).	
Responsable técnico-científico:	Angela Gorgoglione (IMFIA)	
Co-responsable técnico-científico:	Alberto Andrés Castro Casales (InCo-IIE)	
Equipo de investigación:	**IMFIA** Rafael Rodríguez Christian Chreties Mónica Fossati	**InCo** Marcos Pastorini Lorena Etcheverry
Fecha del informe:	14 junio 2021	
Periodo de trabajo:	1 abril 2020 – 31 mayo 2021	

Tabla de contenidos

1. Introducción

1.1. Antecedentes

La calidad del agua tiene un rol esencial en la salud pública y en la conservación de los ecosistemas (Shi et al., 2017). El agua limpia, segura y fresca es un recurso clave para el desarrollo humano, social y económico (Pérez-Gutiérrez et al., 2017). Sin embargo, la degradación de la calidad del agua resultante de fuentes de contaminación puntuales y difusas es un problema ambiental mundial (Shoemaker et al., 2017). Las fuentes puntuales son aquellas a las que se puede atribuir una ubicación física específica, como por ejemplo las tuberías de descarga de aguas residuales domésticas o industriales. En contraste, las fuentes difusas no tienen identificado un origen puntual (Zhen-Gang, 2008). Las principales fuentes no puntuales incluyen lavado de tierras rurales (sedimentos, fertilizantes, microorganismos y pesticidas), lavado de áreas urbanas (aceites, grasas, químicos tóxicos, metales pesados, patógenos y sedimentos), deposición atmosférica (químicos tóxicos, metales pesados, nutrientes y ácidos) y filtración de aguas subterráneas (nutrientes y químicos tóxicos). En los últimos años se ha descubierto que en muchos casos de estudio la fuente dominante de contaminantes son justamente las fuentes difusas (Zhen-Gang, 2008), las cuales a su vez son más difíciles de gestionar que las fuentes puntuales y presentan el mayor desafío en, por ejemplo, la gestión de cuencas (Xu et al., 2019). El transporte de los contaminantes difusos ocurre a través de la superficie terrestre mediante la escorrentía y a través del suelo mediante la percolación. Por esta razón, este tipo de contaminación es intermitente y está fuertemente correlacionada con la escorrentía, los factores climáticos y las características específicas del sitio, como el tipo de suelo, el uso del suelo y la topografía (Ritter and Shirmohammadi, 2001). Como consecuencia, la transformación de suelos naturales a diferentes usos (plantación forestal, cultivos regados, entre otros) aumenta la generación y transmisión de contaminantes a los cuerpos de agua receptores (Miller et al., 2014). Este es un fenómeno que está ocurriendo particularmente en los países en vías de desarrollo, los cuales están experimentando un proceso de expansión e intensificación agrícola, como Uruguay (Goyenola et al., 2015).

A nivel nacional, el aumento de las actividades agrícolas, forestales y lácteas genera problemas de eutrofización y el deterioro de la calidad del agua de los ríos, embalses y zonas costeras. Varios estudios indican que este proceso se está produciendo rápidamente, causando serios problemas de salud en los ecosistemas terrestres y acuáticos (Oyhantçabal and Narbondo, 2014), así como también en el uso de los cuerpos de agua para actividades de pesca, recreación y suministro de agua potable (Rodríguez-Gallego et al., 2017; Arocena et al., 2013; Pacheco et al., 2012). Una de las principales consecuencias de la eutrofización son las floraciones de cianobacterias potencialmente tóxicas. Estos eventos se han registrado en varios cuerpos de agua en las principales cuencas del país (Bonilla et al., 2015; Ferrari et al., 2011; Bonilla, 2009). Distintos

estudios han provisto un análisis exhaustivo del estado actual de la eutrofización de los principales ecosistemas nacionales (AA.VV., 2019; Aubriot et al., 2017). Sin embargo, no hay estudios que profundizan el conocimiento de cómo y en qué cantidad los cambios de uso del suelo afectan la calidad de agua.

Varios estudios internacionales han investigado la asociación entre los cambios en el uso del suelo y la calidad del agua (Namugize et al., 2018; Kändler et al., 2017; Calijuri et al., 2015). Estos estudios se pueden clasificar principalmente en dos grupos: *i)* estudios comparativos, donde se usan datos de series temporales para investigar los impactos que los cambios en el uso del suelo tienen sobre la calidad del agua; y *ii)* estudios de simulación, donde se emplean técnicas estadísticas para establecer relaciones entre los cambios en el uso del suelo y la calidad del agua. Aunque los resultados de los estudios comparativos proporcionan conocimientos esenciales, pueden resultar inadecuados para el diseño de estrategias efectivas que mitiguen los impactos adversos de las actividades antropogénicas en la calidad del agua. La metodología de comparación simple utilizada en los estudios comparativos carece de la capacidad para describir asociaciones causales sólidas entre los indicadores de calidad del agua y sus factores influyentes. Por otro lado, los estudios de simulación utilizan en su mayoría datos relacionados con el uso del suelo y los indicadores de calidad del agua obtenidos en un solo momento para justificar sus hallazgos. Si bien estos datos permiten comparar la influencia de varios factores sobre la calidad del agua en un momento dado, tales datos no facilitan la evaluación de los cambios potenciales en la calidad del agua. En consecuencia, los estudios de simulación no tienen un uso potencial para diseñar estrategias efectivas para reducir la contaminación del agua debido a los cambios en el uso del suelo a largo plazo (Wijesiri et al., 2018).

Asimismo, en los últimos años, la utilización de técnicas de aprendizaje automático en el contexto de las ciencias ambientales viene tomando un fuerte impulso (Mori et al., 2019; Thornhill et al., 2017; Rosecrans et al., 2017; Thornhill et al., 2017; Forio et al., 2015). Principalmente, se han utilizado algoritmos de aprendizaje automático supervisado, en particular predictores y clasificadores. En general, el objetivo ha sido utilizar los datos de las series temporales para crear “estudios de simulación” donde los modelos de cantidad y calidad de agua clásicos no son capaces de resolver problemas con alta complejidad de datos. De todos modos, particularmente a nivel nacional, aun no se han explotado los métodos de aprendizaje automático para buscar las correlaciones subyacentes entre los diferentes factores naturales y antrópicos, y así estudiar los procesos ambientales de una manera holística. Basándonos en lo anterior, este proyecto propone crear un tercer y nuevo enfoque (grupo de estudio), donde se utilizarán datos de series temporales, datos geográficos y técnicas de aprendizaje automático para investigar las correlaciones entre los cambios en el uso y/o cobertura del suelo (LULC, por su sigla en inglés), y los parámetros físico-químicos de calidad del agua.

Se propone llevar adelante este proyecto utilizando la cuenca del río Santa Lucía como cuenca piloto. El río Santa Lucía nace en la Sierra Carapé en el departamento de Lavalleja y recorre una longitud de 230 km hasta su desembocadura en el Río de la Plata. Además del río Santa Lucía, los principales cursos de agua de la cuenca son los ríos Santa Lucía Chico y San José; y los arroyos de la Virgen, Canelón Grande, Canelón Chico, Las Piedras y Colorado (DINOT, 2016). En la cuenca inferior se encuentran los humedales del Santa Lucía, que ingresaron al Sistema Nacional de Áreas Protegidas (SNAP) en febrero de 2015 mediante el Decreto 55/015 (2015). La cuenca del río Santa Lucía, es la fuente de agua bruta para potabilización y abastecimiento de aproximadamente el 60% de la población de Uruguay, destacándose la usina potabilizadora de Aguas Corrientes, que abastece a Montevideo y parte del Área Metropolitana. Además, la cuenca concentra una importante actividad económica que es relevante para el desarrollo del país. Alrededor del 32% de la población rural nacional vive en la cuenca del río Santa Lucía, siendo uno de los principales polos de producción de alimentos. Si bien predomina la actividad agropecuaria, se destacan también diversas industrias como la industria cárnica, láctea, vitivinícola y textil entre otras (DINOT, 2016; MVOTMA, 2017). La actividad antrópica en la cuenca ha generado impactos en la calidad del agua. Según el informe realizado por DINAMA-JICA (JICA-MVOTMA, 2011), las cargas de aporte provenientes de las fuentes difusas en la cuenca del río Santa Lucía corresponden a un 82% para la demanda bioquímica de oxígeno (DBO_5), 82% para nitrógeno total (NT) y 77% para fósforo total (PT); siendo la actividad agrícola-ganadera una de las principales contribuyentes.

En consecuencia, el desafío es desarrollar en la cuenca actividades productivas relevantes para el desarrollo económico del país preservando la calidad de los cuerpos de agua y evitando la afectación de otras actividades como la potabilización de aguas o la preservación de ecosistemas relevantes como los humedales del río Santa Lucía. La investigación propuesta en este proyecto es un aporte relevante para lograr dicho desafío.

1.2. Relevancia del proyecto

En general, el problema de la contaminación hídrica constituye una amenaza ambiental de relevancia. En particular, este problema es muy sentido en la cuenca del río Santa Lucía, ya que es la fuente de agua bruta para potabilización y abastecimiento de aproximadamente el 60% de la población del país. Los procesos que degradan la calidad de las aguas superficiales y subterráneas en la cuenca del río Santa Lucía han sido acumulativos y en caso de superar los umbrales de autodepuración pueden provocar procesos de contaminación irreversibles que afectarían el suministro dela calidad del agua para el consumo humano con consecuencias sociales y económicas que incidirían notoriamente en la calidad de vida de la población. La trasformación de suelos naturales a diferentes usos aumenta la generación y transmisión de contaminantes a los cuerpos de aguas receptores (Miller et al., 2014). Por ejemplo, en las áreas

forestales ganaderas de la cuenca del río Santa Lucía, localizadas en las sierras del este, los procesos de contaminación de los recursos hídricos están asociados al uso de agrotóxicos para el mantenimiento de los sistemas forestales libres de plagas. El lavado realizado por las aguas pluviales dirige los químicos en dilución hacia los cursos fluviales mediante la escorrentía, y como los suelos superficiales tienden a tener buen drenaje, infiltran las aguas contaminadas afectando la calidad hídrica en los acuíferos superficiales y profundos (Achkar et al., 2013). Las prácticas hortofrutícolas y las agrícolas cerealeras orientadas con criterios productivistas en suelos que han sido erosionados y/o han perdido fertilidad natural requieren la utilización de grandes volúmenes de insumos químicos y agrotóxicos para su mantenimiento. No solo resultan insustentables desde el punto de vista energético y económico, sino que, además, actúan como factor desencadenante de procesos de contaminación química de suelos y aguas superficiales y subterráneas. La instalación de industrias y agroindustrias en las periferias urbanas y en las zonas rurales agrícolas – ganaderas y hortofrutícolas, afecta la calidad de las aguas en los niveles freáticos, acuíferos y cursos fluviales (Achkar et al., 2012). El vertido de efluentes industriales sin tratamiento en las aguas constituye un factor de contaminación hídrica de relevancia. En los centros urbanos, las principales causas de la contaminación ambiental y afectación de la calidad de las aguas están vinculadas a la falta o insuficiencia de saneamiento y al manejo y deposición final de los residuos sólidos. En el primer caso, los efluentes domésticos y las infiltraciones desde pozos negros son desencadenantes de la contaminación de las napas freáticas e indirectamente, de los cursos fluviales, mientras que, en el segundo caso, lo es el lixiviado de los residuos en los vertederos sin impermeabilizar.

Basándonos en la importancia del problema de la contaminación hídrica y en su relevancia a nivel nacional para las instituciones y para la población local, este proyecto tiene los objetivos planteados a continuación.

1.3. Objetivos del proyecto

El objetivo general del proyecto es desarrollar una metodología donde se utilizarán datos de series temporales, datos geográficos y técnicas de aprendizaje automático para investigar las correlaciones entre los cambios en el uso del suelo y/o cobertura del suelo (LULC, por su sigla en inglés), y los parámetros físico-químicos de calidad del agua.

Los objetivos específicos (OE) son los siguientes:

- *OE1*: Relevar las principales fuentes de datos, y proceder a la evaluación de su calidad, aplicando herramientas de depuración para obtener datos integrados.

- *OE2*: Evaluar los cambios temporales y espaciales de uso del suelo en la cuenca del río Santa Lucía para el período dado por los años 2000, 2008, 2011 y 2015 en base a los datos disponibles del MVOTMA.
- *OE3*: Analizar la variabilidad temporal y espacial de los parámetros de calidad del agua físico-química en la cuenca hidrográfica del río Santa Lucía a partir de los datos recopilados por DINAMA -MVOTMA.
- *OE4*: Evaluar la relación entre las categorías definidas por uso del suelo con las variables de calidad del agua en sitios críticos de la cuenca del río Santa Lucía

1.4. Descripción de la cuenca de estudio

La cuenca del río Santa Lucía es de importancia estratégica para la sociedad uruguaya, ya que es una de sus principales fuentes de abastecimiento hídrico, dado que provee de agua potable al 60% de la población de todo el país (Achkar et al., 2012). En la subcuenca del río Santa Lucía Chico se localiza la reserva de Paso Severino, un embalse de 15 km^2 de superficie, 20 m de profundidad máxima y con una capacidad de almacenamiento de 65 Hm^3, que recibe agua desde un área de drenaje de 2500 km^2 (departamento de Florida) (Ríos, 2019). El embalse posibilita el control del caudal del río, ya que aguas abajo, se abastece a la planta purificadora de OSE (Obras Sanitarias del Estado), localizada en Aguas Corrientes (Canelones), que abastece a la ciudad de Montevideo (Achkar et al., 2013). El principal destino del agua purificada es el uso doméstico. Contar con un sistema de gestión adecuado de los recursos hídricos en la cuenca es fundamental para disponer de agua en calidad suficiente, para cubrir todos los usos necesarios y evitar conflictos ambientales.

Basándonos en lo anterior, este proyecto utiliza la subcuenca del río Santa Lucía Chico como cuenca piloto (Fig. 1.1). Además, se tuvo en cuenta que: *i)* es una subcuenca de cabecera; *ii)* existe confianza en los datos de caudal; *iii)* hay presencia de varias estaciones de monitoreo de calidad de agua gestionadas por varias instituciones nacionales.

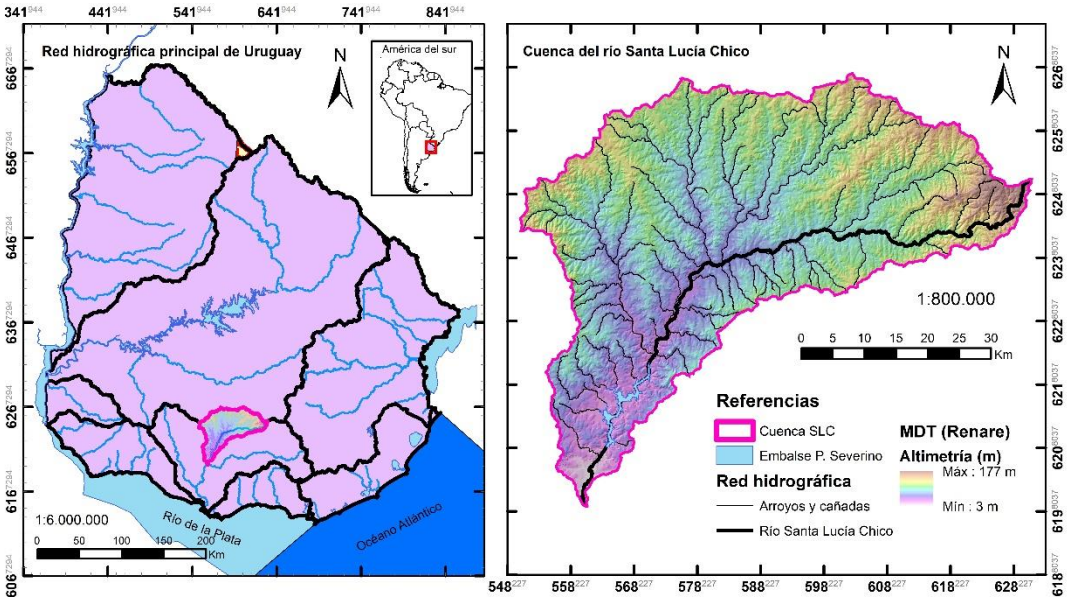


Fig. 1.1. Cuenca de estudio seleccionada: subcuenca del río Santa Lucía Chico.

La cuenca del río Santa Lucía Chico se encuentra comprendida completamente en el departamento de Florida, su área superficial es de aproximadamente 2.570 km^2, el perímetro es de 300 km y el índice de compacidad 1,66. La información de altimetría utilizada para su traza fue tomada del Modelo Digital del Terreno de la Recursos Naturales Renovables (RENARE) del Ministerio de Ganadería, Agricultura y Pesca (MGAP, 2020). La variación en altura se encuentra entre los 177 metros en la zona noreste, 25 metros en el embalse de Paso Severino y 3 metros en la desembocadura del cauce principal. La pendiente media de la cuenca se estimó en 2,68%, la longitud del cauce principal 128,7 km, y el tiempo de concentración mediante el método de de Ramser y Kirpich para flujo concentrado en 1,5 días.

La cuenca se encuentra ubicada sobre gneises del Terreno Piedra Alta (basamento cristalino de la cuenca intracratónica del plata). Los suelos dominantes son Brunosoles con textura franca y ricos en materia orgánica, con mayores proporciones de arcillas y limos en la parte sur. El clima es templado, con cuatro estaciones diferentes, caracterizado por una precipitación anual acumulada que varía entre 1000 mm y 1500 mm, y temperaturas que pueden variar entre los 3°C y los 30°C (Gorgoglione et al., 2020). La Fig. 1.1 presenta la red hidrográfica principal de Uruguay y la cuenca del Santa Lucía Chico, mientras que la Tabla 1.1 resume las principales características de la cuenca del Santa Lucía Chico.

Tabla 1.1. Características físicas de la cuenca del Santa Lucía Chico.

Cuenca SLC	
Área (km^2)	2569,62
Perímetro (km)	300,03
Índice de compacidad	1,66
Máximo desnivel (m)	174
Pendiente media de la cuenca (%)	2,68
Longitud del cauce principal (km)	128,78
Tiempo de concentración (días)	1,52

2. OE1: Relevar las principales fuentes de datos, y proceder a la evaluación de su calidad, aplicando herramientas de depuración para obtener datos integrados.

Para llevar adelante el OE1, se realizó un relevamiento exhaustivo de todas las fuentes de datos disponibles y relevantes. Luego, se aplicaron técnicas de *data profiling* para diagnosticar la calidad de los mismos, buscando detectar valores faltantes, datos inconsistentes, problemas de precisión de los datos, entre otros. Se procedió a continuación a diseñar y aplicar técnicas para el mejoramiento de la calidad de los datos disponibles, las cuales incluyen el uso de métodos de aprendizaje automático (*data imputation*). Por último, se procedió a integrar los datos.

2.1. Datos disponibles en la cuenca

Los datos disponibles para este estudio pueden organizarse en dos grupos. Por un lado, los que refieren al uso y cobertura del suelo, y por otro los que refieren a la cantidad/calidad del agua y clima.

A continuación, se presenta un resumen de los conjuntos de datos fuente que van a ser utilizados en el marco de este proyecto, indicando para cada uno el organismo que lo provee, el periodo de recolección, el formato, y para cada variable contenida su nombre y unidad de medida.

1. Datos hidrológicos:
    - Fuente: DINAGUA
    - Período: 1/1/1971-30/12/2020
    - Formato: planilla
    - Estaciones:
        - Estación de medición de caudal DINAGUA (ver Tabla 2.3)
    - Variables:
        - Caudal (Q). Unidad: m^3/s
        - Nivel de agua (h). Unidad: m
2. Datos meteorológicos:
    - Fuente: INUMET-INIA
    - Período: 1/1/1980-30/6/2020
    - Formato: planilla
    - Estaciones:

- Estación experimental INIA (ver Tabla 2.2)
- Estación pluviométrica convencional INUMET (ver Tabla 2.2)

- Variables:
  - Precipitación (P): Unidad: mm
  - Temperatura del aire (TA): Unidad: ºC
  - Humedad relativa (HR): Unidad: %
  - Radiación solar (RS): Unidad: W/m^2
  - Heliofanía (Hel): Unidad: hs
  - Evapotranspiración Penman (ET): Unidad: mm
  - Velocidad del viento (VV): Unidad: 2m/km/24hs

3. Datos de uso del suelo:
   - Fuente: DINAMA-DINOT
   - Período: 2000 - 2008 - 2011 – 2015 – 2016 – 2018
   - Formato: shape
   - Variable: Uso del suelo
4. Datos de calidad de agua:
   - Fuente: DINAMA
   - Período: 1/1/2004 - 31/8/2020
   - Formato: planilla
   - Estaciones:
     - Estaciones de calidad de agua DINAMA (ver Tabla 2.3)
   - Variables:
     - Fosforo total (PT): Unidad: µg/L
     - Nitrógeno total (NT): Unidad: mg/L
     - Ion nitrato (NO_3^-): Unidad: mg/L
     - Ion nitrito (NO_2^-): Unidad: mg/L
     - Ion amonio (NH_4^+): Unidad: mg/L
     - Ion fosfato (PO_4^{3-}): Unidad: µg/L
     - Glifosato: Unidad: µg/L
     - Sólidos totales (ST), Solidos suspendidos totales (SST): Unidad: mg/L
     - Turbidez: Unidad: NTU
     - Temperatura (T): Unidad: ºC
     - Oxígeno disuelto (OD): Unidad: mg/L
     - Demanda bioquímica de oxígeno (DBO): Unidad: mg/L
     - Clorofila-a (Chl-a): Unidad: µg/L
     - Potencial de hidrógeno (pH)

- Conductividad: Unidad: µS/cm

Los datos que se refieren al uso y cobertura del suelo serán descriptos en el capítulo 4. En este capítulo nos enfocamos en las series temporales (datos hidrológicos, meteorológicos y de calidad de agua).

En algunos casos, los conjuntos de datos están almacenados en más de una planilla, o se observan frecuencias de muestreo variable dependiendo de la variable y/o período. La Tabla 2.1 brinda información detallada en este sentido.

Tabla 2.1. Conjuntos de datos utilizados.

Tipo de datos	Conjunto	Variables	Frecuencia	Período
Datos hidrológicos	florida_RodChao	Q h	3 por día	1971-2020
Datos meteorológicos	INIA_LasBrujas	ET HR Hel RS TA (max, media, min) VV	1 por día	1981-2020 1999-2020 1991-2020 2008-2020
	INUMET_Meteo_Durazno	TA (max) TA (min)		2010-2014
		TA (min)		2010-2016
	INUMET_Meteo_Florida	TA (max) TA (min) Hel		2010-2013
		TA (min) Hel		2010-2016
	INUMET_Precipitacion/1980_1992/*.xlsx	P		1980-1992
	INUMET_Precipitacion/1993_2018/ *.xlsx			1993-2017
Datos de calidad de agua (CA)	SantaLucia_2004_2010	DBO NH_4^+ NO_2^- PO_4^{3-} NO_3^- PT T Turbidez PH NT OD Chl-a ST	1 por mes (para 2 estaciones), entre 2 y 4 meses por año (varía según el año)	2004-2020 2009-2020 2004-2020

	SantaLucia_2011-2018	Chl-a	1 por mes (para 4 estaciones), entre 3 y 6 meses por año (varía según el año)	2011-2020
		Conductividad NH_4^+ NO_3^- NO_2^- FT OD T Turbidez NT	1 por mes (para 5 estaciones), entre 3 y 6 meses por año (varía según el año)	2013-2020
		DBO	1 por mes (para 3 estaciones), entre 3 y 6 meses por año (varía según el año)	
		Glifosato	1 por mes, entre 3 y 6 meses por año (varía según el año)	2014-2018

En la Fig. 2.1 y Tabla 2.2 se presenta la ubicación y coordenadas, respectivamente, de las estaciones meteorológicas y pluviométricas de INUMET y la estación experimental INIA Las Bujas.

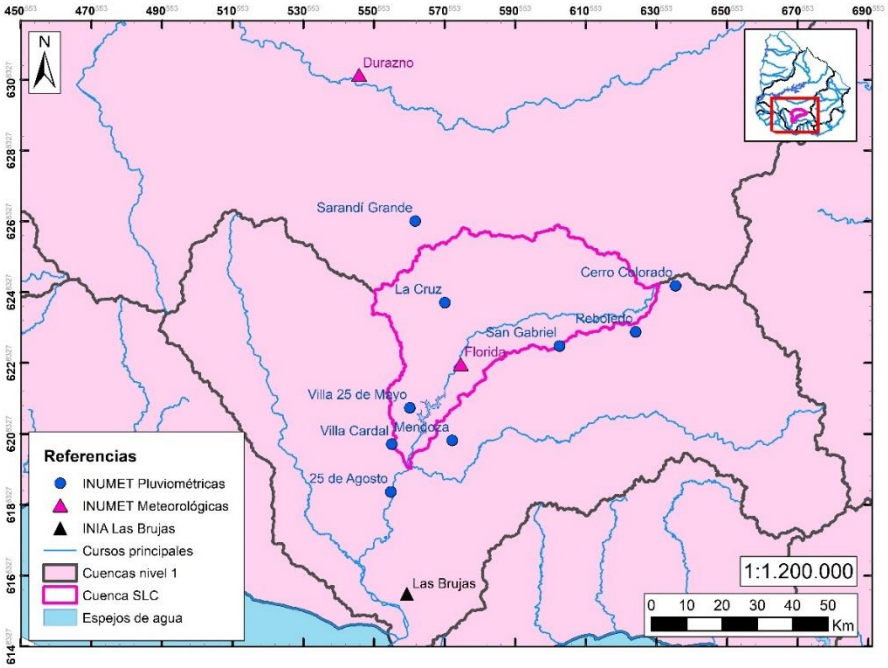


Fig. 2.1. Ubicación de estaciones pluviométricas y meteorológicas de INUMET y estación experimental INIA Las Brujas.

Tabla 2.2. Coordenadas de estaciones pluviométricas y meteorológicas de INUMET y estación experimental INIA Las Brujas.

Estación pluviométrica convencional INUMET				
Nombre	**Departamento**	**Código**	**X (m) UTM 21S**	**Y (m) UTM 21S**
25 de Agosto	Florida	2748A	555324	6191996
Cerro Colorado	Florida	2498A	635948	6250177
La Cruz	Florida	2538A	570632	6245431
Mendoza	Florida	2670	572715	6206515
Reboledo	Florida	2543	624675	6237133
San Gabriel	Florida	2586	603110	6233179
Sarandí Grande	Florida	2395A	562280	6268418
Villa 25 de Mayo	Florida	2669A	560716	6215694
Villa Cardal	Florida	2710A	555588	6205413
Estación meteorológica convencional INUMET				
Nombre	**Departamento**		**X (m) UTM 21S**	**Y (m) UTM 21S**
Durazno	Durazno		546367	6309615
Florida	Florida		575157	6227972
Estación experimental INIA				
Nombre	**Departamento**		**X (m) UTM 21S**	**Y (m) UTM 21S**
Las Brujas	Canelones		560349	6163260

En la Fig. 2.2 y la Tabla 2.3 se presenta la ubicación y coordenadas, respectivamente, de las estaciones de monitoreo de calidad de agua de la DINAMA y la estación de medición de caudal de la DINAGUA.

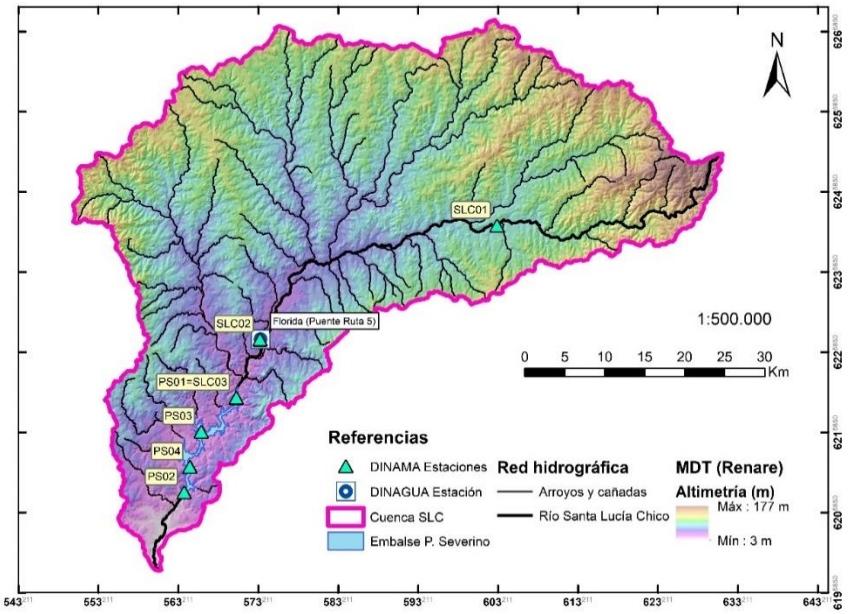


Fig. 2.2. Ubicación de estaciones de monitoreo de calidad de agua de la DINAMA y estación de medición de caudal de la DINAGUA.

Tabla 2.3. Coordenadas de estaciones de monitoreo de calidad de agua de la DINAMA y estación de medición de caudal de la DINAGUA.

Estación de medición de caudal DINAGUA			
Nombre	**Departamento**	**X (m) UTM 21S**	**Y (m) UTM 21S**
Florida (Puente Ruta 5)	Florida	573493	6227455
Estaciones de calidad de agua DINAMA			
Código	**Departamento**	**X (m) UTM 21S**	**Y (m) UTM 21S**
PS02	Florida	563941	6208344
PS04	Florida	564656	6211500
PS03	Florida	566095	6215901
PS01=SLC03	Florida	570470	6220119
SLC02	Florida	573434	6227394
SLC01	Florida	603079	6241618

2.2. Análisis de los datos: técnicas y metodologías utilizadas

2.2.1. Data profiling

Las técnicas de *Data Profiling (DP)* tienen como principal objetivo extraer características relevantes sobre los conjuntos de datos y su calidad (por ejemplo, porcentaje de datos faltantes). Todo el trabajo realizado en esta etapa del proyecto se realizó en un ambiente Python equipado con los paquetes necesarios para la correcta ejecución del código creado. Para la creación de este

ambiente se usó el distribuidor de paquetes *conda*[1] y se creó uno script para Linux que se encarga de preparar dicho ambiente.

El DP se realizó en *notebooks* de Python, utilizando el paquete *pandas_profiling*[2] a fin de mantener los resultados una vez ejecutado el código y generar reportes en formato HTML, los cuales incluyen gráficas que permiten visualizar los resultados obtenidos. Tanto los *notebooks* como los scripts y el código se encuentran disponibles en el repositorio del proyecto[3].

Los conjuntos de datos presentados en la Tabla 2.1 se combinaron agrupando por tipo de datos. En el caso de los datos meteorológicos, se agrupan por un lado los datos de precipitaciones y por otro las demás variables meteorológicas. Se obtienen de esta forma cuatro conjuntos de datos para los cuales se adopta la nomenclatura **[Tipos de datos]_[Fuente de los datos]_[Período]**, los tipos de datos son CA (calidad de agua), MET (meteorología) e HIDRO (hidrología). Antes de realizar el DP, las diferentes planillas que componen cada conjunto de datos fueron fusionadas, y el resultado de estas fusiones se almacenó en formato Apache *parquet*[4] lo cual permite hacer un uso más eficiente de los recursos. A continuación, se presenta para cada uno de estos conjuntos agrupados, los resultados obtenidos y las acciones que se tomaron para solucionar los problemas encontrados.

CA_DINAMA_ 2004_2020

Este conjunto contiene las variables de calidad de agua en el período 2004-2020. Cada una se midió en diferentes estaciones de la cuenca del Santa Lucía, al separar las variables por estación se llegó a una frecuencia de muestreo mensual.

Asimismo, dado que las estaciones **PS01** y **SLC03** identifican una sola estación, se unificaron bajo el identificador **PS01=SLC03** por lo que el conjunto final de estaciones de muestreo es: SLC01, SLC02, PS01=SLC03, PS02, PS03 y PS04.

Los nombres de las variables se unificaron siguiendo la nomenclatura **[Estación] Variable (Abreviación) [Unidad de medida]**, el conjunto final de variables es:

- Clorofila-a (Chl-a) [µg/L]
- Conductividad [µS/cm]
- Demanda bioquímica de oxígeno (DBO) [mg O_2/L]
- Fósforo total (PT) [µg P/L]
- Glifosato [µg/L]

[1] Distribuidor de paquetes *conda* https://docs.conda.io/en/latest/
[2] https://pandas-profiling.github.io/pandas-profiling/docs/master/index.html
[3] Repositorio de datos y código del proyecto https://gitlab.fing.edu.uy/hydroinformatics/fsda-lu-wq
[4] Apache *parquet* https://parquet.apache.org/

- Ion nitrito (NO_2-N) [mg NO_2-N/L]
- Ion amonio (NH_4-N) [mg NH_4-N/L]
- Ion fosfato (PO_4-P) [µg/L]
- Nitrógeno total (NT) [mg N/L]
- Oxígeno disuelto (OD) [mg/L]
- Temperatura del agua (T) [ºC]
- Turbidez [NTU]
- Potencial de hidrógeno (pH) [NA]
- Sólidos totales (ST) [mg/L]
- Sólidos suspendidos totales (SST) [mg/L]

Al realizar el DP se encontraron variables con valores no numéricos, en particular rangos de valores y las etiquetas "<LC" y "<LD", que significan respectivamente límite de cuantificación y límite de detección[5]. Se procedió a reemplazar el valor LC por el valor mínimo de la variable, a reemplazar el valor LD por un valor menor a LC designado oportunamente, a promediar los rangos y a reemplazar los valores registrados como "<x" por $\mathrm{x} - \frac{x}{10}$. En la Tabla 2.4 se muestran los valores detectados para cada variable y los valores calculados de LC y LD siguiendo los criterios antes mencionados. Todas estas modificaciones se documentaron en una nueva columna llamada comentario.

Tabla 2.4. Valores no numéricos detectados para cada variable, LC y LD.

Variable	Valores no numéricos encontrados	LC	LD
Clorofila-a (Chl-a) [µg/L]	'<LD', '<LC', 'LD<x<LC'	0.02	0.01
Demanda bioquímica de oxígeno (DBO) [mg O_2/L]	'<LC', '<LD', 'LD<x<LC'	0.25	0.1
Glifosato [µg/L]	'<LC', '<LD'	0.1	0
Ion amonio (NH_4-N) [mg NH_4-N/L]	'< LC', '<LD'	0.004	0
Ion nitrato (NO_3-N) [mg NO_3-N/L]	'<LD', '<0.060', '0,014 - 0,07'	0.014	0
Ion nitrito (NO_2-N) [mg NO_2-N/L]	'<0.046', '<LD', '<LC', 'LD<X<LC', '0,02 - 0,04'	0.001	0
Sólidos suspendidos totales (SST) [mg/L]	'<LD', '<LC', '<4.6'	9	8

Una vez reemplazados los valores no numéricos se procedió a computar el porcentaje de datos faltantes por variable, estación y período de tiempo.

En la Fig. 2.3, donde en el eje horizontal se presentan los períodos de tiempo para los cuales hay datos para cada variable/estación y el color representa el porcentaje de valores faltantes durante ese período (más oscuro, más datos faltantes), se puede apreciar que el porcentaje de valores

[5] El LC es la cantidad más pequeña del analito en una muestra que puede ser cuantitativamente determinada con exactitud aceptable. El LD se define como la cantidad o concentración mínima de sustancia que puede ser detectada con fiabilidad por un método analítico determinado. Generalmente LC > LD.

faltantes es mayor a 53% para todas las variables, y menos de la mitad cuenta con mediciones antes del año 2011.

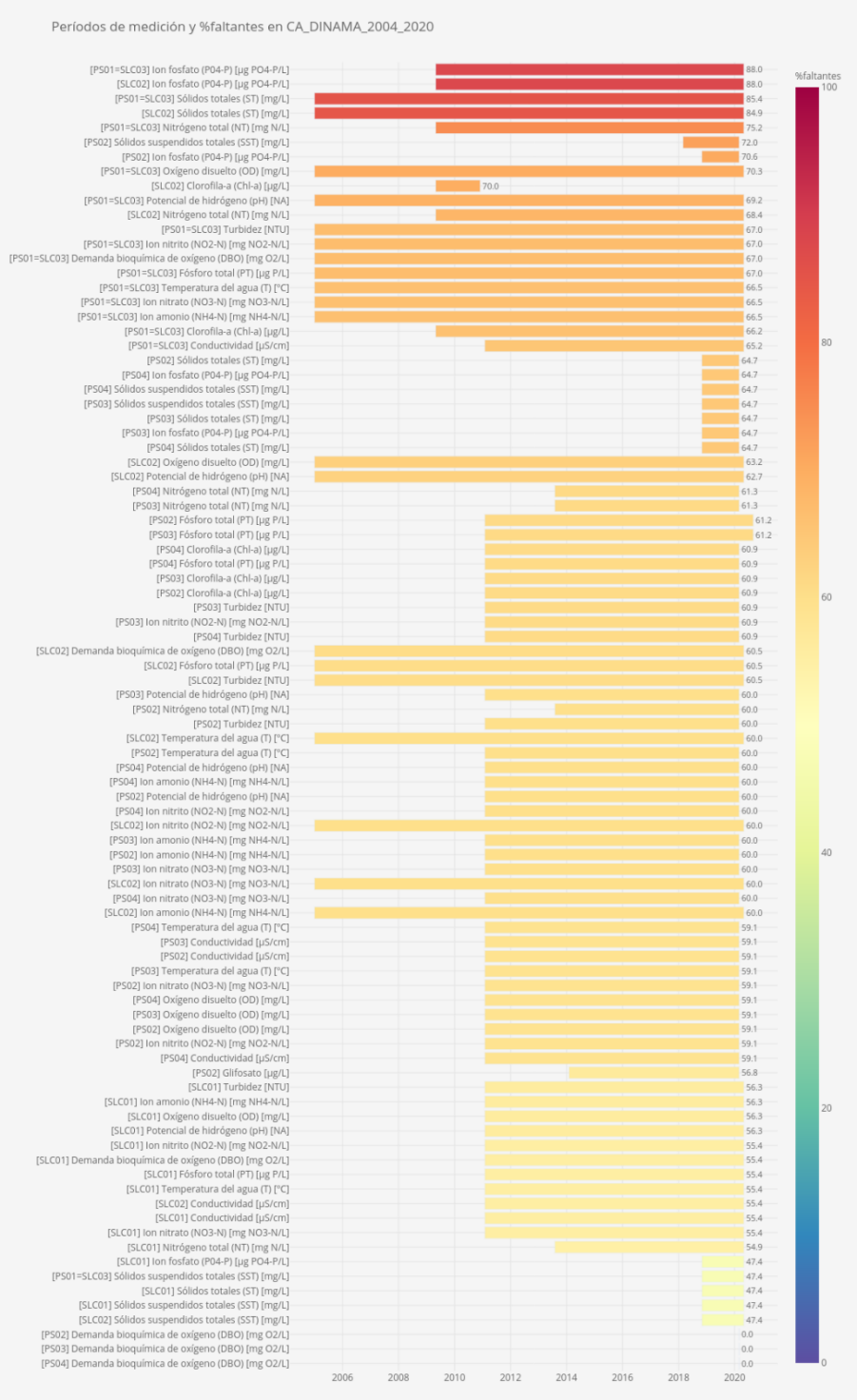


Fig. 2.3. Períodos de medición y % de datos faltantes en el conjunto CA_DINAMA _2004_2020.

HIDRO_DINAGUA_1971_2020

Este conjunto contiene las variables hidrológicas en el período 1971-2020, caracterizadas por una frecuencia de tres muestreos por día.

Los nombres de las variables se unificaron siguiendo la nomenclatura **[Estación] Variable (Abreviación) [Unidad de medida]**, el conjunto final de variables es:

- [Florida (Puente Ruta 5)] Caudal (Q) [m^3/s]
- [Florida (Puente Ruta 5)] Nivel de agua (h) [m]

El proceso de DP no encontró problemas con los valores de las variables de este conjunto.

Se procedió a computar el porcentaje de valores faltantes obteniendo los resultados mostrados en la Fig. 2.4, donde se puede apreciar que el porcentaje de datos faltantes es de sólo 9% para todas las variables.

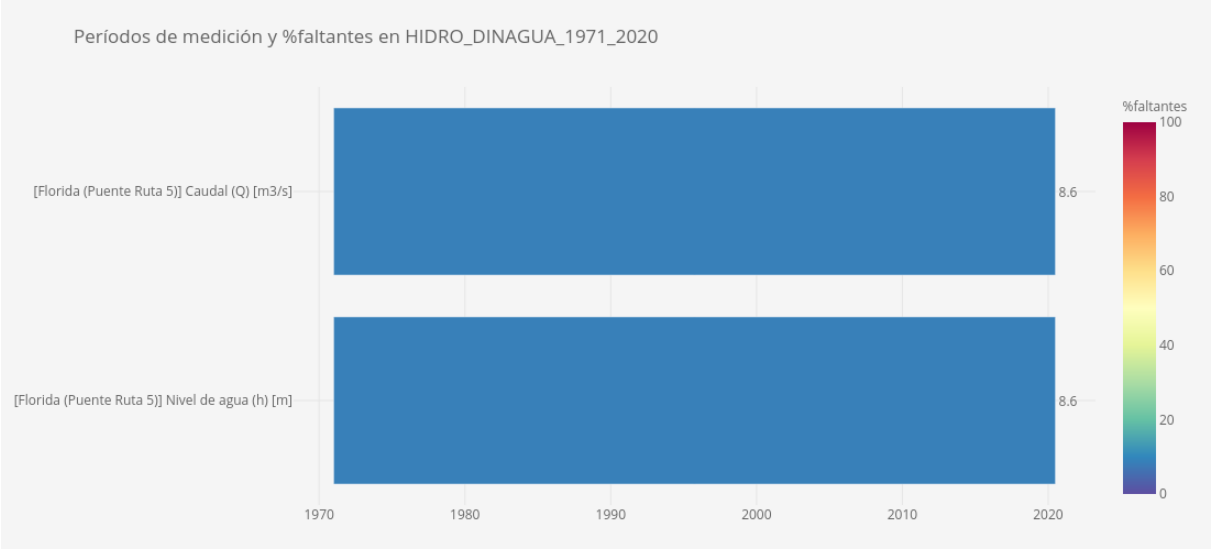


Fig. 2.4. Períodos de medición y % de datos faltantes en el conjunto HIDRO_DINAGUA_1971_2020.

MET_INIA_1980_2020

Este conjunto contiene las variables meteorológicas en el período 1980-2020. Cada una se guardó con frecuencia de muestreo diaria.

Los nombres de las variables se unificaron siguiendo la nomenclatura **[Estación] Variable (Abreviación) [Unidad de medida]**, el conjunto final de variables es:

- [Las Brujas] Evapotranspiración Penman (ET) [mm]
- [Las Brujas] Heliofanía (Hel) [hs]
- [Las Brujas] Humedad relativa media (HR) [%]
- [Las Brujas] Radiación solar (RS) [$cal/cm^2/día$]

- [Las Brujas] Temperatura del aire media (TA) [°C(24hs)]
- [Las Brujas] Temperatura del aire máxima (TA) [°C]
- [Las Brujas] Temperatura del aire mínima (TA) [°C]
- [Las Brujas] Velocidad del viento (VV) [2m/km/24hs]

El proceso de DP no encontró problemas con los valores de las variables de este conjunto.

Se procedió a computar el porcentaje de valores faltantes obteniendo los resultados mostrados en la Fig. 2.5, donde se puede apreciar que el porcentaje de datos faltantes es menor al 1% para todas las variables y que existen variables que no tienen mediciones antes del año 2009.

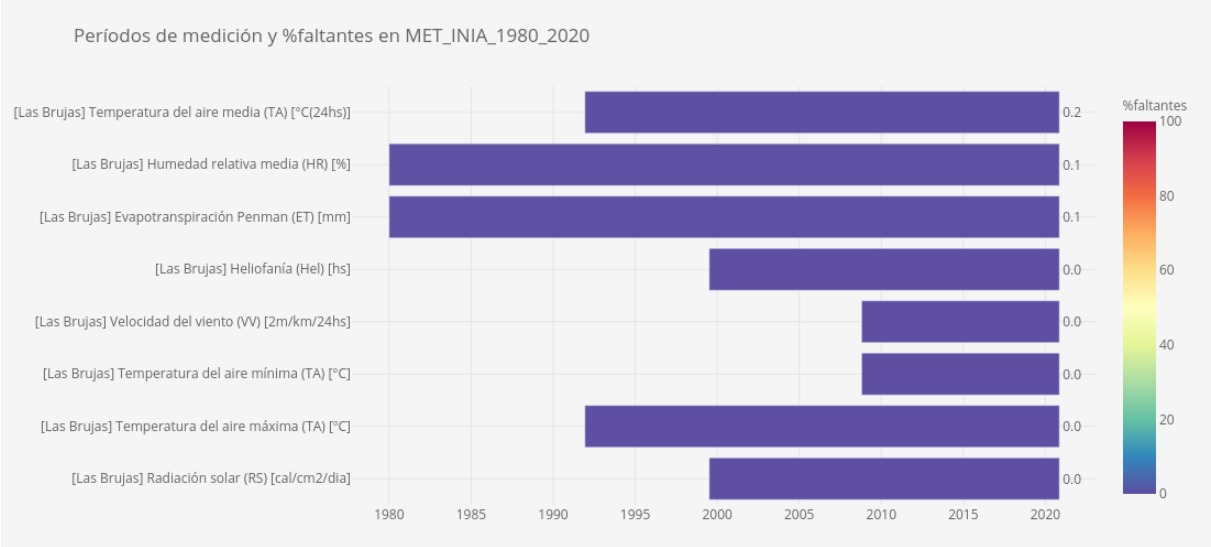


Fig. 2.5. Períodos de medición y % de datos faltantes en el conjunto MET_INIA_1980_2020.

MET_INUMET_1980_2020

Este conjunto contiene la variable meteorológica Precipitación (P) [mm] en el período 1980-2020 con frecuencia de muestreo diaria.

Las variables están listadas por estación, las estaciones **Villa Cardal** y **Villa Cardal AFE** se unificaron bajo el identificador **Villa Cardal** y las estaciones **Reboledo** y **Reboledo AFE** se unificaron bajo el identificador **Reboledo**. El conjunto final de estaciones es:

- Reboledo
- San Gabriel
- Villa 25 de Mayo
- Mendoza
- Cerro Colorado
- Sarandí Grande
- La Cruz

- Villa Cardal
- 25 de Agosto
- Florida

Los nombres de las variables se unificaron siguiendo la nomenclatura **[Estación] Variable (Abreviación) [Unidad de medida]**. Por otro lado, se constató que el conjunto de datos cuenta con una columna comentario donde se incluyen anotaciones en texto que indican datos faltantes o características del fenómeno (por ejemplo, chaparrones). Se decidió utilizar la información de esta columna para corregir los valores de la variable. En la Tabla 2.5 se muestra, para cada comentario, la transformación aplicada.

Tabla 2.5. Transformaciones aplicadas sobre el conjunto MET_INUMET_1980_2020 a partir de los comentarios existentes.

Comentario	Transformación
Dato faltante.	Marcar valor faltante
Sin registro en todo el mes.	
TRAZA	Se asigna valor 0.09
Chaparrones	Nada
Lloviznas	

Luego, el proceso de DP no encontró problemas con los valores de las variables de este conjunto. Respecto al porcentaje de valores faltantes, los resultados obtenidos se muestran en la Fig. 2.6, donde se puede apreciar que el porcentaje de datos faltantes es menor al 9% para todas las variables.

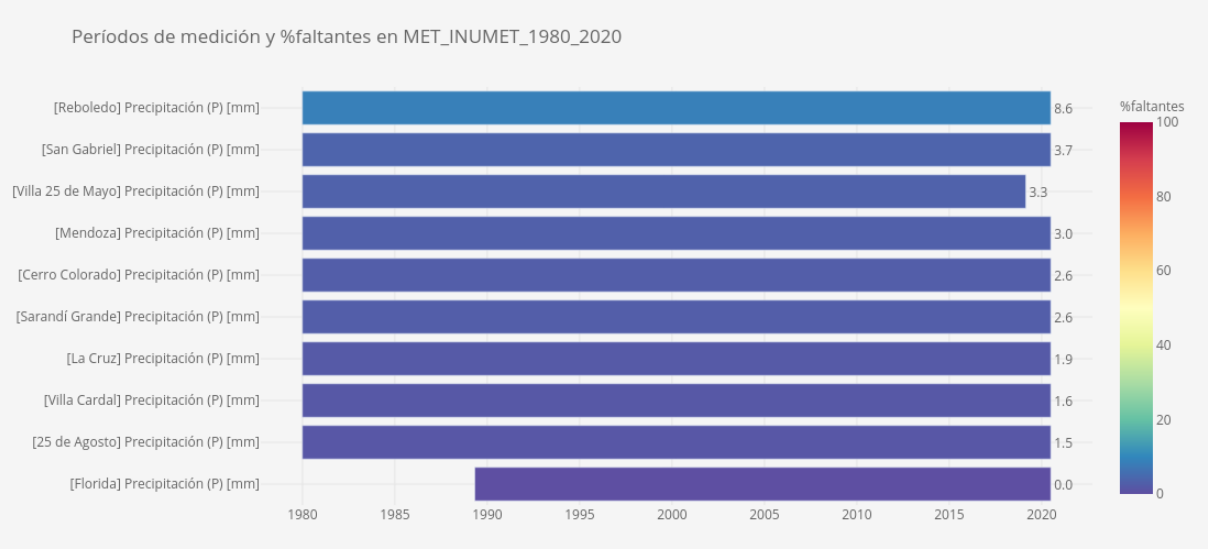


Fig. 2.6. Períodos de medición y % de datos faltantes en el conjunto MET_INUMET_1980_2020.

La alta proporción de datos faltantes en el conjunto de datos sobre calidad de agua hace necesaria la aplicación de técnicas para completar o imputar estos valores. A continuación, se describen las metodologías estudiadas y aplicadas con este fin.

2.2.2. Data imputation (imputación de datos faltantes)

Tal como fue presentado en la sección anterior, los resultados del DP mostraron que el conjunto de datos de calidad de agua (CA_DINAMA_2004_2020) tiene un porcentaje de valores faltantes entre 50% y 70% para todas sus variables, lo cual es muy alto dada la gran importancia que estas variables tienen para el proyecto. Con la finalidad de mejorar su calidad, y en particular la dimensión completitud, se estudian técnicas de imputación de datos faltantes que permitan completar las series temporales.

Antes de decidir qué técnicas explorar y generar una metodología a aplicar, se decidió reducir el período temporal sobre el que trabajar a uno en el que siempre existan mediciones, este período es 2014-2020 y en la Fig. 2.7 se muestran los valores faltantes en todas las variables de los conjuntos dentro de ese período. Todavía, se siguen manteniendo unos altos índices de valores faltantes en el conjunto de variables del grupo de calidad de agua (entre 57% y 66%).

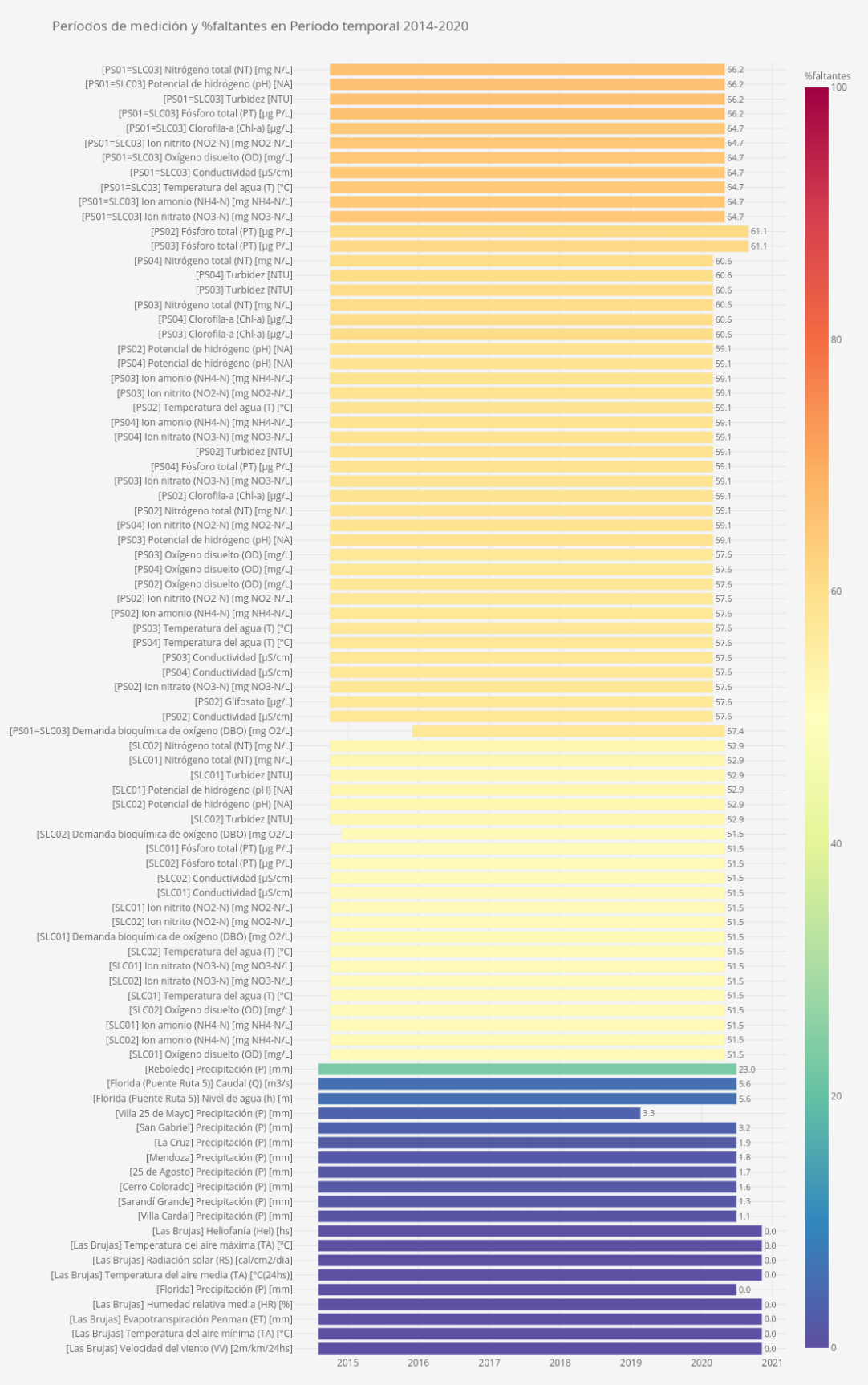


Fig. 2.7. Períodos de medición y % de datos faltantes de todas las variables en el período 2014-2020.

Al momento de realizar la imputación de una variable, varias técnicas y modelos pueden usarse. En esta investigación se consideró fundamental que los métodos seleccionados tuvieran buen poder predictor a partir de una cantidad reducida de datos. Esto es fundamental pues, como se vio anteriormente, el grupo de variables de mayor relevancia cuenta con pocas muestras y muchos datos faltantes.

Los métodos de imputación pueden agruparse en dos categorías: aquellos que no usan información de otras variables para imputar (métodos de imputación univariable) y aquellos que si la usan (métodos de imputación multivariable) (Durbin et al., 2001). Dependiendo de los datos, algunos pueden funcionar mejor que otros, por esta razón, se decidió que no se descartaría ninguno a priori. A continuación, se listan y se describen brevemente los métodos investigados:

Métodos de imputación univariable

En esta categoría se probaron dos métodos que se describen a continuación:

- *Inverse Distance Weighting (IDW)*: modelo que imputa un punto basándose en puntos de la misma variable en otras coordenadas geográficas. Este modelo asigna pesos a los puntos que usa para la imputación dependiendo de su distancia con el punto a imputar (Fortin et al., 2006). Se realizó su implementación en el repositorio del proyecto.
- ARIMA: modelo que imputa un dato basándose en datos anteriores y en estadísticas de toda la serie temporal. Este modelo usa fórmulas estadísticas para analizar los datos anteriores de una serie y generar nuevos datos (Durbin et al., 2001). Se usó la biblioteca *pmdarima*[6] para su implementación. Se descartó su uso en el *pipeline* de imputación pues este tipo de modelo solo funciona bien para series temporales de alta cardinalidad y el tiempo de ajuste es muy alto comparado con el resto de los modelos.

Métodos de imputación multivariable

Los métodos de esta categoría se basan en un conjunto de modelos compuesto por modelos de regresión simple y modelos de regresión que se apoyan en *machine learning*:

- *Random Forest Regressor:* modelo que entrena un conjunto de árboles de decisión en varios subconjuntos de datos y promedia las predicciones para mejorar la capacidad predictiva y controlar el sobreajuste (Breiman, 2001). Se usó la biblioteca *sklearn*[7] para su implementación, en una etapa más tardía del proyecto se descartó este tipo de modelo en favor del *Extremely Randomized Trees Regressor.*
- *Extremely Randomized Trees Regressor:* modelo que entrena un conjunto de árboles de decisión en varios subconjuntos de datos y promedia las predicciones para mejorar la

[6] https://alkaline-ml.com/pmdarima/auto_examples/arima/example_auto_arima.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

capacidad predictiva y controlar el sobreajuste, se diferencia del *Random Forest Regressor* en la forma en la que se decide el punto de división de un nodo, en este caso se hace una división aleatoria mientras que en el anterior se hace de forma óptima, este cambio permite la aceleración del proceso de entrenamiento sin obteniendo resultados iguales o mejores (Geurts, 2006). Se usó la biblioteca *sklearn*[8] para su implementación.

- *Ridge*: modelo que entrena un modelo de regresión donde se busca minimizar la función de mínimos cuadrados ordinarios con un término extra de regularización dada por la suma de los cuadrados de los valores (norma L2) (Farebrother, 1976). Se usó la biblioteca *sklearn*[9] para su implementación.
- *TheilSen Regressor*: se entrena un modelo de regresión que es robusto a *outliers* ya que usa estadísticas de la muestra para el ajuste de parámetros (Dang et al., 2009). Se usó la biblioteca *sklearn*[10] para su implementación.
- *Huber Regressor*: se entrena un modelo de regresión que es robusto a *outliers* ya que cambia la función de pérdida dependiendo de la muestra usada (Owen, 2006). Se usó la biblioteca *sklearn*[11] para su implementación.
- *Bayesian Ridge*: se entrena un modelo de regresión usando distribuciones estadísticas y la fórmula de Bayes para el ajuste de parámetros (Tipping, 2001). Se usó la biblioteca *sklearn*[12] para su implementación.
- *Support Vector Regressor*: modelo que entrena un *Support Vector Machine* donde se busca minimizar la función de suma de los cuadrados de los coeficientes (norma L2) (Chang et al., 2001). Se usó la biblioteca *sklearn*[13] para su implementación.
- *KNeighbors Regressor*: modelo de regresión basado en *k-nearest neighbours* (KNN), las predicciones se calculan con interpolación local de los puntos más cercanos (Mucherino et al., 2009). Se usó la biblioteca *sklearn*[14] para su implementación.
- *Gradient Boosting Regressor*: modelo que entrena iterativamente un conjunto de árboles de decisión agregando modelos mientras se mejore la predicción la cual se calcula mediante el promedio (Ke et al., 2017). Se usó la biblioteca *lightgbm*[15] para su implementación.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.TheilSenRegressor.html
[11] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.HuberRegressor.html
[12] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html
[13] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html
[15] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html

A partir de los modelos anteriores se crearon nuevos modelos que usan la técnica *Multivariate Imputation by Chained Equations (MICE)* para la imputación. En esta técnica se imputan todos los datos faltantes de cada variable con la media, luego, usa uno de los modelos de regresión antes creados para imputar una de las variables del conjunto a partir del resto; después, selecciona otra y usa la imputación anterior para la nueva imputación, repitiendo este proceso hasta que se cumple un criterio de convergencia (Buuren et al., 2011). Se usó la biblioteca *sklearn*[16] para su implementación.

También se agrega la técnica de imputación *Iterative robust model-based imputation (IRMI)*. Esta técnica selecciona una de las variables con valores faltantes, luego crea un modelo de regresión robusto a partir del resto e imputa la variable seleccionada, este proceso es repetido hasta que se cumple un criterio de convergencia (Templ et al., 2011). Se usó la biblioteca *VIM*[17] para su implementación.

Metodología

Para llevar a cabo el proceso de imputación, se empezó realizando una revisión bibliográfica sobre las posibles dependencias entre las variables. Con estas relaciones se realizó un árbol de dependencias donde se ve el orden en que debería imputarse las variables estudiadas (Fig. 2.8).

[16] https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html#sklearn.impute.IterativeImputer
[17] https://rdrr.io/cran/VIM/man/irmi.html

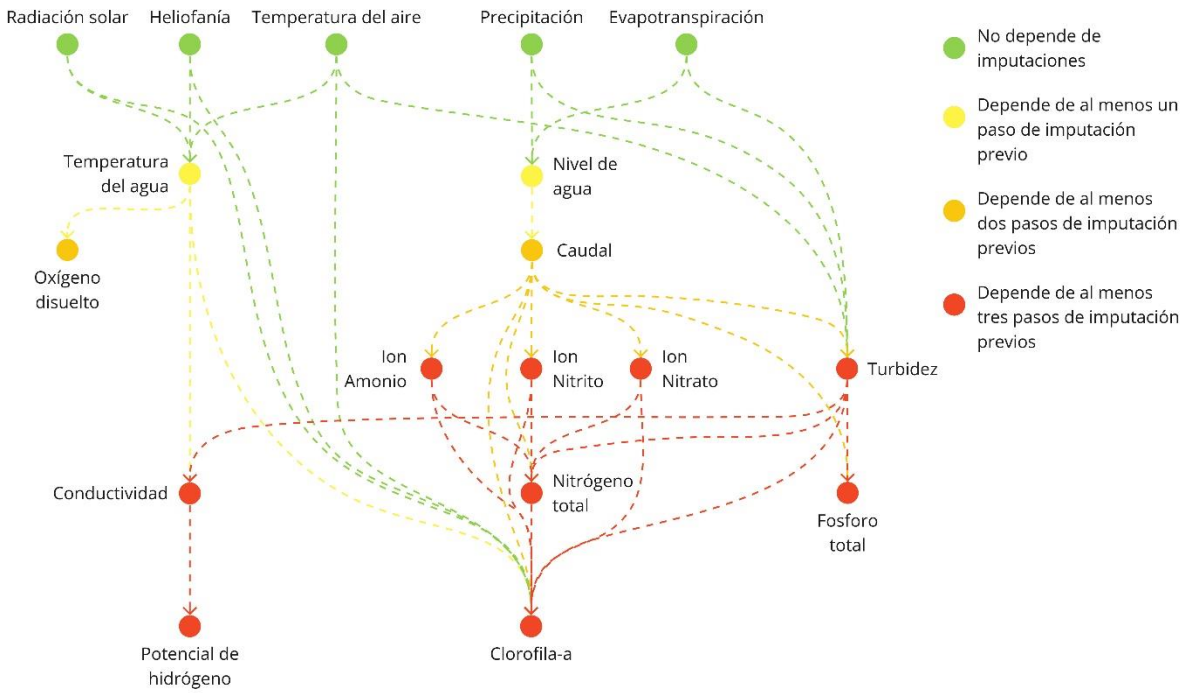


Fig. 2.8 Árbol de dependencias de las imputaciones.

Además, se usó el conocimiento del comportamiento físico de las variables para establecer dependencias espaciales entre las estaciones (aguas arriba/aguas abajo), con esta información se generó otro árbol de dependencias (Fig. 2.9) que, sumado al anterior, se utilizó para decidir el orden de la imputación.

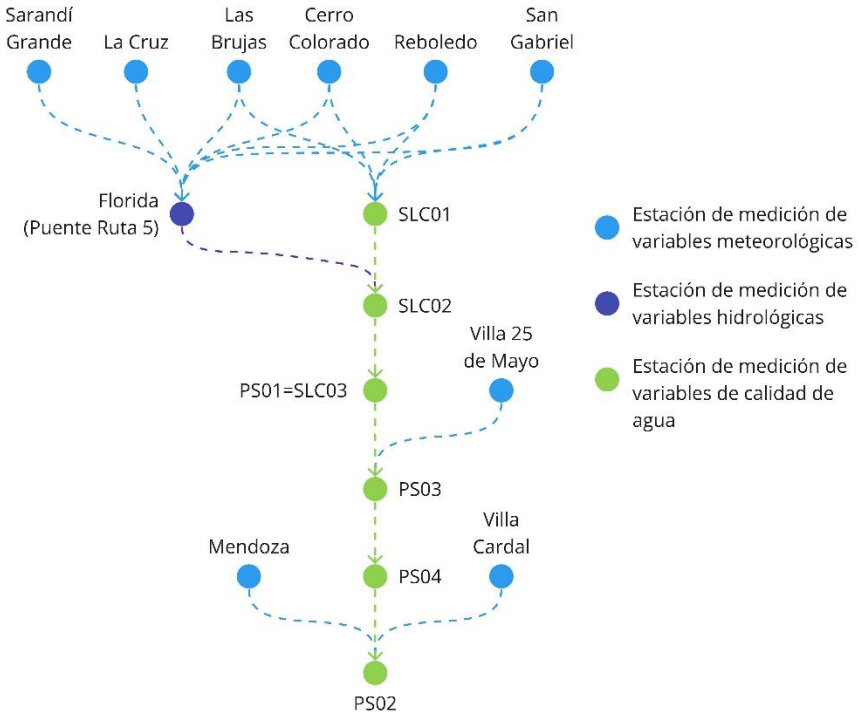


Fig. 2.9 Árbol de dependencias espaciales.

Después, se creó un *pipeline* en el que se comparan los distintos métodos para cada variable y se usan imputaciones ya realizadas para mejorar una imputación actual. El proceso empieza considerando las variables con menor cantidad de datos faltantes (Fig. 2.10).

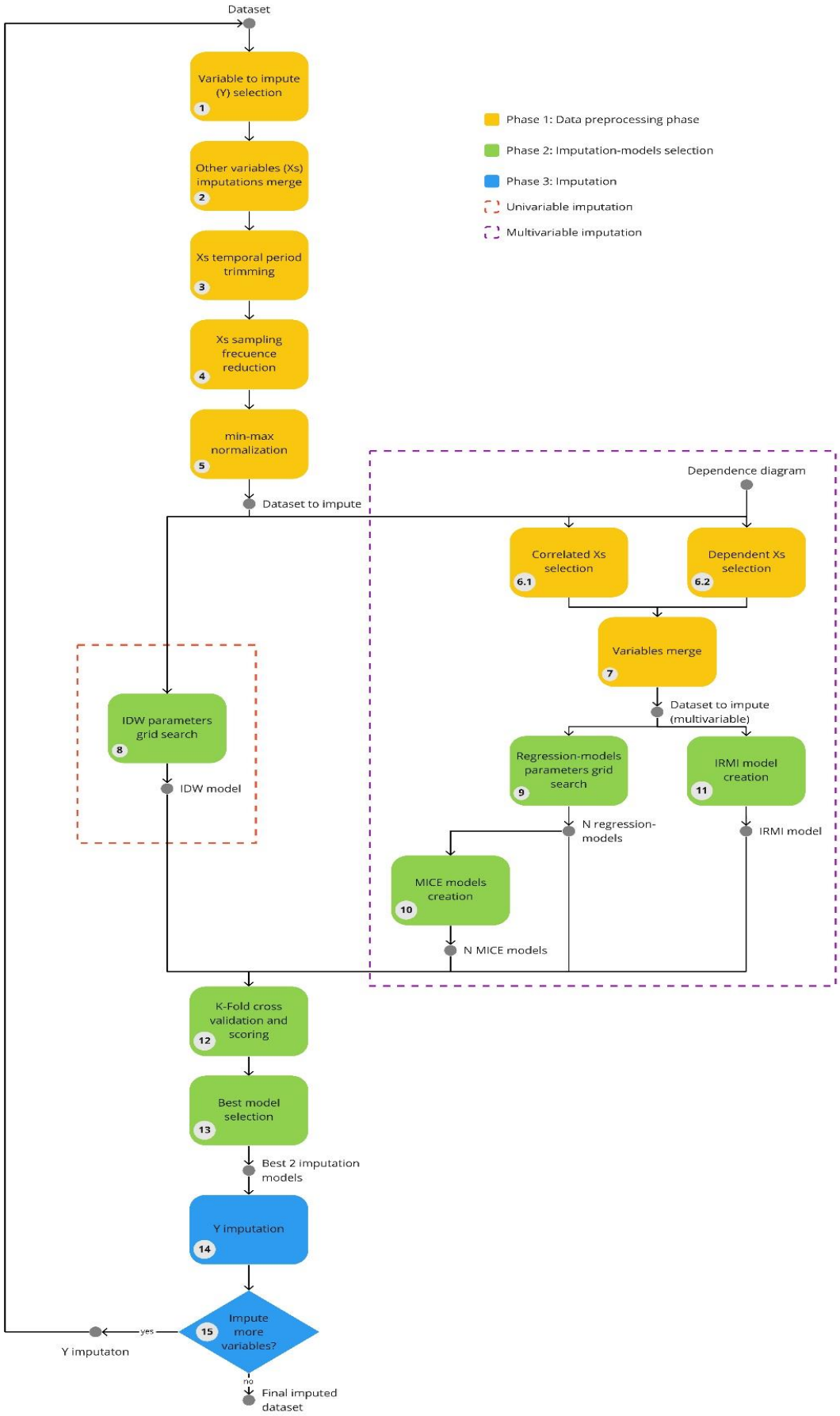


Fig. 2.10. Pipeline de imputación de las variables.

Como se observa en la Fig. 2.10, el *pipeline* se divide en tres fases principales, las cuales se describen a continuación:

Fase 1: Pre-procesamiento de datos

Esta fase tiene como entrada el conjunto de datos de todas las variables y se aplican los siguientes pasos:

1. Se selecciona una variable a imputar (Y) según el orden ya explicado.
2. Para las otras variables (Xs) que tengan una imputación, se reemplazan sus valores por los de la imputación.
3. Se reduce el período temporal de las Xs al período de imputación 2014-2020.
4. Si una variable X tiene frecuencia diaria y la variable a imputar tiene frecuencia mensual, X se reemplaza por las variables X_{max}, X_{mean} y X_{min} reduciendo su frecuencia. En este punto se genera el conjunto de datos para realizar imputación univariable.
5. Se seleccionan los subconjuntos de Xs que se relacionen con Y según los criterios:
    a. Se seleccionan las Xs que presentan una correlación absoluta mayor a 0.5 con Y, para el cálculo de correlación se usa la media de la correlación de *Pearson*, *Kendall* y *Spearman*.
    b. Se seleccionan las Xs de las cuales depende Y según el árbol de dependencias antes detallado.
6. Se combinan los subconjuntos de Xs obtenidos en el paso anterior y se agrega el promedio móvil exponencial ponderado (EWMA) de Y donde la ventana temporal (t) a usar se elige considerando el tiempo de variación de cada variable (Tabla 2.6). Este valor se calcula con la siguiente formula:

$$\begin{cases} \mathrm{EWMA}(y_n) = \dfrac{\sum_{i=0}^{t}(1-\alpha)^i x_{n-1-i}}{\sum_{i=0}^{t}(1-\alpha)^i} \\ \alpha = \dfrac{2}{t+1} \end{cases}$$

7. Si Y es una variable que se mide en varias estaciones, se crean las siguientes versiones del conjunto de datos:
    a. Considerando las dependencias espaciales antes comentadas, se crea una variante del conjunto obtenido donde, para cada estación de Y, solo se consideran las variables de las estaciones de las que esta dependa en el árbol de dependencias.
    b. Se usan las coordenadas geográficas de las estaciones y se crea una nueva variante de los datos donde cada variable se modifica (usando la misma fórmula que le modelo IDW) según su distancia con las estaciones de Y. De esta forma, se agrega de forma implícita información geoespacial en los modelos que de otra forma no la consideran.

El resultado de esta fase son los conjuntos de datos sobre los que se realiza la imputación.

Tabla 2.6. Cantidad de valores usados para el cálculo de EWMA para cada variable imputada.

Variable	Cantidad de datos considerados en la ventana temporal para EWMA
Caudal (Q) [m^3/s]	7
Nivel de agua (h) [m]	
Precipitación (P) [mm]	1
Clorofila-a (Chl-a) [µg/L]	2
Conductividad [µS/cm]	
Demanda bioquímica de oxígeno (DBO) [mg O_2/L]	
Fosforo total (PT) [µg P/L]	
Glifosato [µg/L]	
Heliofanía (Hel) [hs]	
Ion amonio (NH_4-N) [mg NH_4-N/L]	
Ion nitrato (NO_3-N) [mg NO_3-N/L]	
Ion nitrito (NO_2-N) [mg NO_2-N/L]	
Nitrógeno total (NT) [mg N/L]	
Potencial de hidrógeno (pH) [NA]	
Oxígeno disuelto (OD) [mg/L]	
Temperatura del agua (T) [ºC]	
Turbidez [NTU]	

Fase 2: Selección de modelos de imputación

Esta fase tiene como entrada los conjuntos de datos procesados y se divide en varios pasos:

8. Se obtiene la mejor configuración paramétrica de cada modelo de regresión (univariable y multivariable), para realizar la búsqueda, se prueban las distintas variantes los datos procesados. Cómo las configuraciones paramétricas posibles son muy numerosas y también es numerosa la cantidad de modelos a probar, se usa la biblioteca *optuna*[18] que permite una búsqueda eficiente seleccionando estratégicamente las configuraciones a probar, esto es posible mediante el uso del algoritmo TPE (*Tree-structured Parzen Estimator*) (Bergstra et al., 2011).
   Las mejores configuraciones se eligen minimizando la métrica NSE (función objetivo).
   Para evitar valores que no se correspondan con el rango posible de cada variable, se corrigen las salidas de los modelos de modo que, si una modelo predice un valor fuera de

[18] https://optuna.readthedocs.io/en/stable/index.html

rango, se modifica por el valor máximo o mínimo según corresponda, los límites permitidos se ven en la Tabla 2.7.

La salida de este paso son los mejores modelos de regresión encontrados.

9. A partir de las salidas del paso anterior se crea un modelo iterativo de imputación (MICE) por cada modelo de regresión existente. La salida de este paso son los modelos iterativos de imputación.
10. Se crea un modelo de imputación que usa la técnica IRMI, la salida de este paso es el modelo.
11. A partir de los 2N + 1 (siendo N la cantidad de modelos de regresión probados + IRMI) modelos generados anteriormente, se realiza *k-fold cross validation* con $\mathrm{k} = 10$ donde se evalúa el desempeño de cada modelo con las métricas NSE, RMSE, Bias y KGE. En caso de que el conjunto de datos contenga menos de 100 puntos, se *realiza repeated k-fold cross validation*, en esta versión se realiza k-fold con $k = max\left(\frac{N}{10}, 2\right)$ (siendo N la cantidad de elementos del conjunto de datos) $n = \frac{10}{k}$ veces, seleccionando los datos pertenecientes a cada subconjunto de forma aleatoria, n y k se eligen de forma que la cantidad de veces que se mide cada métrica sea igual al caso sin repeticiones.
12. A partir de los resultados de la evaluación del paso anterior, se seleccionan los dos mejores modelos ordenados según el valor de NSE obtenido.

El resultado de esta etapa es el conjunto de los mejores modelos para la imputación de Y.

Tabla 2.7. Rangos de valores posibles para cada variable.

Variable	Valor mínimo posible	Valor máximo posible
Caudal (Q) [m^3/s]	0	$+\infty$
Nivel de agua (h) [m]	2.01	$+\infty$
Precipitación (P) [mm]	0	$+\infty$
Clorofila-a (Chl-a) [µg/L]	0	$+\infty$
Conductividad [µS/cm]	0	$+\infty$
Demanda bioquímica de oxígeno (DBO) [mg O2/L]	0	$+\infty$
Fosforo total (PT) [µg P/L]	0	$+\infty$
Glifosato [µg/L]	0	$+\infty$
Heliofanía (Hel) [hs]	0	$+\infty$
Ion amonio (NH_4-N) [mg NH_4-N/L]	0	$+\infty$
Ion nitrato (NO_3-N) [mg NO_3-N/L]	0	$+\infty$
Ion nitrito (NO_2-N) [mg NO_2-N/L]	0	$+\infty$
Nitrógeno total (NT) [mg N/L]	0	$+\infty$
Potencial de hidrógeno (pH) [NA]	0	14
Oxígeno disuelto (OD) [mg/L]	0	$+\infty$
Temperatura del agua (T) [ºC]	0	$+\infty$
Turbidez [NTU]	0	$+\infty$

Fase 3: Imputación

Esta fase tiene como entrada el conjunto de los mejores modelos para la imputación de Y y se divide en los siguientes pasos:

13. Se realizan las imputaciones de Y con los modelos obtenidos en la fase anterior.
14. Si no hay más variables a imputar se genera el conjunto final; en caso contrario, se agregan al conjunto de datos que es usado como entrada del *pipeline* para la siguiente variable a imputar.

Para aplicar esta metodología se implementó una aplicación cuyo código se encuentra en el repositorio del proyecto.

Es importante remarcar que la información brindada de los dos árboles de dependencias (Fig. 2.8, Fig. 2.9) y de los rangos de variación de las variables consideradas (Tabla 2.7) es puramente física y fue el input de la metodología basada en datos descripta anteriormente. Por lo tanto, en el marco de este proyecto, se desarrolló una metodología híbrida de imputación de datos ambientales que es objeto de publicaciones científicas (ver Capítulo "Actividades de Difusión").

2.3. Resultados y discusión

El *pipeline* antes descrito se ejecutó 78 veces (una iteración por cada una de las variables donde se detectaron valores faltantes), al concluir el proceso de imputación, se obtuvo un conjunto de modelos capaces de imputar de la mejor forma posible cada serie temporal.

Para la selección de las mejores imputaciones se usó la métrica Nash-Sutcliffe Efficiency (NSE) como función objetivo y, para validar los resultados, se usaron las métricas Kling-Gupta efficiency (KGE) y percent bias (PBIAS):

$$NSE = 1 - \frac{\sum_{i=1}^{n}(x_i^o - x_i^c)^2}{\sum_{i=1}^{n}(x_i^o - \overline{x^o})^2}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

$$PBIAS = 100 \times \frac{\sum_{i=1}^{n}(x_i^o - x_i^c)}{\sum_{i=1}^{n}(x_i^o)}$$

Donde x_i^o es el i-ésimo valor observado, x_i^c es el i-ésimo valor calculado (o imputado), $\overline{x^o}$ es la media de los valores observados y n es el tamaño del conjunto de datos de prueba. Siendo (μ^c, σ^c) y (μ^o, σ^o) los dos primeros momentos estadísticos (media y desviación estándar) de x^c y x^o respectivamente, r es la correlación lineal entre observaciones e imputaciones, α es una medida del error de variabilidad de flujo ($\alpha = \sigma^c/\sigma^o$), β es un término de sesgo ($\beta = \mu^c/\mu^o$).

Para cada variable se seleccionaron los mejores modelos (y las imputaciones resultantes de aplicarlos), ordenando la lista según el NSE obtenido en la validación. En la Fig. 2.11, se muestran los valores de las métricas para los mejores modelos (en el eje vertical se listan las variables y en el eje horizontal el valor de la métrica obtenido para el mejor modelo) en forma de boxplots. En la Fig. 2.12, se resumen los valores promedio de las métricas para las variables.

En Fig. 2.12, se puede apreciar que la mayoría de las variables presenta un NSE positivo, o sea, el modelo aquí desarrollado para su imputación tiene un mejor desempeño de la media observada. Además, se puede destacar que más del 75% de las variables presentan un NSE > 0.5. El KGE y el PBIAS confirman los buenos resultados obtenidos con el NSE. En particular, KGE muestra una distribución similar a la obtenida con el NSE y el |PBIAS| presenta valores menores de 15 para el 75% de las variables, confirmando los muy buenos resultados obtenidos.

En la Tabla 2.8, se muestran los modelos usados para las imputaciones ordenándolos según la cantidad de variables en las que se usó cada uno. El modelo *IDW* fue el mejor para 22 variables (lo cual tiene sentido pues las variables tienen dependencia espacial) seguido del modelo de ensemble *Hubber Regressor + SC* usado para 9 variables. En la Tabla 2.9, se muestra la cantidad de veces que se usó una variante de los modelos y/o conjuntos de datos para imputar una variable, también se detalla que significa cada sigla, la mayoría de las veces (30) no se usó una variante, la variante más usada fue SC (25 veces) seguida de la variante SD (13 veces). Esto denota que agregar información espacial de forma implícita es útil para obtener una mejor imputación.

En la Tabla 2.10 se muestra un resumen de los resultados de las imputaciones para todas las variables.

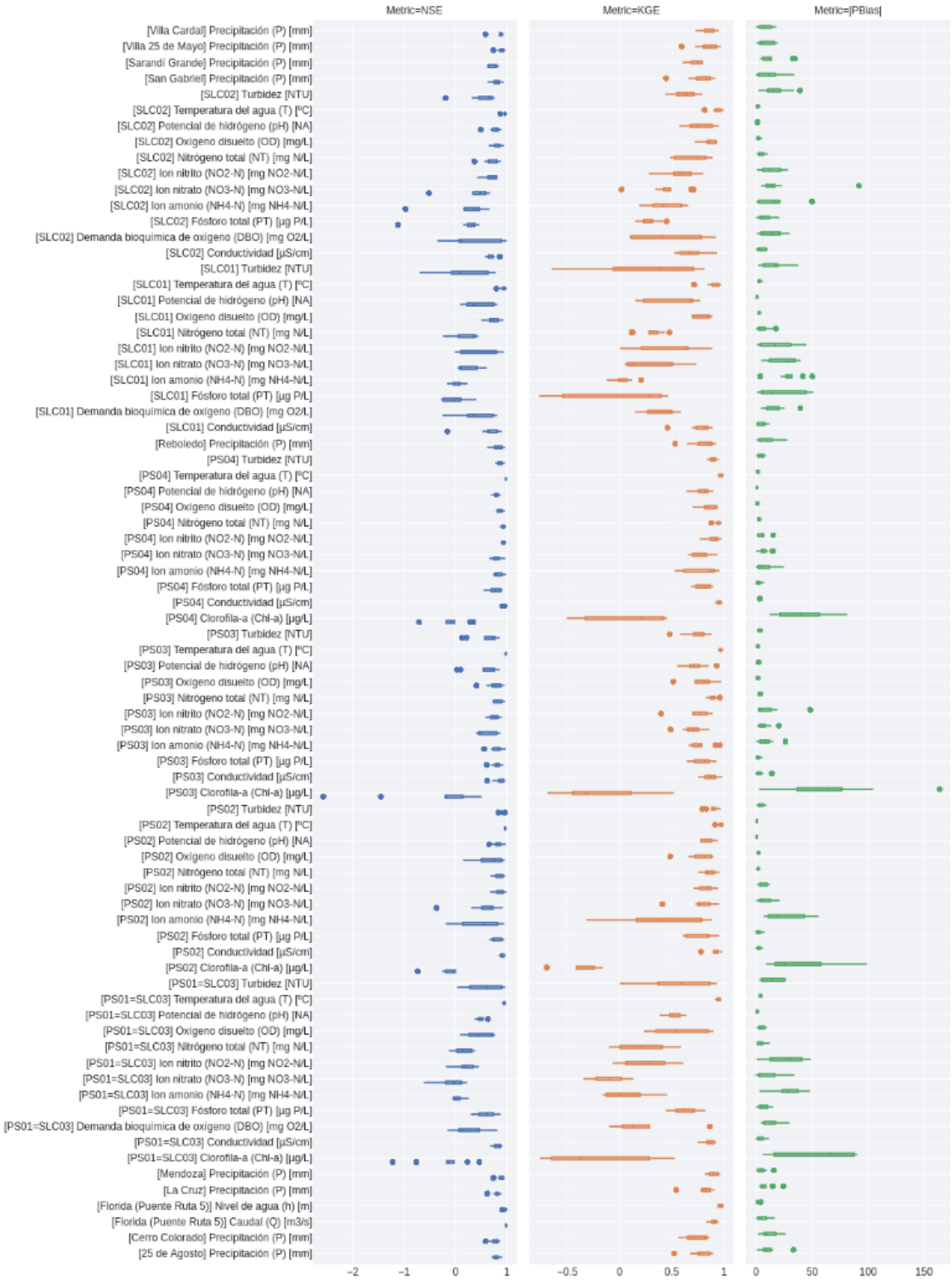


Fig. 2.11. Valores de NSE, KGE y |PBIAS| para el mejor modelo de cada variable.

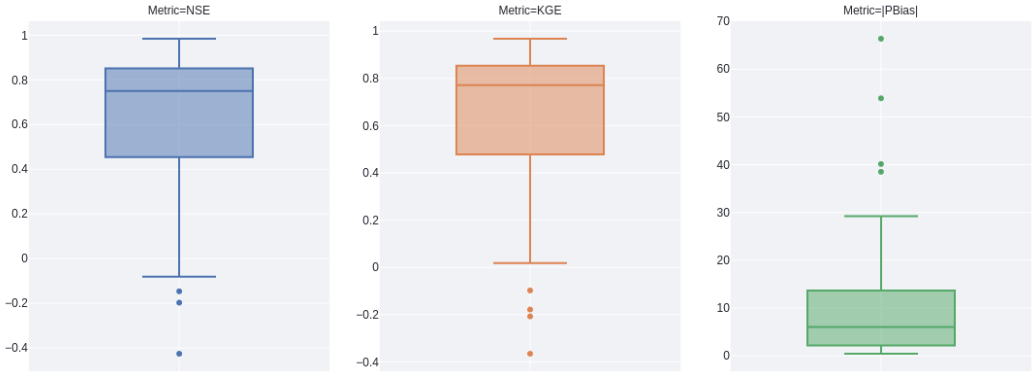


Fig. 2.12. Valores promedio de NSE, KGE y |PBIAS| para el mejor modelo de cada variable.

Tabla 2.8. Cantidad de veces que cada modelo resultó seleccionado como el más apropiado para imputar una variable.

Modelo de imputación	Cantidad de variables imputadas
IDW	22
Hubber Regressor + SC	9
SVR	5
IDW + SD	
KNN + SC	3
SVR + SD	2
Extra Trees Regressor + SC	
Ridge + EWMA + SC	
Hubber Regressor (II)	
Ridge (II)	
Ridge + SC	
Bayesian Ridge (II) + EWMA	
Hubber Regressor + SC + SD	1
Extra Trees Regressor	
SVR (II) + EWMA	
KNN	
Hubber Regressor + EWMA + SC	
KNN + SC + SD	
Extra Trees Regressor + EWMA + SD	
KNN (II)	
Ridge (II) + EWMA	
Hubber Regressor	
Hubber Regressor (II) + EWMA	
Ridge + EWMA	
Bayesian Ridge + SC	

Extra Trees Regressor (II) + EWMA	
Ridge + SC + SD	
Ridge + SD	
SVR (II)	
Hubber Regressor + EWMA + SC + SD	
SVR + EWMA	
TheilSen Regressor + SC	

Tabla 2.9. Cantidad de veces que cada variante de modelo resultó seleccionada para imputar una variable.

Variante	Descripción	Cantidad de variables imputadas
Original	Modelo y datos sin variante	30
SC	Conjunto de datos modificado según la distancia de las estaciones	25
SD	Conjunto de datos reducido según las dependencias espaciales (Fig. 2.9)	13
II	Variante MICE del modelo	12
EWMA	Conjunto de datos con EWMA (Tabla 2.6)	12

Tabla 2.10. Modelos usados en la imputación de las variables junto con sus scores.

Variable	Estación	Modelo de imputación	NSE promedio	NSE std	KGE promedio	KGE std	PBIAS promedio	PBIAS std
Temperatura del agua (T) [ºC]	SLC01	IDW	0,9314	0,0599	0,8983	0,0793	-3,1382	1,4451
	SLC02	Hubber Regressor + SC + SD	0,9573	0,0365	0,9272	0,0512	0,3866	1,8863
	PS01=SLC03	IDW + SD	0,9488	0,0174	0,9473	0,0171	3,7673	0,7890
	PS03	IDW	0,9780	0,0067	0,9672	0,0125	-1,2128	0,9440
	PS04	IDW + SD	0,9792	0,0068	0,9680	0,0178	1,4030	0,9184
	PS02	IDW	0,9717	0,0101	0,9647	0,0238	0,0096	0,7466
Nivel de agua (h) [m]	Florida (Puente Ruta 5)	KNN + SC	0,9756	0,0219	0,9677	0,0189	0,0191	1,7759
Caudal (Q) [m^3/s]	Florida (Puente Ruta 5)	Extra Trees Regressor	0,9842	0,0071	0,9005	0,0356	-1,8522	7,6132
Ion amonio (NH_4-N) [mg NH_4-N/L]	SLC01	SVR (II) + EWMA	0,0304	0,1195	0,0332	0,0945	11,3743	31,3149
	SLC02	KNN	0,2307	0,4769	0,4337	0,1647	4,4340	22,1838
	PS01=SLC03	SVR + SD	0,0463	0,0938	0,0482	0,1985	9,1910	30,3825
	PS03	Hubber Regressor + EWMA + SC	0,8006	0,1233	0,7746	0,1019	3,0274	11,8742
	PS04	KNN + SC + SD	0,8420	0,0868	0,7967	0,1611	3,6923	10,5469

	PS02	IDW	0,4815	0,4098	0,4132	0,4082	19,6069	24,7313
Ion nitrato (NO_3-N) [mg NO_3-N/L]	SLC01	Hubber Regressor + SC	0,2409	0,2201	0,2761	0,2597	22,0784	17,0587
	SLC02	Extra Trees Regressor + EWMA + SD	0,3832	0,3543	0,4538	0,2012	-7,6880	34,2526
	PS01=SLC03	Extra Trees Regressor + SC	-0,0804	0,2568	-0,0980	0,1492	-1,2248	14,8896
	PS03	Hubber Regressor + SC	0,6029	0,1893	0,6898	0,1044	-1,9949	9,2650
	PS04	IDW	0,7987	0,0864	0,7786	0,0900	-4,1429	6,7704
	PS02	IDW	0,5437	0,3894	0,7701	0,1528	4,7196	9,6910
Ion nitrito (NO_2-N) [mg NO_2-N/L]	SLC01	KNN (II)	0,4953	0,3610	0,5062	0,2970	14,6741	19,1182
	SLC02	SVR	0,6902	0,1276	0,5800	0,1601	-10,8878	12,0377
	PS01=SLC03	SVR	0,2075	0,1895	0,2568	0,2253	10,6951	31,0516
	PS03	Hubber Regressor + SC	0,7462	0,1016	0,7355	0,1423	-4,8936	18,7643
	PS04	IDW	0,9340	0,0299	0,8929	0,0660	-1,6042	6,7767
	PS02	IDW	0,8471	0,1027	0,8252	0,0716	-4,0194	7,4823
Oxígeno disuelto (OD) [mg/L]	SLC01	Hubber Regressor + SC	0,7407	0,1263	0,7989	0,0812	-0,2203	2,9298
	SLC02	Ridge (II) + EWMA	0,8113	0,0890	0,8534	0,0731	-0,1235	2,4345
	PS01=SLC03	IDW + SD	0,4749	0,2591	0,5645	0,2578	-2,3090	5,1586
	PS03	Ridge + EWMA + SC	0,7538	0,1649	0,7836	0,1300	0,0739	2,2089
	PS04	Hubber Regressor	0,8715	0,0509	0,8466	0,0849	-0,0073	1,4199
	PS02	IDW	0,6514	0,2692	0,7670	0,1330	-1,7459	1,2409
Precipitación (P) [mm]	25 de Agosto	Hubber Regressor (II) + EWMA	0,7963	0,0570	0,7720	0,1081	9,1749	10,8557
	San Gabriel	Hubber Regressor (II)	0,8024	0,0831	0,7817	0,1419	5,7052	14,8648
	Reboledo	Hubber Regressor (II)	0,8260	0,1007	0,7993	0,1251	-1,5573	12,9362
	Cerro Colorado	Ridge + EMA	0,7584	0,0882	0,7597	0,1134	-1,6796	14,0323
	La Cruz	Bayesian Ridge + SC	0,8018	0,0743	0,8079	0,1030	1,0978	10,6921
	Sarandí Grande	Ridge (II)	0,7250	0,0796	0,7261	0,0656	-3,3671	17,9089
	Villa 25 de Mayo	Hubber Regressor + SC	0,8926	0,0659	0,8454	0,1141	1,7580	11,1761
	Villa Cardal	KNN + SC	0,8575	0,0982	0,8492	0,0656	-0,5492	10,2859
	Mendoza	SVR	0,8917	0,0625	0,8949	0,0477	-1,6310	6,7669
Turbidez [NTU]	SLC01	IDW	0,1626	0,4924	0,2698	0,4746	12,5242	13,4669
	SLC02	Ridge (II)	0,5257	0,3028	0,6337	0,1147	-1,5845	22,1437
	PS01=SLC03	IDW + SD	0,5779	0,3675	0,5751	0,3185	11,3303	14,3568
	PS03	IDW	0,6092	0,2653	0,7363	0,1271	1,0583	3,9356
	PS04	IDW	0,8703	0,0529	0,8954	0,0336	4,2699	2,9450
	PS02	Ridge + SC	0,8927	0,0351	0,8910	0,0531	-0,5808	5,1168
	SLC01	Hubber Regressor + SC	-0,0368	0,2161	0,0181	0,5006	13,2178	26,5851

Fósforo total (PT) [µg P/L]	SLC02	Extra Trees Regressor (II) + EWMA	0,1591	0,4874	0,2920	0,0958	-1,6657	10,4202
	PS01=SLC03	IDW	0,6106	0,1753	0,6262	0,1152	-3,9228	6,9233
	PS03	IDW	0,7985	0,0961	0,7957	0,0995	-1,2744	2,1479
	PS04	Hubber Regressor + SC	0,7815	0,1158	0,8026	0,0762	-0,8959	3,1140
	PS02	Bayesian Ridge (II) + EWMA	0,7857	0,0963	0,7836	0,1260	-0,9484	2,9870
Nitrógeno total (NT) [mg N/L]	SLC01	Extra Trees Regressor + SC	0,2300	0,2294	0,3043	0,1224	1,0497	8,9266
	SLC02	Bayesian Ridge (II) + EWMA	0,6899	0,1551	0,7105	0,1612	-0,3243	5,3730
	PS01=SLC03	Ridge + SC + SD	0,1442	0,1571	0,2184	0,2294	0,4099	5,9802
	PS03	IDW	0,8266	0,0812	0,8962	0,0413	3,3103	1,6380
	PS04	IDW	0,9264	0,0338	0,9400	0,0275	-2,4864	1,0503
	PS02	IDW	0,8558	0,0859	0,8602	0,0644	1,5202	0,9729
Conductividad [µS/cm]	SLC01	IDW	0,6280	0,3172	0,7658	0,1291	-1,2200	6,7769
	SLC02	SVR	0,6928	0,0790	0,6888	0,1274	0,2937	5,8732
	PS01=SLC03	Ridge + SD	0,8239	0,0635	0,8620	0,0516	0,7698	5,2937
	PS03	Ridge + EWMA + SC	0,8519	0,1121	0,8708	0,0741	1,7585	5,1881
	PS04	IDW	0,9660	0,0218	0,9548	0,0195	-1,9677	0,6417
	PS02	Hubber Regressor + SC	0,9162	0,0378	0,9133	0,0556	0,0420	2,7525
Potencial de hidrógeno (pH) [NA]	SLC01	Hubber Regressor + SC	0,4543	0,2704	0,5150	0,2549	-0,1289	0,8340
	SLC02	SVR (II)	0,7635	0,1195	0,7731	0,1281	-0,0269	0,6126
	PS01=SLC03	SVR + SD	0,4887	0,0691	0,5176	0,0758	0,0929	1,0801
	PS03	IDW	0,5704	0,3038	0,7206	0,1166	0,1790	1,2788
	PS04	Ridge + SC	0,7938	0,0470	0,8003	0,0744	0,0009	0,7480
	PS02	IDW	0,8123	0,0995	0,8390	0,0586	-0,4080	0,4485
Clorofila-a (Chl-a) [µg/L]	PS01=SLC03	IDW	-0,1970	0,4764	-0,2076	0,5122	49,3362	42,8180
	PS03	KNN + SC	-0,4256	0,9722	-0,1786	0,3798	1,9494	85,0046
	PS04	Hubber Regressor + EWMA + SC + SD	-0,0773	0,3039	0,0832	0,3693	23,5968	41,6526
	PS02	SVR + EWMA	-0,1464	0,2374	-0,3658	0,2023	14,6096	47,7120
Demanda bioquímica de oxígeno (DBO) [mg O2/L]	SLC01	TheilSen Regressor + SC	0,3770	0,3506	0,3920	0,1449	5,5504	20,5433
	SLC02	IDW + SD	0,3826	0,4970	0,4778	0,3313	-4,6145	17,6020
	PS01=SLC03	SVR	0,2100	0,2920	0,1836	0,2686	-0,3646	16,0429

3. OE2: Evaluar los cambios temporales y espaciales de LULC en la cuenca del río Santa Lucía.

El OE2 planteado en la propuesta del proyecto consiste en evaluar los cambios temporales y espaciales de LULC en la cuenca del río Santa Lucía para el período dado por los años 2000, 2008, 2011 y 2015 en base a los datos disponibles del MVOTMA. Teniendo en cuenta que a la fecha de elaboración de este informe se disponen además de los mapas de LULC para los años 2016 y 2018, se decide incorporarlos en el análisis, y se limita la clasificación de todos los años a la cuenca del río Santa Lucía Chico, seleccionada como cuenca de estudio.

3.1. Recolección y descripción de las capas de cobertura de suelo

La información de clasificación de cobertura de suelo disponible para la cuenca del río Santa Lucia Chico se puede agrupar según su fuente de origen en cuatro categorías:

- Categoría 1: agrupa los mapas de los años 2000, 2008 y 2011 (DINOT 2000, 2008, 2011, 2015)
- Categoría 2: mapa del año 2015 (DINOT 2000, 2008, 2011, 2015)
- Categoría 3: mapa del año 2016 (DINAMA, 2017)
- Categoría 4: mapa del año 2018 (MGAP, 2018)

Según fuentes consultadas de la Dirección Nacional de Ordenamiento Territorial (DINOT), actualmente se está en proceso de elaboración de mapas de cobertura y uso de suelo actualizadas a los años 2019 y 2020 con nuevas metodologías a partir de imágenes SENTINEL. A continuación, se detalla la información que surge de la elaboración de los mapas de las distintas categorías definidas.

3.1.1. Categoría 1: agrupa los mapas de los años 2000, 2008 y 2011

Esta categoría agrupa los mapas del año 2000, 2008 y 2011, en dónde se clasifica la cobertura física y biofísica del Uruguay a escala 1:100.000 en 46 clases, agregables a 17 clases y a 7 u 8 temas. Generada a partir del procesamiento digital de imágenes del satélite LANDSAT 5 TM de los períodos 2000, 2008 y 2011, respectivamente, con el uso de la metodología Land Cover Clasification System (LCCS) de la Global Land Cover Network (GLCN) de la Organización de Naciones Unidas para la Agricultura y Alimentación (FAO) (DINOT 2000, 2008, 2011, 2015; FAO, 2015).

La capa 2008 fue elaborada en 2010, producto de la acción coordinada entre RENARE, DINAMA y DINOT. Las capas 2000 y 2011 fueron elaboradas en 2014 por DINOT. De los trabajos elaborados surge (FAO, 2015), en donde se detalla la metodología de trabajo y los resultados obtenidos.

En (FAO, 2015) se destaca que la generación de la capa 2011 fue uno de los productos principales del proyecto. El acierto global de la clasificación se realizó midiendo la clasificación realizada y la esperada simplemente por azar, es decir si la clasificación ha discriminado las categorías de interés con exactitud significativamente mayor a la que se hubiera obtenido con una asignación aleatoria. El resultado de la validación a partir de la matriz de contingencia mostró que el acierto global de la cartografía asciende el 85% para el año 2011. Para la segmentación e interpretación de imágenes para el año 2000 se utilizó como información de base la capa de cobertura de 2011. Un subproducto adicional obtenido fue la adecuación de la capa 2008 (elaborada en el 2010) a la nueva leyenda de 46 clases y a la simplificada de 17, de manera de poder contar con 3 fechas comparables en el tiempo: 2000, 2008 y 2011.

3.1.2. Categoría 2: mapa del año 2015

La capa del año 2015 fue elaborada por DINOT a partir de la fotointerpretación de las imágenes satelitales Landsat y se complementó con información adicional de otras instituciones e imágenes de Google Earth. Se utilizaron 14 imágenes, libres de nubes, provenientes del satélite Landsat 5 y Landsat 8 (sensores TM y OLI/TIIRS, respectivamente), con una resolución espacial de 30 x 30 metros y una escala de trabajo de 1:100.000 cubriendo toda la superficie nacional. La capa 2015 fue realizada de acuerdo al sistema de clasificación “Land Cover Clasification System” (LCCS), desarrollado por la Global Land Cover Network (GLCN) de la Organización de las Naciones Unidas para la Alimentación y Agricultura (FAO) y el Programa de las Naciones Unidas para el Medio Ambiente (UNEP) (DINOT, 2016 (a) y (b)).

Para la construcción de la capa 2015 se tomó como base, para la fotointerpretación, la capa vectorial 2011 de polígonos segmentados en base a datos espectrales homogéneos. La clasificación consistió en asignar a cada uno de los polígonos alguna de las 16 clases de cobertura del suelo, clasificarlos y extraer datos para el análisis estadístico. Esta nueva capa de cobertura 2015 tiene un acierto global del 88%, se tomaron como verdad terrestre 3.121 puntos para la validación (DINOT, 2016 (b)).

3.1.3. Categoría 3: mapa del año 2016

El mapa de clasificación de usos y coberturas para la cuenca del río Santa Lucía del año 2016 es un producto elaborado en el año 2017 por la División de Información Ambiental de DINAMA, a una escala espacial 1:50.000 (DINOT, 2017). Para cartografiar los usos y coberturas del suelo presentes en la cuenca del río Santa Lucía, se realizó una clasificación supervisada de imágenes provistas por el sensor Operational Land Imager (OLI) a bordo del satélite Landsat 8. Dicho sensor tiene una resolución espacial de 30 x 30 metros y una frecuencia de revisita de 16 días. Se utilizaron dos escenas (223-84 y 224-84) y tres fechas (11/2015 y 01/2016) de forma de captar diferencias fenológicas en la vegetación. Las imágenes fueron corregidas radiométrica y atmosféricamente para lograr que la información espectral sea comparable en tiempo y espacio.

Para generar la clasificación supervisada de cada escena Landsat (223-84 y 224-84) se digitalizaron, mediante fotointerpretación de las imágenes, 142 polígonos de entrenamiento y 83 de control para la escena 224-84 y 274 polígonos de entrenamiento y 135 de control para la escena 223-84. La precisión de la clasificación fue evaluada comparándola con la información de la cobertura real obtenida previamente. Se construyó una matriz de contingencia entre el resultado de la clasificación (filas) y la información de los píxeles correspondientes a los polígonos de control digitalizados, esto permitió calcular el acierto global, el coeficiente Kappa y la precisión del productor y del usuario. La evaluación de las clasificaciones mostró resultados muy satisfactorios, la exactitud global fue de 99.6% y 96.5% para la escena 223-84 y 224-84, respectivamente. Por su parte el coeficiente Kappa fue de 0.9961 y 0.9614 para la escena 223-84 y 224-84, respectivamente (DINOT, 2017).

3.1.4. Categoría 4: mapa del año 2018

Este trabajo resulta de una integración de productos obtenidos por distintas fuentes, en el marco de una iniciativa propiciada por el grupo de trabajo de Infraestructura de datos Espaciales del Ministerio de Ganadería, Agricultura y Pesca (MGAP). Se obtuvo un mapa georreferenciado de clases de cobertura/uso del suelo de Uruguay con énfasis en la producción agropecuaria a partir de imágenes satelitales. Este mapa se realizó en función de las demandas de información actualizada de diferentes áreas del MGAP: conservación de recursos naturales, cambio climático, gestión del riesgo, políticas e investigaciones. Fue realizado por los equipos de Sistemas de Información Geográfica de la Dirección General de Recursos Naturales (DGRN) y de la Oficina de Programación y Política Agropecuaria (OPYPA) (Petraglia et al., 2019).

La definición de cobertura/uso del suelo que se utiliza es la realizada por el NRCS (Natural Resources Conservation Service); es un término que incluye conjuntamente categorías de cobertura y de uso del suelo que permiten representar y clasificar toda la superficie del país. La definición de las clases cobertura/uso se presentan en (Petraglia et al., 2019). La columna valor corresponde al valor de pixel de cada categoría, en el campo clase está el nombre de las categorías y en la descripción se detalla el contenido de la clase.

La escala del mapa es de 1:50.000 acorde a la resolución espacial de las imágenes Sentinel 2, de 10 m de pixel. Se adoptó el criterio de JECAM de 0,25 ha como la unidad mínima identificable para imágenes con resolución espacial de 10-20 m. Además, se utilizaron registros administrativos o se construyeron clases generalizadas que incluyen varios usos, en aquellos casos donde no se pueden discriminar algunas coberturas/usos (pe. los usos hortifrutícolas) (Petraglia et al., 2019).

Para la clasificación de las imágenes se realizaron clasificaciones no supervisadas y supervisadas y correcciones mediante edición visual. También se revisaron y modificaron las clases Pastizal Natural y Bañados. Se usó la plataforma Google Earth Engine para agrupar adecuadamente clases muy similares en cuanto a respuesta espectral, como el caso de los recursos forrajeros y las

categorías cultivos anuales. Se estudió la respuesta fenológica en las zonas de interés con una serie temporal de NDVI de imágenes satelitales MODIS. De acuerdo a las curvas fenológicas, se logró separar las diferentes clases (Petraglia et al., 2019).

3.2. Variabilidad de la cobertura de suelos

3.2.1. Pre-procesamiento de los datos

Se utilizó el software ArcGIS 10.3 para el manejo de la información digitalizada georreferenciada. Todos los archivos de cobertura de suelo se encuentran en el sistema de coordenadas proyectadas WGS84 UTM 21S. Se realizó un recorte de los archivos de cobertura a la geometría de la cuenca del río Santa Lucía Chico, considerando geometría de *Cuencas Hidrográficas Nivel 2* (DINAMA-OAN).

De la Tabla 3.1 a la Tabla 3.4, se presentan las clases y áreas asociadas a la capa de cobertura de suelo del año 2000, 2008, 2011 y 2015, respectivamente. Como se observa, para el año 2000 se tienen 15 clases, mientras que para los años 2008, 2011 y 2015 se tienen 16 clases, 15 de estas clases coinciden con las del año 2000 agregándose una nueva clase (Cultivos Regados > 4-5 has).

Tabla 3.1. Clases de cobertura de suelo para el año 2000.

#	Clase	Área (km^2)	% del total
1	Aguas Artificiales	14,72	0,57%
2	Aguas Naturales	2,47	0,10%
3	Arbustos	5,67	0,22%
4	Área Urbana	9,68	0,38%
5	Áreas Desnudas	1,23	0,05%
6	Áreas Naturales Inundadas	0,19	0,01%
7	Áreas Urbanas Dispersas	6,59	0,26%
8	Canteras, Areneras, Minas a Cielo Abierto	0,33	0,01%
9	Cultivos de Secano > 4-5 has	795,09	30,94%
10	Cultivos Regados y de Secano < 4-5 has	44,24	1,72%
11	Equipamiento Urbano	1,47	0,06%
12	Frutales	1,37	0,05%
13	Herbáceo Natural	1585,77	61,71%
14	Monte Nativo	71,24	2,77%
15	Plantación Forestal	29,56	1,15%
	Total	2569,62	100,00%

Tabla 3.2. Clases de cobertura de suelo para el año 2008.

#	Clase	Área (km^2)	% del total
1	Aguas Artificiales	14,91	0,58%
2	Aguas Naturales	2,46	0,10%
3	Arbustos	0,72	0,03%
4	Área Urbana	9,67	0,38%

5	Áreas Desnudas	1,01	0,04%
6	Áreas Naturales Inundadas	0,19	0,01%
7	Áreas Urbanas Dispersas	4,88	0,19%
8	Canteras, Areneras, Minas a Cielo Abierto	0,27	0,01%
9	Cultivos de Secano > 4-5 has	660,08	25,69%
10	Cultivos Regados > 4-5 has	2,83	0,11%
11	Cultivos Regados y de Secano < 4-5 has	36,40	1,42%
12	Equipamiento Urbano	1,29	0,05%
13	Frutales	0,61	0,02%
14	Herbáceo Natural	1714,81	66,73%
15	Monte Nativo	77,16	3,00%
16	Plantación Forestal	42,32	1,65%
	Total	2569,62	100,00%

Tabla 3.3. Clases de cobertura de suelo para el año 2011.

#	Clase	Área (km^2)	% del total
1	Aguas Artificiales	16,50	0,64%
2	Aguas Naturales	2,43	0,09%
3	Arbustos	3,86	0,15%
4	Área Urbana	9,67	0,38%
5	Áreas Desnudas	1,20	0,05%
6	Áreas Naturales Inundadas	0,19	0,01%
7	Áreas Urbanas Dispersas	6,07	0,24%
8	Canteras, Areneras, Minas a Cielo Abierto	0,29	0,01%
9	Cultivos de Secano > 4-5 has	762,19	29,66%
10	Cultivos Regados > 4-5 has	16,74	0,65%
11	Cultivos Regados y de Secano < 4-5 has	63,25	2,46%
12	Equipamiento Urbano	1,47	0,06%
13	Frutales	0,68	0,03%
14	Herbáceo Natural	1562,76	60,82%
15	Monte Nativo	76,41	2,97%
16	Plantación Forestal	45,91	1,79%
	Total	2569,62	100,00%

Tabla 3.4. Clases de cobertura de suelo para el año 2015.

#	Clase	Área (km^2)	% del total
1	Aguas Artificiales	18,25	0,71%
2	Aguas Naturales	2,47	0,10%
3	Arbustos	11,32	0,44%
4	Área Urbana	9,75	0,38%
5	Áreas Desnudas	1,42	0,06%
6	Áreas Naturales Inundadas	0,19	0,01%
7	Áreas Urbanas Dispersas	7,11	0,28%
8	Canteras, Areneras, Minas a Cielo Abierto	0,37	0,01%
9	Cultivos de Secano > 4-5 has	1087,49	42,32%

10	Cultivos Regados > 4-5 has	15,82	0,62%
11	Cultivos Regados y de Secano < 4-5 has	51,17	1,99%
12	Equipamiento Urbano	1,88	0,07%
13	Frutales	1,47	0,06%
14	Herbáceo Natural	1217,20	47,37%
15	Monte Nativo	70,85	2,76%
16	Plantación Forestal	72,87	2,84%
	Total	2569,62	100,00%

En la Tabla 3.5, se presentan las clases y áreas asociadas a la capa de cobertura de suelo del año 2016. Se destaca que existen dos categorías de clases que no representan la cobertura del suelo, estas son: Nube y Sin Clasificar.

Tabla 3.5. Clases de cobertura de suelo para el año 2016.

#	Clase	Área (km^2)	% del total
1	Agua	13,98	0,54%
2	Área inundable	0,00	0,00%
3	Campo Natural	1460,60	56,86%
4	Cultivos	755,59	29,41%
5	Forestación	48,29	1,88%
6	Monte Nativo	36,82	1,43%
7	Nube	1,10	0,04%
8	Sin Clasificar	0,68	0,03%
9	Suelo Desnudo Agrícola	232,83	9,06%
10	Suelo Desnudo Urbano	18,91	0,74%
	Total	2568,83	100,00%

En la Tabla 3.6, se presentan las clases y áreas asociadas a la capa de cobertura de suelo del año 2018.

Tabla 3.6. Clases de cobertura de suelo para el año 2018.

#	Clase	Área (km^2)	% del total
1	Bañados	12,61	0,49%
2	Bosque nativo	74,85	2,91%
3	Bosque plantado	49,39	1,92%
4	Bosque plantado nuevo, cosecha, rebrote	9,92	0,39%
5	Canteras, Areneras, minas a Cielo abierto	0,17	0,01%
6	Citrus	0,27	0,01%
7	Cuerpos de Agua Naturales	1,53	0,06%
8	Cuerpos de Agua Artificiales	13,38	0,52%
9	Cultivo extensivo con riego por pivote	11,30	0,44%
10	Cultivo extensivo de secano	105,32	4,10%
11	Cultivo extensivo en predios lecheros	31,66	1,23%
12	Mezcla campo natural, pasturas y rastrojos	591,73	23,03%
13	Pastizal natural	1257,62	48,95%

14	Pastizal regenerado	255,32	9,94%
15	Playas, dunas y médanos fijos y semifijos	0,85	0,03%
16	Rastrojo de cultivo de secano	127,47	4,96%
17	Represas para riego	2,97	0,12%
18	Zonas urbanas y urbanizadas	22,99	0,89%
	Total	2569,32	100,00%

Los archivos de clasificación de cobertura de suelo presentan distintas clases por lo que es necesario elaborar categorías comunes que agrupen conjuntos de clases similares, para luego poder comparar los mapas que fueron elaborados bajo distintas metodologías y clasificaciones.

3.2.2. Definición de categorías comunes

La Tabla 3.7 presenta la propuesta de categorías comunes para agrupar las distintas clases encontradas en los mapas de cobertura de suelos de los distintos años.

Tabla 3.7. Categorías comunes para agrupación de clases.

#	***Categorías comunes***
1	Área Desnuda
2	Área Natural inundable
3	Cuerpos de agua
4	Cultivos
5	Forestal
6	Herbáceo Natural
7	Monte Nativo
8	Urbanización

Para realizar el agrupamiento de las clases a las categorías comunes de coberturas de suelo definidas, se recurren a las tablas de clasificación descriptivas de cada uno de los mapas de cobertura de suelo (FAO, 2015; DINOT (a) y (b), 2016; DINOT, 2017; Petraglia et al., 2019). La Tabla 3.8 detalla el agrupamiento propuesto para las distintas clases, y a continuación se realiza una discusión en base a las decisiones consideradas para agrupar las distintas clases.

Tabla 3.8. Agrupamiento de clases propuesto para categorías comunes definidas.

#	Categorías comunes	Años 2000, 2008, 2011 y 2015	Años 2016[2]	Años 2018
1	***Área desnuda***	Áreas Desnudas Canteras, Areneras, Minas a Cielo Abierto		Canteras, Areneras, minas a Cielo abierto Playas, dunas y médanos fijos y semifijos
2	***Área natural inundable***	Áreas Naturales Inundadas	Área Inundable	Bañados
3	***Cuerpos de agua***	Aguas Artificiales Aguas Naturales	Agua	Cuerpos de Agua Naturales Cuerpos de Agua Artificiales Represas para riego
4	***Cultivos***	Cultivos de Secano > 4-5 has Cultivos Regados > 4-5 has[1] Cultivos Regados y de Secano < 4-5 has Frutales	Cultivos Suelo Desnudo Agrícola	Citrus Cultivo extensivo con riego por pivote Cultivo extensivo de secano Cultivo extensivo en predios lecheros Rastrojo de cultivo de secano Mezcla campo natural, pasturas y rastrojos
5	***Forestal***	Plantación Forestal	Forestación	Bosque plantado Bosque plantado nuevo, cosecha, rebrote
6	***Herbáceo natural***	Herbáceo Natural	Campo Natural	Pastizal natural Pastizal regenerado
7	***Monte nativo***	Arbustos Monte Nativo	Monte Nativo	Bosque nativo
8	***Urbanización***	Área Urbana Áreas Urbanas Dispersas Equipamiento Urbano	Suelo Desnudo Urbano	Zonas urbanas y urbanizadas

[1] Para el año 2000 no se tiene definida esta clase.

[2] Para el año 2016 las clases Nube y Sin Clasificar no son tenidas en cuenta para el agrupamiento.

El agrupamiento de las clases para los años 2000, 2008, 2011 y 2015 se considera que la clase *Arbustos*, se debe agrupar con *Monte Nativo* y no en *Herbáceo Natural*.

En cuanto a la agrupación para el año 2016, existen dos clases que, por su definición, no se agrupan dentro de las categorías comunes propuestas, estas son: *Nube* y *Sin Clasificar*. Al no tener en cuenta estas clases para el agrupamiento, los resultados globales podrían ser subestimados, aunque dado el porcentaje que representan estas clases respecto del total (0,07%) se considera que la variación no será significativa. Además, la clase *Suelo Desnudo Agrícola* es agrupada en la categoría *Cultivos*, mientras que la clase *Suelo Desnudo Urbano* es agrupada en la categoría *Urbanización*. De esta forma, la categoría *Área desnuda* no tiene asociada ninguna clase para la capa del año 2016. Estos agrupamientos fueron verificados de forma visual contrastando con los mapas de los años 2000, 2008, 2011 y 2015, entendiendo de esta forma que se logra un agrupamiento razonable frente a la situación alternativa de agrupar la clase *Suelo Desnudo Agrícola* en la categoría *Área desnuda*.

En cuanto a la agrupación para el año 2018, la mayor incertidumbre encontrada se presenta en el agrupamiento de la clase *Mezcla campo natural, pasturas y rastrojos*. La descripción de la clase encontrada en (Petraglia et al., 2019) incluye una variedad importante de coberturas, pero teniendo en cuenta que el uso del suelo está asociado a la ganadería lechera extensiva, se procede a agrupar la clase en la categoría *Cultivos*. Este agrupamiento fue verificado de forma visual contrastando con los mapas de los años 2000, 2008, 2011 2015 y 2016, entendiendo de

esta forma que se logra un agrupamiento razonable frente a la situación alternativa de agrupar la clase *Mezcla campo natural, pasturas y rastrojos* en la categoría *Herbáceo natural*.

3.2.3. Variación espacial de la cobertura de suelo

Para abordar el análisis de la variación espacial de la cobertura de suelo, en Fig. 3.1, se presentan los mapas de cobertura de suelo elaborados a partir del agrupamiento de clases comunes. Es posible identificar dos categorías dominantes: *Herbáceo natural* y *Cultivos*, que en conjunto ocupan aproximadamente el 90% del área total de la cuenca, siendo la distribución espacial de *Cultivos* mayormente concentrados en la zona Sur, centro Sur y Suroeste, mientras que *Herbáceo natural* en la zona Norte, centro Norte y Noreste.

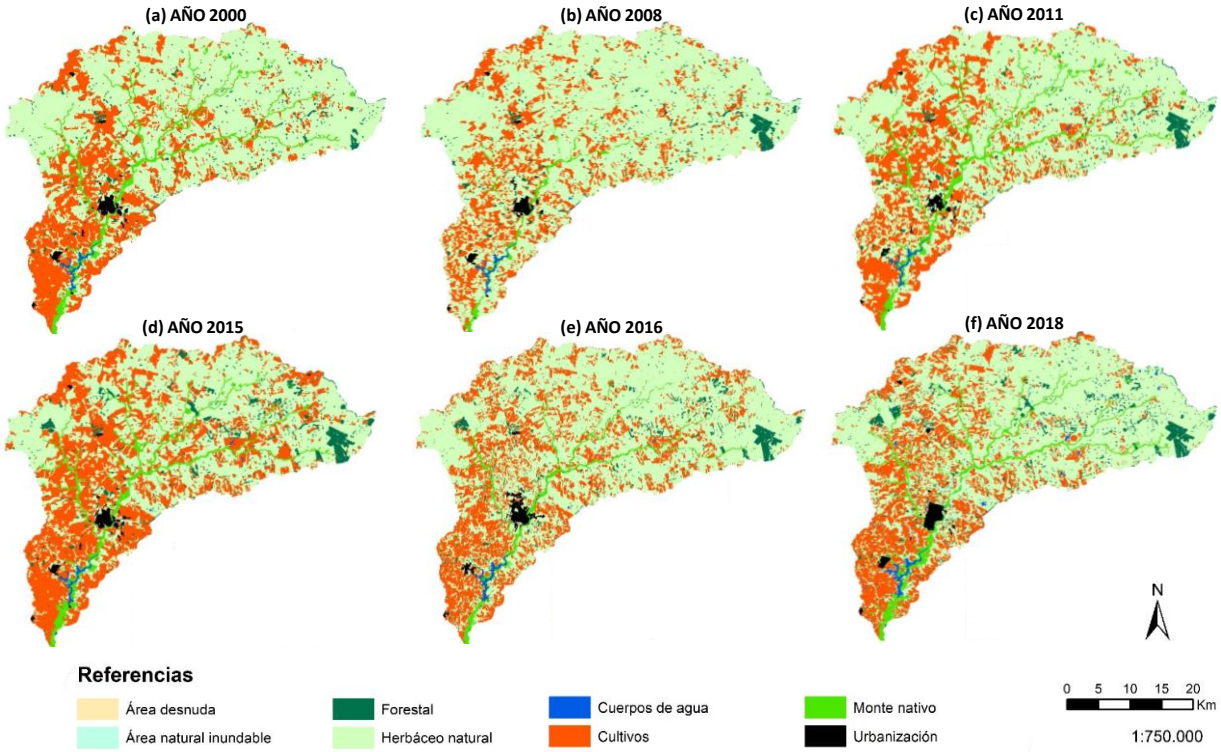


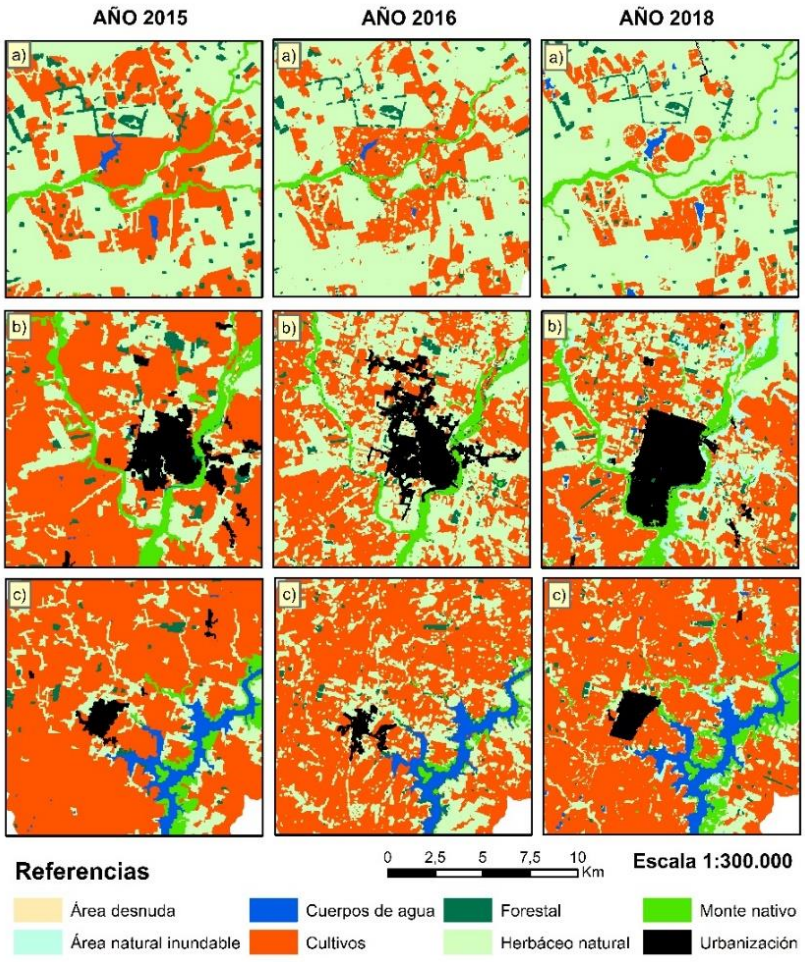
Fig. 3.1. Mapa de cobertura de suelo para los años (a) 2000, (b) 2008, (c) 2011, (d) 2015, (e) 2016, (f) 2018.

Observando los cambios de cobertura de suelo entre los años 2000, 2008 y 2011 (Categoría 1), Fig. 3.1 (a)-(c), se destaca un incremento sostenido en *Forestal* en el sector Noreste de la cuenca. Para el año 2008 existe una diferencia significativa (5% del área total) en *Cultivos* y *Herbáceo natural*, respecto a 2000 y 2011, con un comportamiento espacial que sigue la ubicación de las clases dominantes.

En el año 2015 (Categoría 2), Fig. 3.1 (d), se observa un aumento significativo de *Cultivos* en las zonas del centro Este y Sur de la cuenca, en detrimento de *Herbáceo natural*. Además, el

desarrollo de *Forestal* se produce básicamente en la zona centro Norte de la cuenca, en distintas áreas de la extensión Este – Oeste.

Para la cobertura del año 2016, se observa una mayor concentración de *Cultivos* en la zona centro de la cuenca en comparación con el año 2018. Para este último año mencionado merece señalar que se comienzan a observar geometrías más refinadas, por ejemplo, predios rurales destinados a cultivos por riego mediante pivot central o zonas urbanizadas como polígonos sólidos (ver ejemplo Fig. 3.2).



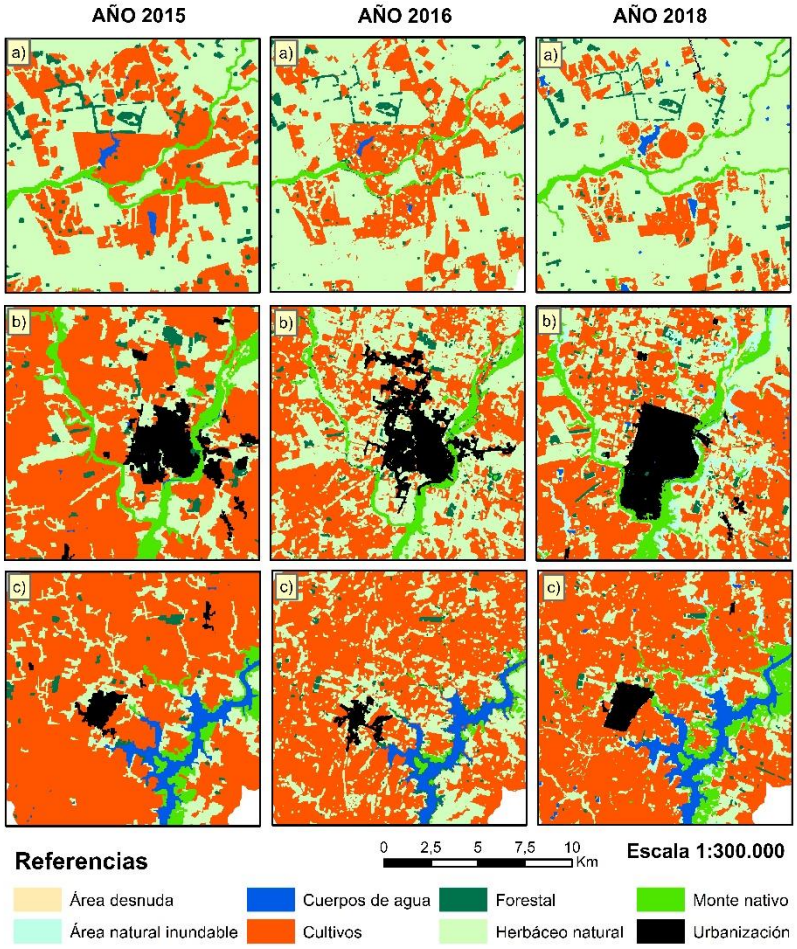


Fig. 3.2. Detalle de comparación de geometrías definidas a partir de las distintas escalas y metodologías para la clasificación de los años 2015, 2016 y 2018.

3.2.4. Variación temporal de la cobertura de suelo

Para abordar el análisis de la variación temporal de la cobertura de suelo, se presentan en la Tabla 3.9, la Tabla 3.10 y en la Fig. 3.3 el área total de las categorías comunes, la distribución porcentual respecto del total y un gráfico de barras para los distintos años, respectivamente. Se observa que las clases dominantes y más dinámicas en todos los años son *Cultivos* y *Herbáceo natural.*

Tabla 3.9. Área total de las categorías comunes para los distintos años.

#	Categorías comunes	Área (km^2)					
		2000	2008	2011	2015	2016	2018
1	***Área desnuda***	1,56	1,27	1,49	1,78	0,00	1,01
2	***Área natural inundable***	0,19	0,19	0,19	0,19	0,00	12,61
3	***Cuerpos de agua***	17,20	17,37	18,94	20,72	13,98	17,87
4	***Cultivos***	840,70	699,94	842,85	1155,95	988,42	867,74
5	***Forestal***	29,56	42,32	45,91	72,87	48,29	59,32
6	***Herbáceo natural***	1585,77	1714,81	1562,76	1217,20	1460,60	1512,94
7	***Monte nativo***	76,90	77,88	80,27	82,17	36,82	74,85
8	***Urbanización***	17,74	15,83	17,21	18,73	18,91	22,99
	Total	2569,62	2569,62	2569,62	2569,62	2567,04	2569,32

Tabla 3.10. Porcentaje respecto del total de las categorías comunes para los distintos años.

#	Categorías comunes	% del total					
		2000	2008	2011	2015	2016	2018
1	***Área desnuda***	0,06	0,05	0,06	0,07	0,00	0,04
2	***Área natural inundable***	0,01	0,01	0,01	0,01	0,00	0,49
3	***Cuerpos de agua***	0,67	0,68	0,74	0,81	0,54	0,70
4	***Cultivos***	32,72	27,24	32,80	44,99	38,50	33,77
5	***Forestal***	1,15	1,65	1,79	2,84	1,88	2,31
6	***Herbáceo natural***	61,71	66,73	60,82	47,37	56,90	58,88
7	***Monte nativo***	2,99	3,03	3,12	3,20	1,43	2,91
8	***Urbanización***	0,69	0,62	0,67	0,73	0,74	0,89
	Total	100,00	100,00	100,00	100,00	100,00	100,00

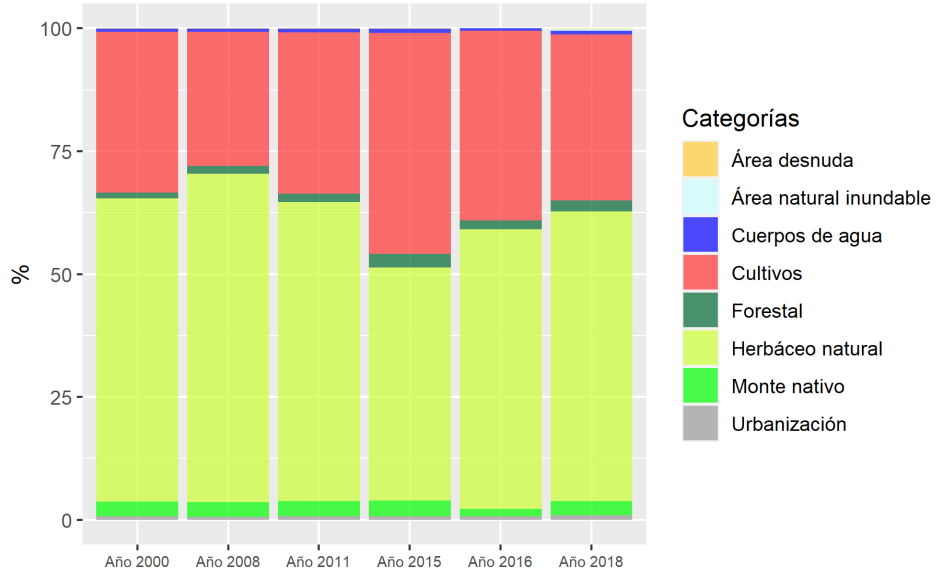


Fig. 3.3. Gráfico de barras de variación temporal de la cobertura de suelo, expresada en porcentaje.

En la Tabla 3.11, se presenta la evolución temporal del cambio de cobertura expresado en porcentaje para años consecutivos en los que se dispone de capa de cobertura. Son llamativos los cambios abruptos que existen entre los años 2011 - 2015 y 2015 – 2016 de las categorías dominantes: *Cultivos* y *Herbáceo natural*. Los cambios que se registran están en el rango de 30.000 a 35.000 has para los 5 años entre 2011 – 2015 y en el rango de 15.000 a 25.000 has para 2 años entre el 2015 – 2016. La magnitud de estos valores da lugar al cuestionamiento de si es físicamente posible la transformación en la cobertura del suelo en los períodos de tiempo observados y, por otro lado, que tan sensibles son estos valores a las metodologías aplicadas para la elaboración de los mapas de cobertura.

Tabla 3.11. Evolución temporal del cambio en la cobertura de suelo expresado en porcentaje.

Categorías comunes	Cambios en la cobertura de suelo (%)				
	2000 - 2008	2008 - 2011	2011 - 2015	2015 - 2016	2016 - 2018
Área desnuda	-0,01	0,01	0,01	-0,01	0,04
Área natural inundable	0,00	0,00	0,00	-0,01	0,49
Cuerpos de agua	0,01	0,06	0,07	-0,27	0,16
Cultivos	-5,48	5,56	12,19	-6,49	-4,73
Forestal	0,50	0,14	1,05	-0,96	0,43
Herbáceo natural	5,02	-5,91	-13,45	9,53	1,98
Monte nativo	0,04	0,09	0,08	-1,77	1,48
Urbanización	-0,07	0,05	0,06	0,01	0,15

La Tabla 3.12 presenta la evolución temporal del cambio en la cobertura del suelo para los periodos 2000 – 2011, 2011 – 2018 y 2000 – 2018, expresada en porcentaje.

Tabla 3.12. Evolución temporal del cambio en la cobertura de suelo expresado en porcentaje, para periodos seleccionados.

Categorías comunes	% de cambio		
	2000 - 2011	***2011 - 2018***	***2000 - 2018***
Área desnuda	0,00	-0,02	-0,02
Área natural inundable	0,00	0,48	0,48
Cuerpos de agua	0,07	-0,04	0,03
Cultivos	0,08	0,97	1,05
Forestal	0,64	0,52	1,16
Herbáceo natural	-0,89	-1,94	-2,83
Monte nativo	0,13	-0,21	-0,08
Urbanización	-0,02	0,22	0,20

Observando los valores de la Tabla 3.12, para las clases dominantes es posible estimar que entre el año 2000 a 2018 hubo un aumento de aproximadamente 2.700 has de la categoría *Cultivos* y una disminución de aproximadamente 7.300 has de la categoría *Herbáceo natural*. Se destaca

también, para el mismo período, el incremento de las categorías *Forestal* y *Urbanización* en aproximadamente 3.000 has y 510 has, respectivamente.

3.3. Generación de mapas con Google Earth Engine

Para poder relacionar las variables de calidad de agua con el uso de suelo, y debido a la poca fidelidad de los mapas de los años anteriores a 2016 junto con las discrepancias de los criterios usados para crear los mapas de 2018 y 2016, se decidió crear mapas clasificados artificialmente a partir de la clasificación del año 2018.

3.3.1. Obtención de mapas

Para generar la clasificación de un mapa, se debe obtener información de un mapa para clasificar, esta información se obtuvo usando la plataforma *Google Earth Engine.* Los datos de cada mapa corresponden a las imágenes obtenidas del satélite Landsat 8, un programa en conjunto de USGS y NASA, que ha estado observando la Tierra desde 1972 hasta el presente. Hoy los satélites Landsat fotografían la superficie de la Tierra a resolución de 30 metros cada dos semanas, incluyendo datos multiespectrales y termales.

Cada mapa cuenta con múltiples capas (llamadas bandas) que contienen distinta información del espectro lumínico. Para generar el mapa de un año, se tomaron todos los mapas generados dentro de la cuenca en el período que incluye dos años hacia atrás y dos años hacia adelante, y se aplicó la mediana en cada banda.

Las bandas usadas se describen a continuación (y se visualizan en la Fig. 3.4):

- Banda 1: Capta ultravioletas.
- Bandas 2, 3 y 4: Captan azul, verde y rojo del espectro visible.
- Banda 5: Capta infrarrojos.
- Bandas 6 y 7: Captan infrarrojos de onda corta.
- Banda 10 y 11: Captan infrarrojos de onda larga.

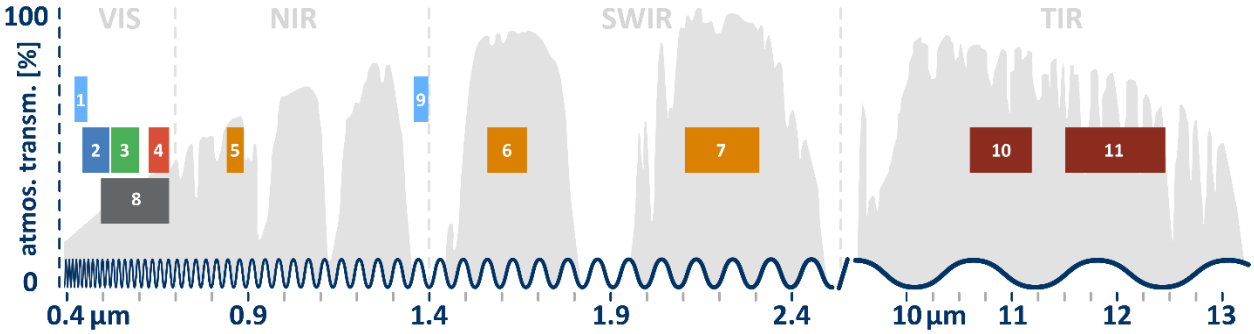


Fig. 3.4 Visualización de las bandas usadas en Landsat 8.

A las bandas antes descritas, se agregaron capas de índices calculados para aportar más información a los algoritmos de clasificación:

- *Normalized Difference Vegetation Index* (NDVI): Índice usado para observar la presencia de vegetación. Se calcula a partir del espectro rojo (R) y el infrarrojo (NIR) con la siguiente formula:

$$NDVI = \frac{NIR - R}{NIR + R} = \frac{B5 - B4}{B5 + B4}$$

  El valor de NDVI varía entre -1 y 1 (Jones, 2010).

- *Normalized Difference Built-up Index* (NDBI): Índice usado para observar la presencia de suelo construido. Se calcula a partir del espectro infrarrojo (NIR) y el infrarrojo de onda corta (SWIR) con la siguiente formula:

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} = \frac{B6 - B5}{B6 + B5}$$

  El valor de NDBI varía entre -1 y 1 (Xu, 2007).

- *Normalized Difference Water Index* (NDWI): Índice usado para observar la presencia de cuerpos de agua. Se calcula a partir del espectro infrarrojo (NIR) y el infrarrojo de onda corta (SWIR) con la siguiente formula (Gao, 1996):

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR} = \frac{B5 - B6}{B5 + B6}$$

  Como el agua pura no refleja ni NIR ni SWIR, se usa la versión modificada (Xu, 2005) y que se calcula a partir del espectro verde (G) y el infrarrojo de onda corta (SWIR) con la siguiente formula:

$$MNDWI = \frac{G - SWIR}{G + SWIR} = \frac{B3 - B6}{B3 + B6}$$

  El valor de MNDWI varía entre -1 y 1.

3.3.2. Etiquetado de mapas

Para generar nuevos mapas, se usaron los datos ya descritos donde cada punto de las imágenes representa un área de 30x30 metros y esa información se divide en las bandas antes comentadas.

El proceso se dividió en dos fases:

Fase 1: Selección de modelos de clasificación

Para etiquetar los mapas se prueban varios tipos de modelo de clasificación, estos tipos de modelo se eligieron de forma de expandir las posibilidades que brinda *Google Earth Engine*, en

esta plataforma solo se cuenta con una versión de parámetros reducidos del *Decision Tree Classifier*[19] sin integración con otras herramientas para buscar los mejores parámetros para el modelo.

Los modelos se describen a continuación:

- *Random Forest Classifier:* modelo que entrena un conjunto de árboles de decisión en varios subconjuntos de datos y promedia las predicciones de la probabilidad de cada clase para mejorar la capacidad predictiva y controlar el sobreajuste (Breiman, 2001). Se usó la biblioteca *sklearn*[20] para su implementación.
- *Extremely Randomized Trees Classifier:* modelo que entrena un conjunto de árboles de decisión en varios subconjuntos de datos y promedia las predicciones de la probabilidad de cada clase para mejorar la capacidad predictiva y controlar el sobreajuste, se diferencia del *Random Forest Regressor* en la forma en la que se decide el punto de división de un nodo, en este caso se hace una división aleatoria mientras que en el anterior se hace de forma óptima, este cambio permite la aceleración del proceso de entrenamiento sin obteniendo resultados iguales o mejores (Geurts, 2006). Se usó la biblioteca *sklearn*[21] para su implementación.
- *KNeighbors Classifier*: modelo de clasificación basado en *k-nearest neighbours* (KNN), las predicciones se calculan con interpolación local de los puntos más cercanos (Mucherino et al., 2009). Se usó la biblioteca *sklearn*[22] para su implementación.
- *Gradient Boosting Classiffier*: modelo que entrena iterativamente un conjunto de árboles de decisión agregando modelos mientras se mejore la predicción la cual se calcula mediante el promedio de las probabilidades de cada clase (Ke et al., 2017). Se usó la biblioteca *lightgbm*[23] para su implementación.

Para cada modelo se realiza la búsqueda de la mejor configuración paramétrica y se usa la biblioteca *optuna*[24] que permite una búsqueda eficiente seleccionando estratégicamente las configuraciones a probar. Esto es posible mediante el uso del algoritmo TPE (*Tree-structured Parzen Estimator*) (Bergstra et al., 2011).
Las mejores configuraciones se eligen maximizando el coeficiente kappa de Cohen.
El conjunto de datos usado para crear y validar los modelos se compone de 80000 puntos seleccionados de forma aleatoria y estratificada, para asegurar que exista aproximadamente la

[19] https://developers.google.com/earth-engine/apidocs/ee-classifier-smilecart
[20] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[21] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
[22] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
[23] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html
[24] https://optuna.readthedocs.io/en/stable/index.html

misma cantidad de puntos para cada categoría posible buscando mejorar la predicción de todas las etiquetas. Para entrenar se usa el 80% de los puntos y para validar el modelo obtenido se usa el restante 20%.

El resultado de esta fase es el mejor modelo de clasificación, caracterizado por el coeficiente kappa más alto sobre el conjunto de validación. En particular, el modelo seleccionado fue *Gradient Boosting Classifier* con coeficiente kappa 0.717.

Fase 2: Creación y corrección de mapa etiquetado

A partir del modelo obtenido en la fase anterior, se genera un mapa usando todos los puntos de unos de los mapas requeridos.

Los mapas generados presentaron zonas donde la clasificación de la mayoría de los puntos contrastaba con algunos dispersos en su interior. Para solucionar este problema y mejorar la calidad de la clasificación, se usó un método de suavizado donde cada punto fue corregido usando la clasificación usada por la mayoría de los puntos del cuadrante directamente circundante. Un ejemplo de este proceso se muestra en la Fig. 3.5.

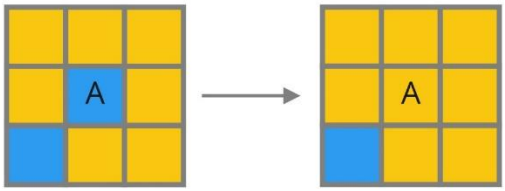


Fig. 3.5 Ejemplo de corrección de la clasificación del punto A.

El resultado de esta fase son los mapas generados para los años 2014, 2016, 2018 y 2020.

3.3.3. Resultados

Para evaluar el desempeño final del mejor modelo de clasificación obtenido anteriormente, se comparó el porcentaje de área ocupada por cada tipo de uso de suelo entre el mapa real, el mapa generado y el mapa corregido para el año 2018. En la Tabla 3.13, se muestran los resultados obtenidos y se nota una mejora en general en el mapa corregido frente al generado. La Tabla 3.14 presenta los resultados de las áreas para todas las clases de los mapas generados del año 2014, 2016, 2018 y 2020. La Tabla 3.15 presenta los resultados expresados en porcentajes del área total. Para todas las clases los cambios temporales en términos porcentuales entre años son pequeñas (0.1 a 2.0% en términos medios), no obstante, en términos de área para las clases más representadas (*Herbáceo natural y Cultivos*) se tienen cambios medios de 5000 ha, por otro parte las clases *Forestal, Monte Nativo, Área natural inundable y Urbanización* tienen cambios medios de 500 ha, las demás clases un orden menos de magnitud. Las variaciones espaciales para los mapas generados se pueden visualizar en la Fig. 3.6. La diferencia más significativa que se observa

a la escala de trabajo es el desarrollo de la forestación en la zona noreste de la cuenca y en menor proporción en la zona centro de la cuenca, esta transformación comienza a desarrollarse en 2016 y a partir del año 2018 se encuentra consolidada.

Tabla 3.13. Comparación de áreas ocupadas según cada uso de suelo para el mapa real y los generados en 2018.

#	Uso de suelo	Área (%)		
		Mapa real	Mapa generado	Mapa corregido
1	***Área desnuda***	0.04	0.50	0.28
2	***Área natural inundable***	0.42	1.10	0.50
3	***Cuerpos de agua***	0.79	0.78	0.79
4	***Cultivos***	37.10	32.10	34.30
5	***Forestal***	3.00	4.40	3.60
6	***Herbáceo natural***	54.20	54.80	55.50
7	***Monte nativo***	3.60	4.50	4.10
8	***Urbanización***	0.86	1.80	0.94

Tabla 3.14. Área para cada categoría de uso de suelo de los mapas generados para los distintos años.

#	Uso de suelo	Área (ha)			
		2014	2016	2018	2020
1	***Área desnuda***	782.20	711.84	711.35	755.20
2	***Área natural inundable***	1600.30	1284.25	1283.61	2200.35
3	***Cuerpos de agua***	2035.82	2003.48	2027.79	1970.37
4	***Cultivos***	97960.99	87909.39	88216.84	91947.77
5	***Forestal***	7434.92	8500.913	9189.45	8924.92
6	***Herbáceo natural***	135644.19	144385.18	142997.88	138202.74
7	***Monte nativo***	10063.33	10658.19	10600.70	10856.80
8	***Urbanización***	1917.06	1985.57	2411.22	2580.51

Tabla 3.15. Porcentaje del área total representado por cada clase de uso de suelo para los distintos años.

#	Uso de suelo	Área respecto del total (%)			
		2014	2016	2018	2020
1	***Área desnuda***	0.30	0.28	0.28	0.29
2	***Área natural inundable***	0.62	0.50	0.50	0.85
3	***Cuerpos de agua***	0.79	0.78	0.79	0.77
4	***Cultivos***	38.05	34.15	34.27	35.72
5	***Forestal***	2.89	3.30	3.57	3.47
6	***Herbáceo natural***	52.69	56.09	55.55	53.68
7	***Monte nativo***	3.91	4.14	4.12	4.22
8	***Urbanización***	0.74	0.77	0.94	1.00

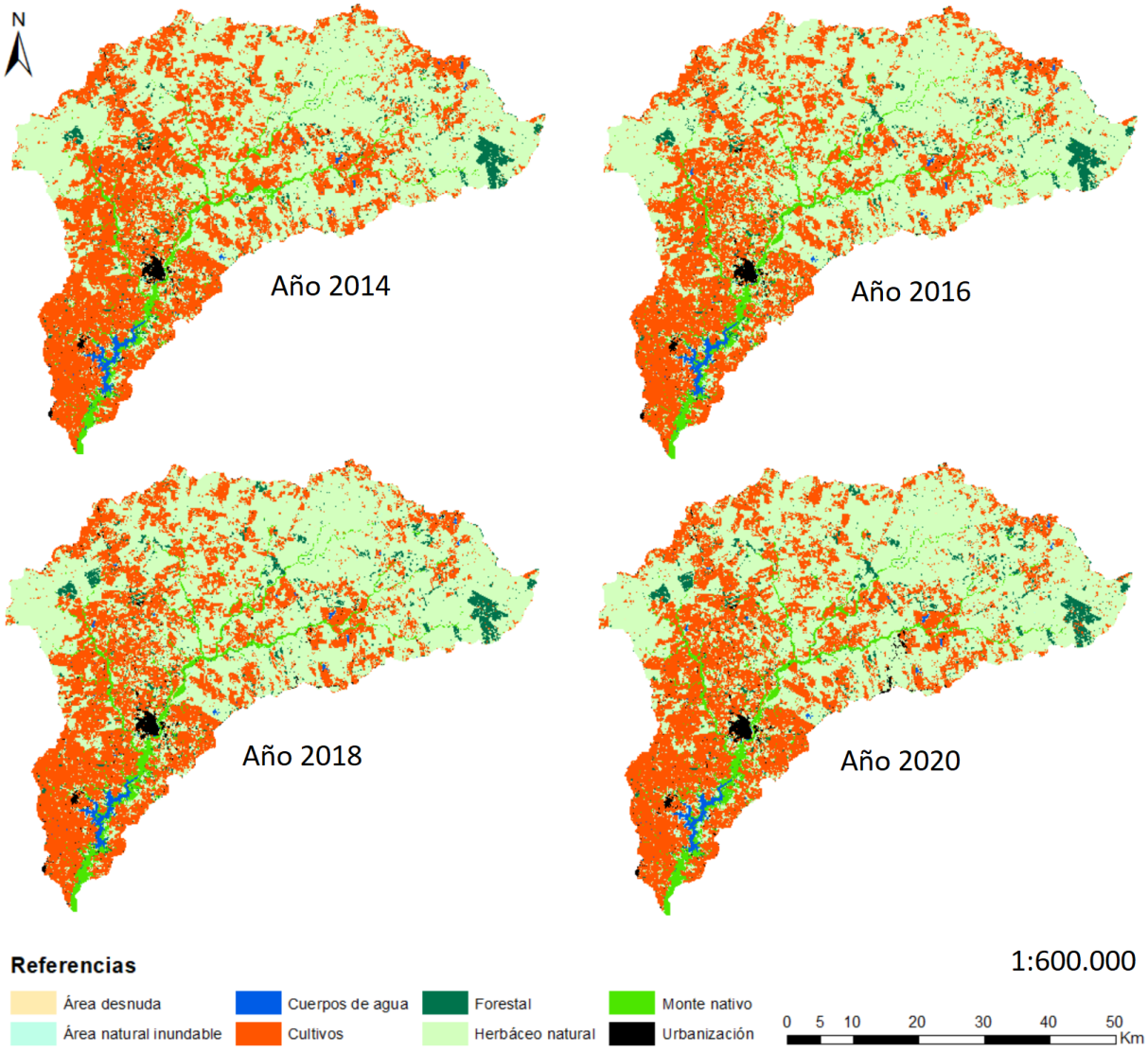


Fig. 3.6. Distribución especial de las clases en los mapas generados.

4. OE3: Analizar la variabilidad temporal y espacial de los parámetros de calidad del agua en la cuenca del río Santa Lucía

El OE3 plantea un análisis de variabilidad tanto temporal como espacial de los datos de calidad de agua aumentados en el OE1 y su comparación con los observados. Para llevar a cabo este OE, se compararon resultados de distintos tests estadísticos entre las variables originales y sus imputaciones, también se calcularon tendencias y se crearon gráficas animadas que muestran la evolución espacio-temporal de cada variable. Para acceder a los distintos análisis se creó una aplicación web usando la biblioteca *streamlit*[25] (Fig. 4.1).

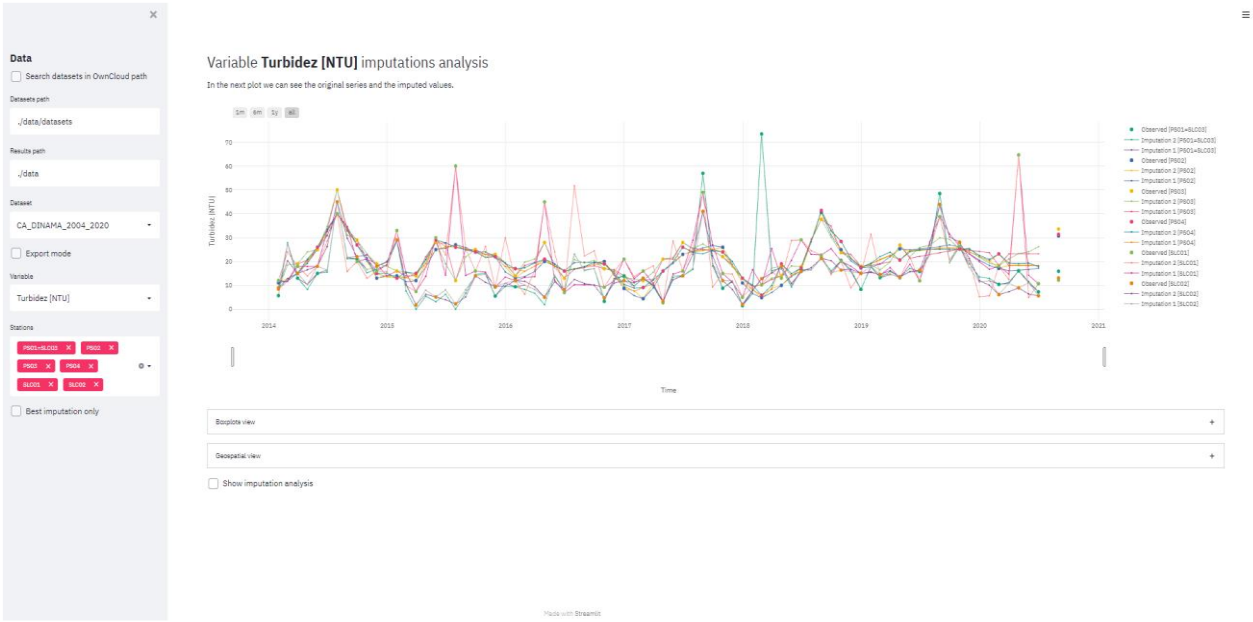


Fig. 4.1. Captura de pantalla de la aplicación desarrollada para la evaluación de la variabilidad espacio-temporal de las variables de calidad de agua.

Es importante destacar que la aplicación fue desarrollada no solo para las variables de calidad de agua, sino para todo el conjunto de variables analizadas para cumplir el OE1 (también variables hidro-meteorológicas).

[25] https://docs.streamlit.io/en/stable/

A la izquierda, es posible seleccionar el dataset (DINAMA, DINAGUA, etc.) al cual la variable pertenece, la variable que se quiere analizar y las estaciones donde esta variable fue registrada. Después de esta selección, en la derecha, es posible visualizar los resultados de diferentes maneras: gráficos de líneas, boxplots, mapas animados y gráficos de descomposición.

A continuación, para ilustrar los resultados, se mostrarán elementos de la aplicación.

4.1. Análisis de variabilidad espacio-temporal

4.1.1. Variabilidad temporal

Para analizar la variabilidad temporal de una variable, se compara de manera gráfica el comportamiento en cada estación de medición (considerando los datos originales e imputados) en todo el período de estudio (2014-2020). Además, se grafican los datos por año en forma de boxplot. En Fig. 4.2 y Fig. 4.3, se muestran las gráficas para la variable Fosforo total (PT) [µg P/L].

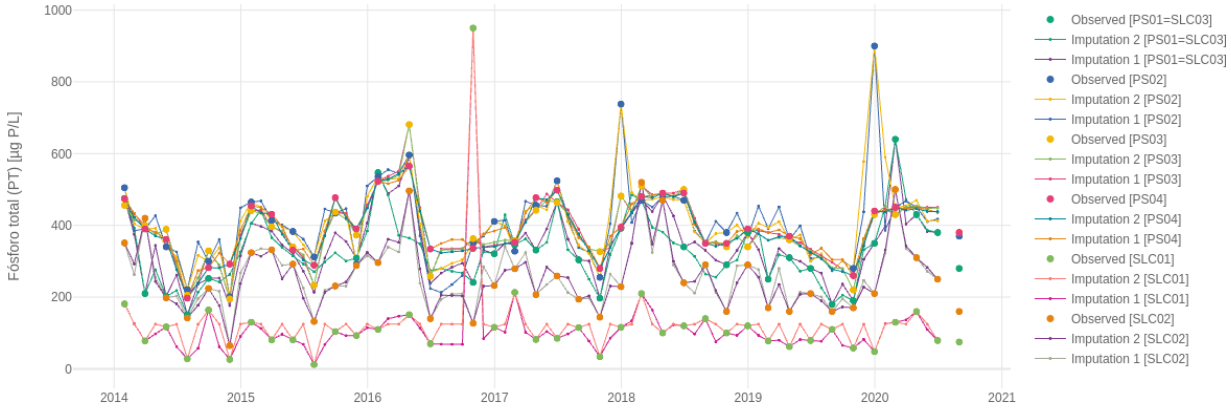


Fig. 4.2 Comportamiento de la variable Fosforo total (PT) [µg P/L] en cada estación (2014-2020). Comparación de observaciones e imputaciones con gráfico de líneas.

Fig. 4.3. Comportamiento de la variable Fosforo total (PT) [µg P/L] por año en cada estación. Comparación de observaciones e imputaciones con boxplots.

En la Fig. 4.2, se puede observar que todas las series imputadas elegidas como las mejores (líneas) incorporan todas las observaciones del PT (puntos). Eso se cumple para todas las variables de calidad de agua en todas las estaciones consideradas.

La Fig. 4.3 muestra como las imputaciones son capaces de mantener la distribución de las variables de calidad de agua a lo largo de los años y en las diferentes estaciones de monitoreo. Considerando esta similaridad, en la figura siguiente (Fig. 4.4), se reportan los boxplots de las observaciones de algunas de las variables de calidad de agua con el fin de evaluar más en detalle su variabilidad temporal y espacial. Mirando los diferentes boxplots, es interesante identificar dos grupos de comportamiento diferentes: los tres sitios de monitoreo ubicados en el embalse de Paso Severino muestran patrones diferentes a los que caracterizan las estaciones ubicadas aguas arriba del embalse. Además, la Temperatura del agua y el Oxígeno Disuelto son los únicos

contaminantes que muestran una fuerte estacionalidad intra e interanual, mientras que no podemos identificar un patrón claro para los otros contaminantes. Es importante remarcar el alto aporte de nutrientes de PS01 (TN, NO_2^-, NO_3^-), donde se ubica la ciudad más grande de la cuenca (Florida, que, con una población de más de 33.000 habitantes, alberga a casi la mitad de los habitantes de la región). Se sabe que las áreas urbanizadas son fuentes de nitrógeno debido a la deposición atmosférica, la aplicación de fertilizantes para el césped, las aguas residuales y la infraestructura de alcantarillado con fugas. La Turbidez muestra un patrón temporal meno marcado a lo largo de los años, con los valores más altos registrados en las estaciones de monitoreo ubicadas aguas arriba del embalse de Paso Severino (SLC01, SLC02 y PS01).

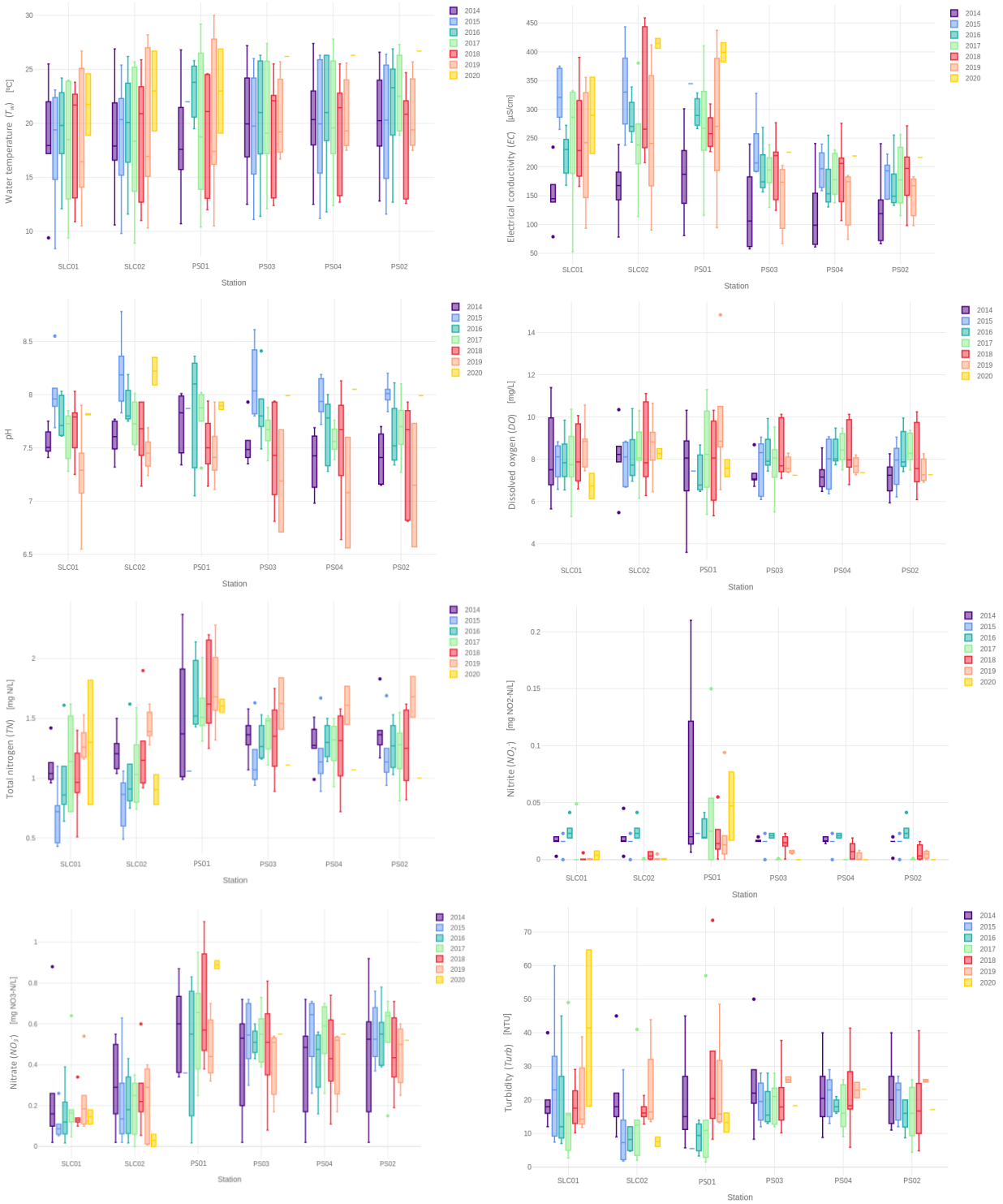


Fig. 4.4. Variabilidad temporal y especial de las observaciones de algunas variables de calidad de agua.

Además, se evaluó la presencia de patrones estacionales para las diferentes variables de calidad de agua. Este analisis se hizo para todas las variables de calidad de agua, considerando cuatro

(verano-otoño-invierno-primavera) y dos (verano-invierno) estaciones. Considerando los cuatros estaciones, no se identificó un patrón claro para la mayoría de las variables, con exclusión de la variable Temperatura del agua (T) [ºC]. Por lo tanto, siguió el analisis considerando dos estaciones. En la Fig. 4.5, se muestran, como ejemplo, los boxplots para la variable Turbidez [NTU].

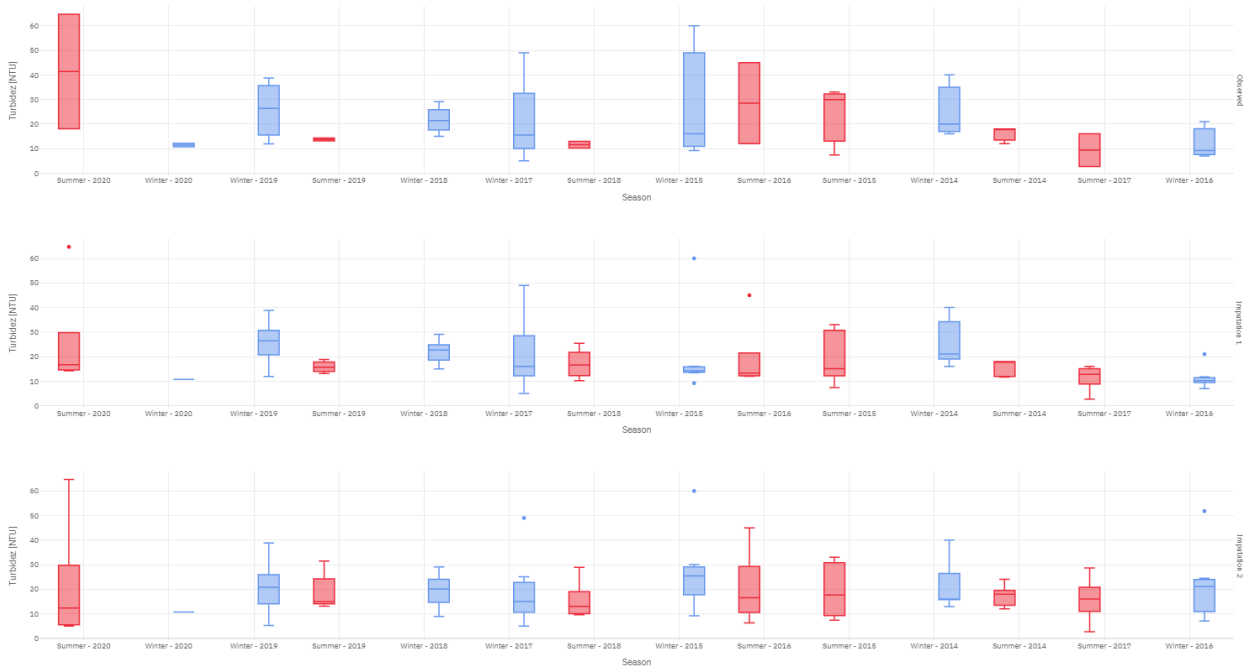


Fig. 4.5. Estacionalidad (invierno-verano) de la variable Turbidez [NTU] en la estación SLC01.

En estos boxplots, el período de "invierno" incluye las temporadas de otoño e invierno (abril-mayo-junio-julio-agosto-septiembre), y la ventana de "verano" considera las temporadas de primavera y verano (octubre-noviembre-diciembre-enero-febrero-marzo). En general, también estacionalmente, las distribuciones de las variables imputadas representan bien las de las observaciones. En particular, para la variable Turbidez, se observa una leve estacionalidad en la cual se pueden observar valores más elevados y extremos más altos en invierno. Eso se puede justificar con el hecho que en invierno hay más eventos lluviosos extremos que hacen que el almacenamiento de agua en el suelo sea mayor. Eso acoplado a las bajas temperaturas, generan una menor evapotranspiración y, por lo tanto, una mayor generación de escorrentía superficial. Dicho proceso, junto a la alta energía de las gotas de lluvia de los eventos extremos, determina una mayor movilización de las partículas del suelo y, por lo tanto, una mayor exportación de dichos sedimentos al cuerpo de agua (Gorgoglione et al., 2020).

Los boxplots de las otras variables de calidad de agua considerando dos y cuatros estaciones están en la aplicación desarrollada en el marco de este proyecto.

4.1.2. Variabilidad espacial

Para realizar un análisis espacial de la evolución de cada variable, se hace una gráfica animada donde se muestran los valores de una variable en cada estación sobre el mapa de la cuenca del río Santa Lucía Chico, para esta gráfica se usan los puntos del mejor modelo de imputación. En la Fig. 4.6 se muestran, como ejemplo, las gráficas para la variable Nitrógeno total (NT) [mg N/L].

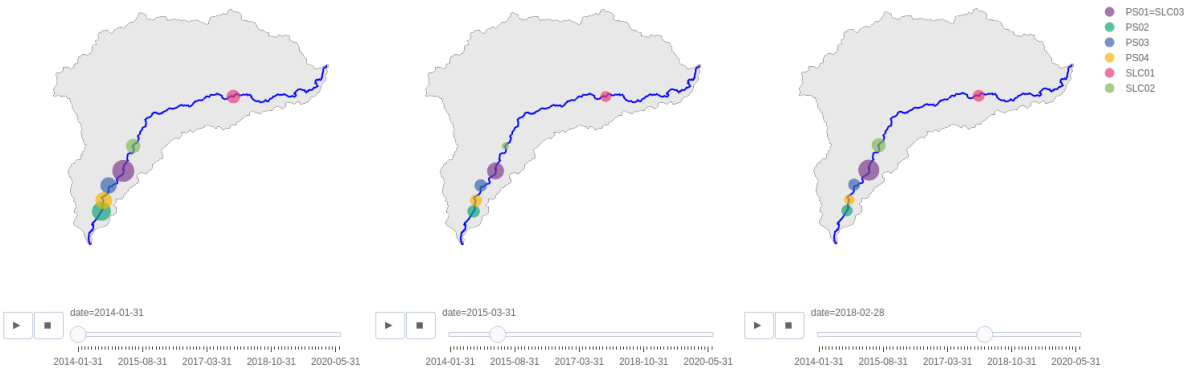


Fig. 4.6 Evolución espacial de la variable Nitrógeno total (NT) [mg N/L] mostrada de forma geográfica por estación.

De la Fig. 4.6 se puede ver que no existe un marcado patrón espacial de la variable NT. Aunque se puede observar que los valores de NT son siempre muy altos en la estación PS01=SLC03, ubicada aguas debajo de la ciudad de Florida. Es confirma los resultados encontrados en la Fig. 4.4. Además, los valores de los nutrientes son siempre más altos en el embalse de Paso Severino (estaciones PS03, PS04 y PS02) respecto a las estaciones ubicadas aguas arriba del embalse, en el curso de agua principal del Santa Lucía Chico (SLC01 y SLC02). Este resultado confirma estudios anteriores en los cuales se detectaron niveles de ipereutrofía en el embalse para el período 2011-2018, mostrando una elevada carga de nutrientes en dicho cuerpo de agua (Gorgoglione et al., 2020).

Las gráficas animadas de todas las otras variables que muestra su variación espacio-temporal están en la aplicación desarrollada en el marco del proyecto.

4.2. Análisis de estacionalidad

Una serie temporal $X = \{X_t\}_{t\geq 1}$ es estacionaria si la media y la variabilidad se mantienen constantes a lo largo del tiempo, es decir:

- $E[X_t] = \mu \forall t$
- $Var(X_t) = \sigma^2 \forall t$
- $Cov(X_t, X_{t+k}) = \gamma_k \forall t, \forall k$

Es interesante que una serie sea estacionaria pues como la media es constante se puede estimar a partir de los datos y usar ese valor para predecir nuevas observaciones.

Series no estacionarias pueden mostrar ciertos fenómenos:

- Existen cambios de varianza.
- Existe una tendencia, es decir, la media aumenta o decrece a lo largo del tiempo.
- Existen efectos estacionales, es decir, el comportamiento de la serie se repite cada cierto período temporal.

Para detectar si la serie temporal de una variable es estacionaria, se realizan tests estadísticos de raíces unitarias. Se dice que una serie temporal X tiene raíz unitaria si se cumple que:

$$X_t = \alpha X_{t-1} + \beta X_e + \epsilon$$

Con $\alpha = 1$ donde X_e es una variable exógena.

La presencia de raíz unitaria es señal de que la serie temporal no presenta estacionalidad, los tests realizados son:

- **Test ADF** (*Augmented Dickey Fuller*): la hipótesis nula (H0) de este test es que $\alpha = 1$, por lo tanto, si se obtiene un p-valor menor al nivel de significancia 0.05 se puede rechazar la H0 e inferir que la serie es estacionaria (Fuller, 1996).
- **Test KPSS** (*Kwiatkowski-Phillips-Schmidt-Shin*): la hipótesis nula (H0) de este test es que la serie es estacionaria alrededor de una tendencia determinística, por lo tanto, si se obtiene un p-valor menor al nivel de significancia 0.05 se puede rechazar la H0 e inferir que la serie no es estacionaria alrededor de una tendencia determinística (Kwiatkowski et al., 1992).

Ambos tests se realizan en la serie temporal original y las imputaciones para detectar posibles cambios. En la Fig. 4.7, se muestra este proceso para la variable Conductividad [µS/cm] en la estación SLC01, en este ejemplo se puede ver que tanto los valores observados como las imputaciones presentan estacionalidad.

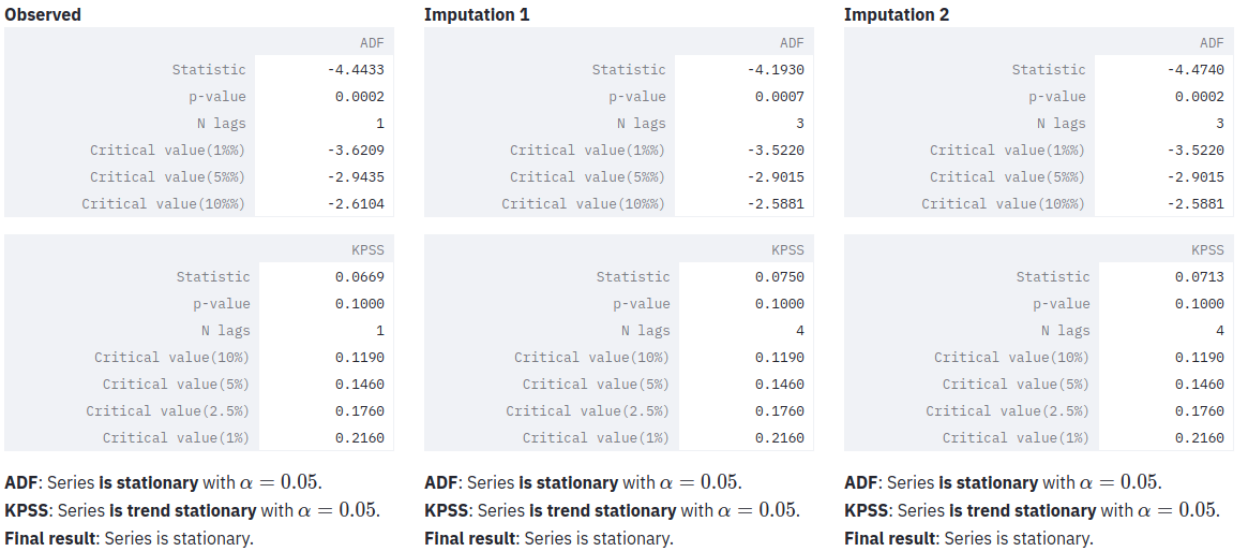


Fig. 4.7 Aplicación de tests estadísticos para determinar estacionalidad de la variable Conductividad [µS/cm] en la estación SLC01. Se comparan los puntos observados y las imputaciones.

Al detectar que una serie temporal es no estacionaria, se puede aplicar un proceso de diferenciación para convertirla en estacionaria, este proceso se realiza aplicando la siguiente transformación:

$$X_t = X_t - X_{t-1}$$

Esta transformación se puede aplicar hasta que los tests indiquen que la serie resultante es estacionaria.

4.3. Descomposición estacional

Una serie temporal se puede descomponer en tres componentes principales (Bowerman, 2007):

- **Tendencia**: describe el comportamiento general de la serie a largo plazo, esta trayectoria puede ser negativa o positiva.
- **Estacionalidad**: describe movimientos oscilatorios de la serie a corto plazo, dentro de un período temporal.
- **Residual**: describe las variaciones aleatorias alrededor de los componentes anteriores.

Una serie se puede descomponer en estos componentes para estudiar su comportamiento. Se prueban dos modelos de descomposición para su estudio:

- **Modelo de descomposición aditiva**: se puede utilizar para descomponer series temporales que muestran una variación estacional constante, se modela con la siguiente formula (Bowerman, 2007):

$$X_t = T_t + E_t + R_t$$

- **Modelo de descomposición multiplicativa**: es útil para descomponer series temporales que manifiestan una variación estacional creciente o decreciente, se modela con la siguiente formula (Bowerman, 2007):

$$X_t = T_t \cdot E_t \cdot R_t$$

Ambos modelos de descomposición se aplican en la serie temporal original y las imputaciones para detectar posibles cambios y para verificar que las imputaciones siguen el mismo comportamiento de las variables originales. Considerando que el modelo de descomposición aditiva es el más común para las variables ambientales, lo consideramos para este estudio. En la Fig. 4.8, se muestra este proceso para la variable Temperatura del agua (T) [ºC].

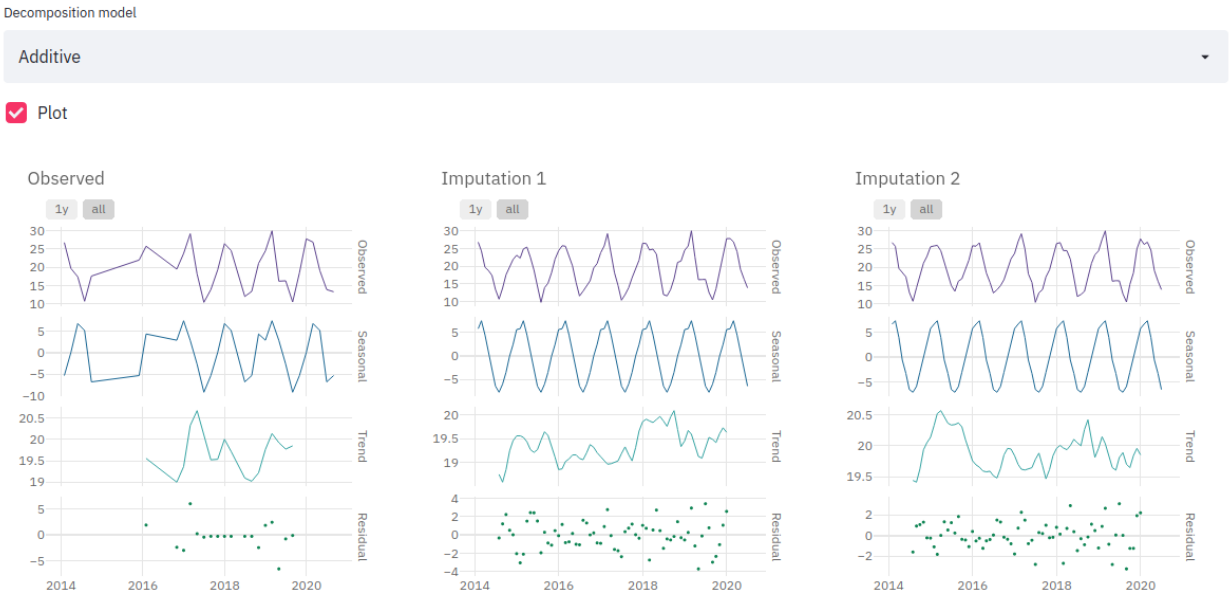


Fig. 4.8 Descomposición estacional para la variable Temperatura del agua (T) [ºC] en la estación PS01=SLC03 usando el modelo de descomposición aditiva. Se comparan los puntos observados y las imputaciones.

Se puede observar claramente que las imputaciones representan bien la tendencia de la variable observada complementándola razonablemente. Este comportamiento es muy claro en la representación de la estacionalidad (Fig. 4.8).

El análisis de descomposición se hizo para todas las variables de calidad de agua y se encuentra en la aplicación desarrollada en el marco de este proyecto.

5. OE4: Evaluar las relaciones entre las categorías definidas por LULC con las variables de calidad del agua en sitios críticos de la cuenca del río Santa Lucía

Para llevar a cabo el OE5, se realizaron dos etapas que consistieron en el procesamiento de los datos obtenidos en los OEs anteriores seguido de la aplicación de varios métodos lineales y no-lineales de aprendizaje no supervisado para encontrar las posibles relaciones entre las variables de calidad de agua y las de uso de suelo. También en este caso, para acceder a los distintos análisis se creó una aplicación web usando la biblioteca *streamlit*[26] (Fig. 5.1), esto permite una gran capacidad de variabilidad en los estudios realizados.

En la izquierda, la aplicación permite seleccionar tanto los datos de entrada como de salida para los distintos modelos que se usan para mostrar las relaciones entre las variables. En particular, se pueden seleccionar las estaciones de monitoreo, el tipo de contaminante, los agregadores estadísticos que se quieren considerar en el análisis (min, Q1, media, mediana, etc.), los escenarios de uso del suelo (entera subcuenca y diferentes zonas buffers), las clases de uso del suelo a considerar y las diferentes métricas de uso del suelo. Después de esta selección, en la derecha, es posible visualizar los resultados: los estadísticos de cada contaminante en cada estación por año, la variabilidad temporal de los usos del suelo seleccionados en los diferentes escenarios y los resultados de los análisis de relación uso del suelo-calidad de agua a diferentes escalas espaciales (escala de clase de uso del suelo y escala de paisaje).

[26] https://docs.streamlit.io/en/stable/

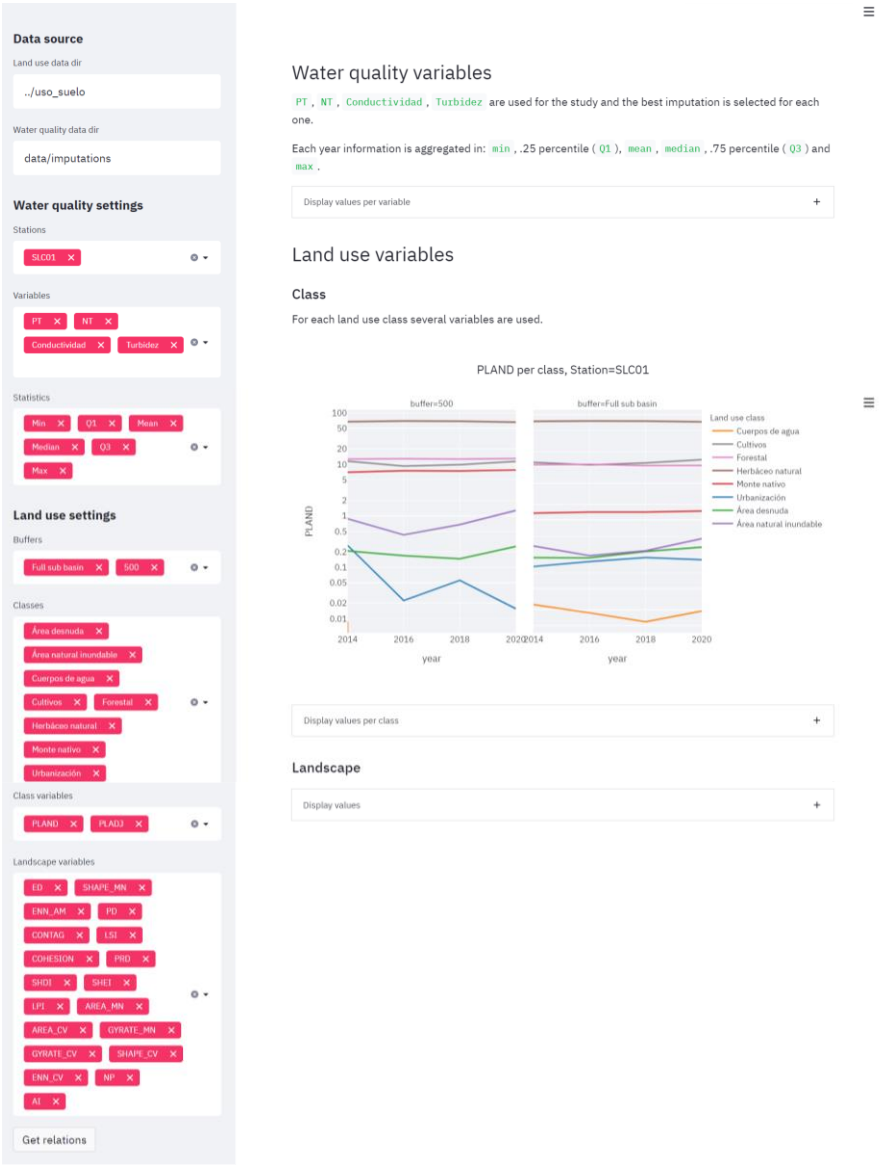


Fig. 5.1. Captura de pantalla de la aplicación desarrollada para la evaluación de las relaciones entre uso del suelo y calidad del agua.

5.1. Procesamiento de datos

Para cada conjunto de datos se aplicó una metodología distinta que se explica a continuación.

5.1.1. Variables de uso de suelo

A partir de los mapas generados en el OE2, se calcularon diferentes métricas de paisaje con el fin de describir la composición y configuración del paisaje para cada cuenca. Una métrica de paisaje es un valor numérico (escalar) que resume de cierta forma algún componente de la estructura

del paisaje. En general, el uso común del término “métricas de paisaje” se refiere exclusivamente a índices desarrollados para patrones de mapas categóricos, por ejemplo, mapas de uso del suelo. En este sentido, las métricas de paisaje son algoritmos que cuantifican características espaciales específicas de parches, clases de parches o mosaicos de paisajes completos. Los parches individuales poseen relativamente pocas características espaciales fundamentales (por ejemplo, tamaño, perímetro y forma), la agregación de parches tiene una variedad de propiedades dependiendo de si la agregación es de una sola clase o de varias clases, dentro de una subregión específica de un paisaje o en todo el paisaje.

Las métricas a nivel de clase representan la cantidad y la distribución espacial de un solo tipo de parche y pueden interpretarse como índices de fragmentación. Las métricas a nivel del paisaje representan el patrón espacial de todo el mosaico del paisaje y pueden interpretarse de manera más amplia como índices de heterogeneidad del paisaje porque miden la estructura general del paisaje.

Se ha desarrollado una gran cantidad de métricas para cuantificar los patrones del paisaje en mapas categóricos (McGarigal, K., 2014; McGarigal, K., 2015). Estas métricas se dividen en dos categorías: las que cuantifican la composición del mapa sin referencia a los atributos espaciales y las que cuantifican la configuración espacial del mapa.

La composición se refiere a las características asociadas con la variedad y abundancia de tipos de parches dentro del paisaje, pero sin considerar el carácter espacial, la locación o la ubicación de los parches. La composición requiere la integración de todos los tipos de parches. Las principales medidas de composición son:

- Abundancia proporcional de cada clase,
- Riqueza, representa el número de diferentes tipos de parches,
- Uniformidad, abundancia relativa de diferentes tipos de parches,
- Diversidad, es una medida compuesta de riqueza y uniformidad.

La configuración espacial se refiere al carácter espacial y la disposición, posición u orientación de los parches dentro de la clase o el paisaje. Algunos aspectos de la configuración, como el aislamiento o el contagio, son medidas de la ubicación de los tipos de parches en relación con otros parches u otras características de interés. Los aspectos principales de la configuración y una muestra de métricas representativas son:

- Área y perímetro,
- Complejidad de la forma,
- Agregación,
- Subdivisión,
- Aislamiento.

Existen numerosos desafíos para el uso y la interpretación adecuada de las métricas del paisaje, que incluyen: 1) definir un paisaje relevante para el fenómeno de estudio, 2) obtener una comprensión teórica y empírica adecuada del comportamiento de las métricas para ayudar en la interpretación de cada métrica, 3) comprender las redundancias teóricas y empíricas entre las métricas para asegurar su uso parsimonioso, y 4) desarrollar un marco de referencia adecuado para interpretar ecológicamente el valor calculado de cada métrica.

Para el estudio de las métricas de paisaje, se definieron 20 escalas espaciales distintas, correspondientes a 4 subcuencas dentro de la cuenca del río Santa Lucía Chico y 4 zonas buffer dentro de cada subcuenca, de 500 m, 1000 m, 1500 m y 2000 m. Las subcuencas quedan definidas por los puntos de cierre presentados en la Tabla 5.1 y pueden visualizarse en la Fig. 5.2. El trazado de las mismas se realizó utilizando el modelo digital de elevaciones del RENARE (MGAP, 2020) e IDEuy (IDE) las herramientas de análisis espacial del programa ArcGis 10.3. La subcuenca 4 contiene a todas las demás subcuencas y se considera el cierre de la misma en el punto de muestreo más cercano a la represa de Paso Severino.

Tabla 5.1. Definición de subcuencas según punto de cierre.

Puntos de cierre			
Nombre subcuenca	**Código estación DINAMA**	**X (m) UTM 21S**	**Y (m) UTM 21S**
Subcuenca 1	SLC01	603079	6241618
Subcuenca 2	SLC02	573434	6227394
Subcuenca 3	PS01=SLC03	570470	6220119
Subcuenca 4	PS02	563941	6208344

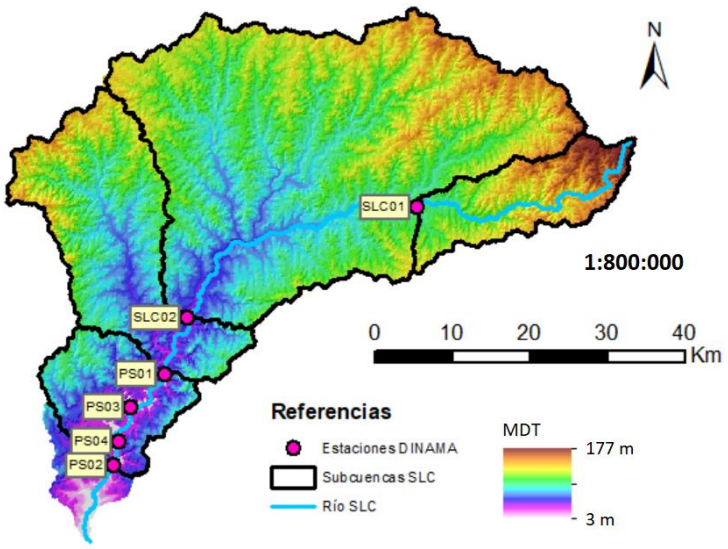


Fig. 5.2. Visualización de las subcuencas definidas.

Para la selección de las métricas de paisaje a ser utilizadas, se realizó, en primera instancia, un análisis de correlación de diversas métricas calculadas a nivel de paisaje, para comparar los resultados obtenidos del mapa original 2018 con el mapa generado para el mismo año. Se realizaron cálculos de las métricas de correlación lineal de Pearson y las no lineales de Spearman y Kendall, entre el conjunto *X* y el conjunto *Y*, siendo $X = \{X_1, X_2, X_3..., X_{20}\}$ una métrica de paisaje calculada con el mapa original del año 2018 y los subíndices 1 a 20 los distintos niveles espaciales analizados (4 subcuencas y 4 zonas buffer). El conjunto *Y* representa la misma métrica, pero calculada para el mapa generado del año 2018. Los resultados de este análisis de regresión son presentados en forma gráfica en la Fig. 5.3. Todas las métricas fueron calculadas con el programa FRAGSTATS v4.2 (ver referencia McGarigal, K. et al., 2012, para la formulación de cada métrica), con la configuración establecida por defecto (i.e., regla de 8 vecinos cercanos, sin estrategia de muestreo, etc). En base al análisis de correlación y las publicaciones de (Li, N.X. et al., 2020; Sen, X. et al., 2020; Lee, S.-W. et al., 2009) se realizó una selección de métricas que permiten describir la composición y configuración del paisaje, las cuales son presentadas en las Tabla 5.2 y Tabla 5.3.

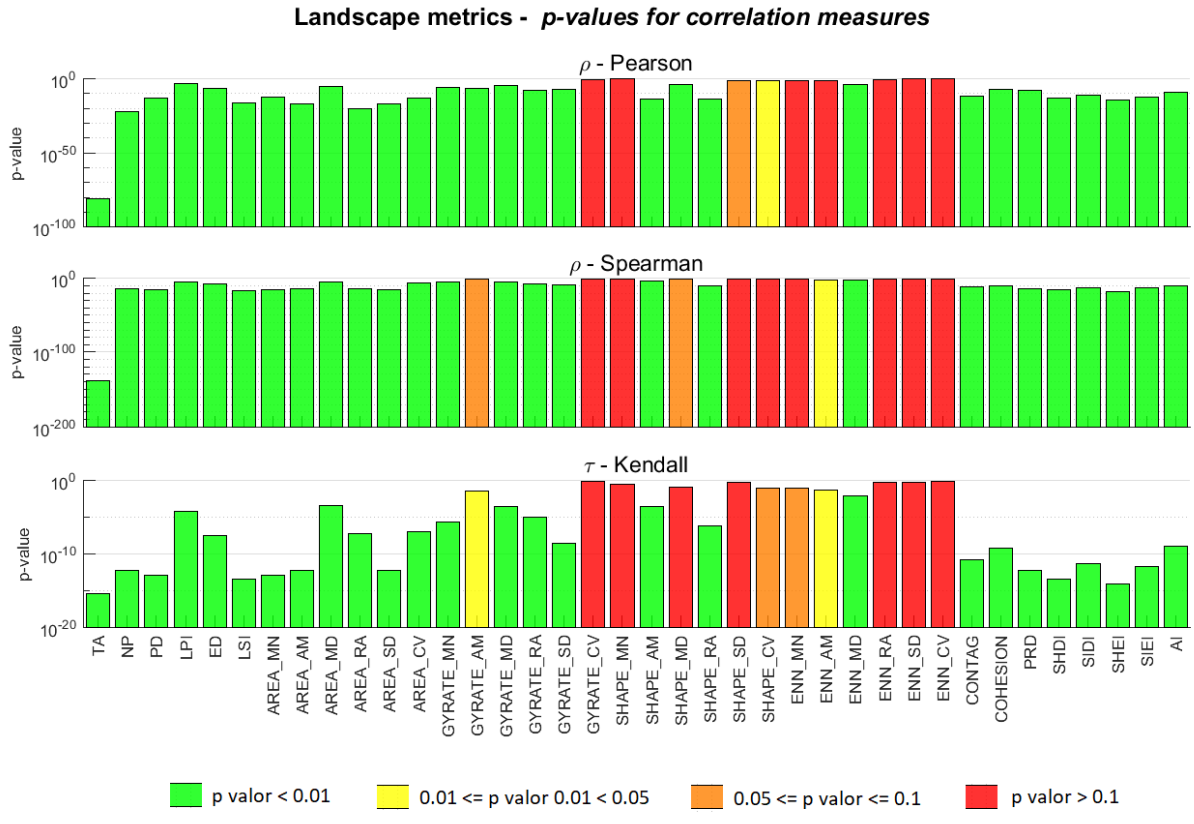


Fig. 5.3. Resultados del análisis de regression para la selección de las métricas de uso del suelo.

Tabla 5.2. Métricas a nivel de clase.

Nombre FRAGSTATS	Nombre	Unidades	Rango	Descripción
PLAND	Porcentaje de área de cada clase	%	0 < PLAND <= 100	Cuantifica la abundancia relativa de cada clase en el paisaje.
AI	Índice de agregación	%	0 <= AI <= 100	Permite cuantificar el nivel de agregación de la clase en el patrón espacial.
COHESION	Índice de cohesión	Adimensional	0 <= COHESION <= 100	Indica la conexión física de la correspondiente clase o uso de suelo.

Tabla 5.3. Métricas a nivel de paisaje.

Nombre FRAGSTATS	Nombre	Unidades	Rango	Descripción
PD	Densidad de parches	#/100 ha	PD > 0	Es el número de parches en el paisaje por unidad de área.
LPI	Índice de parche más largo	%	0 < LPI <= 100	Cuantifica el porcentaje del área total del paisaje ocupada por el parche más grande.
ED	Densidad de perímetro	m/ha	ED >= 0	Representa la suma de todas las longitudes de los segmentos de perímetros dividido el área total del paisaje.
LSI	Índice de forma del paisaje	Adimensional	LSI >= 1	Es una medida estandarizada de la densidad de perímetro ajustada al tamaño del paisaje.
AREA_MN	Valor medio del área	ha	AREA_MN > 0	Valores medio del área de los parches que componen el paisaje.
GYRATE_MN	Valor medio del radio de giro	m	GYRATE_MN >= 0	El radio de giro es una medida de la extensión del parche a lo largo del paisaje.
CONTAG	Índice de contagio	%	0 < CONTAG <= 100	Tendencia de las clases o los tipos de uso de suelo a estar agregados.
SHDI	Índice de diversidad de Shannon	Adimensional	SHDI >= 0	Índice de diversidad de parches en el paisaje, basado en teoría de la información.
SHEI	Índice de homogeneidad de Shannon	Adimensional	0 <= SHEI <= 1	Es una medida de la distribución del área entre cada parche.
SIDI	Índice de diversidad de Simpson	Adimensional	0 <= SIDI < 1	Representa la probabilidad de que dos pixeles elegidos de forma aleatoria sean de tipos de parches diferentes.
SIEI	Índice de homogeneidad de Simpson	Adimensional	0 <= SIEI <= 1	Es una medida de la distribución del área entre cada parche.

5.1.2. Variables de calidad de agua

A partir del conjunto de datos imputado y aumentado en el OE1, se decidió enfocar este análisis en los contaminantes Nitrógeno total (NT) [mg N/L], Fósforo total (PT) [µg P/L] y Turbidez [NTU]. Esto se justifica a partir del estudio realizado por DINAMA-JICA (JICA-MVOTMA, 2011), lo cual calcula que las cargas de aporte provenientes de las fuentes difusas en la cuenca del río Santa Lucía corresponden a un 82% para NT y 77% para PT; siendo la actividad agrícola-ganadera una de las principales contribuyentes. Además, se considera también la Turbidez como *proxy* de los sedimentos, para evaluar la porción particulada de los nutrientes que estamos estudiando.

La variabilidad anual de cada contaminante se muestra en la Fig. 5.4.

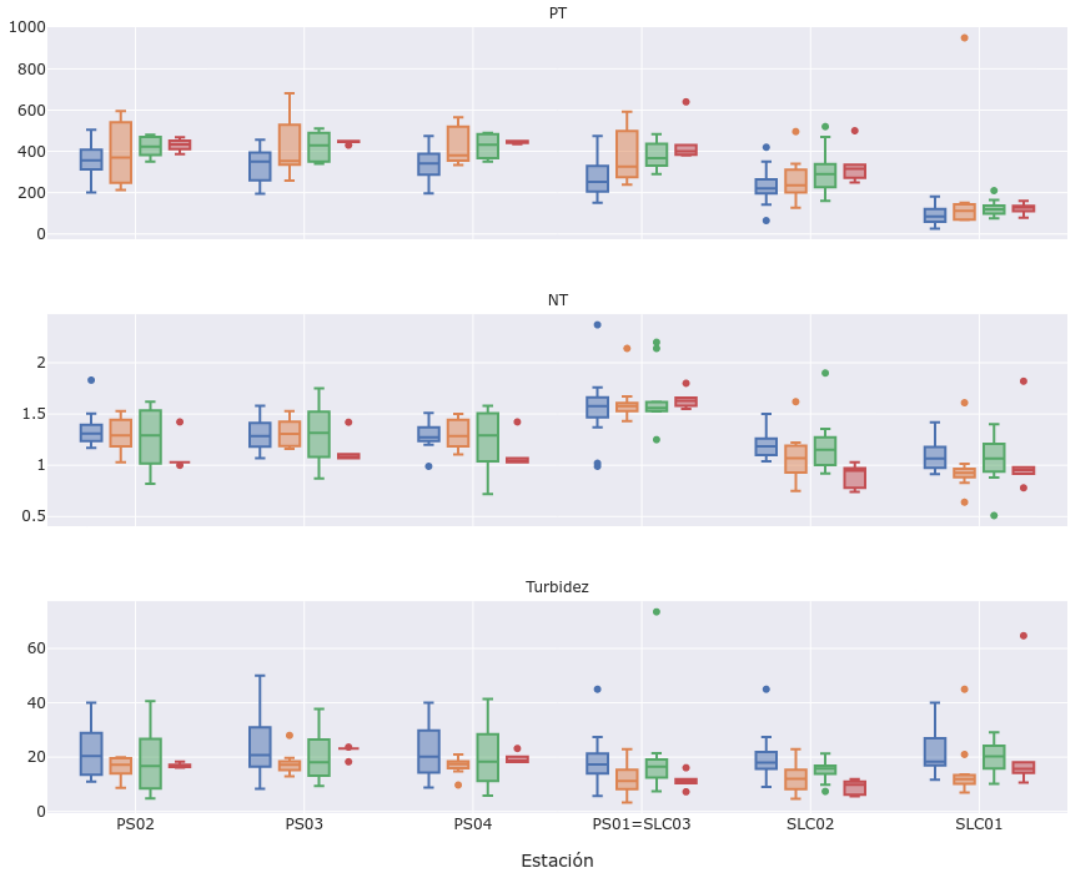


Fig. 5.4 Variabilidad anual de cada contaminante.

Las series temporales que se presentan con frecuencia mensual fueron agregadas de forma que toda la información anual quedara resumida en un conjunto de indicadores estadísticos que pudiese ser relacionado con los índices de uso de suelo descritos anteriormente. Los años usados para el análisis fueron 2014, 2016, 2018 y 2020; la información de cada año se resumió en el valor mínimo, el valor máximo, el promedio, la mediana (Q2) y los cuartiles 25 (Q1) y 75 (Q3) de cada

serie temporal. Se usaron las estaciones de medición SLC01, SLC02, PS01=SLC03 y PS02. A modo de ejemplo, en la Tabla 5.4 se muestran los estadísticos para la variable Nitrógeno total en el año 2014.

Tabla 5.4. Estadísticos calculados para representar la serie temporal de Nitrógeno total en el año 2014.

Año	Estación	Mínimo	Q1	Q2	Promedio	Q3	Máximo
2014	SLC01	0.81	0.91	0.99	1.00	1.03	1.42
	SCL02	1.04	1.10	1.19	1.19	1.21	1.50
	PS01=SLC03	0.99	1.48	1.60	1.58	1.72	2.37
	PS02	1.17	1.24	1.31	1.35	1.39	1.83

5.2. Evaluación de relaciones

Para un estudio en particular se usan 4 observaciones (una por cada año) las cuales se seleccionan como se explica a continuación.

El análisis entre las relaciones de calidad de agua y uso de suelos se realizó en 6 escalas espaciales distintas, correspondientes a 3 subcuencas de la cuenca del SLC (subcuenca 1, subcuenca 3 y subcuenca 4) y sus zonas buffer de 500 m. En todos los casos, las métricas de clase se incorporaron con la condición de que el PLAND de la clase sea mayor al 1%.

Las variables de entrada (X) se crean a partir del conjunto de índices de uso de suelo antes descrito. Para el estudio, se usan tres conjuntos:

- X_{class}: el uso de suelo es descrito por los índices de clase para cada una de las ocho clases en las que se clasifica cada mapa, en total hay 272 variables (34 por cada clase).
- $X_{landscape}$: el uso de suelo es descrito por los índices de paisaje, estos índices corresponden a todo el territorio y no dependen de cada clase, en total hay 38 variables.

En un estudio se pueden incluir en forma conjunta variables de los distintos conjuntos o trabajar con las de uno en particular.

Las variables de salida (Y) se crean a partir del conjunto de variables de calidad de agua antes descrito. Para un estudio, se pueden seleccionar uno o varios estadísticos creados para representar estas variables. En cada caso, se crea un modelo particular para cada contaminante.

Cada estudio incluye varios modelos que se entrenan sobre los subconjuntos de datos elegidos. Dichos modelos son:

- *Partial Least Squares* (PLS): es un método de regresión lineal rápido y eficiente basado en la covarianza. Se recomienda su uso en casos donde el número de variables de entrada es muy alto y donde se supone que las variables de entrada pueden estar correlacionadas.

La idea de la regresión PLS es crear, a partir de un conjunto de m observaciones y n variables de entrada, un conjunto de h componentes que describen la relación entre las variables de entrada y las de salida (Pirouz, 2006).
En este caso, se crean 2 componentes y se muestra el peso asignado de las variables de entrada a cada componente en forma gráfica como se muestra en la Fig. 5.5. También se mide la distancia euclidiana a los puntos (1, 1), (-1, -1), (1,-1) y (-1,1) para obtener de forma rápida aquellas variables a las que se les asigna una mayor importancia, en otras palabras, identificar los índices de uso de suelo que más influencian la variable de calidad de agua bajo estudio; luego del cálculo se ordenan las variables según la menor distancia.

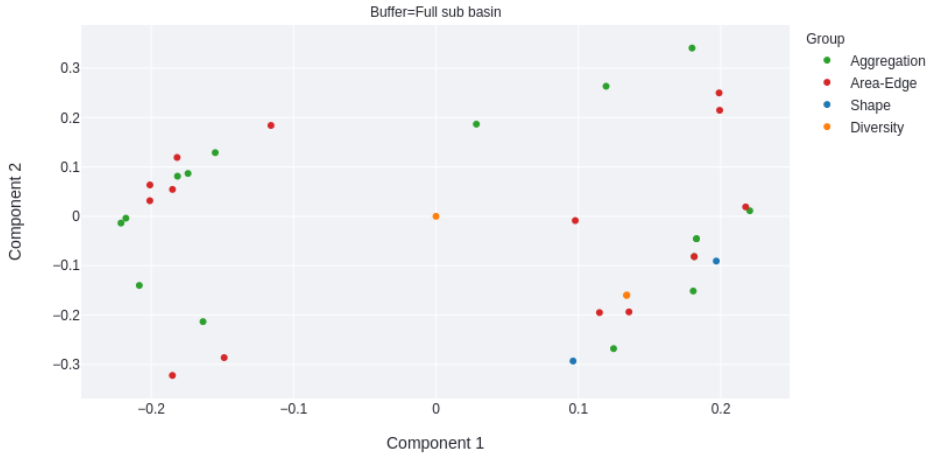


Fig. 5.5 Ejemplo de asignación del peso de cada variable de entrada en cada componente.

- *Random Forest Trees* (RF): es un modelo que entrena un conjunto de árboles de decisión en varios subconjuntos de datos y promedia las predicciones para mejorar la capacidad predictiva y controlar el sobreajuste (Breiman, 2001). En este caso, se espera que el modelo se sobreajuste a los datos y sea capaz de mostrar las relaciones no lineales de las variables.

Para visualizar las relaciones encontradas se usan los valores SHAP.

- *Shapely Additive Explanations* (SHAP): es un algoritmo creado usando la teoría de juego donde se cuantifica la contribución de cada variable de entrada en cada predicción de un modelo. Para esto, se selecciona iterativamente un subgrupo de variables de entrada que se mantiene inalterado y se introduce ruido en el resto, de esta forma se puede medir la influencia que genera la variación de cada entrada en el modelo y su predicción, a este valor de variación se le llama valor *Shapely* (Lundberg et al., 2017).

Para determinar el subconjunto de variables de entrada que mayor importancia tienen en las variables predichas, se suma el valor absoluto de los valores *Shapely* de cada punto para cada variable y se las ordena según magnitud.

Tanto para los modelos PLSR como para los modelos RF se aplica el algoritmo SHAP, los resultados se muestran en forma gráfica de tal forma que se vea al mismo tiempo el valor de salida de cada punto (según el color) como la importancia dada por los valores *Shapely* (el eje X), un ejemplo se muestra en la Fig. 5.6.

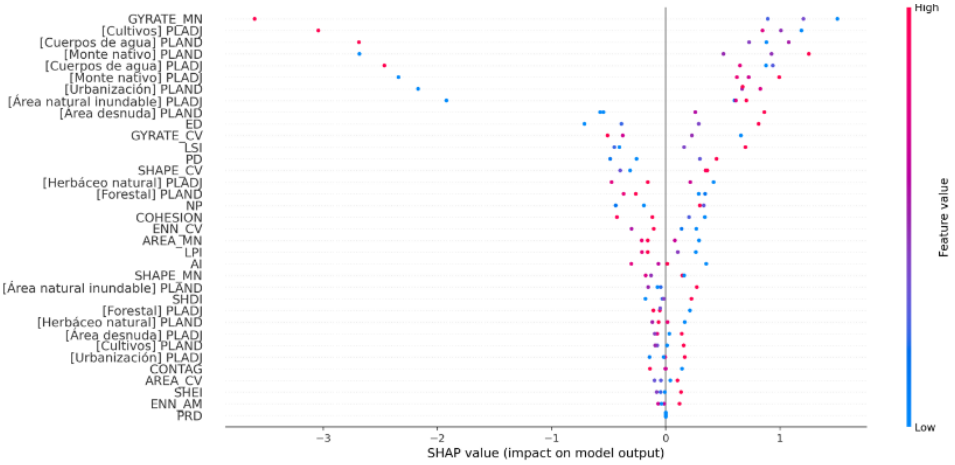


Fig. 5.6 Ejemplo de gráfica de importancia de las variables de entrada según SHAP.

5.3. Análisis de resultados

A continuación, se presenta el análisis de los resultados obtenidos para el modelo PLSR y RF, y los resultados obtenidos del SHAP. Los resultados se presentan por subcuenca y por zona buffer (500 m) para los contaminantes PT, NT y Turbidez. Se eligió analizar la zona buffer de 500 m entre las otras (1000 m, 1500 m, 2000 m) porque es la que más se diferencia del escenario “subcuenca”, sobre todo en caso que dicha subcuenca no es muy grande.

5.3.1. Subcuenca 1 – Cierre: estación SLC01

En la Fig. 5.7, se presenta la evolución temporal de la representación de cada clase para la subcuenca 1, considerando únicamente aquellas clases que verifican PLAND > 1%.

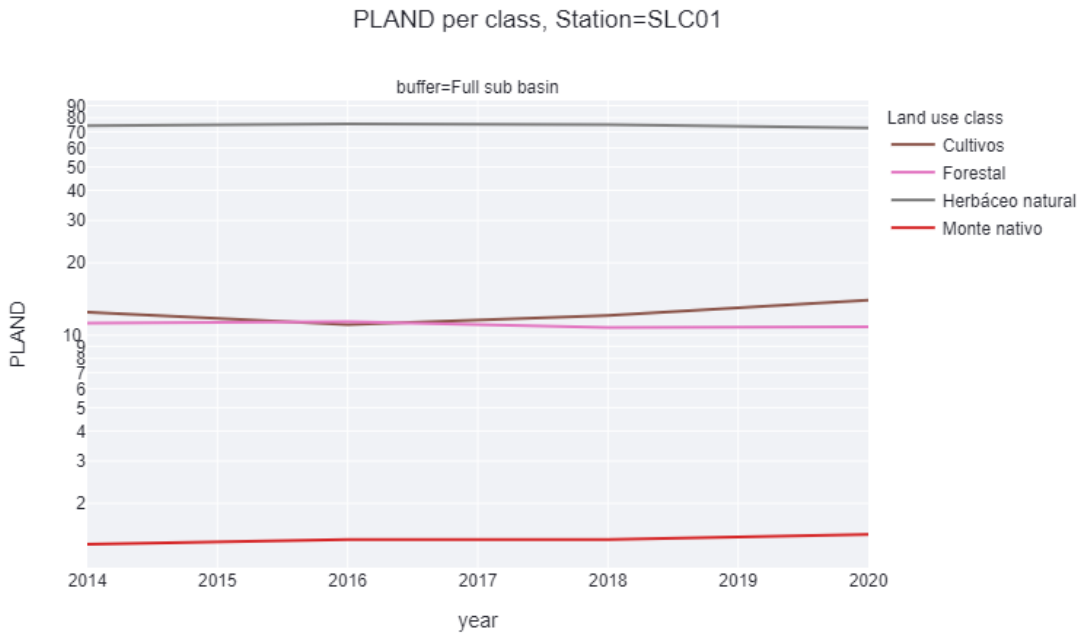


Fig. 5.7. Evolución temporal de la representación de cada clase para la subcuenca 1, para PLAND > 1%.

En la subcuenca 1, solo cuatro usos del suelo ocupan más del 1% de la cuenca (*Cultivos*, *Forestal*, *Herbáceo natural* y *Monte nativo*). En Fig. 5.7 se observa que *Herbáceo natural* es el uso del suelo dominante de la cuenca, seguido por *Forestal* y *Cultivos*. También se observa un leve incremento del área cultivada a partir del 2016.

Fósforo total

La Fig. 5.8 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.5 presenta la varianza explicada por cada componente. La Fig. 5.9 y Fig. 5.10 presentan los valores del SHAP para el modelo PLSR y RF, respectivamente.

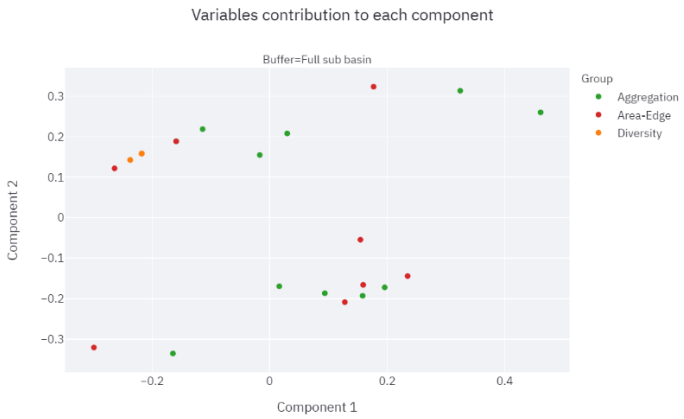


Fig. 5.8. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para PT *subcuenca 1.*

Tabla 5.5. Varianza explicada por cada componente del modelo PLSR para PT *subcuenca 1.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	47.9772	44.1541	7.8688	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	65.0566	29.1821	5.7613	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.30 a 0.46, mientras que en el componente 2 en -0.33 a 0.32. La componente 1 explica el 48.0% de la varianza de las variables de uso de suelo y un 65.1% de la varianza del valor medio del contaminante. La componente 2 explica el 44.2% de la varianza de las variables de uso de suelo y un 29.2% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 92.2% de la varianza de las variables de uso de suelo y un 94.3% de la varianza del valor medio del contaminante.

El valor medio del radio de giro (*GYRATE_MN*) y el índice de agregación para la clase Cultivos (*[Cultivos] AI*) tienen la mayor influencia negativa, mientras que el índice de cohesión y el índice de agregación de la clase Monte Nativo (*[Monte nativo] COHESION y [Monte nativo] AI*) tienen la mayor influencia positiva. El porcentaje de la clase Monte Nativa (*[Monte nativo] PLAND*) tiene el mayor peso positivo sobre la componente 2 y se agrupa cerca de *[Monte nativo] COHESION y [Monte nativo] AI*.

Además, se observa que el porcentaje de la clase Cultivo (*[Cultivo] PLAND*) tiene un alto peso negativo sobre el componente 1 y se agrupa cerca de los índices de diversidad de Simpson que prácticamente se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*). El porcentaje de la clase Herbáceo Natural (*[Herbáceo natural] PLAND)* y se agrupa cerca del índice de Contagio (*CONTAG*).

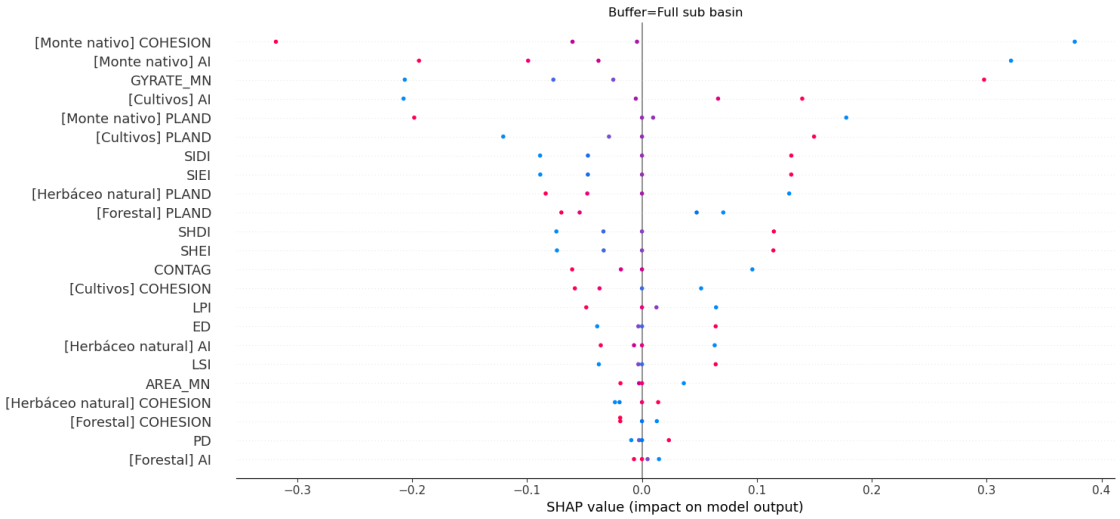


Fig. 5.9. Valores del SHAP para el modelo PLSR para PT *subcuenca 1.*

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.35 a 0.40. El índice *[Monte nativo] COHESION* es el índice más sensible a la salida del modelo, valores bajos generan impactos positivos altos en la salida del modelo, mientras que valores altos generan impactos negativos altos en la salida del modelo. Valores bajos de *[Monte nativo] AI* y altos de *GYRATE_MN* generan impactos positivos altos similares en la salida del modelo. Valores altos de *[Monte nativo] AI* y *[Cultivos] PLAND*, y valores bajos de *GYRATE_MN y [Cultivos] AI* generan impactos negativos similares en la salida del modelo. El valor medio de radio de giro *GYRATE_MN* genera impactos altos en la salida del modelo, mientras que valores bajos generan impactos negativos similares a los generados por valores bajos de *[Cultivos] AI* y valores altos de *[Monte Nativo] PLAND*. Valores bajos de *[Cultivos] PLAND* genera un impacto positivo alto en la salida del modelo.

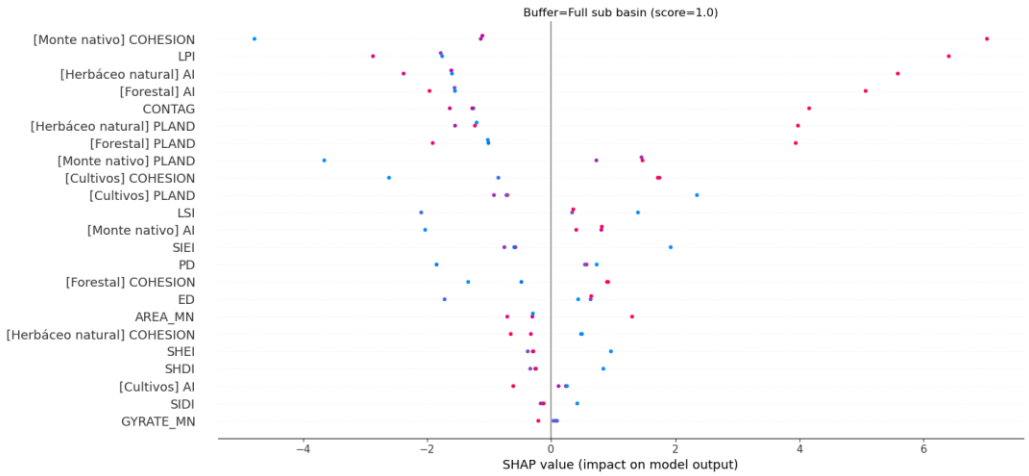


Fig. 5.10. Valores del SHAP para el modelo RF para PT *subcuenca 1.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -5 a 7. Al igual que en el modelo de PLSR el índice *[Monte nativo] COHESION* es el índice más sensible a la salida del modelo, pero con valores bajos generando impactos negativos altos y valores altos generando impactos positivos altos. Valores *[Monte nativo] PLAND* genera impactos negativos altos en la salida del modelo.

Para los índices *LPI*, *[Herbáceo natural] AI, [Forestal] AI, CONTAG y [Herbáceo natural] PLAND* se observa que valores bajos y medios generan impactos negativos en la salida del modelo, mientras pequeñas variaciones en valores altos generan grandes impactos en la salida del modelo. Estas observaciones sugieren comportamiento no-lineal de las variables en el modelo.

Nitrógeno total

La Fig. 5.11 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.6 presenta la varianza explicada por cada componente. La Fig. 5.12 y Fig. 5.13 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

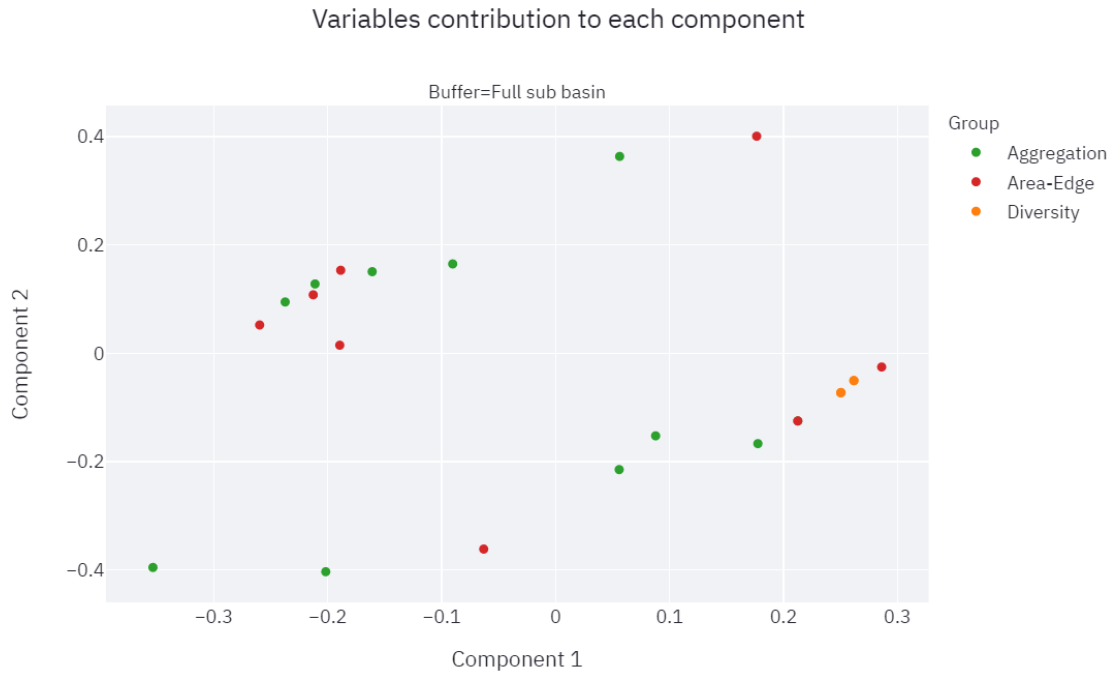


Fig. 5.11. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para NT *subcuenca 1.*

Tabla 5.6. Varianza explicada por cada componente del modelo PLSR para NT *subcuenca 1.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	65.0366	27.1122	7.8512	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	56.4162	38.3677	5.2161	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.35 a 0.28, mientras que en el componente 2 en -0.40 a 0.40. La componente 1 explica el 65.0% de la varianza de las variables de uso de suelo y un 56.4% de la varianza del valor medio del contaminante. La componente 2 explica el 27.1% de la varianza de las variables de uso de suelo y un 38.4% de la varianza del valor medio del contaminante.

Las dos primeras componentes explican 92.1% de la varianza de las variables de uso de suelo y un 94.8% de la varianza del valor medio del contaminante.

La mayor influencia negativa viene dada por el índice de cohesión y el índice de agregación de la clase Monte Nativo (*[Monte nativo] COHESION y [Monte nativo] AI*), mientras que el valor medio del radio de giro (*GYRATE_MN*) y el índice de agregación para la clase Cultivos (*[Cultivos] AI*) tienen la mayor influencia positiva. El porcentaje de la clase Monte Nativo (*[Monte nativo] PLAND*) tiene un peso alto negativo sobre la componente 2 y se agrupa cerca de *[Monte nativo] COHESION y [Monte nativo] AI*.

Además, se observa que el porcentaje de la clase Cultivo (*[Cultivo] PLAND*) tiene el mayor peso positivo sobre el componente 1 y se agrupa cerca de los índices de diversidad de Simpson que prácticamente se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*). El porcentaje de la clase Herbáceo Natural (*[Herbáceo natural] PLAND)* y se agrupa cerca del índice de Contagio (*CONTAG*).

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.35 a 0.30. El índice *[Monte nativo] COHESION* es el índice más sensible a la salida del modelo, valores bajos generan impactos positivos bajos en la salida del modelo, mientras que valores altos generan impactos positivos altos en la salida del modelo. Valores bajos de *[Monte nativo] AI* y altos de *GYRATE_MN* generan impactos positivos altos similares en la salida del modelo. Valores altos de *[Monte nativo] AI* y *[Cultivos] PLAND*, y valores bajos de *GYRATE_MN y [Cultivos] AI* generan impactos negativos similares en la salida del modelo.

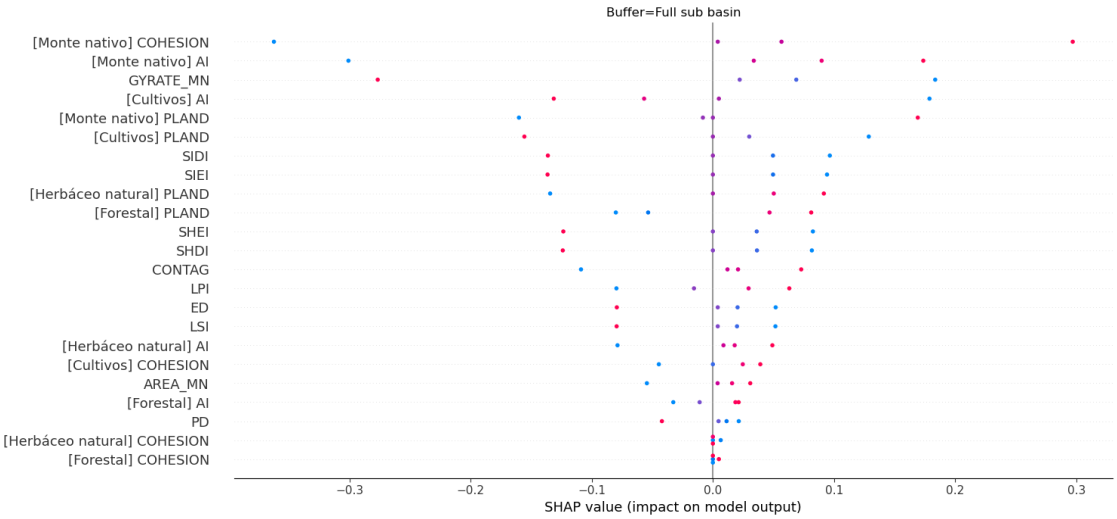


Fig. 5.12. Valores del SHAP para el modelo PLSR para NT *subcuenca 1.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.012 a 0.04. El índice de diversidad SHEI es el más sensible a la salida del modelo, pequeñas variaciones de los valores bajos generan grandes impactos, indicando un comportamiento no lineal. El *[Cultivos] PLAND* presenta un comportamiento similar a SHEI pero de menor impacto en la salida del modelo. Pequeñas variaciones en valores altos de *[Herbáceo natural] AI y LPI,* y pequeñas variaciones en valores bajos de *PD y LSI*, generan variaciones amplias en la salida del modelo.

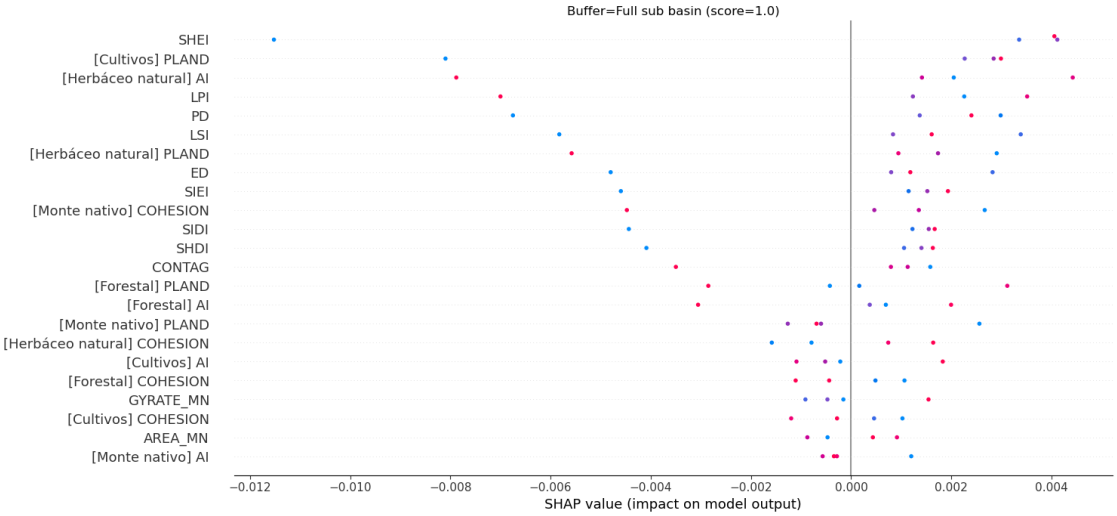


Fig. 5.13. Valores del SHAP para el modelo RF para NT *subcuenca 1.*

Turbidez

La Fig. 5.14 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.7 presenta la varianza explicada por cada componente. La Fig. 5.15 y Fig. 5.16 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

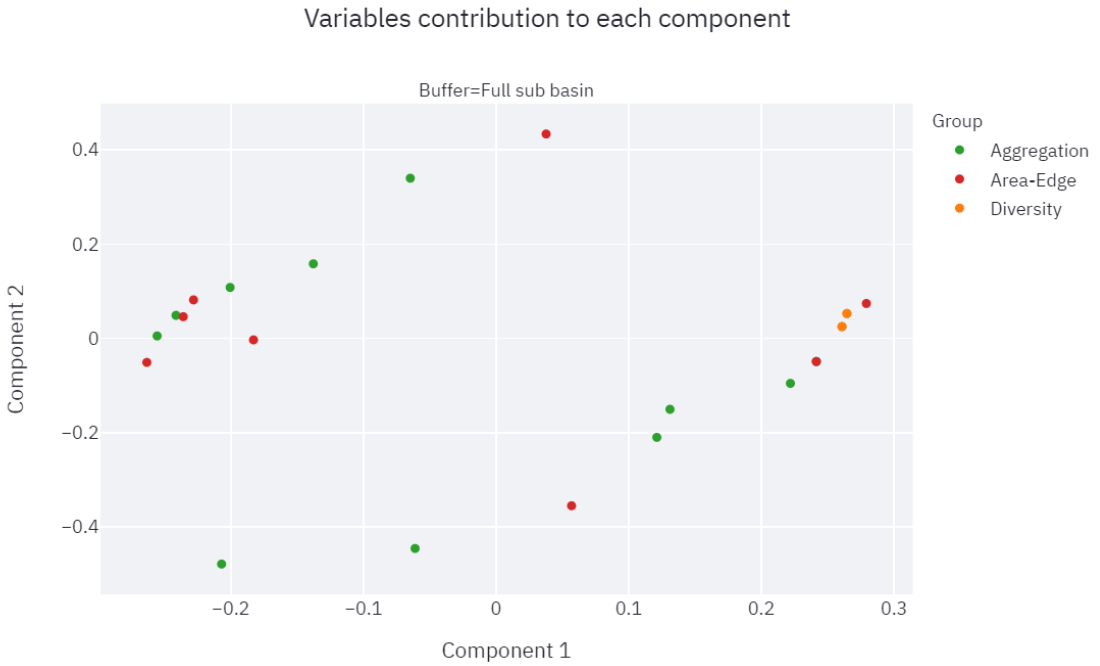


Fig. 5.14. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para Turbidez *subcuenca 1.*

Tabla 5.7. Varianza explicada por cada componente del modelo PLSR para Turbidez *subcuenca 1.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	72.0446	20.3722	7.5833	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	71.0671	26.7370	2.1959	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.26 a 0.28, mientras que en el componente 2 en -0.47 a 0.46. La componente 1 explica el 72.0% de la varianza de las variables de uso de suelo y un 71.1% de la varianza del valor medio del contaminante. La componente 2 explica el 20.4% de la varianza de las variables de uso de suelo y un 26.7% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 92.4% de la

varianza de las variables de uso de suelo y un 97.8% de la varianza del valor medio del contaminante.

La mayor influencia negativa viene dada por el índice de cohesión de la clase Monte Nativo (*[Monte nativo] COHESION).* El *GYRATE_MN* tiene el peso positivo más alto sobre la componente 2, mientras que *[Monte nativo] AI* tiene un peso negativo importante sobre esta componente. El porcentaje de la clase Cultivo (*[Cultivo] PLAND*) tiene el mayor peso positivo sobre el componente 1 y se agrupa cerca de los índices de diversidad de Simpson que prácticamente se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*). El porcentaje de la clase Herbáceo Natural (*[Herbáceo natural] PLAND)* tiene el mayor peso negativo sobre la componente 1 y se agrupa cerca del índice de Contagio (*CONTAG*).

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.30 a 0.25. El índice *[Monte nativo] COHESION* es el índice más sensible a la salida del modelo, valores bajos generan impactos negativos altos en la salida del modelo, mientras que valores altos generan impactos positivos altos en la salida del modelo. Valores bajos de *[Monte nativo] AI* y altos de *GYRATE_MN* generan impactos negativos altos similares en la salida del modelo. Valores altos de *[Monte nativo] AI* y valores bajos de *GYRATE_MN y [Cultivos] PLAND* generan impactos positivos similares en la salida del modelo.

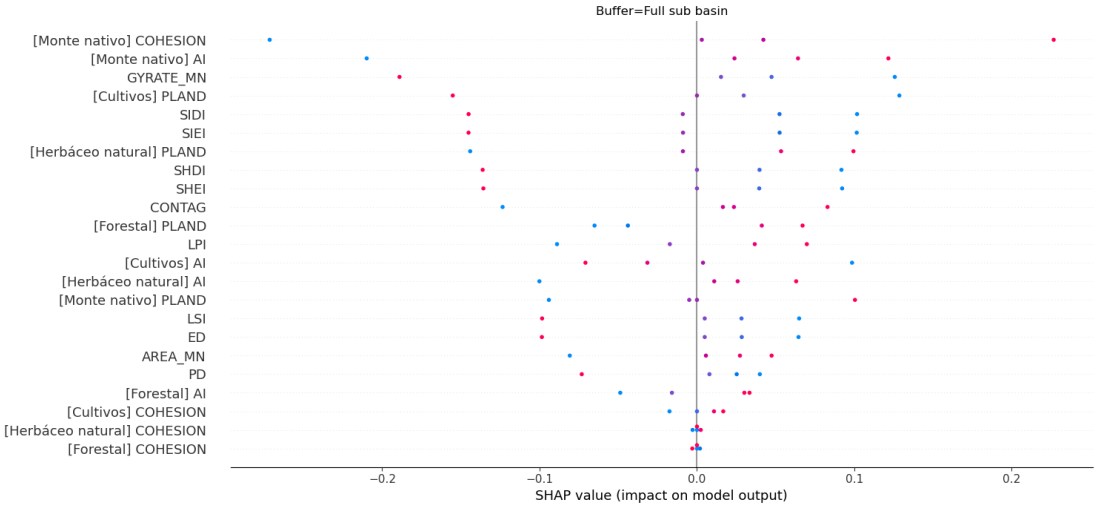


Fig. 5.15. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 1.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.7 a 0.4. El índice de diversidad SHEI es el más sensible a la salida del modelo, pequeñas variaciones de los valores bajos generan grandes impactos, los valores bajos generan impactos negativos altos y los valores altos generan impactos positivos altos. El *[Cultivos] PLAND* presenta un comportamiento similar a *SHEI* pero de menor impacto en la salida del modelo.

Pequeñas variaciones en valores altos de *[Herbáceo natural] AI y LPI,* y pequeñas variaciones en valores bajos de *LSI*, generan variaciones amplias en la salida del modelo. Los valores más altos de *[Herbáceo natural] AI y LPI,* y más bajos de *LSI* generan impactos negativos.

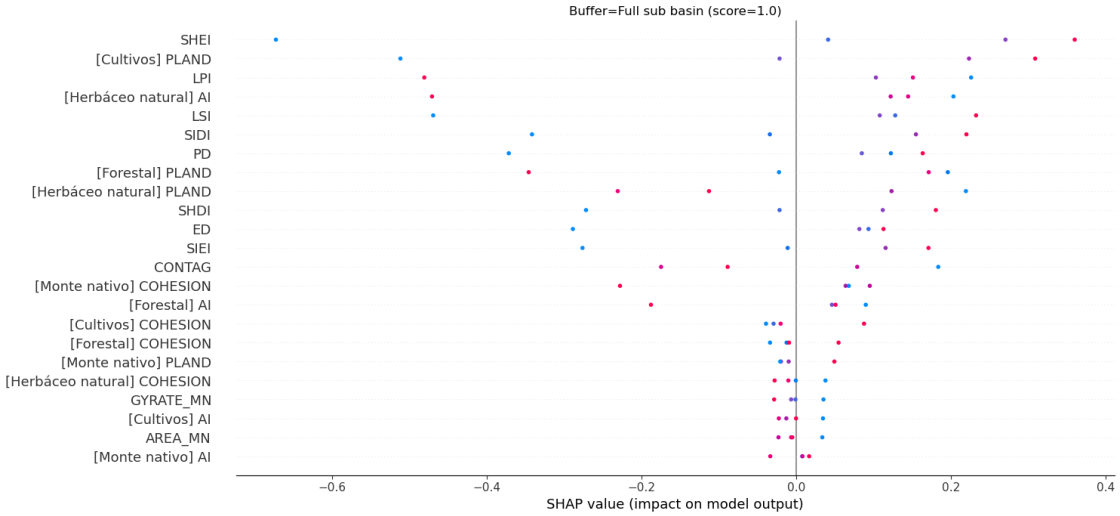


Fig. 5.16. Valores del SHAP para el modelo RF para Turbidez *subcuenca 1.*

5.3.2. Subcuenca 1 zona buffer 500 – Cierre: estación SLC01

En la Fig. 5.17 se presenta la evolución temporal de la representación de cada clase para la subcuenca 1 zona buffer 500 m, considerando únicamente aquellas clases de verifican PLAND > 1%.

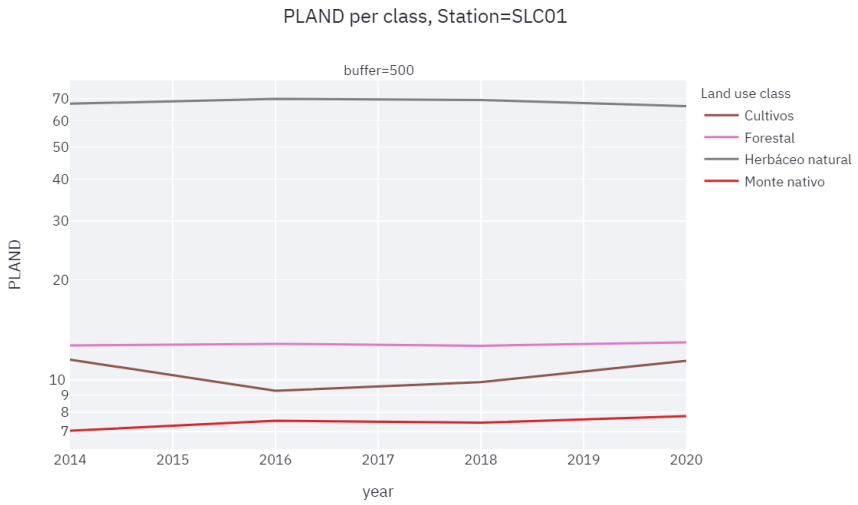


Fig. 5.17. Evolución temporal de la representación de cada clase para la subcuenca 1 zona buffer 500 m, para PLAND > 1%.

En la zona buffer de la subcuenca 1, solo cuatro usos del suelo ocupan más del 1% de la cuenca (*Cultivos*, *Forestal*, *Herbáceo natural* y *Monte nativo*). Estos son los mismos del escenario

"subcuenca 1". En Fig. 5.17 se observa que *Herbáceo natural* es el uso del suelo dominante de la zona buffer, seguido por *Forestal* y *Cultivos*. El 2016 representa un punto de quiebre para casi todas las tendencias, en cuanto el *Herbáceo natural* muestra un muy leve disminución, *Cultivos* empieza a aumentar así como *Monte Nativo*.

Fósforo total

La Fig. 5.18 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.8 presenta la varianza explicada por cada componente. La Fig. 5.19 y Fig. 5.20 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

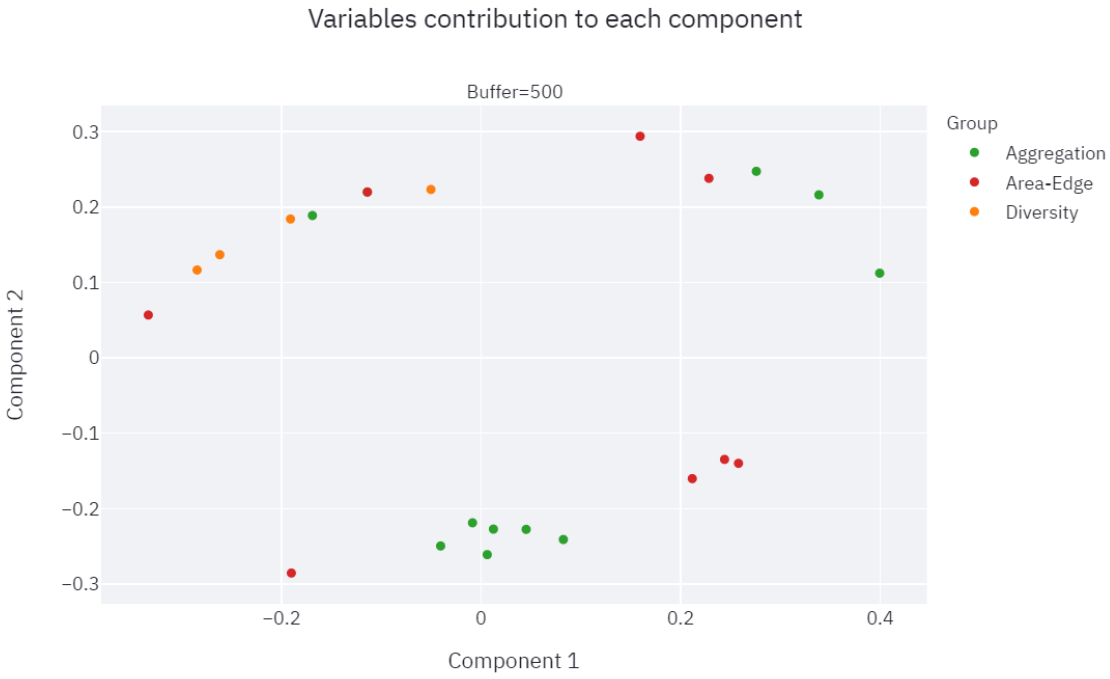


Fig. 5.18. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para PT *subcuenca 1 zona buffer 500 m.*

Tabla 5.8. Varianza explicada por cada componente del modelo PLSR para PT *subcuenca 1 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	39.0011	56.9425	4.0564	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	85.7838	12.6671	1.5491	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.33 a 0.40, mientras que en el componente 2 en -0.28 a 0.29. La componente 1 explica el 39.0% de la varianza de las variables de uso de suelo y un 85.8% de la varianza del valor medio del contaminante. La componente 2 explica el 56.9% de la varianza de las variables de uso de suelo y un 12.7% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 95.9% de la varianza de las variables de uso de suelo y un 98.5% de la varianza del valor medio del contaminante.

[Herbáceo natural] COHESION, *[Monte nativo] COHESION*, [*Monte nativo*] *AI*, y [*Monte nativo*] *PLAND*, son las variables de mayor influencia positiva sobre las componentes. El [*Monte nativo*] *PLAND* y *LPI* son los índices con mayor peso negativo sobre la componente 1 y 2, respectivamente.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.30 a 0.30. El índice *[Herbáceo natural] COHESION* es el índice más sensible a la salida del modelo, valores altos generan impactos negativos altos en la salida del modelo, mientras que valores bajos generan impactos positivos altos en la salida del modelo.

Valores altos de *[Monte nativo] COHESION* generan impactos negativos altos. Valores bajos de los índices *[Monte nativo] AI, [Monte nativo] PLAND y [Monte nativo] COHESION* generan impactos positivos altos en la salida del modelo.

Valores bajos de *[Cultivos] PLAND* generan impactos negativos en la salida del modelo, mientras que valores altos generan impactos altos en la salida del modelo. Impactos similares se observan para el valor de *LPI* pero con inversión de valores de las variables.

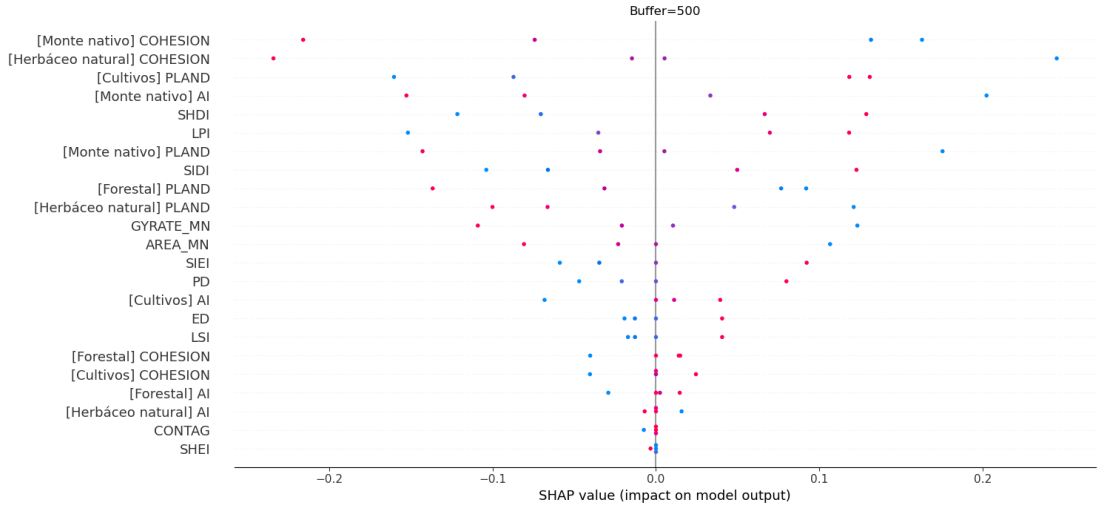


Fig. 5.19. Valores del SHAP para el modelo PLSR para PT *subcuenca 1 zona buffer 500 m.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -6.0 a 8.5. El *GYRATE_MN*, *[Monte nativo] COHESION, [Herbáceo natural] COHESION* y *CONTAG* son más sensible a la salida del modelo, pequeñas variaciones de los valores altos generan grandes impactos, los valores altos generan impactos positivos altos en la salida del modelo. Los valores bajos de *[Herbáceo natural] COHESION* y *[Monte nativo] PLAND* generan los mayores impactos negativos sobre la salida del modelo.

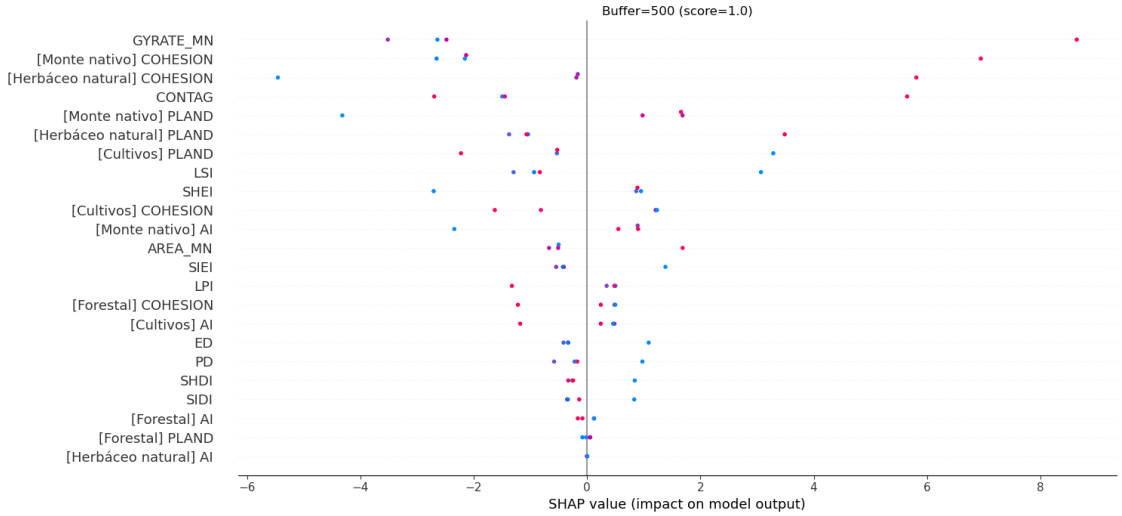


Fig. 5.20. Valores del SHAP para el modelo RF para PT *subcuenca 1 zona buffer 500 m*

Nitrógeno total

La Fig. 5.21 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.9 presenta la varianza explicada por cada componente. La Fig. 5.22 y Fig. 5.23 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

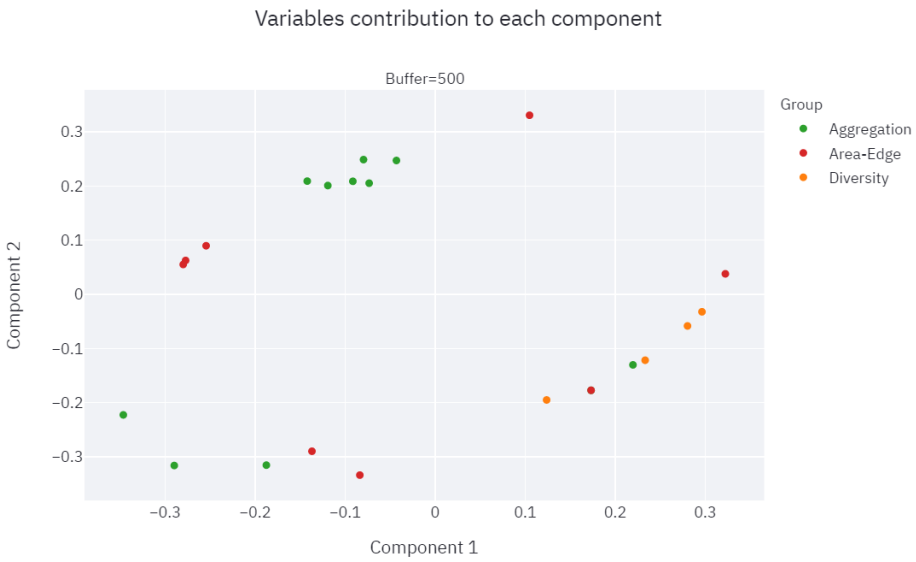


Fig. 5.21. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para NT *subcuenca 1 zona buffer 500 m.*

Tabla 5.9. Varianza explicada por cada componente del modelo PLSR para NT *subcuenca 1 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	55.0077	40.9004	4.0919	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	73.9619	22.8482	3.1899	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.34 a 0.32, mientras que en el componente 2 en -0.33 a 0.33. La componente 1 explica el 55.0% de la varianza de las variables de uso de suelo y un 74.0% de la varianza del valor medio del contaminante. La componente 2 explica el 40.9% de la varianza de las variables de uso de suelo y un 22.8% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 95.9% de la varianza de las variables de uso de suelo y un 96.8% de la varianza del valor medio del contaminante.

[Herbáceo natural] COHESION, *[Monte nativo] COHESION*, [*Monte nativo*] *AI*, y [*Monte nativo*] *PLAND*, son las variables de mayor influencia negativa sobre las componentes. El [*Cultivos*] *PLAND y [Herbáceo natural] COHESION* son los índices con mayor peso positivo y negativo sobre la

componente 1, respectivamente. El *LPI y* [*Cultivos*] *PLAND* son los índices con mayor peso positivo y negativo sobre la componente 2, respectivamente.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.25 a 0.25. El índice *[Herbáceo natural] COHESION* es el índice más sensible a la salida del modelo, valores altos generan impactos positivos altos en la salida del modelo, mientras que valores bajos generan impactos negativos altos en la salida del modelo.

Valores altos de *[Monte nativo] COHESION* generan impactos positivos altos. Valores bajos de los índices *[Monte nativo] AI, [Monte nativo] PLAND y [Monte nativo] COHESION* generan impactos positivos altos en la salida del modelo. Valores bajos de *[Cultivos] PLAND* generan impactos positivos en la salida del modelo, mientras que valores altos generan impactos altos en la salida del modelo.

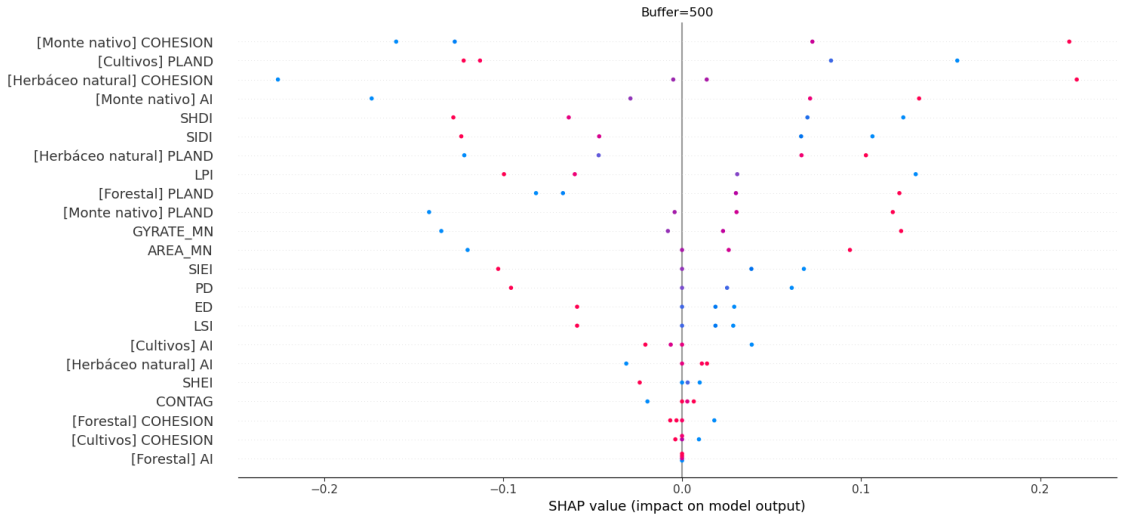


Fig. 5.22. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 1 zona buffer 500 m.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.015 a 0.005. El índice de diversidad SHEI es el más sensible a la salida del modelo, pequeñas variaciones de los valores bajos generan grandes impactos, indicando un comportamiento no lineal. El *LSI* presenta un comportamiento similar a SHEI pero de menor impacto en la salida del modelo. A su vez, GYRATE_MN presenta un comportamiento similar en cuanto a impacto en la salida del modelo, pero invirtiendo la relación de valores de las variables.

Pequeñas variaciones en valores bajos de *[Cultivos] PLAND y* generan variaciones amplias en la salida del modelo, con valores bajos generando impactos negativos, y valores altos generando impactos positivos. Valores bajos de *[Herbáceo natural] COHESION* generan impactos positivos altos en la salida del modelo, mientras que valores altos generan impactos negativos.

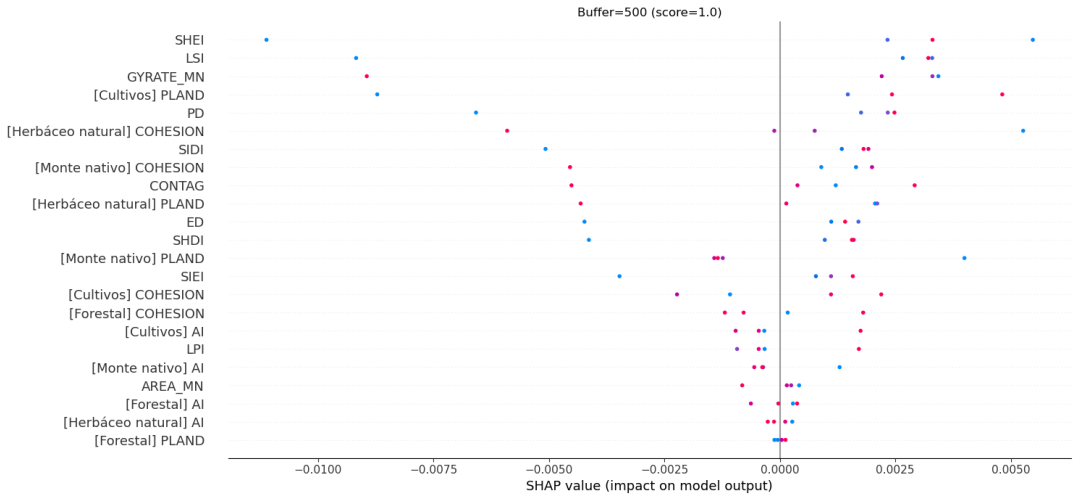


Fig. 5.23. Valores del SHAP para el modelo RF para Turbidez *subcuenca 1 zona buffer 500 m.*

Turbidez

La Fig. 5.24 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.10 presenta la varianza explicada por cada componente. La Fig. 5.25 y Fig. 5.26 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

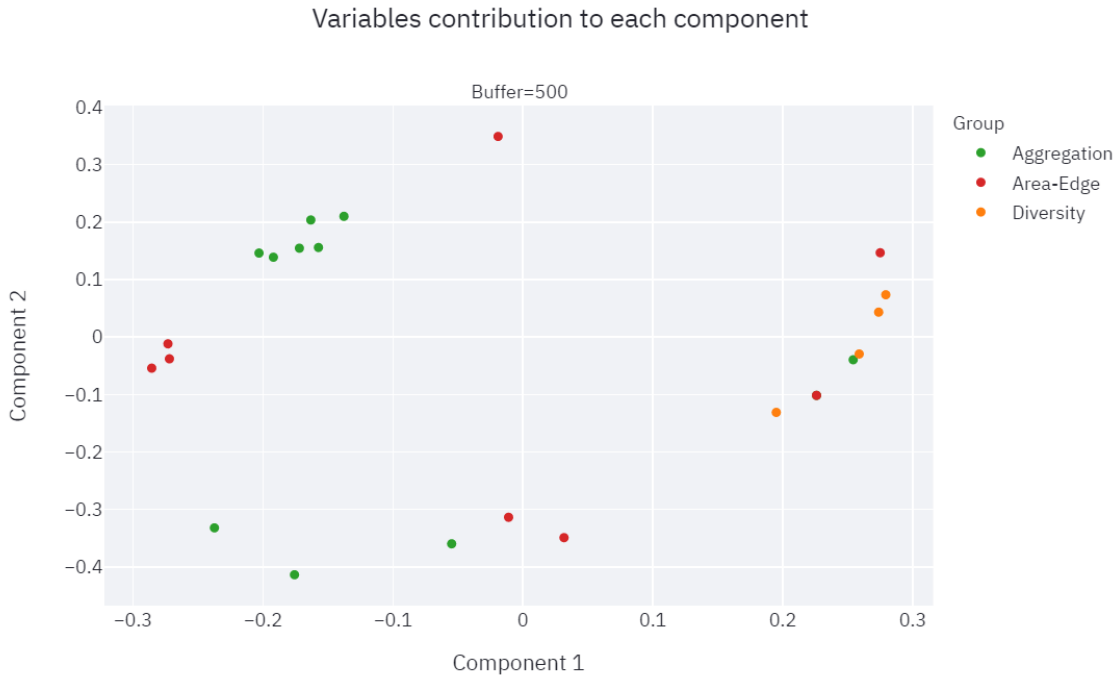


Fig. 5.24. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para Turbidez *subcuenca 1 zona buffer 500 m.*

Tabla 5.10. Varianza explicada por cada componente del modelo PLSR para Turbidez *subcuenca 1 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	66.7269	29.1398	4.1334	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	77.6421	19.2033	3.1546	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.28 a 0.28, mientras que en el componente 2 en -0.41 a 0.34. La componente 1 explica el 66.7% de la varianza de las variables de uso de suelo y un 77.6% de la varianza del valor medio del contaminante. La componente 2 explica el 29.1% de la varianza de las variables de uso de suelo y un 19.2% de la varianza del valor medio del contaminante.

Las dos primeras componentes explican 95.8% de la varianza de las variables de uso de suelo y un 96.8% de la varianza del valor medio del contaminante.

[Herbáceo natural] COHESION y *[Monte nativo] COHESION* son las variables de mayor influencia negativa sobre las componentes. GYRATE_MN, *[Herbáceo natural] PLAND y AREA_MN* se encuentran agrupadas con los mayores pesos negativos sobre la componente 1. El SHDI, SIDI y *[Cultivos] PLAND* se encuentran relativamente cerca y tienen los mayores pesos positivos sobre la componente 1. El *LPI* tiene el mayor peso positivo sobre la componente 2, respectivamente.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.20 a 0.20. El índice *[Herbáceo natural] COHESION* es el índice más sensible a la salida del modelo, valores altos generan impactos positivos altos en la salida del modelo, mientras que valores bajos generan impactos negativos altos en la salida del modelo.

Valores altos de *[Monte nativo] COHESION* generan impactos positivos altos y valores bajos generan impactos negativos. Valores bajos de los índices *[Cultivos] PLAND y SHDI,* generan impactos positivos altos mientras que valores altos generan impactos negativos.

Valores bajos de *GYRATE_MN* generan impactos negativos en la salida del modelo, mientras que valores altos generan impactos positivos en la salida del modelo.

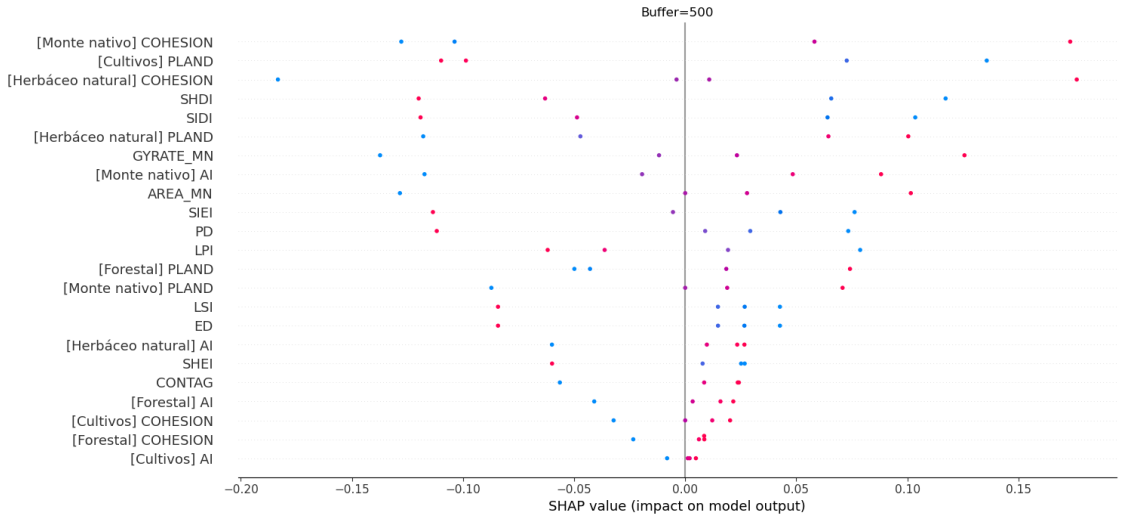


Fig. 5.25. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 1 zona buffer 500 m.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.8 a 0.3. El índice de diversidad SHEI es el más sensible a la salida del modelo, pequeñas variaciones de los valores bajos generan grandes impactos, indicando un comportamiento no lineal. El *LSI* presenta un comportamiento similar a *SHEI* pero de menor impacto en la salida del modelo. Valores altos de estos índices generan impactos positivos altos.

Pequeñas variaciones en valores bajos de *[Cultivos] PLAND, SIDI y PD,* generan variaciones amplias en la salida del modelo, con valores bajos generando impactos negativos, y valores altos generando impactos positivos.

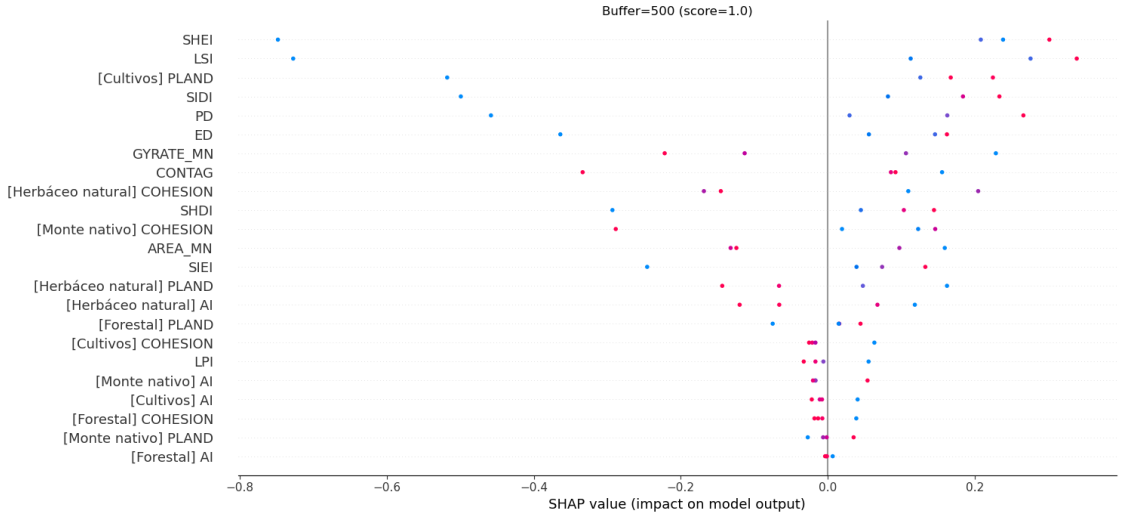


Fig. 5.26. Valores del SHAP para el modelo RF para Turbidez *subcuenca 1 zona buffer 500 m.*

5.3.3. Subcuenca 3 – Cierre: estación PS01=SLC03

En la Fig. 5.27 se presenta la evolución temporal de la representación de cada clase para la subcuenca 3, considerando únicamente aquellas clases de verifican PLAND > 1%.

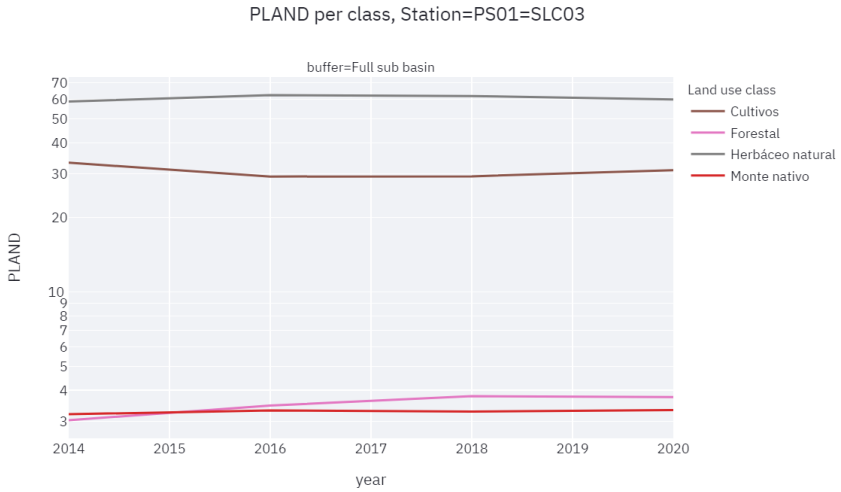


Fig. 5.27. Evolución temporal de la representación de cada clase para la subcuenca 3, para PLAND > 1%.

También en la subcuenca 3, solo cuatro usos del suelo ocupan más del 1% de la cuenca (*Cultivos*, *Forestal*, *Herbáceo natural* y *Monte nativo*). En Fig. 5.27 se observa que *Herbáceo natural* y *Cultivos* son los usos del suelo dominantes de la cuenca. A partir del 2016, también se observa un leve incremento del área cultivada con una consecuente disminución de *Herbáceo natural*.

Fósforo total

La Fig. 5.28 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.11 presenta la varianza explicada por cada componente. La Fig. 5.29 y Fig. 5.30 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

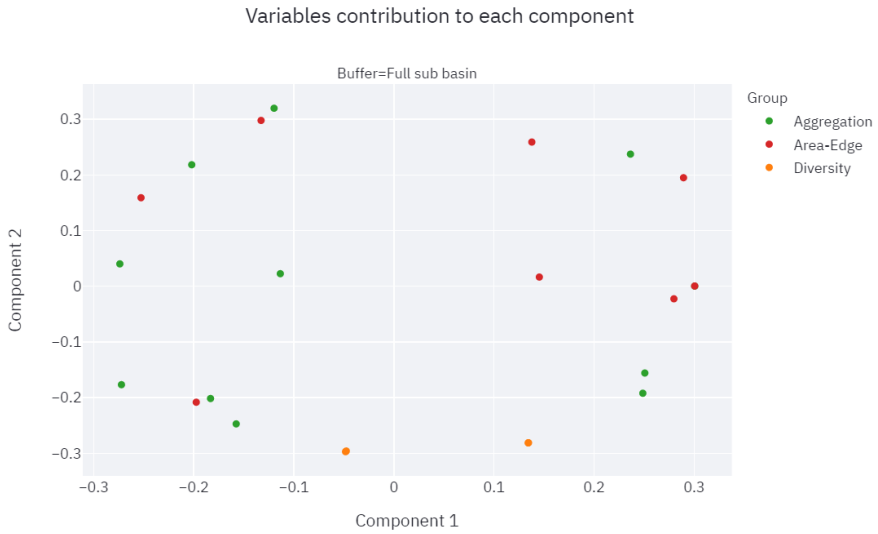


Fig. 5.28. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para PT *subcuenca 3.*

Tabla 5.11. Varianza explicada por cada componente del modelo PLSR para PT *subcuenca 3.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	47.9935	41.2348	10.7717	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	97.9172	1.3867	0.6961	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.33 a 0.33, mientras que en el componente 2 en -0.29 a 0.29. La componente 1 explica el 48.0% de la varianza de las variables de uso de suelo y un 97.9% de la varianza del valor medio del contaminante. La componente 2 explica el 41.2% de la varianza de las variables de uso de suelo y un 1.4% de la varianza del valor medio del contaminante.

Las dos primeras componentes explican 89.2% de la varianza de las variables de uso de suelo y un 99.3% de la varianza del valor medio del contaminante.

Los índices *[Monte nativo] PLAND y [Forestal] AI* tienen una mayor influencia positiva sobe los componentes, mientras que el índice *[Cultivos] AI* tiene la mayor influencia negativa sobre los componentes. Además, se observa que *ED* y *LSI* tienen el mayor peso positivo sobre la componente 1, agrupado cerca de estos índices se encuentra *[Forestal] PLAND*. El *[Herbáceo natural] PLAND* presenta un peso negativo importante sobre la componente 1.

Además, se observa los índices de diversidad de Simpson que prácticamente se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*).

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.15 a 0.20. Valores bajos de *[Monte nativo] PLAND y altos de [Cultivos] AI* generan impactos positivos altos en la salida del modelo. Valores bajos de *[Forestal] PLAND, ED*, *LSI y [Forestal] AI,* generan impactos positivos altos en la salida del modelo. En cambio, valores altos de *ED*, *LSI* y valores bajos de *[Herbáceo natural] AI* generan impactos negativos altos en la salida del modelo.

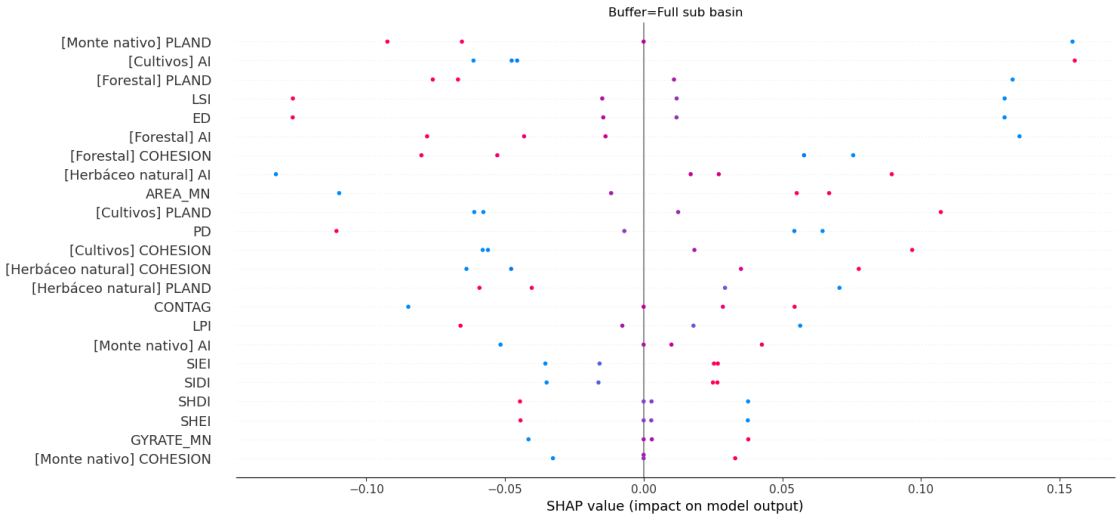


Fig. 5.29. Valores del SHAP para el modelo PLSR para PT *subcuenca 3.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -12.0 a 8.0. El *[Herbáceo natural] AI y PD* son los índices más sensibles a la salida del modelo. Pequeñas variaciones en los valores altos de *[Herbáceo natural] AI* generan grandes impactos, y los valores altos generan impactos negativos altos mientras que los valores bajos generan impactos positivos altos. En el caso de *PD* pequeñas variaciones en de los valores bajos generan grandes impactos, los valores más bajos generan impactos negativos. Los valores altos de *[Cultivos] PLAND* y bajos de *LSI y [Forestal] COHESION* generan impactos negativos en la salida del modelo.

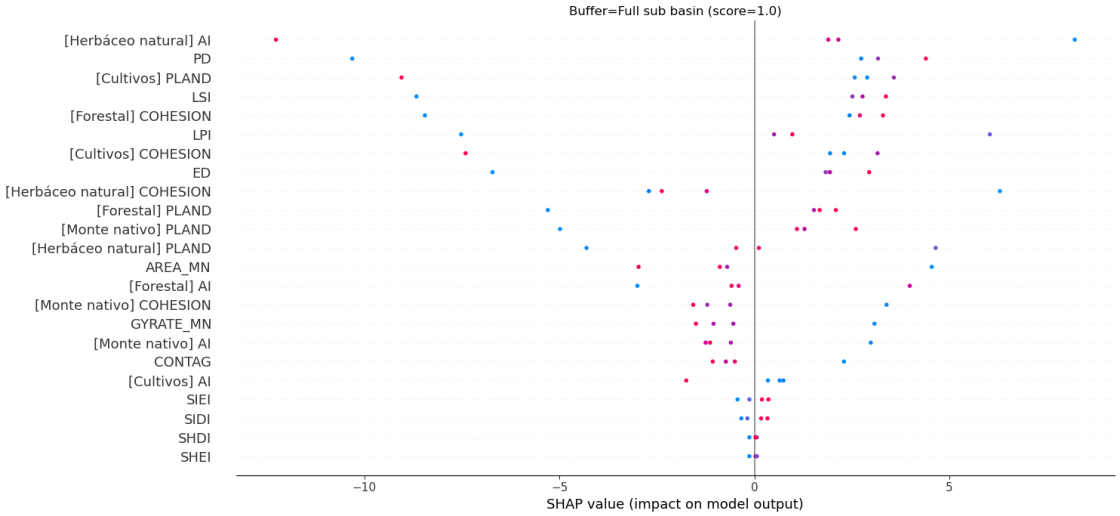


Fig. 5.30. Valores del SHAP para el modelo RF para PT *subcuenca 3.*

Nitrógeno total

La Fig. 5.31 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.12 presenta la varianza explicada por cada componente. La Fig. 5.32 y Fig. 5.33 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

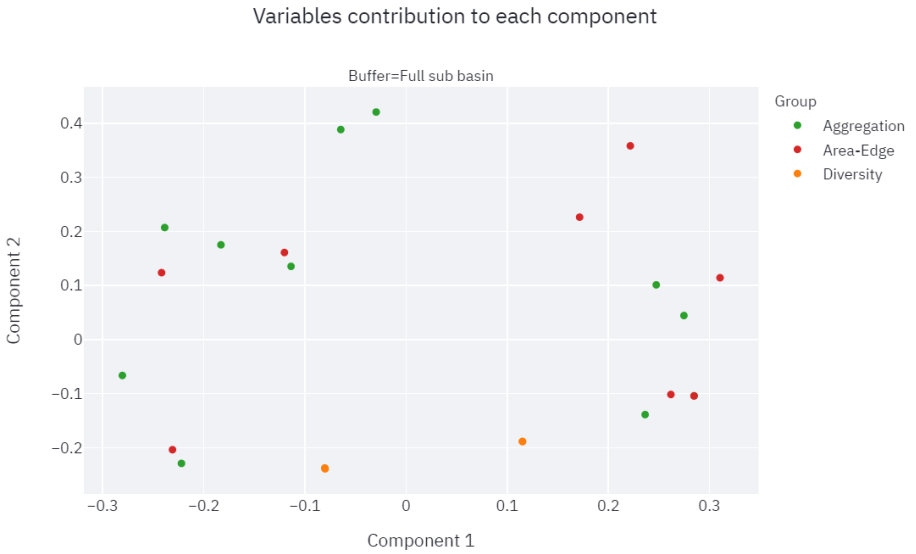


Fig. 5.31. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para NT *subcuenca 3.*

Tabla 5.12. Varianza explicada por cada componente del modelo PLSR para NT *subcuenca 3.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	46.9632	41.4174	11.6195	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	96.8729	1.9160	1.2111	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.28 a 0.31, mientras que en el componente 2 en -0.23 a 0.42. La componente 1 explica el 47.0% de la varianza de las variables de uso de suelo y un 96.9% de la varianza del valor medio del contaminante. La componente 2 explica el 41.4% de la varianza de las variables de uso de suelo y un 1.9% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 88.4% de la varianza de las variables de uso de suelo y un 98.8% de la varianza del valor medio del contaminante.

El índice LPI tienen una mayor influencia positiva sobe los componentes, mientras que los índices *[Cultivos] COHESION y [Cultivos] PLAND* tienen la mayor influencia negativa sobre los componentes. Además, se observa que *[Forestal] PLAND* tiene el mayor peso positivo sobre la componente 1. El *[Cultivos] AI* presenta el mayor peso negativo sobre la componente 1. Además, se observa los índices de diversidad de Simpson que prácticamente se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*).

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.12 a 0.15. Valores bajos de *[Forestal] PLAND y altos de [Cultivos] AI* generan impactos positivos altos en la salida del modelo. Valores altos de *[Cultivos] PLAND y [Cultivos] COHESION* generan impactos positivos altos en la salida del modelo, mientras que valores bajos generan impactos negativos en la salida del modelo. Valores bajos de *[Forestal] AI* generan impactos positivos altos en la salida del modelo. Valores altos de *LPI*, *LSI y ED,* generan impactos negativos altos en la salida del modelo. Valores altos de *[Forestal] PLAND y [Forestal] COHESION* generan impactos negativos en la salida del modelo. Valores bajos de *LPI*, *LSI, ED y [Monte nativo] PLAND* generan impactos positivos altos en la salida del modelo.

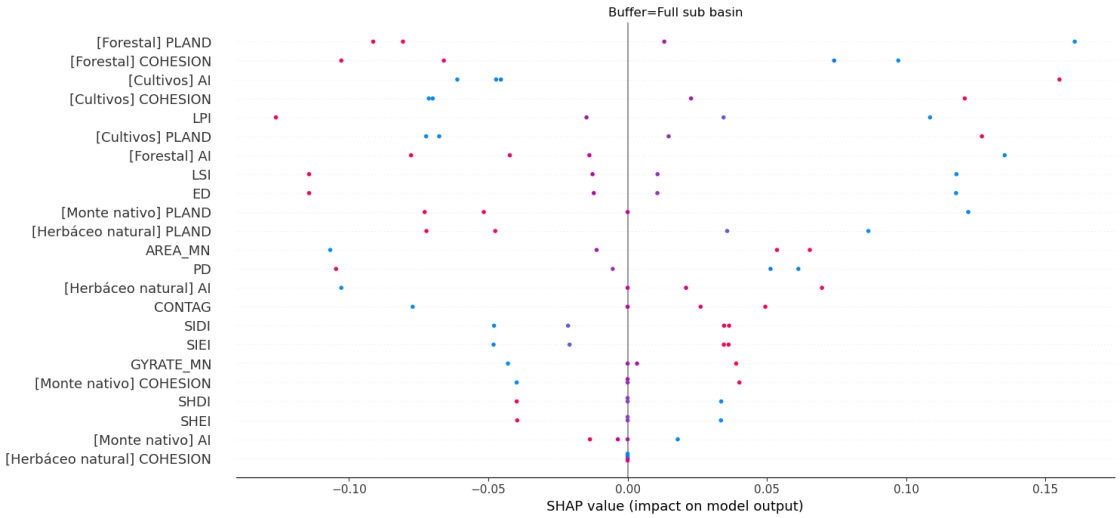


Fig. 5.32. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 3.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.007 a 0.003. El *[Herbáceo natural] AI y PD* son los índices más sensibles a la salida del modelo. Pequeñas variaciones en os valores altos de *[Herbáceo natural] AI* generan grandes impactos, y los valores altos generan impactos negativos altos mientras que los valores bajos generan impactos positivos altos. En el caso de *PD* pequeñas variaciones en los valores bajos generan grandes impactos, los valores más bajos generan impactos negativos altos. Los índices *[Forestal] COHESION, [Forestal] PLAND y LSI* presentan comportamientos similares a *PD* pero de menor impacto.

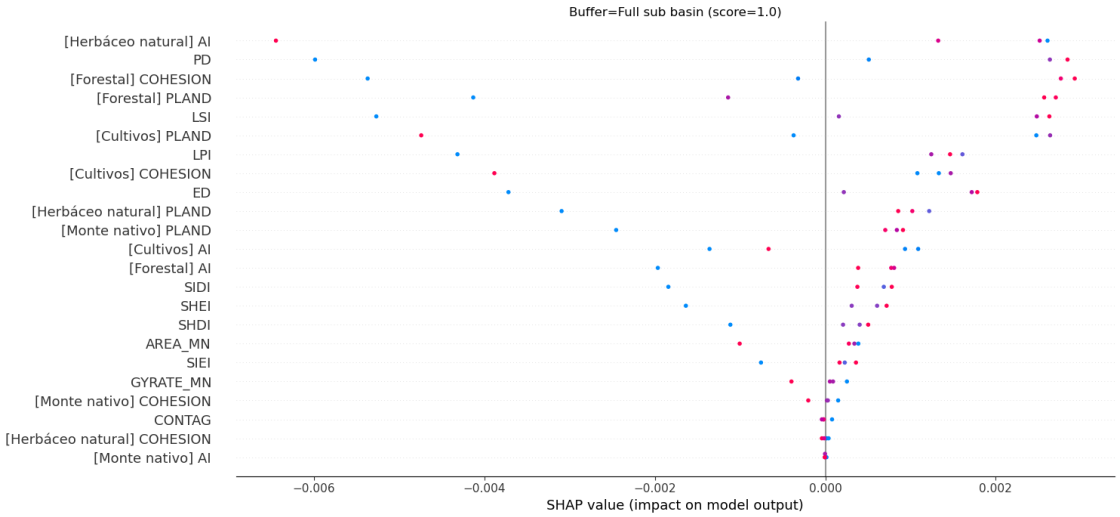


Fig. 5.33. Valores del SHAP para el modelo RF para Turbidez *subcuenca 3.*

Turbidez

La Fig. 5.34 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.13 presenta la varianza explicada por cada componente. La Fig. 5.35 y Fig. 5.36 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

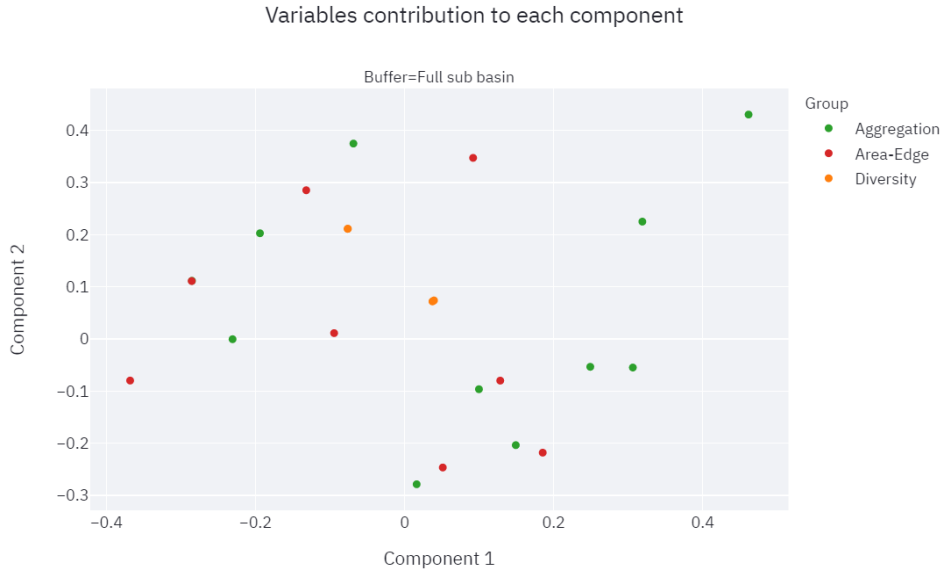


Fig. 5.34. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para Turbidez *subcuenca 3.*

Tabla 5.13. Varianza explicada por cada componente del modelo PLSR para Turbidez *subcuenca 3.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	45.2925	15.8056	38.9018	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	55.7266	41.6272	2.6461	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.37 a 0.46, mientras que en el componente 2 en -0.27 a 0.43. La componente 1 explica el 45.3% de la varianza de las variables de uso de suelo y un 55.7% de la varianza del valor medio del contaminante. La componente 2 explica el 15.8% de la varianza de las variables de uso de suelo y un 41.6% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 61.1% de la varianza de las variables de uso de suelo y un 97.3% de la varianza del valor medio del contaminante.

El *[Herbáceo natural] COHESION* tiene la mayor influencia positiva sobe los componentes, seguida por *[Monte nativo] AI.* El *[Monte nativo] PLAND* tienen el mayor peso negativo sobre la componente 1, agrupado cerca de estos índices se encuentra *[Forestal] PLAND*. El *[Herbáceo natural] PLAND* presenta un peso negativo importante sobre la componente 1. Además, se observa los índices de diversidad de Simpson que prácticamente se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*).

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.50 a 0.45 Valores bajos de *[Herbáceo natural] COHESION y [Monte nativo] AI* generan impactos positivos altos en la salida del modelo. Valores altos de *[Herbáceo natural] COHESION y [Monte nativo] AI* generan impactos negativos altos en la salida del modelo. Valores bajos de *[Monte nativo] PLAND* genera impactos negativos en la salida del modelo, mientas que valores altos genera impactos positivos.

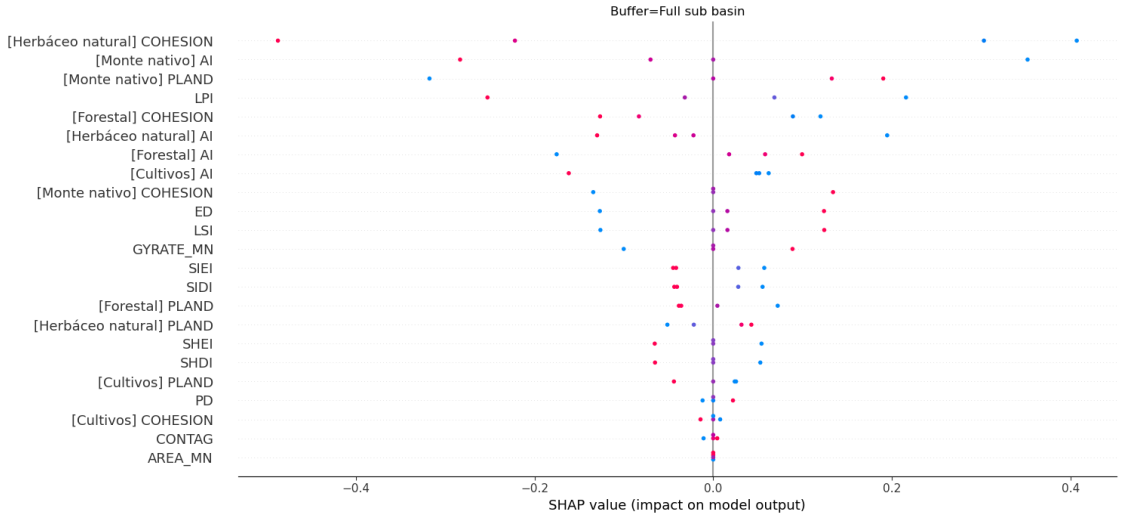


Fig. 5.35. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 3.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -1.0 a 1.0. El *[Herbáceo natural] COHESION y [Monte Nativo] PLAND* son los índices más sensibles a la salida del modelo. Valores altos de *[Herbáceo natural] COHESION* generan impactos positivos altos mientas que valores bajos generan impactos negativos altos. Valores bajos de *[Monte Nativo] PLAND* generan impactos positivos mientas que valores altos generan impactos negativos altos.

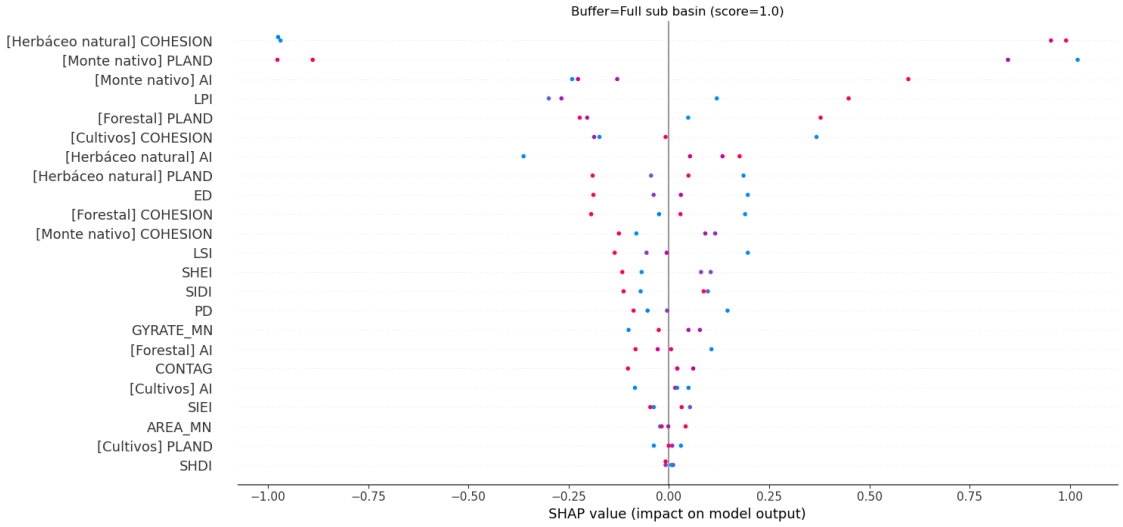


Fig. 5.36. Valores del SHAP para el modelo RF para Turbidez *subcuenca 3.*

5.3.4. Subcuenca 3 zona buffer 500 m – Cierre: estación PL01=SLC03

En la Fig. 5.37, se presenta la evolución temporal de la representación de cada clase para la subcuenca 3 zona buffer 500 m, considerando únicamente aquellas clases de verifican PLAND > 1%.

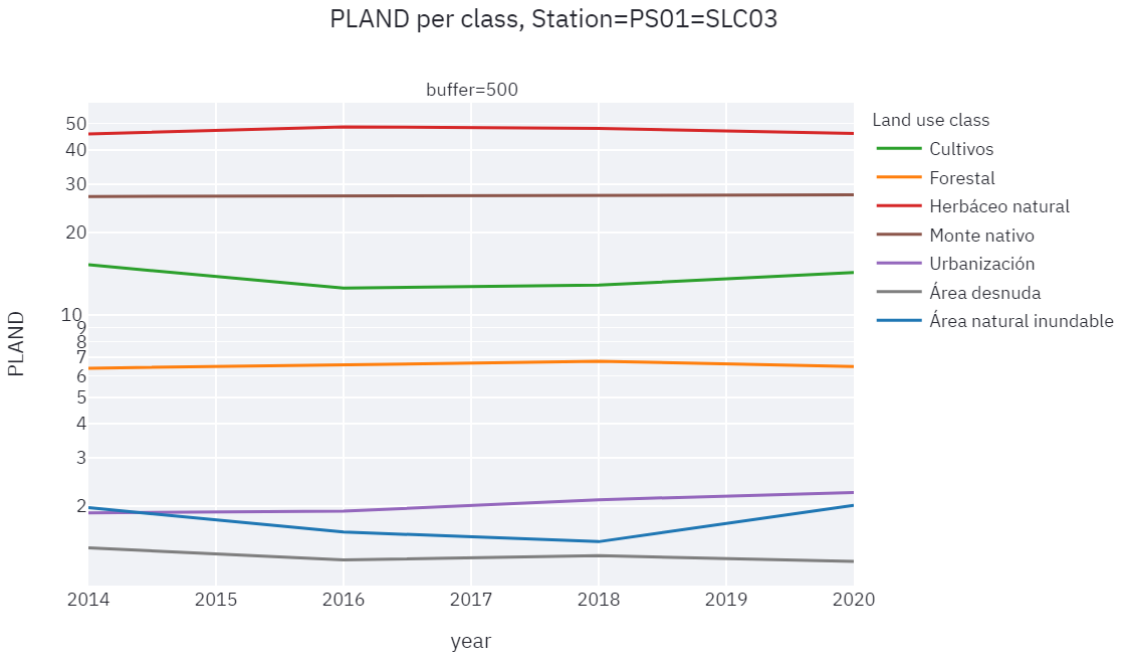


Fig. 5.37. Evolución temporal de la representación de cada clase para la subcuenca 3 zona buffer 500 m, para PLAND > 1%.

En la zona buffer de la subcuenca 3, casi todos los usos del suelo ocupan más del 1% de la cuenca, con excepción del uso *Cuerpo de agua*. En Fig. 5.37 se observa que *Herbáceo natural,* seguido por *Monte nativo* y *Cultivos,* son los usos del suelo dominantes de la zona buffer. A partir del 2016, también se observa un leve incremento del área cultivada con una consecuente disminución de *Herbáceo natural*. En los usos del suelo con menor porcentaje, se observa un incremento de la *Urbanización* a partir del 2016 y un incremento del *Área natural inundable* a partir del 2018.

Fósforo total

La Fig. 5.38 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.14 presenta la varianza explicada por cada componente. La Fig. 5.39 y Fig. 5.40 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

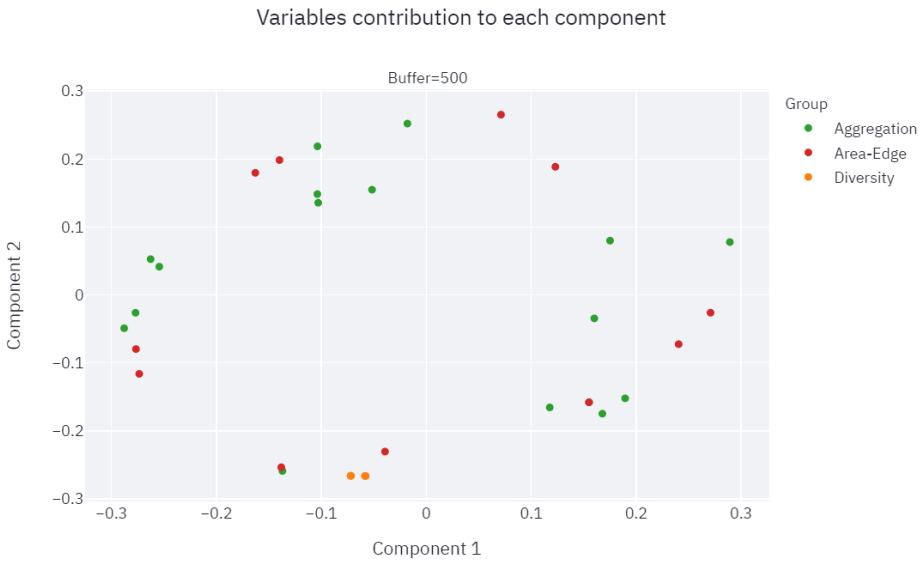


Fig. 5.38. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para PT *subcuenca 3 zona buffer 500 m.*

Tabla 5.14. Varianza explicada por cada componente del modelo PLSR para PT *subcuenca 3 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	41.5389	44.9814	13.4797	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	94.8514	5.0031	0.1455	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.29 a 0.29, mientras que en el componente 2 en -0.29 a 0.29. La componente 1 explica el 41.5% de la varianza de las variables de uso de suelo y un 94.9% de la varianza del valor medio del contaminante. La componente 2 explica el 45.0% de la varianza de las variables de uso de suelo y un 5.0% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 86.5% de la varianza de las variables de uso de suelo y un 99.9% de la varianza del valor medio del contaminante.

Los índices *[Herbáceo natural] COHESION, [Monte nativo] PLAND y [Urbanización] PLAND* tienen los mayores pesos positivos en la componente 1, mientras que el índice *[Forestal] COHESION, [Forestal] AI, LPI, [Área desnuda] PLAND, [Cultivos] COHESION y [Cultivos] AI* tienen los mayores pesos negativos sobre la componente 1. Además, se observa los índices de diversidad se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*), y se agrupan cerca Simpson y Shannon.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.12 a 0.15. Valores altos de *[Forestal] AI, [Área desnuda] PLAND, LPI, [Forestal] COHESION* y bajos de *[Herbáceo natural] COHESION* generan los mayores impactos positivos en la salida del modelo. Valores altos de *[Monte nativo] PLAND* y bajos de *[Cultivos] COHESION y [Cultivos] AI*, generan los mayores impactos negativos en la salida del modelo. Valores bajos de *[Forestal] AI, [Área desnuda] PLAND, LPI y [Forestal] COHESION* generan impactos negativos en la salida del modelo. Valores altos de *[Herbáceo natural] COHESION y [Urbanización] PLAND* generan impactos negativos en la salida del modelo.

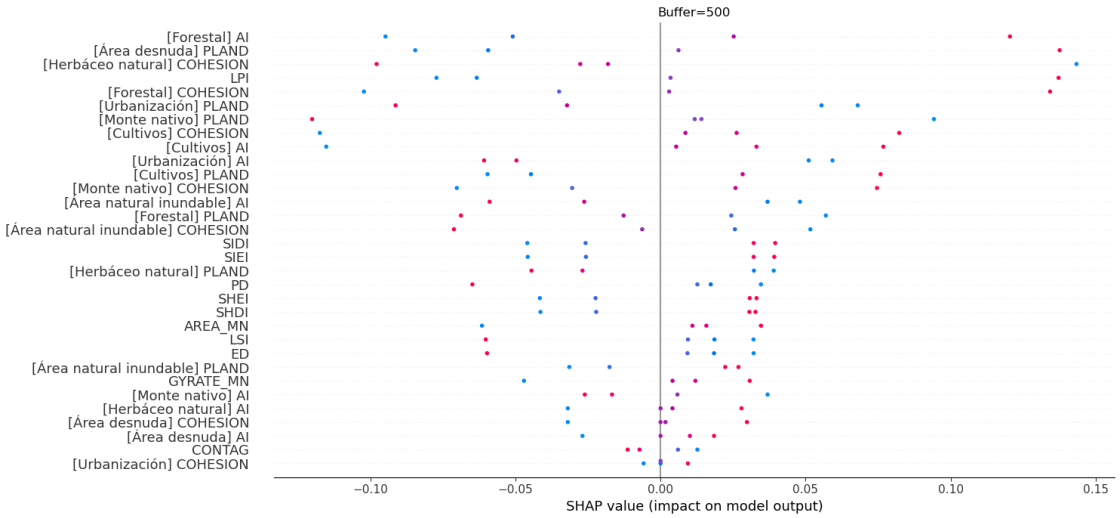


Fig. 5.39. Valores del SHAP para el modelo PLSR para PT *subcuenca 3 zona buffer 500 m.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -10.0 a 7.5. El *LPI* es el índice más sensible a la salida del modelo, pequeñas variaciones en de los valores medios a altos generan grandes impactos, valores altos generan impactos negativos altos mientras que los valores bajos generan impactos positivos altos.

Valores altos de *[Cultivos] COHESION* genera grandes impactos negativos en la salida del modelo, mientras que valores bajos grandes impactos positivos. El *[Forestal] COHESION* presenta un comportamiento en cuento al impacto levemente menor que *[Cultivos] COHESION.* Los índices de diversidad SHDI y SIEI presentan un comportamiento no linean, pequeñas variaciones en valores altos generan un amplio impacto en los valores de salida del modelo. Valores bajos de *[Herbáceo natural] COHESION* generan impactos negativos relativamente altos en la salida del modelo.

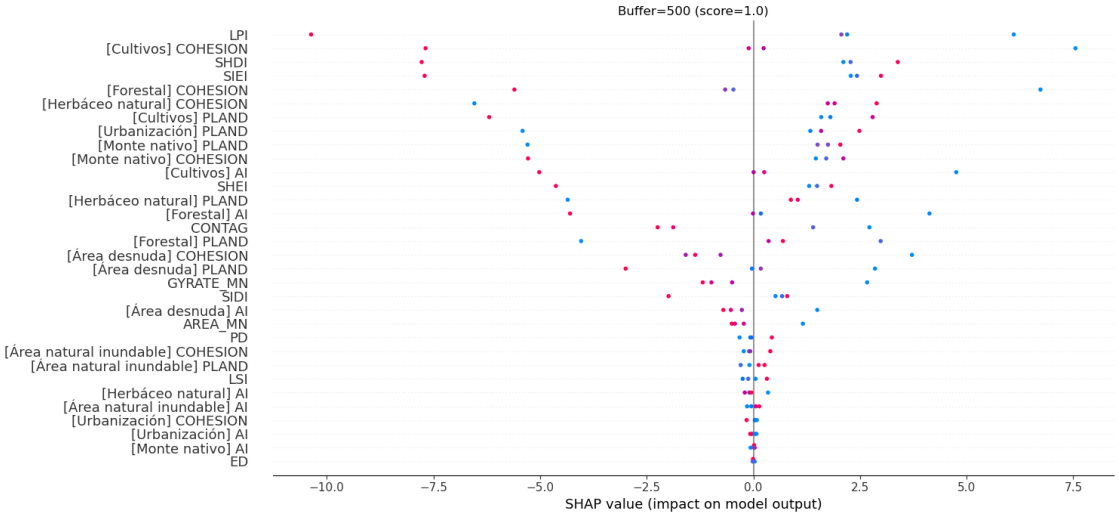


Fig. 5.40. Valores del SHAP para el modelo RF para PT *subcuenca 3 zona buffer 500 m.*

Nitrógeno total

La Fig. 5.41 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.15 presenta la varianza explicada por cada componente. La Fig. 5.42 y Fig. 5.43 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

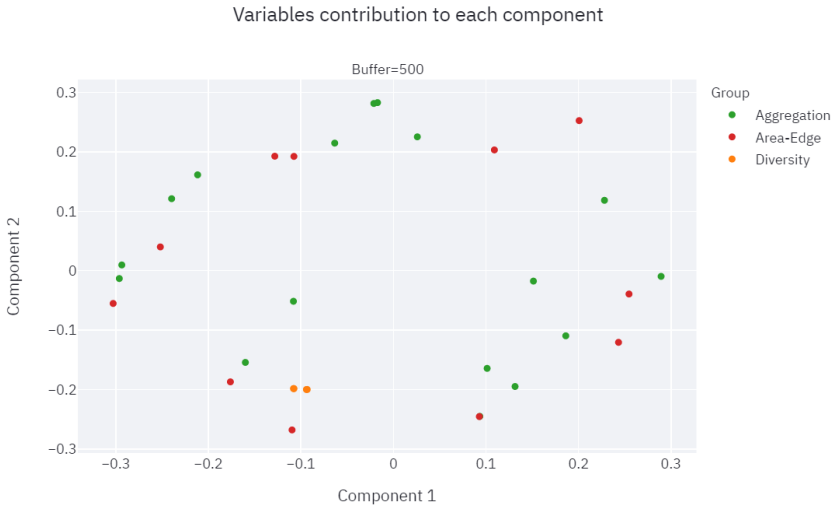


Fig. 5.41. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para NT *subcuenca 3 zona buffer 500 m.*

Tabla 5.15. Varianza explicada por cada componente del modelo PLSR para NT *subcuenca 3 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	37.4082	47.7081	14.8837	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	95.4176	3.7424	0.8400	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.30 a 0.29, mientras que en el componente 2 en -0.27 a 0.28. La componente 1 explica el 31.4% de la varianza de las variables de uso de suelo y un 95.4% de la varianza del valor medio del contaminante. La componente 2 explica el 47.7% de la varianza de las variables de uso de suelo y un 3.7% de la varianza del valor medio del contaminante.

Las dos primeras componentes explican 79.1% de la varianza de las variables de uso de suelo y un 99.1% de la varianza del valor medio del contaminante.

Los índices *[Herbáceo natural] COHESION, [Monte nativo] PLAND, [Urbanización] PLAND y [Área natural inundable] AI* tienen los mayores pesos positivos en la componente 1, mientras que el *LPI, [Forestal] AI, [Forestal] COHESION y [Área desnuda] PLAND* tienen los mayores pesos negativos sobre la componente 1.

Además, se observa los índices de diversidad se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*), y se agrupan cerca Simpson y Shannon.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.12 a 0.15. Valores altos de *[Forestal] AI, [Área desnuda] PLAND, LPI, [Forestal] COHESION* y bajos de *[Herbáceo natural] COHESION* generan los mayores impactos positivos en la salida del modelo. Valores altos de *[Monte nativo] PLAND* y bajos de *[Cultivos] COHESION y [Cultivos] AI*, generan los mayores impactos negativos en la salida del modelo.

Valores altos de *[Forestal] AI, LPI, [Forestal] COHESION y [Área desnuda] PLAND,* y valores bajos de *[Herbáceo natural]* generan impactos positivos altos en la salida del modelo. Valores bajos de *[Herbáceo natural] COHESION y [Urbanización] PLAND* generan impactos negativos en la salida del modelo.

Valores bajos de *[Forestal] AI, [Forestal] COHESION y [Cultivos] COHESION,* generan impactos negativos altos en la salida del modelo. Valores altos de *[Área natural inundable] AI,*

[Urbanización] PLAND, [Herbáceo natural] COHESION, [Forestal] PLAND y [Monte nativo] PLAND generan impactos negativos altos en la salida del modelo.

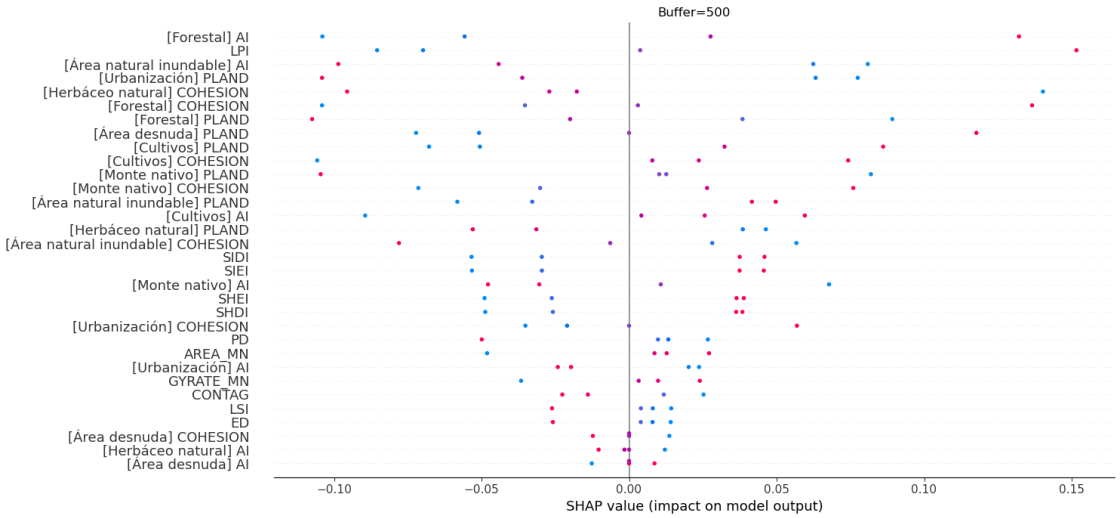


Fig. 5.42. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 3 zona buffer 500 m.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.006 a 0.003. El *LPI* es el índice más sensible a la salida del modelo, pequeñas variaciones en de los valores medios a altos generan grandes impactos en la salida del modelo, valores altos generan impactos negativos altos mientras que los valores bajos generan impactos positivos altos. Valores bajos de *[Urbanización] PLAND y [Herbáceo natural] COHESION*, y valores altos de *[Cultivos] COHESION* generan grandes impactos negativos en la salida del modelo. Valores altos de *[Urbanización] PLAND y [Herbáceo natural] COHESION,* y bajos altos de *[Cultivos] COHESION* generan impactos positivos en la salida del modelo. Los índices de diversidad *SIEI y SHDI* presentan un comportamiento no linean, pequeñas variaciones en valores altos generan un amplio impacto en los valores de salida del modelo. Valores altos de *[Forestal] COHESION y [Cultivos] AI,* generan impactos negativos en la salida del modelo similares a los generados por valores bajos de *[Herbáceo natural] PLAND y [Forestal] PLAND.*

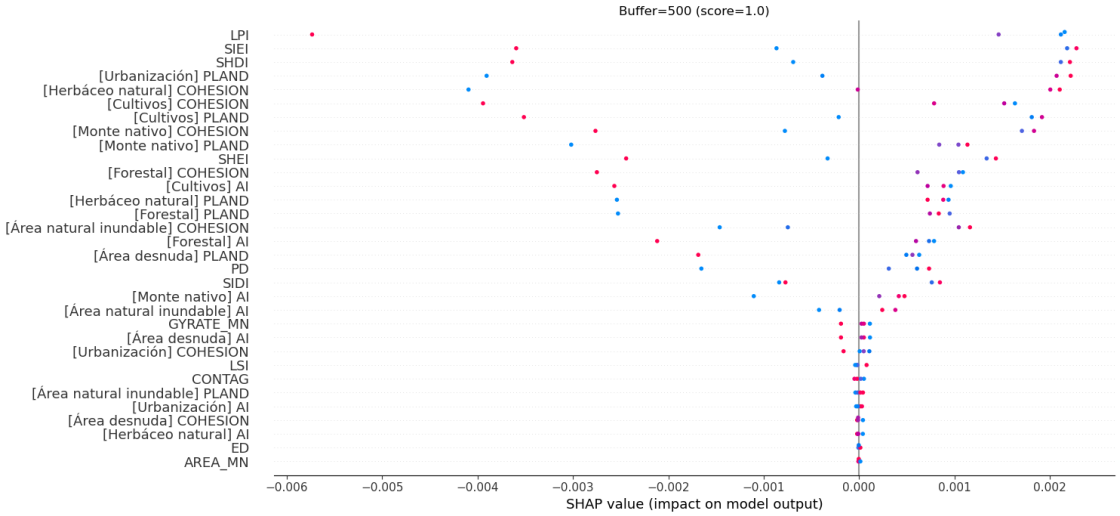


Fig. 5.43. Valores del SHAP para el modelo RF para Turbidez *subcuenca 3 zona buffer 500 m.*

Turbidez

La Fig. 5.44 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.16 presenta la varianza explicada por cada componente. La Fig. 5.45 y Fig. 5.46 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

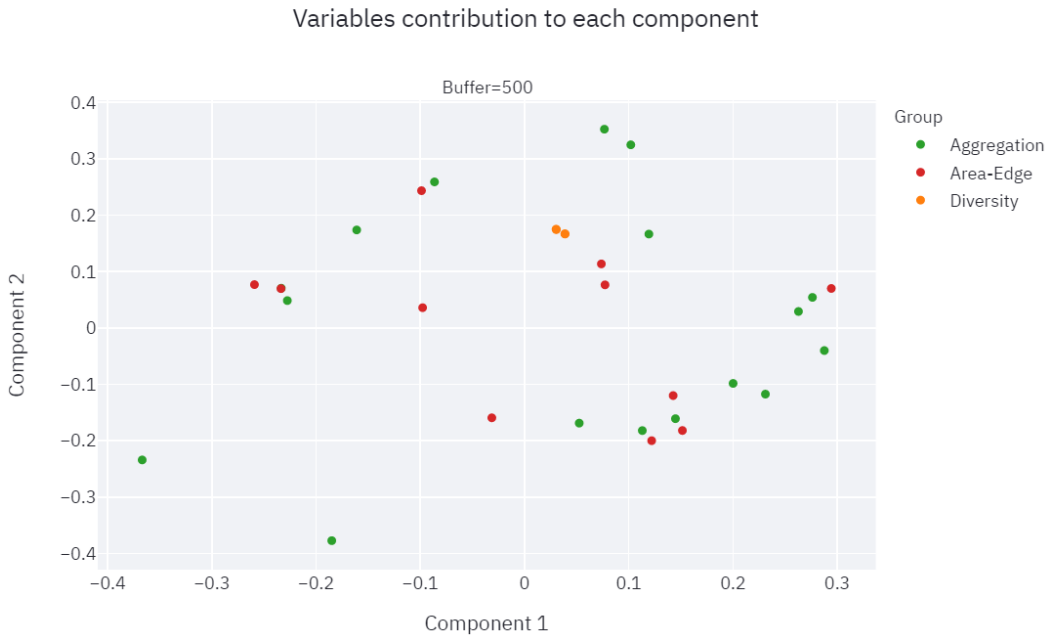


Fig. 5.44. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para Turbidez *subcuenca 1 zona buffer 500 m.*

Tabla 5.16. Varianza explicada por cada componente del modelo PLSR para Turbidez *subcuenca 3 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	41.3998	29.8969	28.7033	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	70.7700	25.2431	3.9869	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.37 a 0.29, mientras que en el componente 2 en -0.38 a 0.35. La componente 1 explica el 41.4% de la varianza de las variables de uso de suelo y un 70.8% de la varianza del valor medio del contaminante. La componente 2 explica el 29.9% de la varianza de las variables de uso de suelo y un 25.2% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 71.3% de la varianza de las variables de uso de suelo y un 96.0% de la varianza del valor medio del contaminante.

Los índices *[Urbanización] COHESION y [Urbanización] AI* tienen la mayor influencia negativa sobre en las dos primeras componentes. Por otro lado *[Área desnuda] PLAND, [Área desnuda] COHESION, [Herbáceo natural] AI* y *[Cultivos] AI* tienen los mayores pesos positivos sobre la componente 1. El *[Monte nativo] AI* y *[Área natural indudable] AI* tienen los mayores pesos positivos sobre la componente 2.

Además, se observa los índices de diversidad se superponen homogeneidad (*SIEI*) y diversidad (*SIDI*), y se agrupan cerca Simpson y Shannon.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.20 a 0.20. Valores altos de *[Urbanización] COHESION y [Urbanización] AI* generan los mayores impactos positivos en la salida del modelo. Los valores bajos de estos índices generan impactos negativos altos a medios. Valores altos de *[Área desnuda] PLAND* genera impactos negativos altos en la salida del modelo. Valores bajos de *[Área desnuda] COHESION, [Cultivos] AI y [Herbáceo natural] AI*, generan impactos positivos en la salida del modelo. Valores altos de estos índices generan impactos negativos.

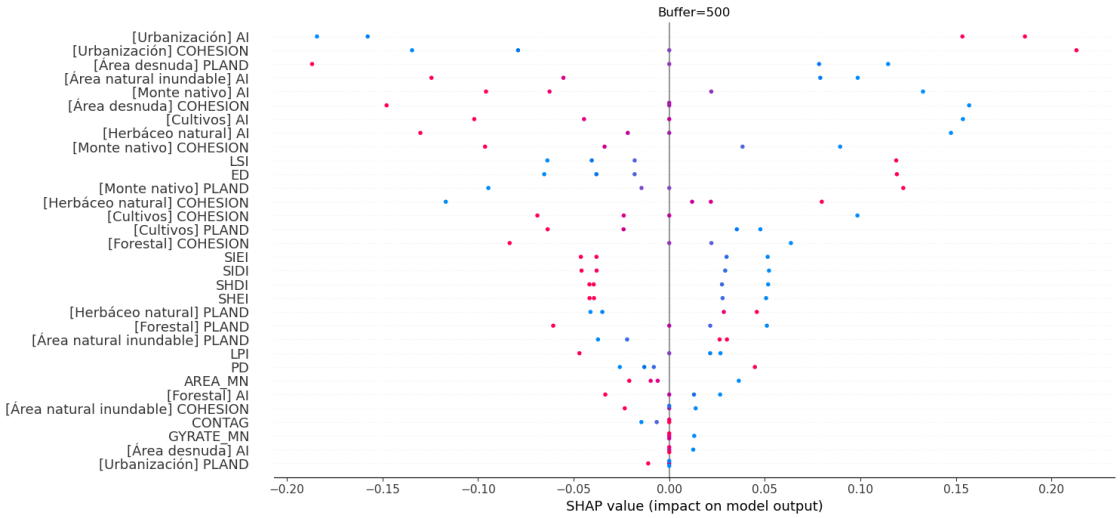


Fig. 5.45. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 3 zona buffer 500 m.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -0.6 a 0.6. El *[Urbanización] AI* es el índice más sensible a la salida del modelo, valores altos generan impactos negativos altos mientras que los valores bajos generan grandes impactos positivos. El índice *[Monte Nativo] PLAND* presenta un comportamiento similar a cuanto el impacto. Valores bajos de *[Área desnuda] PLAND y [Cultivos] AI* generan grandes impactos negativos en la salida del modelo, mientras que valores altos generan grandes impactos positivos.

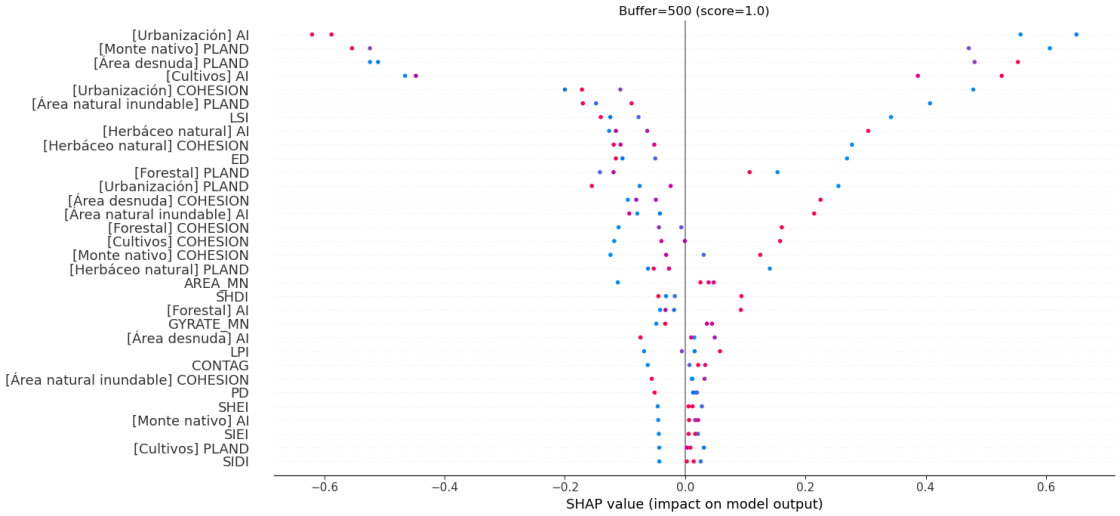


Fig. 5.46. Valores del SHAP para el modelo RF para Turbidez *subcuenca 3 zona buffer 500 m.*

5.3.5. Subcuenca 4 – Cierre: estación PS02

En la Fig. 5.47, se presenta la evolución temporal de la representación de cada clase para la subcuenca 4, considerando únicamente aquellas clases de verifican PLAND > 1%.

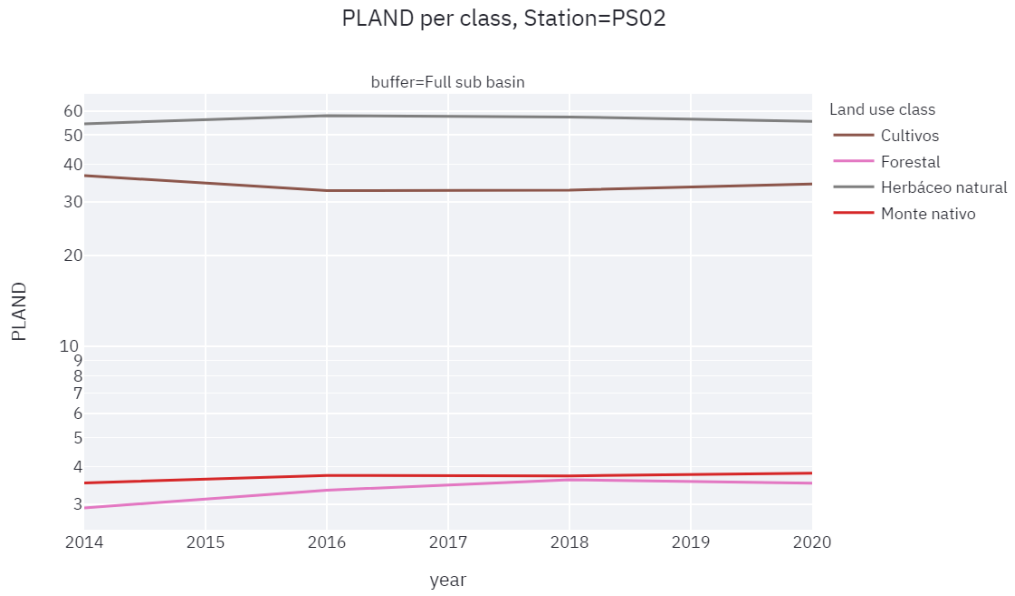


Fig. 5.47. Evolución temporal de la representación de cada clase para la subcuenca 4, para PLAND > 1%.

También en la subcuenca 3, solo cuatro usos del suelo ocupan más del 1% de la cuenca (*Cultivos*, *Forestal*, *Herbáceo natural* y *Monte nativo*). En Fig. 5.47 se observa que *Herbáceo natural,* seguido por *Cultivos,* son los usos del suelo dominantes de la subcuenca. A partir del 2016, también se observa un leve incremento del área cultivada con una consecuente disminución de *Herbáceo natural*.

Fósforo total

La Fig. 5.48 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.17 presenta la varianza explicada por cada componente. La Fig. 5.49 y Fig. 5.50 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

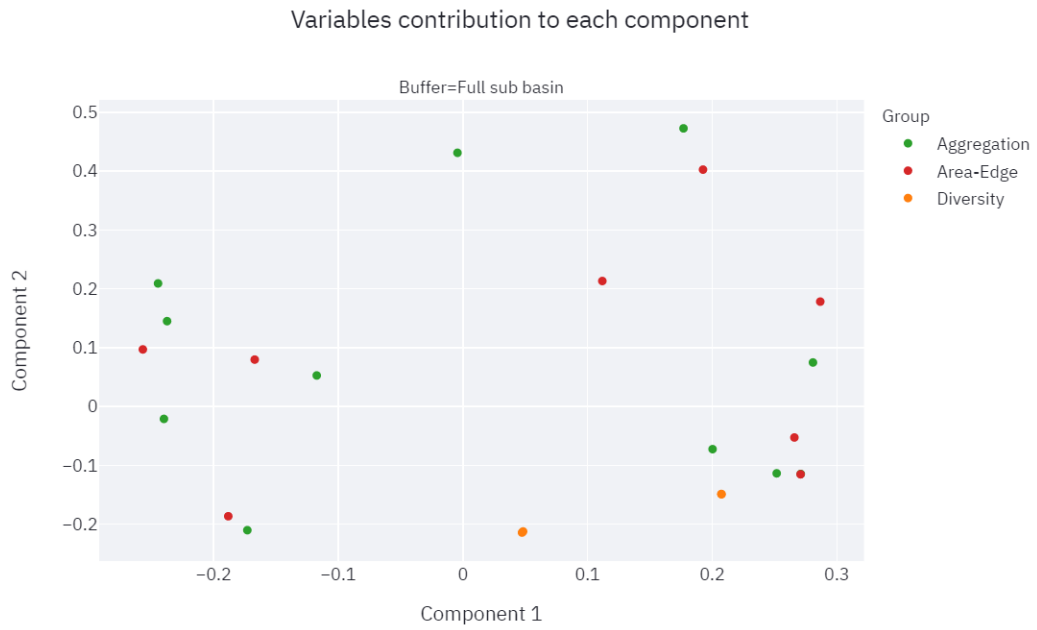


Fig. 5.48. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para PT *subcuenca 4.*

Tabla 5.17. Varianza explicada por cada componente del modelo PLSR para PT *subcuenca 4.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	51.5818	29.4328	18.9855	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	96.2016	2.7229	1.0755	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.25 a 0.30, mientras que en el componente 2 en -0.22 a 0.49. La componente 1 explica el 51.6% de la varianza de las variables de uso de suelo y un 96.2% de la varianza del valor medio del contaminante. La componente 2 explica el 29.4% de la varianza de las variables de uso de suelo y un 2.7% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 81.0% de la varianza de las variables de uso de suelo y un 98.9% de la varianza del valor medio del contaminante.

Los índices *[Monte nativo] AI y LPI* tienen una mayor influencia positiva sobre los dos primeros componentes, mientras que el índice *[Cultivos] PLAND y [Cultivos] COHESION* tiene la mayor influencia negativa sobre los dos primeros componentes. Además *[Forestal] PLAND y [Forestal]*

COHESION tienen el mayor peso positivo sobre la componente 1. El índice *AREA_AM* tiene el mayor peso negativo sobre la componente 2 y se encuentra agrupado cerca de *CONTAG*, *[Herbáceo natural] AI* y *[Cultivos] AI.* Además, se observa los índices de diversidad homogeneidad y diversidad prácticamente se superponen, pero se distancian según su formulación Simpson y Shannon.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.15 a 0.15. Se destacan los valores bajos de *[Monte nativo] PLAND* generan impactos positivos altos en la salida del modelo. Valores altos de *[Monte nativo] AI y LPI* generan impactos negativos altos en la salida del modelo.

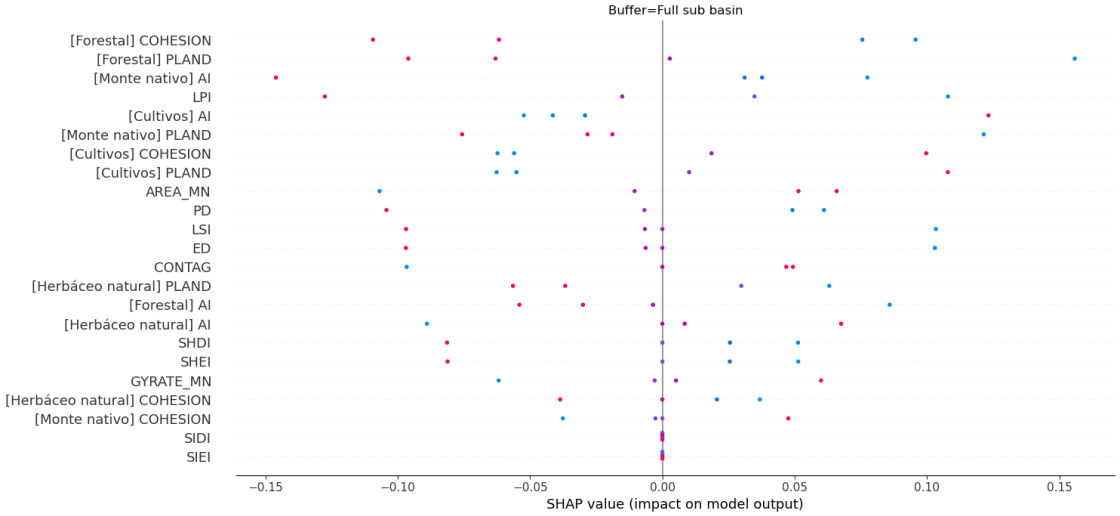


Fig. 5.49. Valores del SHAP para el modelo PLSR para PT *subcuenca 4.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente entre -5.0 a 4.0. El valor de *SHEI* es el índice más sensible a la salida del modelo, valores altos generan impactos positivos altos mientras que los valores bajos generan grandes impactos negativos. Los índices *LSI*, *[Forestal] COHESION y PD* generan impactos similares a los de *SHEI*.

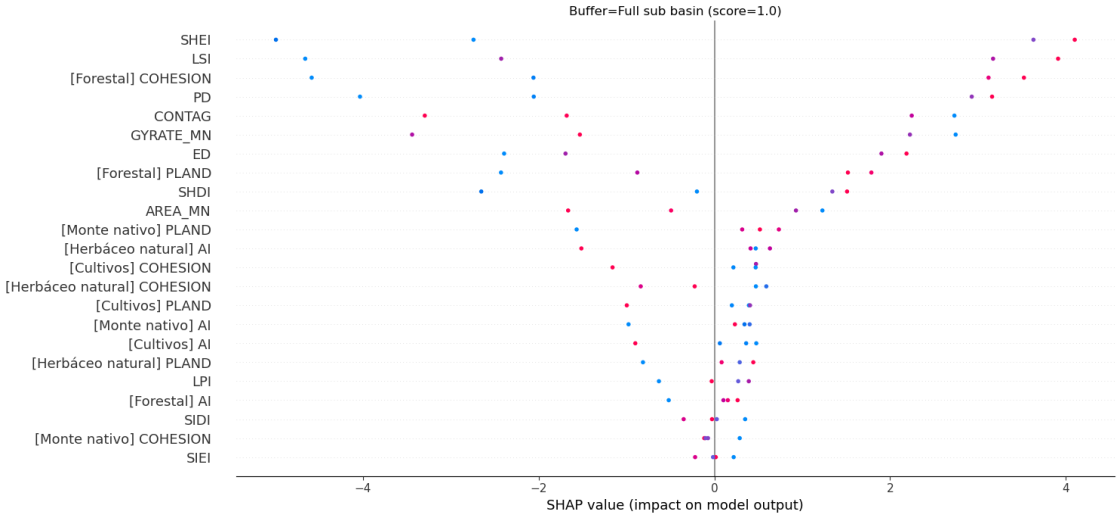


Fig. 5.50. Valores del SHAP para el modelo RF para PT *subcuenca 4.*

Nitrógeno total

La Fig. 5.51 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.18 presenta la varianza explicada por cada componente. La Fig. 5.52 y Fig. 5.53 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

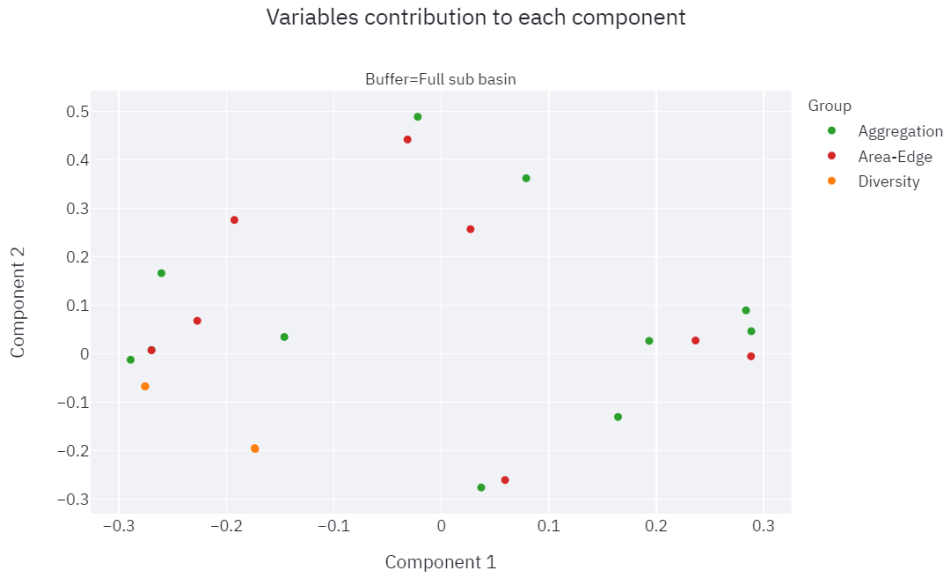


Fig. 5.51. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para NT *subcuenca 4.*

Tabla 5.18. Varianza explicada por cada componente del modelo PLSR para NT *subcuenca 4.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	51.9872	31.1897	16.8231	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	98.6382	0.9692	0.3925	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.29 a 0.28, mientras que en el componente 2 en -0.27 a 0.49. La componente 1 explica el 52.0% de la varianza de las variables de uso de suelo y un 98.6% de la varianza del valor medio del contaminante. La componente 2 explica el 31.2% de la varianza de las variables de uso de suelo y un 1.0% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 83.1% de la varianza de las variables de uso de suelo y un 99.6% de la varianza del valor medio del contaminante.

Los índices *CONTAG, AREA_AM* y *[Herbáceo natural] AI* tienen el mayor peso positivo sobre la componente 1. Por otro lado, los índices *PD, SHDI, SHEI, ED y LSI* tienen los mayores pesos negativos sobre la componente 1. Además, se observa los índices de diversidad homogeneidad y diversidad de Shannon prácticamente se superponen, lo mismo sucede con los índices *ED y LSI*.

El rango de variación del valor del SHAP se encuentra aproximadamente en -0.10 a 0.15. Se destacan los valores bajos de *AREA_AM, CONTAG, [Herbáceo natural] AI* y altos de *PD, SHDI, SHEI* que generan impactos positivos altos en la salida del modelo. Por otro lado, valores bajos de *ED y LSI* generan los mayores impactos negativos en la salida del modelo, en menor medida valores bajos de *[Monte nativo] PLAND*.

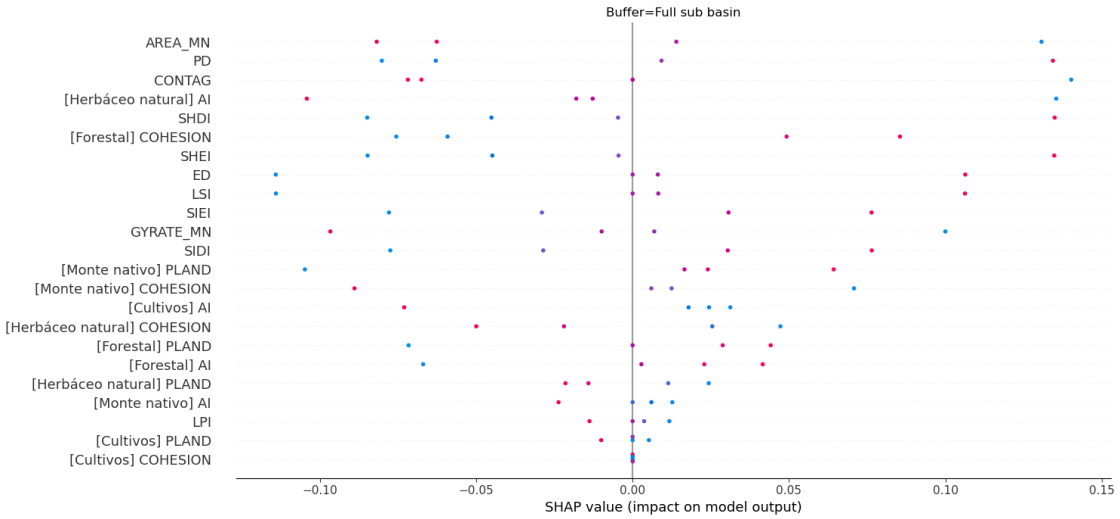


Fig. 5.52. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 4.*

Los resultados del SHAP aplicado al modelo RF indica los valores bajos de los índices *GYRATE_MN*, *[Herbáceo natural] AI*, *[Monte nativo] COHESION y CONTAG* generan impacto negativo altos en la salida del modelo, y en menor medida los valores altos de los índices SHEI, *[Forestal] COHESION, [Monte nativo] PLAND* y LSI.

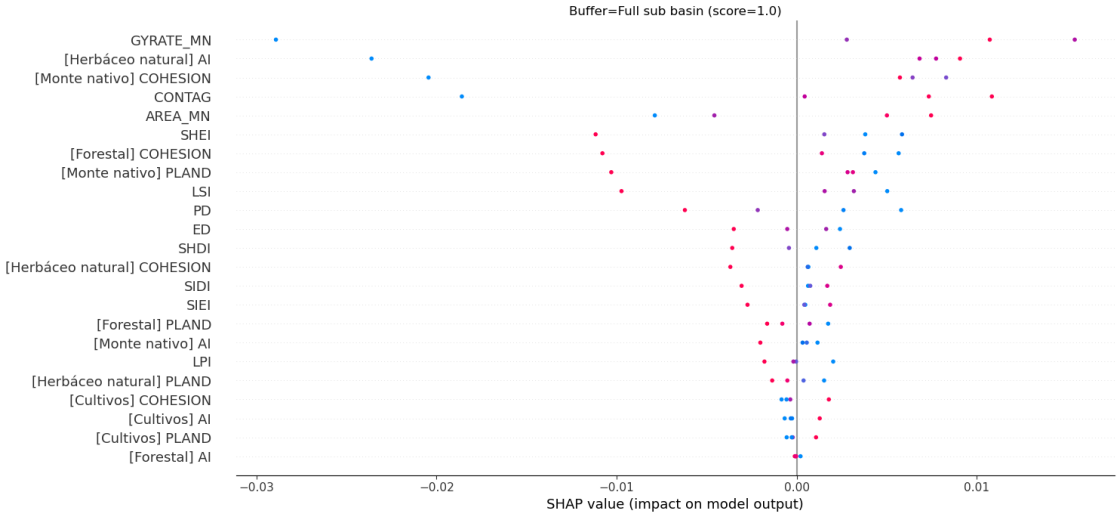


Fig. 5.53. Valores del SHAP para el modelo RF para Turbidez *subcuenca 4.*

Turbidez

La Fig. 5.54 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.19 presenta la varianza explicada por cada componente. La Fig. 5.55 y Fig. 5.56 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

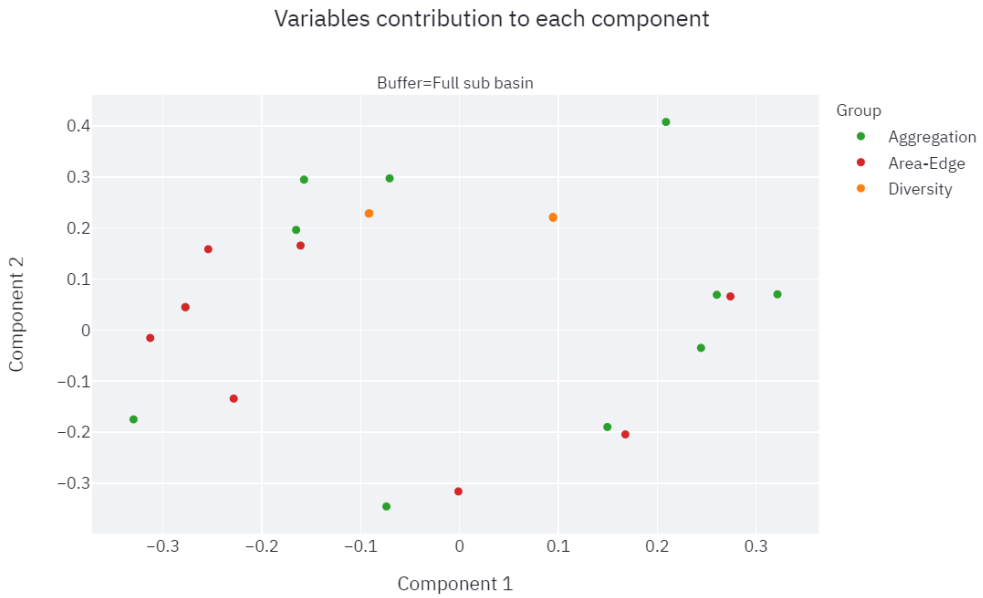


Fig. 5.54. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para Turbidez *subcuenca 4.*

Tabla 5.19. Varianza explicada por cada componente del modelo PLSR para Turbidez *subcuenca 4.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
Full sub basin	47.0804	35.7250	17.1946	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
Full sub basin	88.6988	7.6797	3.6215	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.17 a 0.32, mientras que en el componente 2 en -0.34 a 0.40. La componente 1 explica el 47.1% de la varianza de las variables de uso de suelo y un 88.7% de la varianza del valor medio del contaminante. La componente 2 explica el 35.7% de la varianza de las variables de uso de suelo y un 7.7% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 82.8% de la varianza de las variables de uso de suelo y un 96.4% de la varianza del valor medio del contaminante.

El índice *[Herbáceo natural] COHESION* tiene la mayor influencia positiva sobre las primeras dos componentes, mientas que el índice *[Forestal] AI* tiene la mayor influencia negativa sobre las primeras dos componentes. El índice *[Cultivos] AI* tiene el mayor peso positivo sobre la componente 1, mientras que el índice *[Monte nativo] PLAND,* tiene un peso negativo importante

sobre la componente 2. Además, se observa los índices de diversidad homogeneidad y diversidad prácticamente se superponen, pero se distancian según su formulación Simpson y Shannon.

El rango de variación del valor del SHAP para el modelo PLSR se encuentra aproximadamente en -0.20 a 0.15. Se destacan los valores bajos de *[Herbáceo natural] COHESION* y los valores altos de *[Forestal] AI* que generan impactos positivos altos en la salida del modelo. Por otro lado, los valores bajos de *[Forestal] AI y [Monte nativo] PLAND y valores altos de [Cultivos] AI* generan impactos negativos altos en la salida del modelo.

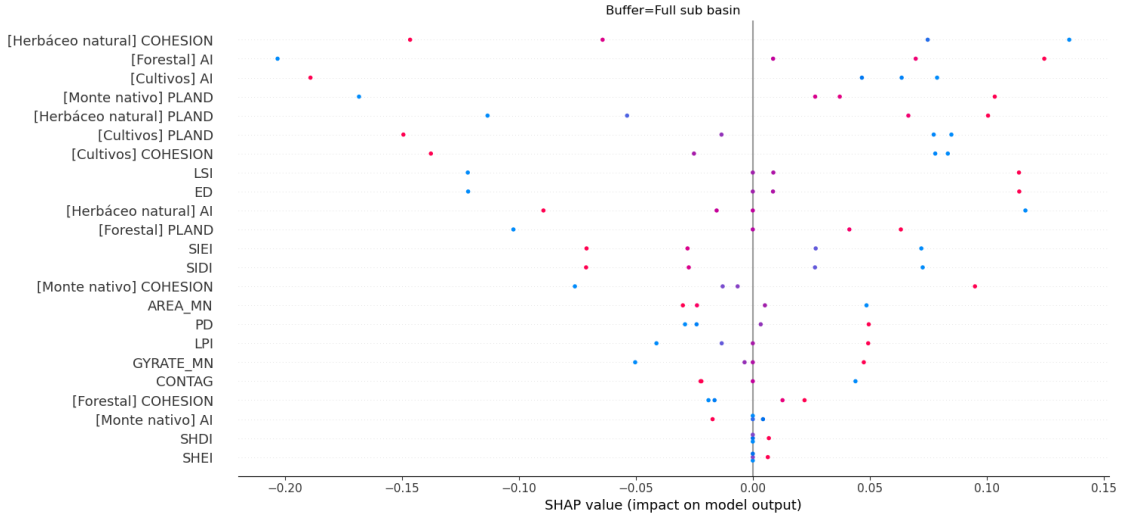


Fig. 5.55. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 4.*

El rango de variación del valor del SHAP para el modelo RF se encuentra aproximadamente en -0.20 a 0.40. Se destacan gran sensibilidad de la salida del modelo, con comportamiento no lineal, a los índices de una cantidad importante de índices. El *[Herbáceo natural] COHESION* parece seguir un comportamiento lineal generando para valores bajos impactos negativos en la salida del modelo y para valores altos impactos positivos.

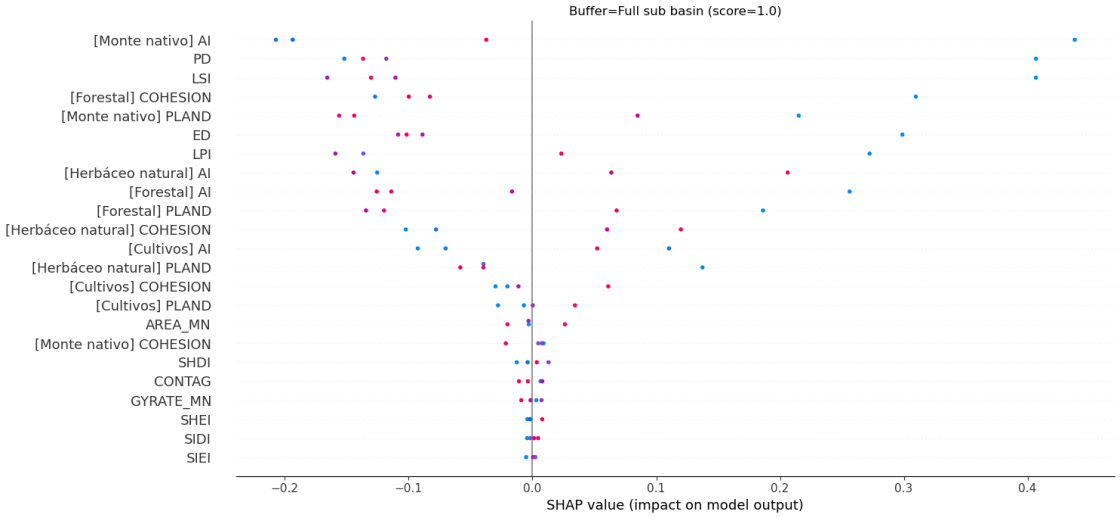


Fig. 5.56. Valores del SHAP para el modelo RF para Turbidez *subcuenca 4.*

5.3.6. Subcuenca 4 zona buffer 500 m – Cierre: estación PS02

En la Fig. 5.57, se presenta la evolución temporal de la representación de cada clase para la subcuenca 4 zona buffer 500 m, considerando únicamente aquellas clases de verifican PLAND > 1%.

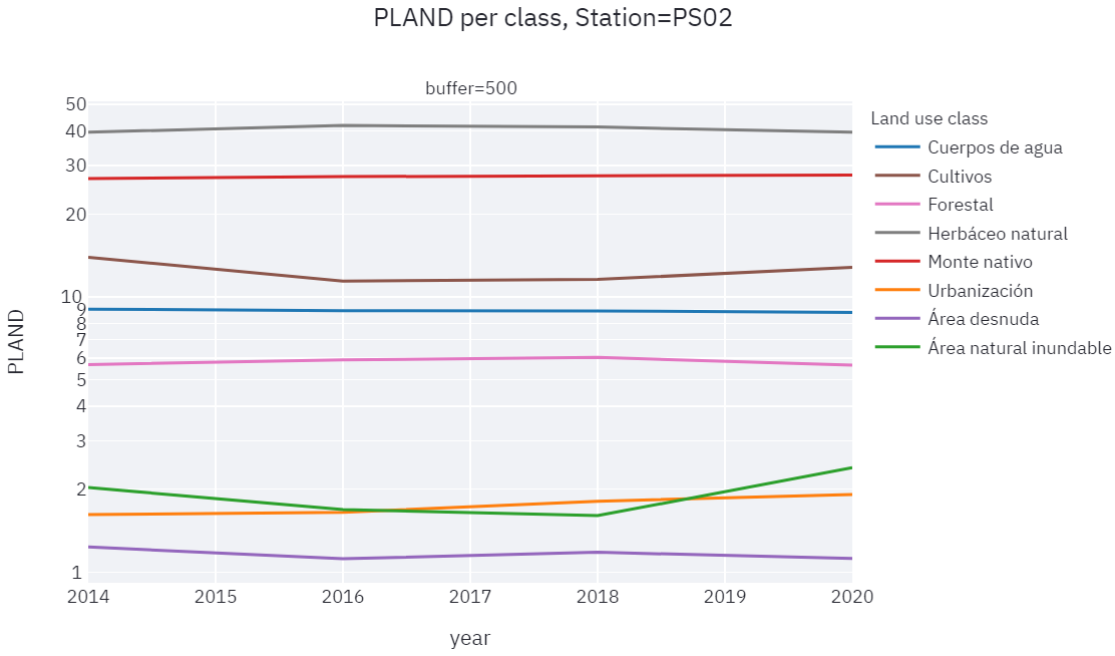


Fig. 5.57. Evolución temporal de la representación de cada clase para la subcuenca 4 zona buffer 500 m, para PLAND > 1%.

En la zona buffer de la subcuenca 4, todos los usos del suelo ocupan más del 1% de la cuenca, también *Cuerpos de agua* en cuanto esta subcuenca incluye el embalse Paso Severino. En Fig.

5.57 se observa que *Herbáceo natural,* seguido por *Monte nativo,* son los usos del suelo dominantes de la zona buffer. A partir del 2016, también se observa un leve incremento del área cultivada con una consecuente disminución de *Herbáceo natural*. En los usos del suelo con menor porcentaje, se observa un incremento de la *Urbanización* a partir del 2016 y un incremento del *Área natural inundable* a partir del 2018.

Fósforo total

La Fig. 5.58 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.20 presenta la varianza explicada por cada componente. La Fig. 5.59 y Fig. 5.60 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

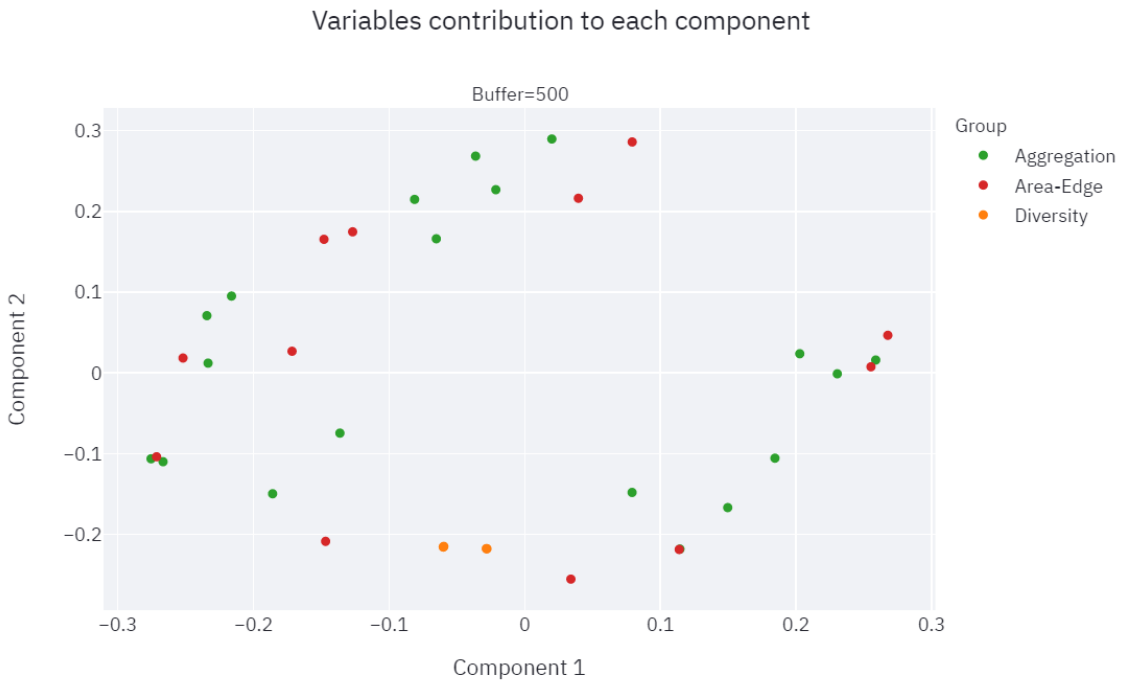


Fig. 5.58. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para PT *subcuenca 4 zona buffer 500 m.*

Tabla 5.20. Varianza explicada por cada componente del modelo PLSR para PT *subcuenca 4 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	44.9505	41.8008	13.2487	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	90.3184	7.9442	1.7374	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.27 a 0.26, mientras que en el componente 2 en -0.25 a 0.29. La componente 1 explica el 45.0% de la varianza de las variables de uso de suelo y un 90.3% de la varianza del valor medio del contaminante. La componente 2 explica el 41.8% de la varianza de las variables de uso de suelo y un 7.9% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 86.8% de la varianza de las variables de uso de suelo y un 98.2% de la varianza del valor medio del contaminante.

Los índices *[Forestal] AI, LPI y [Forestal] COHESION* se agrupan, aportando el mayor peso negativo sobre la componente 1. Los índices *[Monte nativo] PLAND, [Herbáceo natural] COHESION y [Urbanización] PLAND* se agrupan aportando el mayor peso positivo sobre la componente 1. Además, se observa los índices de diversidad homogeneidad y diversidad prácticamente se superponen y se agrupan cerca según su formulación Simpson y Shannon.

El rango de variación del valor del SHAP aplicado al PLSR se encuentra aproximadamente en -0.10 a 0.15. Se destacan los valores altos de *[Forestal] AI*, LPI y *[Forestal] COHESION* generan impactos positivos altos en la salida del modelo. Valores altos de *[Urbanización] PLAND,* y bajos de *[Cuerpos de agua] PLAND* y *[Cultivos] COHESION* generan los mayores impactos negativos en la salida del modelo.

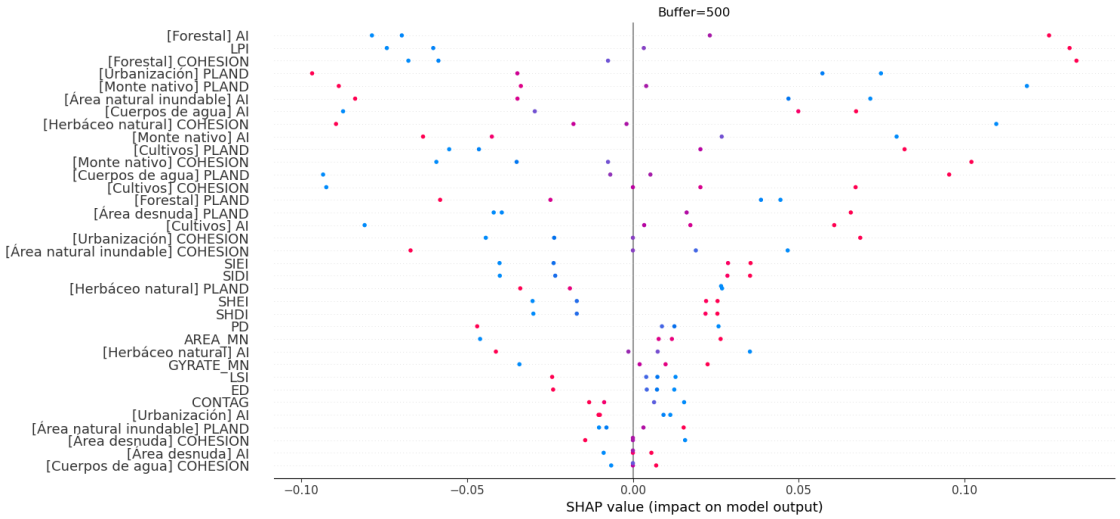


Fig. 5.59. Valores del SHAP para el modelo PLSR para PT *subcuenca 4 zona buffer 500 m.*

El rango de variación del valor del SHAP aplicado al RF se encuentra aproximadamente en -4 a 3. Algunas métricas presentan comportamiento no lineal como son *[Área natural inundable] COHESION y [Urbanización] COHESION,* de mayor sensibilidad a la salida del modelo. Se observa que valores altos de los índices *[Cuerpos de agua] PLAND, LPI, [Forestal] COHESION y [Forestal] AI,* y valores bajos de *[Urbanización] PLAND y [Herbáceo natural] COHESION* generan impactos negativos en el modelo.

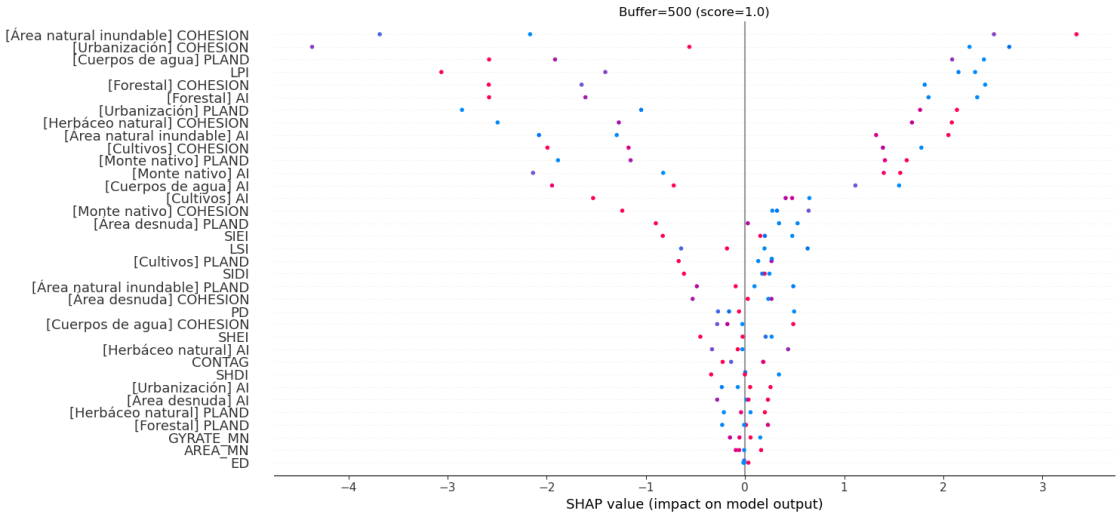


Fig. 5.60. Valores del SHAP para el modelo RF para PT *subcuenca 4 zona buffer 500 m.*

Nitrógeno total

La Fig. 5.61 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.21 presenta la varianza explicada por cada componente. La Fig. 5.62 y Fig. 5.63 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

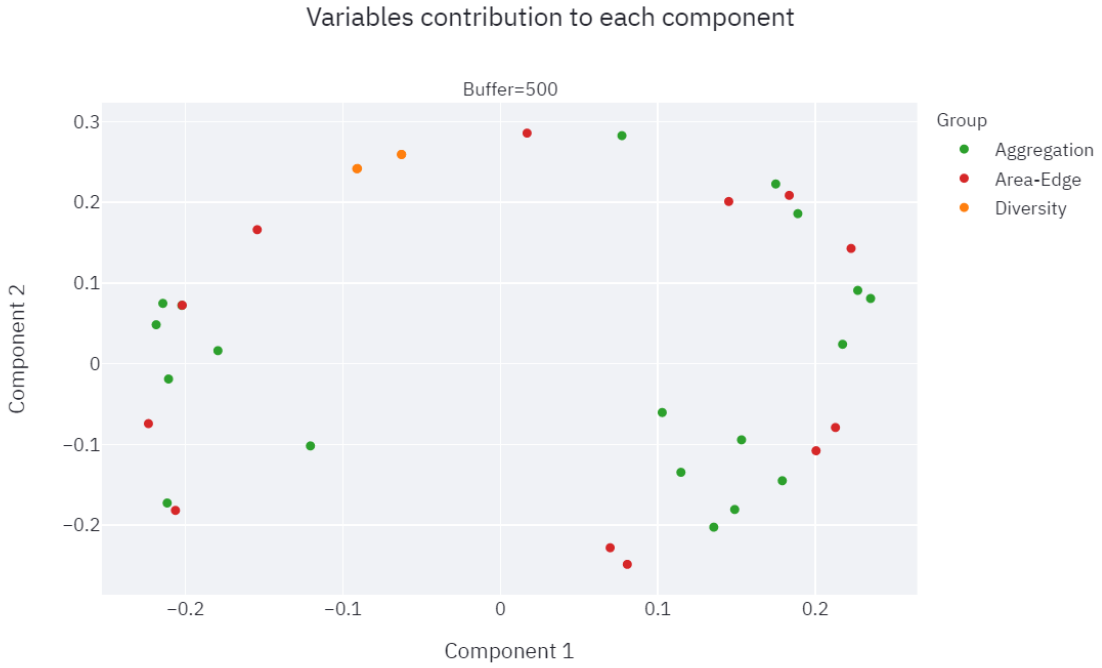


Fig. 5.61. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para NT *subcuenca 4 zona buffer 500 m.*

Tabla 5.21. Varianza explicada por cada componente del modelo PLSR para NT *subcuenca 4 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	53.5693	34.6434	11.7872	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	96.8804	3.1126	0.0070	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.22 a 0.23, mientras que en el componente 2 en -0.25 a 0.29. La componente 1 explica el 53.7% de la varianza de las variables de uso de suelo y un 96.9% de la varianza del valor medio del contaminante. La componente 2 explica el 34.6% de la varianza de las variables de uso de suelo y un 3.1% de la

varianza del valor medio del contaminante. Las dos primeras componentes explican 88.2% de la varianza de las variables de uso de suelo y un 100.0% de la varianza del valor medio del contaminante.

Los índices *[Urbanización] PLAND* aportando el mayor peso negativo sobre la componente 1. El índice *[Cultivos] COHESION* aportando el mayor peso positivo sobre la componente 1. Además, se observa los índices de diversidad homogeneidad y diversidad prácticamente se superponen y se agrupan cerca según su formulación Simpson y Shannon.

El rango de variación del valor del SHAP aplicado al PLSR se encuentra aproximadamente en -0.080 a 0.10. Se destacan los valores bajos de *[Cultivos] AI* y *[Cultivos] COHESION* generan impactos positivos altos en la salida del modelo. Valores bajos de *[Monte nativo] PLAND* y *[Herbáceo natural] COHESION* generan impactos negativos altos en la salida del modelo. Por otro lado, los valores altos de *[Forestal] AI, [Forestal] COHESION, LPI y [Cuerpos de agua] PLAND* generan impactos negativos altos en la salida del modelo.

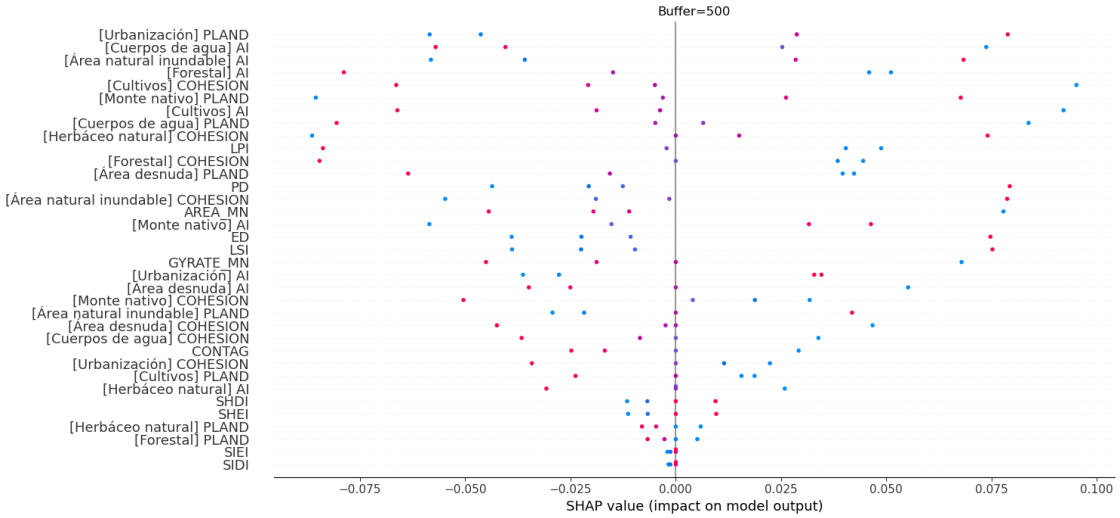


Fig. 5.62. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 4 zona buffer 500 m.*

El rango de variación del valor del SHAP aplicado al RF se encuentra aproximadamente en -0.015 a 0.008. Los resultados indican que muchos de los índices presentan comportamiento no lineal y de gran sensibilidad para los valores bajos.

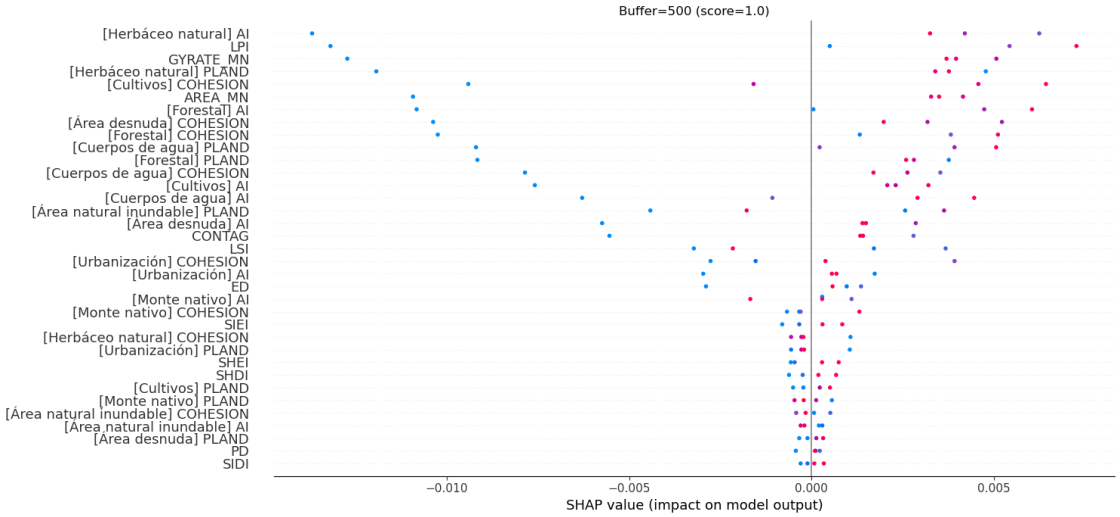


Fig. 5.63. Valores del SHAP para el modelo RF para Turbidez *subcuenca 4 zona buffer 500 m.*

Turbidez

La Fig. 5.64 presenta el gráfico de dispersión para los pesos de los dos primeros componentes del modelo PLSR, mientras que la Tabla 5.22 presenta la varianza explicada por cada componente. La Fig. 5.65 y Fig. 5.66 presenta los valores del SHAP para el modelo PLSR y RF, respectivamente.

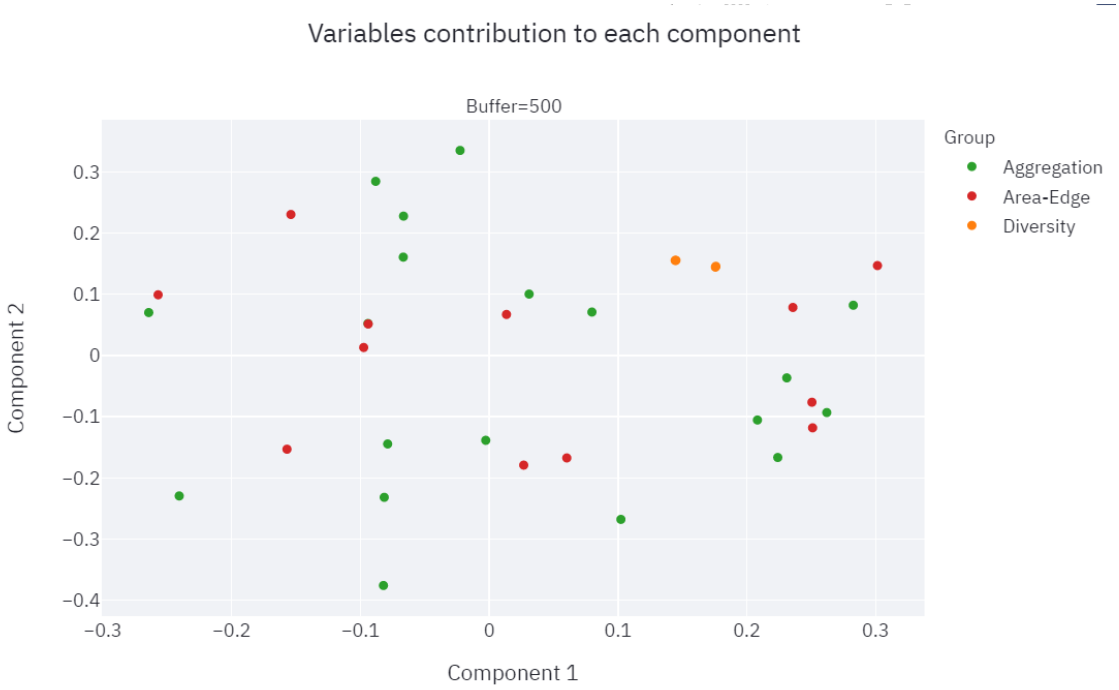


Fig. 5.64. Dispersión para los dos pesos de los dos primeros componentes del modelo PLSR para Turbidez *subcuenca 4 zona buffer 500 m.*

Tabla 5.22. Varianza explicada por cada componente del modelo PLSR para Turbidez *subcuenca 4 zona buffer 500 m.*

Components variance

Landuse variables

	Component 1	Component 2	Component 3	Component 4
500	34.7922	45.3794	19.8284	0

Contaminant

	Component 1	Component 2	Component 3	Component 4
500	91.2828	5.3402	3.3770	0

El rango de variación de los pesos según el componente 1 se encuentra en -0.26 a 0.30, mientras que en el componente 2 en -0.37 a 0.33. La componente 1 explica el 34.8% de la varianza de las variables de uso de suelo y un 91.3% de la varianza del valor medio del contaminante. La componente 2 explica el 45.4% de la varianza de las variables de uso de suelo y un 5.3% de la varianza del valor medio del contaminante. Las dos primeras componentes explican 80.2% de la varianza de las variables de uso de suelo y un 96.6% de la varianza del valor medio del contaminante.

El índice *[Urbanización] AI* representa la mayor influencia negativa en las dos primeras componentes. Los índices *[Monte nativo] PLAND, [Herbáceo natural] COHESION* se agrupan, aportando el mayor peso negativo sobre la componente 1. El *[Área desnuda] PLAND* representa la mayor influencia positiva sobre las dos primeras componentes, el índice *[Monte nativo] COHESION* se agrupa cerca de este último aportando un peso alto positivo sobre la componente 1. Además, se observa los índices de diversidad homogeneidad y diversidad prácticamente se superponen y se agrupan cerca según su formulación Simpson y Shannon.

El rango de variación del valor del SHAP aplicado al PLSR se encuentra aproximadamente en -0.15 a 0.10. Se destacan los valores altos de *[Área desnuda] PLAND* y *[Monte nativo] COHESION* generan impactos negativos altos en la salida del modelo. Valores bajos de *[Monte nativo] AI y [Cuerpos de agua] PLAND* generan los mayores impactos negativos. Valores altos de *[Urbanización] AI, [Herbáceo natural] COHESION* y *[Monte Nativo] PLAND* generan impactos positivos en la salida del modelo.

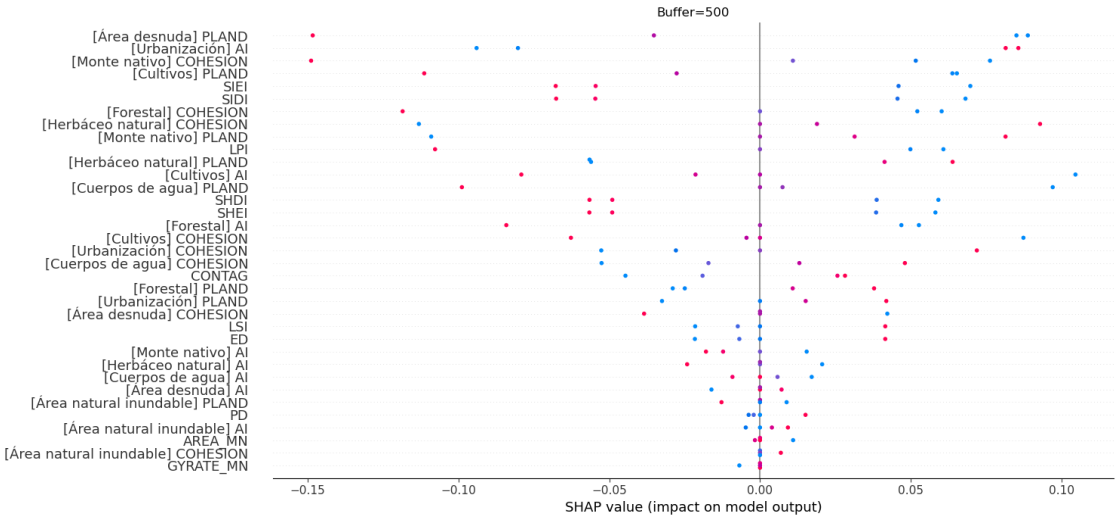


Fig. 5.65. Valores del SHAP para el modelo PLSR para Turbidez *subcuenca 4 zona buffer 500 m.*

El rango de variación del valor del SHAP aplicado al RF se encuentra aproximadamente en -0.2 a 0.5. Los resultados indican que muchos de los índices presentan comportamiento no lineal y de gran sensibilidad para los valores altos y en menor medida para valores bajos.

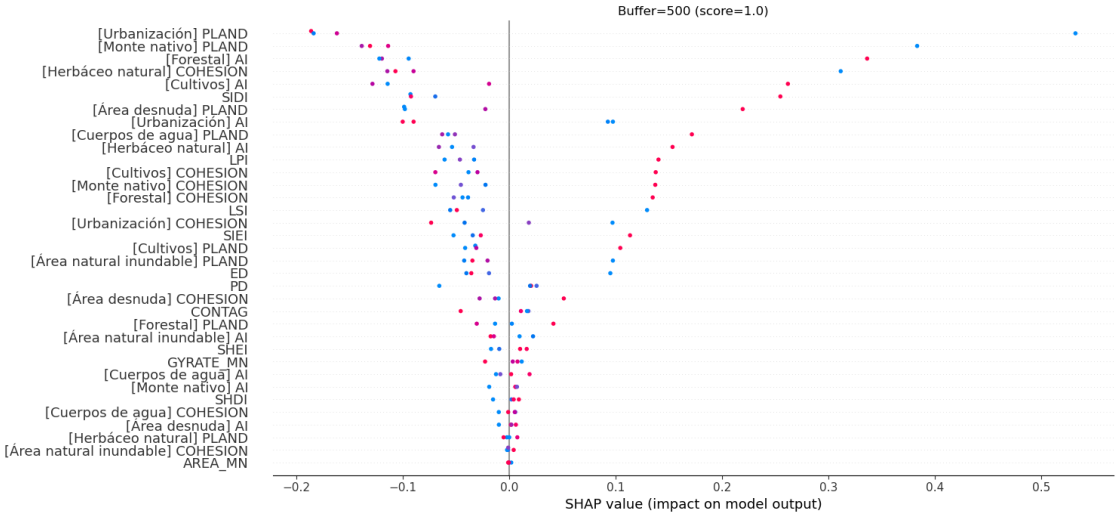


Fig. 5.66. Valores del SHAP para el modelo RF para Turbidez *subcuenca 4 zona buffer 500 m.*

5.4. Síntesis e interpretación de resultados

En esta sección se presenta una síntesis del análisis realizado y una interpretación de los resultados, diferenciando el modelo PLSR y RF. Dentro de cada modelo se presentan los resultados agregados por estación. Para el caso del modelo PLSR se realiza una comparación entre el nivel de cuenca y zona buffer. Al comprar los modelos de RF y PLSR una de las principales

diferencias en cuanto a los resultados es la no linealidad para ciertas situaciones que se observan en el RF. Las no linealidades del modelo RF, suceden en general independiente del número de variables de entrada comparando entre contaminantes y a distintas escalas (i.e., PT estación SLC01 subcuenca, Turbidez estación PS02 buffer 500) ya que es una propiedad del modelo. Estos comportamientos le agregan complejidad al problema en cuanto a la interpretación física de los resultados. En el otro extremo, hay situaciones para un mismo contaminante a una misma escala donde el RF y el PLSR devuelven resultados muy similares de linealidad para los índices más sensibles (i.e., Turbidez estación PS01=SLC03 subcuenca).

5.4.1. Modelo PLSR

Estación SLC01

Para el caso de la subcuenca 1 (estación SLC01), la composición y configuración del paisaje de la cuenca de aporte tienen efectos importantes en el valor medio de las variables de calidad de agua analizadas. A nivel de paisaje, el índice más influyente es el valor medio del radio de giro de los parches (*GYRATE_MN*) correlacionando de manera positiva con PT y negativa con NT y Turbidez. A nivel de clase, las variables de calidad de agua *PT, NT y Turbidez* se encuentran relacionadas con la clase *Monte nativo*. La conectividad, agregación y representación de los parches de la clase *Monte nativo* correlaciona de forma inversa con el PT, y directa con NT y Turbidez. En menor medida, la agregación de la clase *Cultivos* explica relaciones directas con el *PT e* inversas con *NT*, en cambio la representación *Cultivos* sigue una relación inversa con *Turbidez*.

A nivel de la zona buffer, 500 metros, de la subcuenca 1 (estación SLC01), las variables de calidad de agua *PT, NT y Turbidez* se encuentran fuertemente relacionado con la configuración y composición del paisaje respecto al *Monte nativo, Herbáceo natural* y *Cultivos.* La agregación, conectividad física y representación del *Monte nativo* correlaciona negativamente con los valores de *PT* y positivamente con *NT* y *Turbidez*. La representación de la clase *Cultivos* en el paisaje correlaciona de forma directa con *PT* e inversa con *NT y Turbidez*. Por otro lado, la conectividad física de *Herbáceo natural* se relaciona de forma inversa con el valor de *PT*, y directa con *NT y Turbidez.*

Comparando nivel de cuenca con zona buffer, en términos generales, las métricas con mayor explicación de los valores de calidad de agua están asociadas a índices a nivel de clase, en particular a nivel de zona buffer. No obstante, a nivel de subcuenca el radio medio de giro de los parches es un índice que tiene una mayor influencia en la explicación de los valores de calidad de agua. A nivel de cuenca la composición y configuración de la clase *Monte nativo* tiene influencias muy marcadas en los valores de calidad de agua, y en menor medida la clase *Cultivos*. Por otro lado, a nivel de zona buffer, la composición y configuración del paisaje en cuanto a las clases

Monte nativo, *Cultivos* y *Herbáceo natural* son las que más explican influencias en los valores de calidad de agua.

Estación PS01=SLC03

Para el caso de la subcuenca 3 (estación PLS01=SLC03), de los resultados del modelo de PLSR, se observa que a nivel de paisaje las métricas de paisaje ED y LSI correlacionan de forma directa con PT y NT, y LPI de forma directa con NT y Turbidez. A nivel de clase, las variables de calidad de agua *PT, NT y Turbidez* se encuentran relacionadas con la composición y configuración en el paisaje de las clases *Monte nativo, Cultivos, Herbáceo Natural y Forestal*. La conectividad, agregación y representación de los parches de la clase *Forestal* correlaciona de forma directa con el PT, e inversa con NT y Turbidez. Además, la agregación de la clase *Cultivos* explica relaciones directas con el *PT e* inversas con *NT*, en cambio la representación *Cultivos* sigue una relación inversa con *Turbidez*. La representación y agregación del *Monte nativo* correlaciona de forma inversa con *PT*, y de forma directa con *NT*. Por otro lado, la *Turbidez* y tiene una relación inversa con la agregación de la clase *Monte nativo*, mientras que la conectividad física de la clase *Herbáceo natural* correlaciona de forma inversa con la Turbidez.

A nivel de la zona buffer, 500 metros, de la subcuenca 3 (estación PS01=SLC03), los resultados del modelo PLSR indican que a nivel de paisaje se encuentra una relación directa entre LPI y los parámetros de calidad de agua *PT y NT*. A nivel de clase, las variables de calidad de agua *PT, NT y Turbidez* se encuentran relacionadas con la composición y configuración en el paisaje de las clases *Área desnuda, Urbanización, Herbáceo Natural, Cultivos, Forestal* y en menor medida *Área natural inundable*. En este sentido, la representación del *Área desnuda* tiene una correlación directa con PT y NT e inversa con Turbidez, además la conectividad de los parches de esta clase se relaciona de forma inversa con la Turbidez. La representación de *Urbanización* correlaciona inversa con PT y NT, mientras que la agregación y conectividad de esta clase tiene una relación directa con la Turbidez. La conectividad de la clase *Herbáceo natural* muestra relación inversa con PT y NT, y la agregación de esta clase una relación inversa con la Turbidez. La conectividad física y agregación de la clase *Cultivos* correlaciona negativamente con PT, positivamente con NT, y la agregación de esta clase correlaciona negativamente con la Turbidez. La agregación y conectividad de la clase *Forestal* presenta relación directa con PT y NT. Por otro lado, la representación *Monte nativo* una relación inverso PT y NT, y la agregación del *Área natural inundable* muestra una relación inversa con NT.

Comparando nivel de cuenca con zona buffer, en términos generales, las métricas con mayor explicación de los valores de calidad de agua están asociadas a índices a nivel de clase. No obstante, a nivel de cuenca y zona buffer los índices *ED*, *LSI* y *LPI*, medidas de desagregación y dominancia de los parches en el paisaje, tiene una influencia en la explicación de los valores de calidad de agua. Tanto a nivel de cuenca como de zona buffer la composición y configuración de las clases que tiene una representación mayor al 1% del área total del paisaje tiene influencias

muy marcadas en los valores de calidad de agua. Cuando se analiza la zona buffer, el paisaje se ve representados por una distribución más uniforme de las clases, pero con dominancias marcadas por *Herbáceo natural* y *Monte nativo*, y en menor medida por *Cultivos*.

Estación PS02

Para el caso de subcuenca 4 (estación PS02), a nivel de subcuenca, los resultados del modelo de PLSR indican que las métricas de paisaje *SHDI*, *SHEI*, *ED*, *LSI* y *PD* que indican relaciones de homogeneidad, dominancia y desagregación de los parches en el paisaje presentan una relación directa con el NT, mientras que el *LPI* lo hace con el PT. Además, los índices *AREA_MN* y *CONTAG* guardan una relación inversa con el NT. A nivel de clase las variables de calidad de agua PT, NT y Turbidez se encuentran relacionadas con la composición y configuración en el paisaje de las clases *Monte nativo*, *Herbáceo Natural*, *Cultivos y Forestal*. En este sentido, la representación de la clase *Monte nativo* sigue una relación inversa con PT y Turbidez, y directa con NT. Además, la agregación de esta clase presenta relación negativa con el PT. La conectividad física y agregación de los parches de la clase *Herbáceo natural* presenta relación inversa con el NT, en cambio la conectividad de esta clase guarda una relación inversa con la Turbidez. Por otro lado, la Turbidez presenta relación positiva con la agregación de los parches de las clases *Forestal* y *Cultivos*.

A nivel de la zona buffer, 500 metros, de la subcuenca 4 (estación PS02) los resultados del modelo de PLSR indican que a nivel de paisaje el índice *LPI* presenta relación directa con el PT e inversa con el NT. A nivel de clase las variables de calidad de agua PT, NT y Turbidez se encuentran relacionadas con la composición y configuración en el paisaje de las clases *Cuerpos de agua*, *Área desnuda*, *Urbanización*, *Herbáceo Natural*, *Cultivos*, *Forestal*. En este sentido, la agregación y conectividad física de la clase Forestal presenta relación directa con PT e inversa con NT. La representación de la clase *Urbanización* correlación inversa con PT, mientras que la agregación de los parches de esta clase una relación positiva con la Turbidez. La representación de la clase Cuerpos de agua presenta una relación directa con el PT e inversa NT y Turbidez. La conectividad física de las clases *Cultivos* una relación directa con el PT, mientras que la agregación y conectividad de esta clase una relación inversa con el NT. La representación de la *Monte nativo* correlaciona positivamente con el NT, la agregación de esta clase negativamente con la Turbidez, mientras que la conectividad física positivamente con la Turbidez. Además, la conectividad *Herbáceo natural* guarda una relación directa NT y Turbidez, mientras que la representación de la clase *Área desnuda* una relación directa con Turbidez.

Comparando nivel de cuenca con zona buffer, en términos generales, a nivel de cuenca donde tanto las métricas de paisaje como las de clase explican de forma similar las relaciones con las variables de calidad de agua, particularmente para el NT, y en menor medida para PT y Turbidez. A nivel de zona buffer las métricas con mayor explicación de los valores de calidad de agua están asociadas a índices a nivel de clase. Tanto a nivel de cuenca como de zona buffer la composición y configuración de las clases que tiene una representación mayor al 1% del área total del paisaje

tiene influencias en los valores de calidad de agua. Cuando se analiza la zona buffer, el paisaje se ve representados por una distribución más uniforme de las clases, pero con dominancias marcadas por *Herbáceo natural y Monte nativo.*

Comparando entre las subcuencas 1 y 4, se hace notar que la cantidad de variables de entrada a los modelos es la misma pero la distribución de la representación de las clases (con PLAND > 1%) es distinta. No obstante, al aumentar la extensión del análisis se está más cerca de representar el total del mapa categórico elaborado, en donde fueron definidas las clases a ser representadas por el algoritmo de clasificación. Los resultados indican que cuanto más grande es la extensión a nivel de cuenca mayor peso tienen las métricas de paisaje sobre el comportamiento de las variables de calidad de agua, comparando con el caso de las zonas buffer.

En las zonas buffer (500 m) de las subcuencas 1 y 4, la cantidad de variables de entrada es mayor para la zona buffer de la subcuenca 4, debido a que se tienen mayor representación de todas las clases en general, además de una distribución distinta de esas representaciones. Para estas situaciones se observa que las métricas de clase explican en mayor medida los valores medios del contaminante.

5.4.2. Modelo RF

Estación SLC01

Para el caso de la subcuenca 1 (estación SLC01) los resultados del modelo de RF destacan la no linealidad para ciertas métricas, tanto a nivel de paisaje como a nivel de clase. No obstante, pueden realizarse algunas interpretaciones para un conjunto de métricas con comportamiento lineal. La conectividad física de la clase *Monte nativo* correlaciona directamente con PT. El *NT* correlaciona de forma directa con la representación de la clase Cultivos, y de forma inversa con la agregación de los parches de la clase *Herbáceo natural*. La Turbidez presenta comportamientos similares a los observados para el *NT*.

A nivel de la zona buffer, 500 metros, de la subcuenca 1 (estación SLC01), se destaca una no linealidad fuerte entre algunos índices de paisaje, *SHEI, LSI, GYRATE_MN*, y el *NT y* la *Turbidez*. A nivel de clase se destaca una relación directa entre la conectividad física de la clase *Monte nativo* y el PT. La conexión física de *Herbáceo natural* correlaciona positiva con *PT* y negativa con *NT.* La representación de la clase *Cultivos* se relación de forma directa con NT y Turbidez.

Estación PS01=SLC03

Para el caso de subcuenca 3 (estación PLS01=SLC03) los resultados del modelo de RF destacan la correlación directa de la densidad de parche a nivel de paisaje (PD) con PT y NT, y la relación inversa entre LPI y PT, NT y Turbidez. En cuanto a las métricas de clase se destacan relaciones negativas entre agregación de la clase *Herbáceo natural* y representación de la clase *Cultivos* con PT y NT. Además, se observa correlación negativa entre la conectividad física y la representación

de la clase Forestal con el PT. Por otro lado, la Turbidez presenta correlación directa con la conexión física de la clase *Herbáceo natural* y una relación inversa con la representación de la clase *Monte Nativo*.

A nivel de la zona buffer, 500 metros, de la subcuenca 3 (estación PS01=SLC03) los resultados del modelo de RF indican un comportamiento no lineal para los índices de paisaje *SHDI* y *SIEI* para PT y NT. Además, a nivel de paisaje el LPI correlaciona negativamente con el PT y NT. A nivel de clase, las variables de calidad de agua *PT, NT y Turbidez* se encuentran relacionadas con la composición y configuración en el paisaje de las clases *Área desnuda, Urbanización, Herbáceo Natural, Cultivos, Monte nativo y Forestal.* En este sentido la conectividad física de la clase *Cultivos* presenta una relación negativa con el PT y NT. La agregación de los parches de la clase *Cultivos* también presenta una relación negativa con el NT y una relación positiva con la Turbidez. La conectividad física de la clase *Forestal* presenta una relación negativa con PT y NT, mientras que la representación de esta clase sigue una relación directa con el NT. La conectividad física de la clase *Herbáceo natural* sigue una relación positiva con el PT, mientras que la representación de esta clase una relación directa NT. La representación de la clase *Urbanización* muestra una relación directa con NT, mientras que la agregación de esta clase tiene una relación inversa con la Turbidez. La representación de la clase *Monte Nativo* tiene una relación inversa con la *Turbidez*, mientras que para *Área desnuda* se observa una relación directa con la *Turbidez*.

Estación PS02

Para el caso de subcuenca 4 (estación PS02) los resultados del modelo RF destacan para los índices de paisaje relación directa entre *SHEI,* LSI y PD con PT, GYRATE_MN con NT y relaciones inversas entre CONTAG y LPI con NT. Para la turbidez se observan no linealidades de varios índices. A nivel de clase, el modelo señala relaciones positivas entre la conectividad física de los parches de la clase *Forestal* con el PT, y negativas con el NT; relaciones negativas entre la agregación de los parches de la clase *Herbáceo natural*, y positivas para la conectividad de los parches de esta clase con la Turbidez. Para la clase *Monte nativo*, se observan relaciones negativas entre la representación de la clase y la conectividad de los parches con el NT.

A nivel de la zona buffer, 500 metros, de la subcuenca 4 (estación PS02) los resultados del modelo indican comportamiento no lineal entre varios índices a nivel de clase y paisaje con NT y Turbidez. Para el caso de PT, se observan no linealidades para la conectividad de la clase *Área natural inundable y Urbanización*. A nivel de paisaje se destaca una relación negativa entre LPI y el PT. A nivel de clase, se destaca la relación negativa del PT y la representación de la clase *Cuerpos de agua*, la agregación y conectividad de la clase *Forestal* y *Herbáceo natural*. Por otro lado, se observan una relación positiva entre la representación de la clase *Urbanización* y el PT.

5.5. Discusión de los resultados

Considerando que los resultados obtenidos de los análisis no presentan una significancia estadística y, por tanto, no permiten realizar generalizaciones y deben ser evaluados con mucho cuidado, la discusión se limitará a señalar algunos puntos de contacto entre los resultados obtenidos mediante el modelo PLSR con los resultados señalados en la literatura.

Los análisis realizados muestran que la calidad del agua está relacionada con la proporción y la configuración del tipo de uso de suelo en la cuenca. Los resultados observados del PLSR a nivel de paisaje permiten señalar que los índices de relaciones de homogeneidad, dominancia y desagregación de los parches en el paisaje presentan relaciones marcas con las variables de calidad de agua, estas observaciones están alineadas con lo señalado en (Lee, S.-W. et al., 2009).

Para la subcuenca 1, se observó que la clase la composición y configuración de la clase *Monte nativo* es un factor significativo para explicar los menores valores de PT lo que muestra cierta similitud con lo señalado por Lee, S.-W. et al. (2009) y Li, N.X. et al. (2020) respecto a que los patrones de uso de suelo forestal que parecen correlacionar mejor con una mejor calidad del agua son aquellos menos fragmentados, bien agregados y conectados físicamente.

Si se observan los resultados obtenidos del PLSR para la subcuenca 3 a nivel de la zona buffer 500 m, la representación de *Urbanización* correlaciona inversa con PT y NT, mientras que la agregación y conectividad de esta clase tiene una relación directa con la Turbidez. Estos resultados están parcialmente alineados con las observaciones realizadas en Lee, S.-W. et al. (2009), donde se señala que áreas urbanas compactas y mayormente agregadas pueden minimizar los impactos adversos en las áreas naturales. Por otro lado, teóricamente se esperaría que la representación de la *Urbanización* correlaciones positivamente con el PT y NT si se tiene en cuenta que el punto de muestreo se encuentra a la altura de la ciudad de Florida, que si bien cuenta con una buena cobertura del sistema colector de saneamiento (mayor a 75%) no cuenta con una planta de tratamiento terciaria. Esta observación da lugar a cuestionamientos de si las fuentes puntuales de contaminación pueden ser vinculadas con métricas de clase o paisaje ajustando las escalas del análisis (por ejemplo, en este caso, zonas buffer de cuencas no incrementales).

Si bien los resultados de los análisis desarrollado entre la relación de métricas de paisaje y calidad de agua no permiten ser concluyentes ya que no presentan una significancia estadística que sustente las relaciones encontradas, algunos resultados encontrados son alentadores ya que se condicen con la literatura, y además el proceso realizado deja varias enseñanzas y aprendizajes para seguir avanzando y delinear una metodología que permita orientar la búsqueda de las relaciones entre la estructura del paisaje y la calidad de agua.

Algunos autores han abordado la búsqueda de parsimonia en las métricas de paisaje (Cushman, S.A. et al., 2008) y en la selección de un conjunto de métricas reducidas que representen la heterogeneidad espacial (Plexida, S.G. et al., 2014). Sin embargo, muchas de las métricas de paisaje se han concebido bajo la óptica de la ecología del paisaje, por lo que no todas las métricas podrán capturar o ser fáciles de relacionar con el fenómeno de estudio. Por tanto, en primera instancia se debe lograr una profunda y correcta interpretación de las métricas de paisaje y sus relaciones, a distintas escalas espaciales a nivel de clase y paisaje. Un paso más en este sentido, sería considerarse métricas combinadas que contemplen aspectos vinculados con la topografía (a través de los modelos digitales del terreno) y por ende el movimiento del agua dentro de las cuencas hidrológicas, como por ejemplo se realiza en (Staponites, L.R. et al., 2019).

Para lograr el cometido anterior, un paso importante es la automatización del cálculo de métricas, para eso se sugiere adoptar la librería implementada en el lenguaje R, la cual se base en el programa FRAGSTATS pero se encuentra en constante proceso de actualización debido a la amplia comunidad científica que lo utiliza en las ciencias biológicas y naturales, además de ser libre y de código abierto. Para incorporar métricas combinadas se puede utilizar los modelos digitales del terreno de IDEuy y la plataforma de análisis de datos geoespaciales de acceso libre y código abierto WhiteboxTools (Lindsay, J.B., 2021; Lindsay, JB., 2014).

Otro aspecto positivo y distintivo que deja el análisis realizado es el de utilizar una técnica lineal (PLSR) y otra no lineal (RF), aplicando sobre cada modelo un enfoque de teoría de juegos (SHAP) para explicar el resultado de los mismos. Sin embargo, estas herramientas pueden ser automatizadas en un nivel más profundo, en la medida de lo posible, para la extracción de resultados para su interpretación, por ejemplo, fijando valores umbrales del SHAP que se consideran de relevancia en cuanto al impacto de salida del modelo y tendencias de las variables que indiquen correlaciones entre las mismas de forma de lograr obtener un set de variables depurado para el análisis.

Otro aspecto importante a destacar es la posibilidad de incorporar la interdisciplinaridad, en particular en el dominio de la ecología, para lograr una comprensión funcional y sistémica más profunda.

6. Resumen de los resultados y Conclusiones

6.1. Resumen de los resultados

Los resultados más importantes del proyecto se pueden sintetizar como sigue:

- OE1: Se logró desarrollar una metodología hibrida (con técnicas de aprendizaje automático alimentadas de información físicamente basada) para la imputación de las diferentes variables ambientales consideradas en este proyecto. Dichas variables, además de las variables de calidad de agua, incluyen también las variables hidrológicas y meteorológicas, las cuales se utilizaron como "variables de ayuda" para la imputación de las variables de calidad de agua basándonos en las correlaciones calculadas y en los procesos físicos a escala de cuenca. El diagrama de flujo de la metodología desarrollada, con su descripción detallada, se encuentra en el Informe Técnico (Fig. 2.10). Dicha metodología se ejecutó 78 veces (una iteración por cada una de las variables donde se detectaron valores faltantes); al concluir el proceso de imputación, se obtuvo un conjunto de modelos capaces de imputar de la mejor forma posible cada serie temporal. Para la selección de las mejores imputaciones se usó la métrica NSE como función objetivo y, para validar los resultados, se usaron las métricas KGE y PBIAS. La mayoría de las variables presenta un NSE positivo, o sea, el modelo aquí desarrollado para su imputación tiene un mejor desempeño de la media observada. Además, se puede destacar que más del 75% de las variables presentan un NSE > 0.5. El KGE y el PBIAS confirman los buenos resultados obtenidos con el NSE. En particular, KGE muestra una distribución similar a la obtenida con el NSE y el |PBIAS| presenta valores menores de 15 para el 75% de las variables, confirmando los muy buenos resultados obtenidos. El modelo IDW fue el mejor para 27 variables, seguido del modelo de ensamble Hubber Regressor + SC (conjunto de datos modificado según la distancia de las estaciones) usado para 9 variables.

- OE2: Los cambios temporales y espaciales de LULC en la cuenca del río Santa Lucía Chico fueron abordados por diversos mapas generados por fuentes distintas para los años 2000, 2008, 2011, 2015, 2016 y 2018. Estas diferencias en fuentes, en algunos casos, implicaron diferencias en metodologías de elaboración de mapas. Una primera aproximación para salvar la diferencia metodológica de los mapas fue definir categorías de uso de suelo comunes para unificar las distintas categorías definidas en los mapas, de esta forma se definieron las clases: *Área desnuda, Área natural inundable, Cuerpos de agua, Cultivos, Forestal, Herbáceo natural, Monte nativo y Urbanización.* Se constató que la diferencia entre metodologías de generación de mapas dificulta sobre todo evaluar los cambios temporales en el uso del suelo. No obstante, las tendencias espaciales generales pueden ser bien capturadas por estos mapas. En ese sentido, se observa que la clase *Herbáceo*

natural (55%) y *Cultivos (35%)* son los que dominan el paisaje de la cuenca del Santa Lucía Chico, representando en conjunto en términos medios más del 85% del área total de la cuenca. La clase *Cultivos* se encuentra mayormente concentrada hacia la zona suroeste, cercana a la mayor concentración y representación de las clases *Urbanización* y *Cuerpos de agua*, mientras que el *Herbáceo natural* predomina en la zona centro este y como "tapiz" de fondo sobre todo el paisaje. El desarrollo de *Forestal* se produce básicamente en la zona centro oeste con mayor consolidación en los últimos años y el *Monte nativo* tiene un desarrollo marcado siguiendo la principal red de drenajes de la cuenca. Para de salvar las diferencias metodológicas y contar con mapas bases consistentes entre sí para investigar las relacionar de los usos del suelo con la calidad del agua, se desarrolló una metodología para la generación de mapas de uso de suelo con *Google Earth Engine*. Estos mapas generados para los años 2014, 2016, 2018 y 2020 indicaron buena precisión global (coeficiente kappa 0.72), siendo las clases menos representadas las que aportan mayores componentes ruidosas. Para todas las clases de los mapas generados, los cambios temporales en términos medios porcentuales son pequeñas (0,1 a 2,0%), no obstante, en términos de área para las clases más representadas (*Herbáceo natural y Cultivos*) se tienen cambios medios de 5000 has, mientras que las clases *Forestal, Monte Nativo, Área natural inundable y Urbanización* tienen cambios medios de 500 has. Las variaciones espaciales para los mapas generados indican desarrollo de la forestación en la zona noreste de la cuenca y en menor proporción en la zona centro de la cuenca, esta transformación comienza a desarrollarse en 2016 y a partir del año 2018 se encuentra consolidada.

- OE3: Considerando el alto número de variables a considerar en las seis estaciones de monitoreo, se desarrolló una aplicación donde se muestran gráficamente los resultados de los diferentes análisis que se corrieron para evaluar la variabilidad espacio-temporal de las diferentes variables de calidad del agua. Esta aplicación nos permitió comparar los resultados y sacar conclusiones con más facilidad. En particular, desde el punto de vista espacial, se puede afirmar que existen dos grupos de comportamiento diferentes: los tres sitios de monitoreo ubicados en el embalse de Paso Severino muestran patrones diferentes a los que caracterizan las estaciones ubicadas aguas arriba del embalse. Además, las estaciones ubicadas en el embalse muestran valores de nutrientes más elevados de las estaciones que se encuentran en el cauce principal del río Santa Lucía, justificando el nivel hipereutrófico detectado en el embalse por estudios anteriores. Asimismo, en la estación PS01=SLC03, ubicada aguas abalo de la ciudad de Florida, se detectaron valores elevados de Nitrógeno Total, justificando el hecho que las áreas urbanizadas son caracterizadas por diferentes fuentes de nitrógeno como, por ejemplo, deposición atmosférica, la aplicación de fertilizantes para el césped, las aguas residuales

y la infraestructura de alcantarillado con fugas. Desde el punto de vista temporal, la Temperatura del Agua y el Oxígeno Disuelto son los únicos contaminantes que muestran una fuerte estacionalidad intra e interanual, mientras que no podemos identificar un patrón claro para los otros contaminantes. Del análisis de estacionalidad, la Turbidez presentó una muy leve tendencia en mostrar valores más altos en invierno, debido a la mayor escorrentía y a los eventos extremos de lluvias más intensos que determinan un desprendimiento y una exportación más importantes de las partículas del suelo al cuerpo de agua.

- OE4: Considerando el alto número de escenarios a considerar, también en este caso se desarrolló una aplicación que nos permitió comparar los resultados y sacar conclusiones con más facilidad. Los resultados de los análisis desarrollados entre la relación de métricas de paisaje y calidad de agua no permiten ser concluyentes ya que no presentan significancia estadística que sustente las relaciones encontradas. No obstante, algunos resultados parecen alentadores ya que se condicen con la literatura, además, el proceso realizado deja varias enseñanzas y aprendizajes para seguir avanzando en establecer una metodología que permita orientar la búsqueda de las relaciones entre la estructura del paisaje y la calidad de agua. En este sentido, en primera instancia, se debe lograr una profunda y correcta interpretación de las métricas de paisaje y sus relaciones, a distintas escalas espaciales agregadas a nivel de clase y paisaje. Además de las métricas que surgen de la disciplina de la ecología del paisaje, es necesario explorar métricas combinadas que contemplen aspectos de la estructura del paisaje con la topografía, y, por ende, con el movimiento del agua dentro de las cuencas hidrológicas. Un aspecto positivo y distintivo que deja el análisis realizado es la utilización de técnicas lineales y no lineales (PLSR y RF, respectivamente) sobre las cuales se aplican un enfoque de la teoría de juegos (SHAP) para una mejor explicación y más fácil interpretación de los resultados. Otro aspecto importante a destacar es la posibilidad de incorporar la interdisciplinaridad, en particular, en el dominio de las ciencias biológicas y naturales, para lograr una comprensión funcional y sistémica del paisaje más profunda.

6.2. Conclusiones

A partir del análisis de los datos disponibles, el objetivo principal de este proyecto fue la evaluación de cómo influyen los cambios en el uso del suelo en la calidad del agua de la cuenca del río Santa Lucía a lo largo del tiempo. Los resultados obtenidos contribuirán principalmente al conocimiento de los cambios en el uso del suelo en una cuenca de rápido desarrollo con un acelerado deterioro de la calidad del agua, lo cual es usual en muchos países en vías de desarrollo como Uruguay.

Este proyecto pretende ser un paso fuerte en la consolidación de la línea de investigación y desarrollo local sobre hidroinformática, donde converjan los esfuerzos de los grupos del IMFIA y del InCo. En la medida de consolidar dicha línea, los avances en este proyecto piloto que considera una cuenca prevalentemente agropecuaria con diferentes peculiaridades hidrológicas e hidráulicas permitirán prepararse de la mejor manera posible para lograr el máximo aprovechamiento para las otras cuencas semejantes del país que presentan problemáticas similares.

Los usuarios directamente beneficiados de los resultados de dicha línea de investigación serán todos los productores instalados en la cuenca del río Santa Lucia, agricultores, productores lecheros, población residente, turistas, entre otros. Además, las instituciones que tienen en sus competencias aspectos de gestión vinculadas a la cuenca o recursos hídrico también se beneficiarán directamente (MGAP, DINAGUA, DINAMA, OSE, principalmente). Además del beneficio económico-social, otro aspecto importante será el aporte a la preservación y salvaguardia de los ecosistemas naturales (praderas, humedales y bosques nativos). En este sentido la sociedad en su conjunto es el beneficiario directo. El conocimiento de qué categoría de uso de suelo tiene el mayor efecto sobre la calidad de agua y de cuáles son los principales compuestos que identifican categorías específicas de uso del suelo, podrá permitir a las autoridades competentes (MVOTMA y MGAP) armar un plan de acción para formular y ejecutar las operaciones principales para controlar, detener y revertir el proceso de deterioro de la calidad de agua en la cuenca hidrográfica del río Santa Lucía.

7. Actividades de difusión

7.1. Publicaciones en revistas científicas

El primer trabajo sobre la metodología basada en técnicas de aprendizaje automático desarrollada para la imputación de los datos de calidad de agua en la cuenca de estudio fue publicado en la revista Sustainability (MDPI), indexada en Scopus y Web of Science y con factor de impacto de 2.576.

El artículo se puede encontrar online en: https://www.mdpi.com/2071-1050/13/11/6318

Además, el equipo de trabajo está redactando los siguientes artículos que serán enviados en dos revistas científicas indexadas con alto factor de impacto:

- Un artículo científico sobre la metodología híbrida (físicamente basada y basada en datos) desarrollada para la imputación de los datos hidrológicos, meteorológicos y de calidad de agua (resultado del OE1). Este artículo será enviado a la revista Environmental Modelling and Software, indexada en Scopus y Web of Science con un factor de impacto de 4.8. El probable título del artículo será "*Missing data imputation in environmental science: a multi-domain and hybrid approach*".
- Un artículo científico sobre las relaciones entre las categorías de uso de suelo y las variables de calidad de agua (resultado del OE4). Este artículo será enviado a la revista Science of the Total Environment, indexada en Scopus y Web of Science con un factor de impacto de 6.5. El probable título del artículo será "*Influence of land-use change on surface-water quality*".

8. Actividades de divulgación

8.1. Entrevista “La Diaria”

El 2 de octubre de 2020, la responsable científica del proyecto fue entrevistada por el periódico "La Diaria" sobre la temática central del proyecto: correlación entre calidad de agua y uso/cobertura del suelo en la cuenca del Santa Lucía.

Link al artículo: https://ladiaria.com.uy/ciencia/articulo/2020/11/santa-lucia-encuentran-una-fuerte-correlacion-entre-altas-cantidades-de-fosforo-total-y-la-presencia-de-agricultura-y-ganaderia-en-su-cuenca/

8.2. Participación al webinar organizado por el CTAgua

El 21 de octubre de 2020, la responsable científica del proyecto fue invitada a participar como expositora al webinar organizado por CTAgua sobre “Monitoreo y fuentes difusas de contaminación en la cuenca del Santa Lucía”. La charla impartida fue sobre la temática del proyecto “*Estudio de las fuentes difusas de contaminación en la cuenca del río Santa Lucía*”, enfocándose en el impacto del uso del suelo sobre la calidad de agua a nivel de cuenca.

El artículo del evento se encuentra en: https://ctagua.uy/2021/01/08/webinar-monitoreo-y-fuentes-difusas-de-contaminacion-en-la-cuenca-del-santa-lucia-3/

8.3. Audiovisual de proyecto

Una parte del equipo de trabajo participó en la producción de un audiovisual de proyecto, que estará disponible y pronto para la difusión en agosto 2021. Pensamos al audiovisual como una herramienta dinámica y atractiva para llegar con contenido de divulgación científica importante, como el tratado en este proyecto, a la comunidad y a todos los actores beneficiaros de los resultados de nuestro trabajo.

9. Referencias

AA. VV. Water Quality in the Americas - Risks and opportunities. ISBN: En trámite, 2019.

Achkar, M., Dominguez, A., Pesce, F. Cuenca del Río Santa Lucía – Uruguay, 2012.

Achkar, M., Dominguez, A., Pesce, F. Cuencas hidrográficas del Uruguay, 2013.

Arocena, R., Chalar, G., Perdomo, C., Fabián, D., Pacheco, J.P., González, M., Olivero, V., Silva, M., Etchebarne, V. Impacto de la producción lechera en la calidad de los cuerpos de agua. Augmdomus, 5, 42-63, 2013.

Aubriot, L., Delbene, L., Haakonsson, S., Somma, A., Hirsch, F., Bonilla, S. Evolución de la eutrofización en el Río Santa Lucía: influencia de la intensificación productiva y perspectivas. INNOTEC, 14, 7-16, 2017.

Bergstra, James & Bardenet, R. & Kégl, Balázs & Bengio, Y. Algorithms for Hyper-Parameter Optimization, 2011.

Bonilla, S. Cianobacterias Planctónicas del Uruguay; Manual para la identificación y medidas de gestión. Montevideo: UNESCO, S. Bonilla, 2009.

Bonilla, S., Haakonsson, S., Somma, A., et al. Cianobacterias y cianotoxinas en ecosistemas límnicos de Uruguay. INNOTEC, 10(9-22), 2015.

Bowerman, B., O'Connell, R., Pronósticos, Series de Tiempo Y Regresión: Un Enfoque Aplicado, 2007.

Breiman, L. Random Forests. Machine Learning, 45(1), 5-32, 2001.

Buuren, S., Karin Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45: 1-67, 2011.

Calijuri, M.L., Castro, J. De S., Costa, L.S., Assemany, P.P., Mattos Alves, J.E. Impact of land use/land cover changes on water quality and hydrological behavior of an agricultural subwatershed. Environ. Earth Sci., 74, 5373-5382, 2015.

Chang, C., Lin, C., LIBSVM: A Library for Support Vector Machines, 2001.

Choi, S., Kim, T., Yu, W., Performance Evaluation of RANSAC Family, 2009. Disponible online: http://www.bmva.org/bmvc/2009/Papers/Paper355/Paper355.pdf (último acceso: 8 de octubre 2020).

Dang, X., Peng, H., Wang X., Zhang, H. Theil-Sen Estimators in a Multiple Linear Regression Model, 2009. Disponible online: http://home.olemiss.edu/~xdang/papers/MTSE.pdf (último acceso: 8 de octubre 2020).

Decreto 55/015. Aprobación de la selección del área natural protegida denominada "humedales de Santa Lucía. 2015.

DINAMA - División de Información Ambiental, DINAMA. Mapa de uso y cobertura del suelo del Uruguay, año 2016. 2017.

DINAMA-OAN. Cuencas Hidrográficas - Nivel 2. Disponible online: https://www.dinama.gub.uy/geoservicios/ (último acceso: 8 de octubre 2020).

DINOT (a). Avances en el sistema de clasificación de la cobertura del suelo en la Cuenca del río Santa Lucía, 2016.

DINOT (b). Está disponible un nuevo mapa satelital de cobertura de nuestro territorio, 2016.

DINOT. Clasificación de usos y coberturas para la Cuenca del Río Santa Lucía. División de Información Ambiental, DINAMA-MVOTMA, 2017.

DINOT. Mapas de cobertura del suelo del Uruguay de los años 2000, 2008, 2011 y 2015. Disponible online: https://sit.mvotma.gub.uy/websdatos/cobertura.html (último acceso: 8 de octubre 2020).

Drucker, H. Improving Regressors using Boosting Techniques, 1997.

Durbin, J. and Koopman, S. J. Time Series Analysis by State Space Methods, Oxford University Press, 2001.

FAO. Atlas de cobertura del suelo de Uruguay Cobertura de suelo y cambios 2000-2011, 2015. Disponible online: http://www.fao.org/family-farming/detail/es/c/288333/ (último acceso: 8 de octubre 2020).

Farebrother, R. W. Further results on the mean square error of ridge regression, Journal of the Royal Statistical Society, Series B (Methodological), 38, 248-250., 1976. Disponible online: https://www.jstor.org/stable/2984971 (último acceso: 8 de octubre 2020).

Ferrari, G., Pérez, M.C., Dabezies, M., Míguez, D., Saizar, C. Planktic Cyanobacteria in the Lower Uruguay River, South America. Fottea, 11, 225–234, 2011.

Forio, M.A.E., Landuyt, D., Bennetsen, E., Lock, K., Nguyen, T.H.T., Ambarita, M.N.D., Musonge, P.L.S., Boets, P., Everaert, G., Dominguez-Granda, L., Goethals, P.L.M. Bayesian belief network models to analyse and predict ecological water quality in rivers. Ecological Modelling, 312, 222-238, 2015.

Fortin, M. -J. and Dale, M. Spatial Analysis: A guide for Ecologists, Cambridge: Cambridge University Press, 2006.

Fuller, W. A. Introduction to Statistical Time Series, second ed., New York: John Wiley and Sons, 1996.

Gao, B.C. NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space. Remote Sensing of Environment: 58, 257-266, 1996.

Geurts, P.; Ernst D.; and Wehenkel L. Extremely randomized trees. Machine Learning, 63(1), 3-42, 2006.

Gorgoglione, A.; Alonso, J.; Chreties, C.; Fossati, M. IOP Conf. Ser.: Earth Environ. Sci., 2020, 612 012002.

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, 4692. https://doi.org/10.3390/su12114692

Goyenola, G., Meerhoff, M., Teixeira-de Mello, F., González-Bergonzoni, I., Graeber, D., Fosalba, C., Vidal, N., Mazzeo, N., Ovesen, N.B., Jeppesen, E., Kronvang, B. Phosphorus dynamics in lowland streams as a response to climatic, hydrological and agricultural land use gradients. Hydrol. Earth Syst. Sci. Discuss., 12, 3349 – 3390, 2015.

IDE. Modelo digital del terreno IDEuy. Disponible en: https://visualizador.ide.uy/

JICA-MVOTMA. Proyecto sobre el Control de la Contaminación del Agua y la Gestión de la Calidad del Agua en la Cuenca del Río Santa Lucía. 2011.

Jones H., Vaughan R., Remote sensing of vegetation: principles, techniques, and applications. Oxford University Press, New York, p 353, 2010.

Kändler, M., Blechinger, K., Seidler, C., Pavlů, V., Šanda, M., Dostál, T., Krása, J., Vitvar, t., Štich, M. Impact of land use on water quality in the upper Nisa catchment in the Czech Republic and in Germany. Sci. Tot. Env., 586, 1316-1325, 2017.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, NIPS, 2017.

Kevin McGarigal. FRAGSTATS HELP v4.2. Sole proprietor, LandEco Consulting. Professor, Department of Environmental Conservation. University of Massachusetts, Amherst, 2015. Disponible en: https://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf. Último acceso 06/06/2021.

Kwiatkowski, D.; Phillips, P. C. B.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54 (1-3): 159-178, 1992.

Li NX, Xu JF, Yin W, Chen QZ, Wang J, Shi ZH. Effect of local watershed landscapes on the nitrogen and phosphorus concentrations in the waterbodies of reservoir bays. Sci Total Environ, 2020, 716:137132.

Lindsay, JB. 2014. The Whitebox Geospatial Analysis Tools project and open-access GIS. Proceedings of the GIS Research UK 22nd Annual Conference. The University of Glasgow, 16-18 April, DOI: 10.13140/RG.2.1.1010.8962

Lindsay, JB. 2021. WhiteboxTools Version 1.5.0. University of Guelph, Guelph, Canada. Disponible en: https://jblindsay.github.io/wbt_book/intro.html . Consultado por última vez 12/06/2020.

Lundberg, S & Lee, S. A Unified Approach to Interpreting Model Predictions, 2017.

McGarigal, K. Landscape Pattern Metrics. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels), 2014. https://doi.org/10.1002/9781118445112.stat07723

McGarigal, K., SA Cushman, E Ene. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst, 2012. Available at the following web site: http://www.umass.edu/landeco/research/fragstats/fragstats.html

MGAP. Mapa integrado de cobertura/uso del suelo del Uruguay año 2018. Disponible online: https://www.gub.uy/ministerio-ganaderia-agricultura-pesca/comunicacion/publicaciones/mapa-integrado-coberturauso-del-suelo-del-uruguay-ano-2018 (último acceso: 8 de octubre 2020).

MGAP. Modelo Digital del Terreno de la Recursos Naturales Renovables (RENARE). Obtenido a partir de comunicación con IDEuy (maria.morales@ide.gub.uy) el 18/09/2020.

MGAP. Modelo digital del terreno RENARE, 2020. Disponible en: https://www.gub.uy/ministerio-ganaderia-agricultura-pesca/tramites-y-servicios/servicios/modelo-digital-terreno

Miller, J.D., Kim, H., Kjeldsen, T.R., Packman, J., Grebby, S., Dearden, R. Assessing the impact of urbanization on storm runoff in a peri-urban catchment using historical change in impervious cover. J. Hydrology, 515, 59 – 70, 2014.

Montgomery, D. C., Peck, E. A., Vining, G. G. Introduction to linear regression analysis (4th ed.), New York: Wiley, 2012.

Mori, N., Debeljak, B., Škerjanec, M., Simčič, T., Kanduč, T.c, Brancelj, A. Modelling the effects of multiple stressors on respiration and microbial biomass in the hyporheic zone using decision tres. Water Research, 149, 9-20, 2019.

Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Soil Water Div. Asabe 2007, 50, 885–900.

Mucherino A., Papajorgji P.J., Pardalos P.M. *k*-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY, 2009.

MVOTMA. Atlas de la cuenca del Santa Lucía. DINOT, 2016.

MVOTMA. Plan Nacional de Aguas. 2017.

Namugize, J.N., Jewitt, G., Graham, M. Effects of land use and land cover changes on water quality in the Umngeni river catchment, South Africa. Physics and Chemistry of the Earth, 105, 247-264, 2018.

Owen, A. B. A robust hybrid of lasso and ridge regression. 2006. Disponible online: https://statweb.stanford.edu/~owen/reports/hhu.pdf (último acceso: 8 de octubre 2020).

Oyhantçabal, G., Narbondo, I. Radiografía del agronegocio sojero uruguayo. Alternativa. Revista de Estudios Rurales, 1, 1-30, 2014.

Pacheco, J.P., Arocena, R., Chalar, G., García, P., González, Piana, M., Fabián, D., Olivero, V. Evaluación del estado trófico de arroyos de la cuenca de Paso Severino (Florida, Uruguay) mediante la utilización del índice biótico TSI-BI. Augmdomus, 4, 80-91, 2012.

Pérez-Gutiérrez, J.D., Paz, J.O., Tagert, M.L.M. Seasonal water quality changes in on-farm water storage systems in a south-central U.S. agricultural watershed. Agric. Water Manag., 187, 131 – 139, 2017.

Petraglia, C., Dell'Acqua, M., Pereira, G., Yussim, E. Anuario OPYPA 2019. Mapa integrado de cobertura / uso del suelo del Uruguay, año 2018, pp. 523 - 531. Disponible online: https://www.gub.uy/ministerio-ganaderia-agricultura-pesca/comunicacion/publicaciones/mapa-integrado-coberturauso-del-suelo-del-uruguay-ano-2018 (último acceso: 8 de octubre 2020).

Pirouz, D. An Overview of Partial Least Squares. SSRN Electronic Journal, 2006. DOI: 10.2139/ssrn.1631359

Plexida, S.G., Sfougaris, A.I., Ispikoudis, I.P., Papanastasis, V.P. Selecting landscape metrics as indicators of spatial heterogeneity—A comparison among Greek landscapes, International Journal of Applied Earth Observation and Geoinformation, 2014, 26, 26-35.

Ríos, A. Implementación de un modelo hidrodinámico tridimensional en el embalse de Paso Severino. Aportes para la modelación de calidad de agua. Tesis de Maestría en Ingeniería Ambiental. Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA), Facultad de Ingeniería (FIng), Universidad de la República (UdelaR), 2019.

Ritter, W.F., Shirmohammadi, A. Agricultural nonpoint source pollution. ISBN: 1-56670-222-4, 2001.

Rodríguez-Gallego, L., Achkar, M., Defeo, O., Vidal, L., Meerhoff, E., Conde, D. Effects of land use changes on eutrophication indicators in five coastal lagoons of the Southwestern Atlantic Ocean. Estuar. Coast. Shelf Sci., 188, 116-126, 2017.

Rosecrans, C.Z., Nolan, B.T., Gronberg, J.M. Prediction and visualization of redox conditions in the groundwater of Central Valley, California. Journal of Hydrology, 546, 341-356, 2017.

Samuel A. Cushman, Kevin McGarigal, Maile C. Neel. Parsimony in landscape metrics: Strength, universality, and consistency. Ecological Indicators, 2008, 8(5), 691-703.

Sang-Woo Lee, Soon-Jin Hwang, Sae-Bom Lee, Ha-Sun Hwang, Hyun-Chan Sung. Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics. Landscape and Urban Planning, 2009, 92(2), 80-89.

Sen Xu, Si-Liang Li, Jun Zhong, Cai Li. Spatial scale effects of the variable relationships between landscape pattern and water quality: Example from an agricultural karst river basin, Southwestern China. Agriculture, Ecosystems & Environment, 2020, 300, 106999.

Shi, P., Zhang, Y., Li, Z.B., Li, P., Xu, G.C. Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales. Catena, 151, 182 – 190, 2017.

Shoemaker, C.M., Ervin, G.N., Diorio, E.W. Interplay of water quality and vegetation in restored wetland plant assemblages from an agricultural landscape. Ecol. Eng., 108, 255 – 262, 2017.

Staponites, L.R., Barták, V., Bílý, M. et al. Performance of landscape composition metrics for predicting water quality in headwater catchments. Sci. Rep. 2019, 9, 14405.

Templ, M., Kowarik, A., Filzmoser, P. Iterative stepwise regression imputation using standard and robust methods. Journal of Computational Statistics and Data Analysis. Vol. 55, pp. 2793-2806. 2011.

Thornhill, I., Batty, L., Death, R.G., Friberg, N.R., Ledger, M.E. Local and landscape scale determinants of macroinvertebrate assemblages and their conservation value in ponds across an urban land-use gradient. Biodiversity and Conservation, 26, 1065-1086, 2017.

Thornhill, I., Ho, J.G., Zhang, Y., Li, H., Ho, K.C., Miguel-Chinchilla, L., Loiselle, S.A. Prioritising local action for water quality improvement using citizen science; a study across three major metropolitan areas of China. Sci. Tot. Env., 584-585, 1268-1281, 2017.

Tipping, M. E. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research, Vol. 1, 2001.

Uuemaa, E., Antrop, M., Roosaare, J., Marja, R., Mander, Ü. Landscape Metrics and Indices: An Overview of Their Use in Landscape Research. Living Rev. Landscape Res., 2009, 3, 1.

Wijesiri, B., Deilami, K., Goonetilleke, A. Evaluating the relationship between temporal changes in land use and resulting water quality. Env. Pollut., 234, 480-486, 2018.

Xu, G., Li, P., Lu, K., Tantai, Z., Zhang, J., Ren, Z., Wang, X., Yu, K., Shi, P., Cheng, Y. Seasonal changes in water quality and its main influencing factors in the Dan River basin. Catena, 173, 131 – 140, 2019.

Xu, H. A Study on Information Extraction of Water Body with the Modified Normalized Difference Water Index (MNDWI). Journal of Remote Sensing, Vol. 9, pp. 589-595, 2005.

Xu, H. Extraction of urban built-up land features from Landsat imagery using a thematic-oriented index combination technique. Photogrammetric Engineering & Remote Sensing, 73: 1381-1391, 2007.

Zhen-Gang Ji. Hydrodynamics and Water Quality: Modeling Rivers, Lakes, and Estuaries. ISBN: 978-0-470-13543-3, 704 pages, 2008.