Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Análisis de patrones temporales en humedales del Delta del Paraná mediante el estudio de series de tiempo

Tesis de Licenciatura en Ciencias Matemáticas

Jesica Maia Numerosky

Director: Rafael Grimson

Buenos Aires, septiembre 2022

ABSTRACT EN CASTELLANO

En esta tesis analizamos mediante algoritmos de aprendizaje automático no supervisado los diferentes patrones de vegetación y de inundación que ve el sensor MODIS a bordo del satélite Aqua. El área de estudio elegida es el Sitio Ramsar Delta del Paraná y nuestra fuente de datos es la serie de tiempo del Índice Diferencial Normalizado de Vegetación (NDVI). Utilizamos, para nuestro análisis, técnicas de reducción de la dimensión en primer lugar, de *clustering* en segundo lugar y, por último, de posprocesamiento. Los resultados son comparados con los mapas de unidades de humedal de alta precisión realizados con anterioridad de manera de constatar qué información se puede extraer de esta serie de tiempo y comparamos distintas técnicas de reducción de la dimensionalidad. Pretendemos hacer un aporte en términos de la variedad de metodologías y mejoras en su uso con respecto a los trabajos anteriores en estos mismos temas, además de su aplicación en un área distinta.

Palabras claves: Humedales, aprendizaje automático, clustering, teledetección.

ABSTRACT IN ENGLISH

In this thesis we analyze, using unsupervised learning techniques, vegetation and flood patterns detected by the MODIS sensor, on board the Aqua satellite. The chosen study area is the Paraná Delta Ramsar Site and our data source is the Normalized Difference Vegetation Index (NDVI) time series. For the analyisis, we use dimensionality reduction, clustering and post-processing techniques. The results are compared with previously-made high-resolution wetland unit maps in order to verify what information can be extracted from this time series, and we also compare different dimensionality-reduction techniques. We intend to diversify and improve methodologies that are already used in previous works, and also apply them on a different area.

Keywords: Wetlands, machine learning, clustering, remote sensing.

AGRADECIMIENTOS

A Rafa, mi director, por ayudarme a canalizar mis ganas de estudiar matemática hace muchos años; por mostrarme que la matemática y la ecología pueden cruzarse de maneras hermosas y por su acompañamiento en la elaboración de esta tesis, que cierra un círculo que comienza en la UNSAM, pasa por la UBA y termina en ambos lugares.

A Patricia y Daniela, por ofrecerme su mano amorosa y su ojo de experta cada vez que lo necesité.

A Agus, Maite y Celeste, que me apoyaron sin dudarlo cada vez que lo necesité; durante los momentos de entusiasmo y de crisis.

A mis colegas de Eryx, con quienes aprendo todos los días. Esta tesis fue una excusa para aprender a programar y ustedes me acompañaron y me enseñaron muchísimo en el proceso. Gracias especialmente a Agus, Facu, Chou, Caro, Mati y Laski.

A mis compañeres de Exactas, que hicieron inolvidable mi paso por la universidad pública. Desde las tardes de estudio hasta los debates y asambleas; todas esas experiencias me formaron como matemática, como militante, como cooperativista, como docente y como persona.

A toda mi familia y en particular a mi abuelo, de quien heredé su gusto por la matemática; y a mi hermana, compañera incondicional.

A Manuela y Jean-Philipe, que enseñan matemática con amor.

Para abordar los problemas ambientales es necesario lograr una verdadera articulación de las diversas disciplinas involucradas. (...) No se trata de aprender "más cosas", sino de "pensar de otra manera" los problemas que se presentan en la investigación.

Sería más correcto decir que "la realidad no es disciplinaria" entendiendo por tal que la realidad no presenta sus problemas cuidadosamente clasificados en correspondencia con las disciplinas que han ido surgiendo en la historia de la ciencia.

– ***Rolando García, "Interdisciplinariedad y sistemas complejos"*** *[10]*

Índice general

1. INTRODUCCIÓN

América del Sur es un continente que tiene una proporción enorme de su superficie cubierta por humedales. Los mismos representan más del 20 % de su superficie, mucho más que el porcentaje estimado a nivel mundial (6,2–7,6 %). Además de eso, la heterogeneidad ambiental del continente genera una gran variedad de tipos de humedales, gracias a su extensión a nivel territorial, su amplio rango de altitudes y su diversidad geológica y climática, influida por los océanos Pacífico y Atlántico. La abundancia de humedales asociados a planicies de inundación de grandes ríos como el Amazonas, el Orinoco y el Paraná-Paraguay llevó incluso a que Sudamérica sea nombrado "el continente fluvial". Existe también una miríada de humedales alimentados por aguas subterráneas, lluvias y nieve, además de miles de kilómetros de humedales costeros [18].

Los humedales juegan un rol clave en los ciclos hidrológicos y biogeoquímicos, alojando una gran parte de la biodiversidad mundial, y proveen múltiples servicios a la humanidad. Sin embargo, según el Grupo de Examen Científico y Técnico de la Convención de Ramsar, durante el Siglo XX la extensión total de los humedales a nivel global decreció entre un 64 % y un 71 % [22]. Más aún: las pérdidas y degradación de los humedales continúan a lo largo y ancho del planeta a causa de presiones que toman la forma de modificaciones en el uso del suelo, cambios hidrológicos, explotación intensiva de recursos y contaminación. A pesar de que se consideraba que la mayor parte de los humedales de Sudamérica se encontraban en buenas condiciones relativas hasta hace algunas décadas, en años recientes su integridad comenzó a ser amenazada por el cambio climático y presiones en el uso del suelo, fundamentalmente para la agricultura y la ganadería [18].

A pesar de la importancia de los humedales, Argentina aún no cuenta con un inventario con información detallada acerca de su extensión, estado de conservación y tipos de humedales; y tampoco existen planes de monitoreo a implementarse en el mediano o largo plazo. Más aún, existen "huecos de información" a este respecto para los países del hemisferio sur [15].

Es crítico tener mediciones sistemáticas y consistentes de las variables biogeofísicas en los humedales, y la teledetección es una herramienta clave para abordar este tema. El uso de la misma para mejorar el conocimiento, monitoreo e inventariado de los humedales siempre fue atractivo por el hecho de que ayuda a superar dificultades en el acceso y cobertura de estos ecosistemas [18]. La teledetección es menos costosa que la investigación basada únicamente en trabajo de campo y provee información sobre un rango más amplio de escalas temporales y espaciales.

Hay disponible un espectro enorme de datos satelitales: los sensores remotos a bordo de satélites y vehículos aéreos nos proveen variables físicas muy valiosos de la Tierra. La teledetección nos provee un conjunto de herramientas para determinar la estructura y algunos aspectos de la función de paisajes heterogéneos [7].

Por este mismo motivo es que, en el marco de la elaboración de un Inventario Nacional de Humedales, la teledetección juega un papel clave, y es en este sentido que nos proponemos hacer un aporte.

A pesar de que el sistema terrestre es observado continuamente desde el espacio desde hace más de cuarenta años, la densidad de sistemas de observación satelital operativos en las últimas dos décadas es mayor que la de todos los años anteriores juntos. Por esto

último, el volumen de datos ofrecidos es muy grande. Sin embargo, la capacidad para utilizarlos es bastante limitada, llegando al punto que solo el 10 % de los datos adquiridos es requerido por algún usuario [23]. Creemos que la capacidad de los algoritmos de aprendizaje automático (*machine learning*) para ayudarnos a analizar grandes volúmenes de datos, así como también el poder de cómputo creciente de nuestra infraestructura puede colaborar en un mejor aprovechamiento de los datos que los sensores nos ofrecen.

Para esta tesis nos basamos en trabajos anteriores realizados en los humedales de nuestro país en general y en la región del río Paraná en particular por miembros del Laboratorio de Ecología, Teledetección y Ecoinformática del Instituto de Investigaciones e Ingeniería Ambiental de la Universidad Nacional de San Martín, además de otras publicaciones y tesis. Elegimos aquellas que emplean sensores remotos de resolución media para poder hacer aportes en cuanto a las técnicas de reducción de dimensionalidad y aprendizaje no supervisado (*clustering*). Nuestro objetivo es comparar dos formar de reducir la dimensionalidad en conjuntos de datos grandes (puesto que se trata de series de tiempo de imágenes), analizando las ventajas y desventajas de cada una, además de su capacidad de preservar información. En el proceso, lidiamos con los desafíos de conjugar múltiples resoluciones espaciales y la interpretación biológica de los resultados matemáticos.

Creemos que, en ecosistemas tan estratégicos como estructurantes de los paisajes, el abordaje interdisciplinario es fundamental. La combinación de la teledetección y el aprendizaje automático, de la ecología y la matemática, y de todas las ciencias involucradas, son herramientas fundamentales para el entendimiento del estado de los humedales, los beneficios que proveen, su evolución y los peligros que enfrentan. Además de eso, son insumos fundamentales para la elaboración de políticas públicas que apunten a su mejor aprovechamiento y conservación.

2. HUMEDALES

2.1. Características generales: definición y beneficios

Para poder llevar a cabo aportes matemáticos en el campo de la teledetección y el análisis de los humedales, primero es necesario definirlos. Al ser ecosistemas que pueden presentar tanta variabilidad, necesitamos definiciones que permitan capturar su esencia y caracterizarlos. Sin embargo, buena parte de las definiciones existentes son más bien de carácter enumerativo. Por eso es que nos quedaremos con la definición surgida por consenso en el marco del taller "Hacia un Inventario Nacional de Humedales" organizado por el Ministerio de Ambiente y Desarrollo Sustentable de la Nación en 2016:

> Se entiende por humedal a un ambiente en el cual la presencia temporaria o permanente de agua superficial o subsuperficial causa flujos biogeoquímicos propios y diferentes a los ambientes terrestres y acuáticos. Rasgos distintivos son la presencia de biota adaptada a estas condiciones, comúnmente plantas hidrófitas, y suelos hídricos o sustratos con rasgos de hidromorfismo. [4]

Esta definición apela a más a las cuestiones funcionales (régimen hidrológico, flujos biogeoquímicos) como carácter determinante (o mejor aún, elementos diagnósticos [15]) de estos ambientes más que a su fisonomía, que puede variar mucho de región a región.

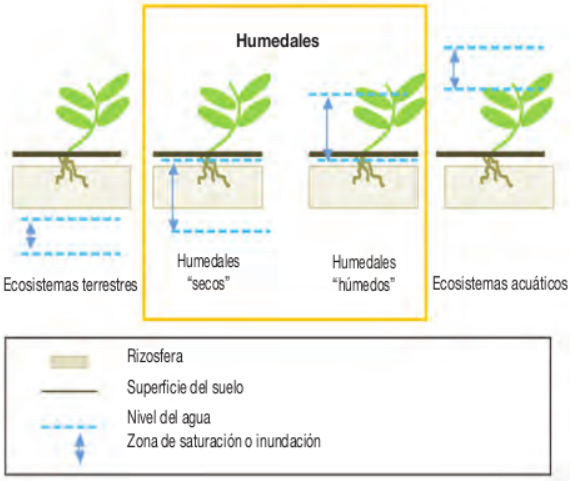


Fig. 2.1: Esquema que muestra las principales diferencias entre los ecosistemas terrestres, los acuáticos y los humedales en relación con la variación del nivel de agua. [3]

Gracias a estas funciones que les son propias y distintivas, estos ecosistemas se destacan por la gran cantidad y diversidad de beneficios (bienes y servicios ambientales) que aportan a la sociedad. Tal como se resume en [4],

> El abastecimiento de agua, la amortiguación de las inundaciones, la reposición de aguas subterráneas, la estabilización de costas, la protección contra las tormentas, la retención y exportación de sedimentos y nutrientes, la retención de

contaminantes y la depuración de las aguas son algunos de los servicios derivados de las funciones de regulación de estos ecosistemas. Los humedales proveen hábitat, alimento y refugio para el sostén de la diversidad biológica y de ellos se obtienen numerosos productos, entre los que se incluyen pescado, animales silvestres, maderas, forraje, plantas medicinales, etc. Ofrecen ambientes de interés paisajístico, cultural y educativo. Son ecosistemas de importancia respecto al cambio climático, tanto para los procesos de mitigación (algunos intervienen en el secuestro y almacenamiento de carbono), como para los procesos de adaptación dado que actúan como "infraestructura natural" para reducir el riesgo de fenómenos extremos como tormentas, inundaciones y sequías.

Además, albergan una excepcional biodiversidad y se estima que constituyen entre el 12,8 % y el 21,5 % de la superficie del país (este número depende de las bases de datos que se consulten [15]), mucho más que el porcentaje a nivel global, estimado entre un 5 y un 8 % [3]. Por esto podemos decir que los humedales son ecosistemas estratégicos (más información se puede consultar en [14] páginas 42-43).

Sin embargo, hoy en día vemos muy comprometida la existencia de los humedales en nuestro planeta: se estima que su extensión global disminuyó entre un 64 y un 71 % en el siglo XX. ¿A qué se debe este deterioro? Fundamentalmente un aumento de la población y su consecuente incremento del desarrollo económico carente de criterios de sustentabilidad ambiental. Para ser más precisos, algunos agentes de la degradación de estos ambientes son el desarrollo de infraestructura, la conversión de las tierras para diferentes usos (agrícola, ganadero, forestal, urbano, etc.), la extracción de agua, la contaminación, la sobreexplotación y la introducción de especies exóticas invasoras.

2.2. La necesidad de un Inventario Nacional de Humedales

A esta altura es pertinente mencionar la Convención sobre los Humedales, de la cual Argentina es Parte Contratante por la Ley 23.919 desde 1992. Es conocida como Convención de Ramsar por la ciudad iraní homónima en la que se celebró en 1971. Su misión es "la conservación y el uso racional de los humedales mediante acciones locales y nacionales y gracias a la cooperación internacional, como contribución al logro de un desarrollo sostenible en todo el mundo" [8]. En el artículo 3 se marca la importancia del desarrollo de inventarios, evaluaciones y monitoreos de los humedales como herramientas para su conservación y uso racional.

Este tipo de instrumentos son esenciales para mejorar el conocimiento de estos ecosistemas y brindar información adecuada para la adopción de medidas de conservación y pautas para el monitoreo de sus cambios, su funcionamiento y su valoración en términos de los bienes y servicios que brindan a la sociedad [3].

A pesar de eso, es bastante reciente el desarrollo de inventarios de humedales en el mundo. En ese marco, en lo que va de este siglo se realizaron múltiples talleres para establecer bases y metodologías para el desarrollo de la enorme tarea que implica la elaboración de este documento. Su objetivo fue la puesta en común de definiciones, marcos conceptuales, alcances, enfoques y escalas de análisis, además de la capacitación de profesionales de los ámbitos académicos y de gestión.

3. ÁREA DE ESTUDIO

3.1. El corredor fluvial Paraná-Paraguay

Dentro de las múltiples regiones de humedales de nuestro país se encuentra la Región Humedales del corredor fluvial Chaco-Mesopotámico, en la que se encuentra el corredor fluvial Paraná-Paraguay. El mismo es el principal colector de las aguas superficiales de la Cuenca del Plata, y se destaca por presentar grandes extensiones de humedales. Sus flujos de agua integran regiones y funcionan como corredores térmicos, geoquímicos, biogeográficos, de transporte humano y de distintas modalidades de vida.

Como se indica en [4], que constituye una regionalización de los humedales de nuestro país, la misma está compuesta por varias subregiones, entre las cuales se encuentra la subregión Ríos, esteros, bañados y lagunas del río Paraná (ver 3.1).

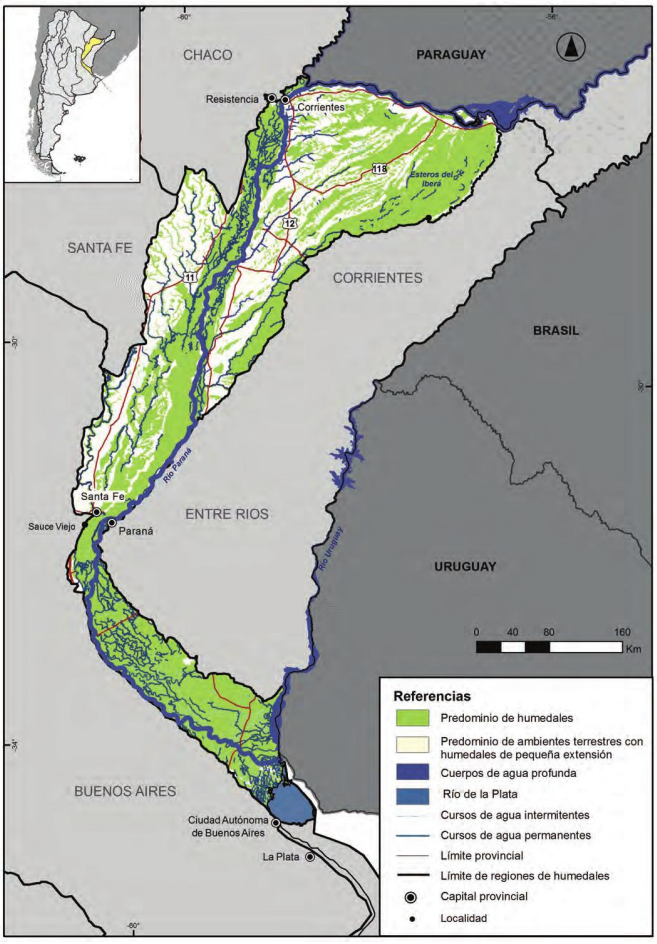


Fig. 3.1: Subregión Ríos, esteros, bañados y lagunas del río Paraná.[4]

Al ser tan heterogénea en términos de los paisajes que presenta, dentro de esta misma subregión se identificaron diez sistemas de paisajes de humedales (ver 3.2). Estos paisajes constituyen unidades ecológicas que se interconectan de manera temporal o permanente

(fundamentalmente dependiendo del grado de inundación o sequía) intercambiando flujos horizontales de información (nutrientes, sedimentos, semillas, huevos) y son resultado de los procesos de modelado fluvial de los dos grandes ríos mencionados. Los sistemas de paisaje son territorios que presentan un origen geológico, climático y geomorfológico común, donde la acción del agua de distintas fuentes (lluvia, escorrentía superficial o subterránea) ha generado modelos de drenaje y permanencia del agua distintivos. Estas características interactúan con la vegetación y los usos del suelo [3].

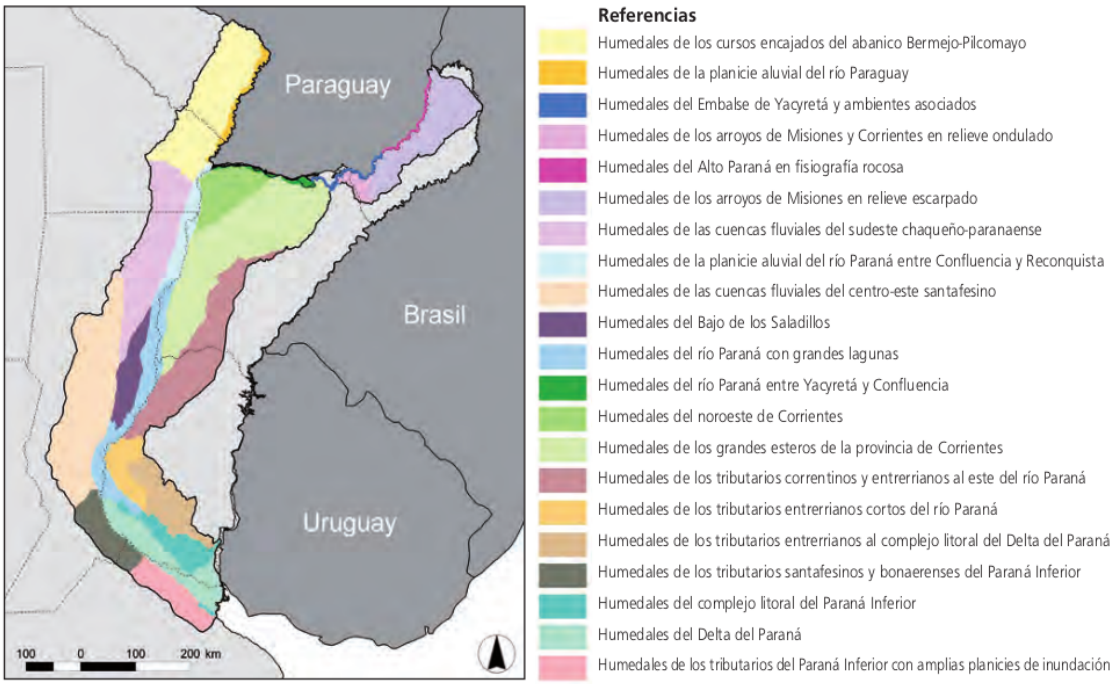


Fig. 3.2: Subregión Ríos, esteros, bañados y lagunas del río Paraná.[3]

Pero afinemos un poco la lupa. Para esta tesis nos focalizaremos en un área perteneciente a dos sistemas: Humedales del río Paraná con grandes lagunas y Humedales del Delta del Paraná. Podemos decir que pertenece en parte a la parte final del Paraná Medio y a los inicios del Paraná Inferior o Delta.

3.2. Sitio Ramsar Delta del Paraná

3.2.1. Designación

El área en la que haremos foco para esta tesis es el Sitio Ramsar Delta del Paraná (a partir de ahora SRDP o simplemente Sitio Ramsar). El Delta sufrió grandes incendios en 2008, que comprometieron unas doscientas mil hectáreas. A partir de estos hechos, se puso en evidencia que en estos suelos hacía falta una herramienta de ordenamiento ambiental debido a los cambios en su uso que afectaban negativamente a la población y a los ecosistemas. Consecuencia de esto fue la elaboración de un *Plan integral estratégico para la conservación y el aprovechamiento sostenible de la región del delta del Paraná*, elaborado conjuntamente entre el Gobierno Nacional y los de las provincias de Buenos Aires, Santa Fe y Entre Ríos [11].

A partir de ese momento se llevaron a cabo diversas acciones tendientes a implementar este plan, además de la publicación de varios documentos, tanto académicos o científicos como gubernamentales, de ordenamiento territorial y pesquero o planes estratégicos. En 2015 se logró la designación internacional del Sitio Ramsar Delta del Paraná.

A partir de que un área es designada como Sitio Ramsar, la misma se incluye en la lista de Humedales de Importancia Internacional. Actualmente existen veintitrés en nuestro país, y son aquellos que cumplen una serie de criterios de importancia, ya sea a nivel de la biodiversidad que alberga, las especies en peligro que contiene, etcétera (consultar en [11]).

En este caso, en 2019 se elaboró un *Plan de manejo* con objetivos de conservación y protección, además de un comité interdepartamental e interjurisdiccional, con participación del Estado, académicos y organizaciones no gubernamentales para su efectivización.

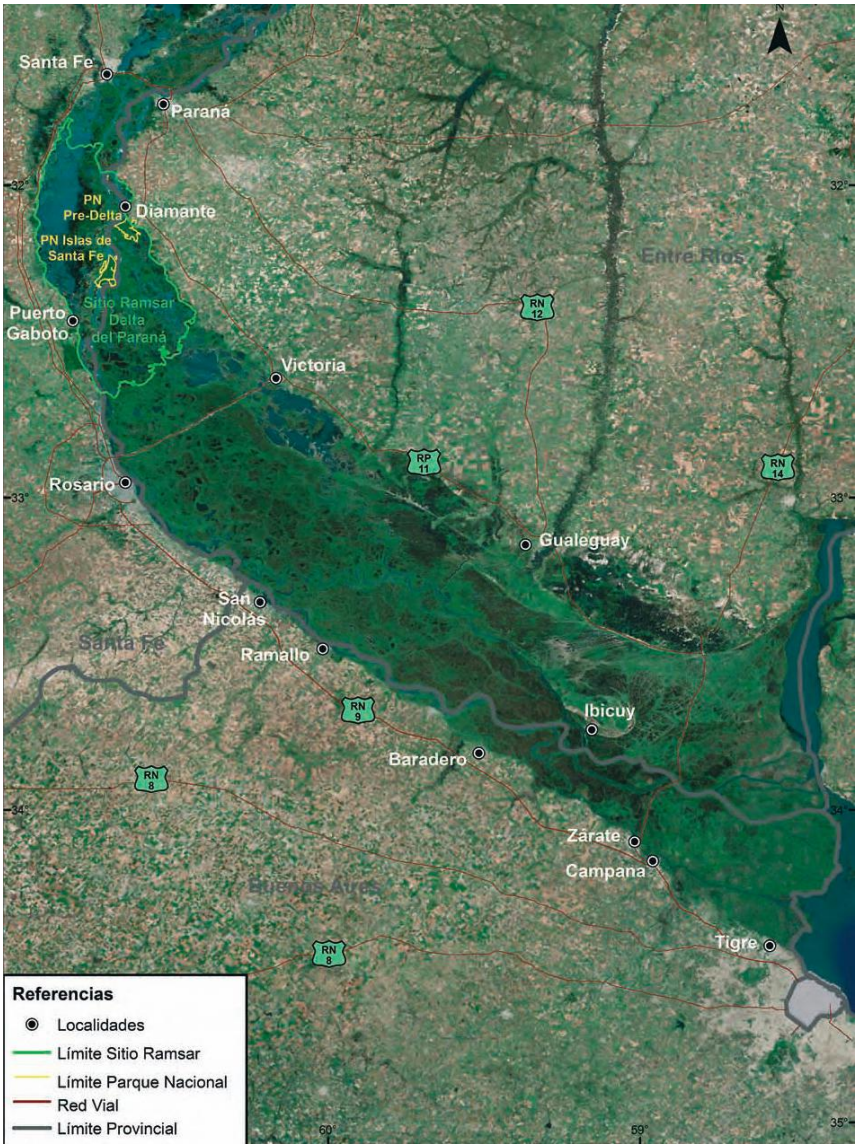


Fig. 3.3: Sitio Ramsar Delta del Paraná. [11]

3.2.2. Algunas características

En este trabajo solo nombraremos algunas características que nos interesan del Sitio Ramsar (geomorfología, régimen hidrológico, vegetación), dado que su descripción en profundidad (fauna, flora, actividades económicas, clima, calidad del agua, etc.) figuran con detalle en [11], que también servirá como principal fuente para esta sección.

El paisaje de nuestra área de interés se encuentra moldeado por el régimen pulsátil del Paraná. El mismo, como se explica en [23], está determinado principalmente por las precipitaciones de las regiones tropicales y subtropicales de su alta cuenca. Presenta un período de ascenso a partir del mes de septiembre, culminando con un máximo en el mes de marzo. Luego comienza a descender alcanzando las bajantes más pronunciadas en el mes de agosto. Sin embargo, pueden producirse repuntes excepcionales en junio y en octubre. La época de precipitaciones se da de octubre a abril. Contiene una red compleja de cuerpos de agua lóticos -arroyos y ríos- y lénticos -lagunas- que cambian de forma y área de acuerdo a la profundidad. Estos cuerpos se alternan con bañados, albardones e islas que evolucionan de formas simples hacia formas más complejas. Son características las grandes lagunas, de escasa profundidad, varias decenas de kilómetros cuadrados y gran elasticidad.

La vegetación se caracteriza por un mosaico de comunidades arbóreas, arbustivas y herbáceas, que se disponen en el espacio de acuerdo a la duración del período en que permanecen ocupadas por las aguas de las crecientes. Los bosques con fisonomías arbóreas ocupan las porciones altas (albardones) donde el agua permanece menos tiempo. En sectores intermedios (medias lomas) aparecen tanto comunidades leñosas (boscosas o arbustivas) como herbáceas. Donde el agua de inundación permanece por más tiempo (bajos) se desarrollan fundamentalmente comunidades herbáceas, destacándose extensos pajonales. Las plantas acuáticas ocupan las porciones más bajas del gradiente topográfico, donde prácticamente toda el área está permanentemente inundada. Su importancia relativa es mayor en las zonas internas y/o bajas de las islas y menores en los márgenes de los cursos de agua de alta energía y transporte de sedimentos.

De esta forma, gran parte de la vegetación fluvial se distribuye en un amplio rango de condiciones de hábitats, y esto redunda en una gran plasticidad ante variaciones ambientales. Las comunidades herbáceas se regeneran rápidamente luego de disturbios catastróficos como inundaciones extraordinarias, por lo que en su conjunto son resilientes[1] a este tipo de eventos. Tanto la frecuencia como la permanencia del agua en los diferentes sectores del gradiente topográfico, como las adaptaciones de las diferentes especies a los ciclos hidrológicos, determinan en gran parte la presencia y los límites de cada comunidad. Las mismas, en algunos casos, contribuyen a la formación y estabilidad de los sitios que colonizan, aportando a la dinámica espacio-temporal de los humedales y generando nuevos hábitats para el resto de la biota.

3.2.3. Régimen de disturbios

Además de las inundaciones que ocurren regularmente cada uno a tres años, cada cierto tiempo ocurren algunas extraordinarias. Por ejemplo, en 2007 (febrero a abril), 2009-2010 (noviembre a julio) y 2016 el Paraná se vio afectado por crecientes extraordinarias (por fuera de las inundaciones ordinarias que ocurren cada uno a tres años) asociadas al

[1] Según [7], la resiliencia es la capacidad de un sistema socioecológico de mantener similares su estructura, función y mecanismos de retroalimentación a pesar de los *shocks* y perturbaciones que pueda llegar a sufrir.

fenómeno El Niño/Oscilación del Sur (ENOS)[2]. En el gráfico 3.4 podemos ver la evolución a lo largo del tiempo de los niveles hidrométricos, obtenidos en el Portal de datos del Sistema de Información Hidrológica de la Cuenca del Plata ([21]).

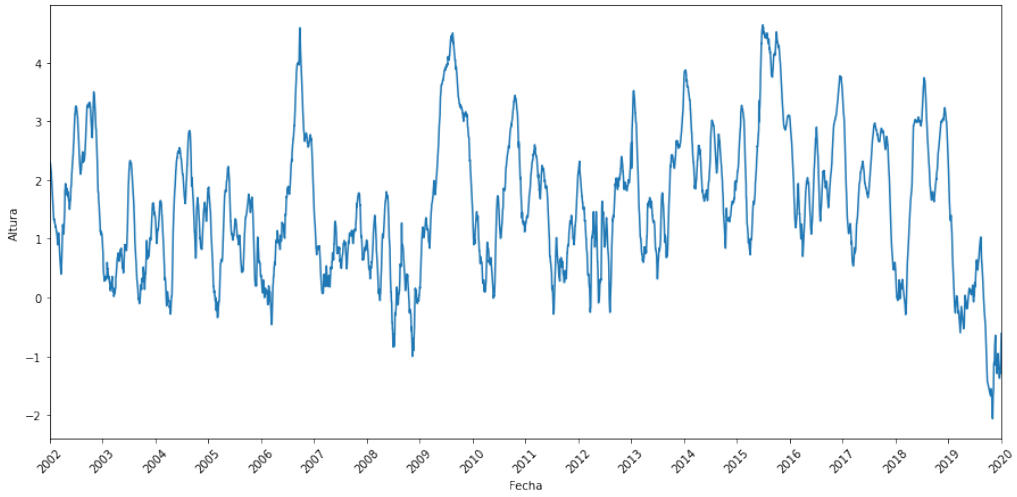


Fig. 3.4: Niveles hidrométricos en la zona del nacimiento del río Coronda.

Las inundaciones constituyen disturbios, que, tal como se define en [7], son eventos relativamente discretos o puntuales en el tiempo y en el espacio que remueven la biomasa de las plantas o que alteran la estructura de las poblaciones, comunidades y ecosistemas y ocasionan cambios en la disponibilidad de recursos o el ambiente físico. Estos disturbios son una de las causas principales de las fluctuaciones a largo plazo en la estructura y función de los ecosistemas, es decir que disparan procesos de sucesión. En este caso, las inundaciones extraordinarias (sobre todo las de mayor duración) interrumpen el ciclo fenológico de la vegetación debido a que el agua la arrastra. El impacto global de los disturbios en un ecosistema depende de los patrones de paisaje que gobiernan la resiliencia y la renovación del mismo, así como también su régimen de disturbios. Con esto último nos referimos a la frecuencia y la interacción de múltiples tipos de disturbios (incendios, inundaciones, etcétera) así como también la naturaleza de cada uno de ellos (su alcance, locación, magnitud, frecuencia).

3.2.4. Beneficios y amenazas

Entre las principales amenazas que afectan a los humedales del corredor fluvial Paraná-Paraguay se encuentra la modificación de la dinámica hidrológica con fines de riego, almacenamiento y desvío de agua. La misma está dada por la expansión de la frontera agropecuaria y el avance de las urbanizaciones así como los terraplenes para la construcción de

[2] El fenómeno denominado "El Niño" consiste en un calentamiento anómalo de las aguas superficiales del Océano Pacífico Ecuatorial Central y Oriental. A pesar de registrarse en el Pacífico Sur, el ENOS es un fenómeno global, ya que lleva asociado la interrupción de la circulación general atmosférica y de los patrones climáticos globales. Ocasiona alteraciones climáticas en muchas regiones del planeta, produciendo sequías en algunas (Este de Australia) y precipitaciones muy abundantes en otras. Particularmente, la región sudeste de América del Sur (sur de Brasil, nordeste de Argentina, Uruguay y sur de Paraguay) presenta una fuerte respuesta al fenómeno con un incremento considerable en las precipitaciones. Este aumento tiene influencia directa sobre la cuenca del Plata debido al aumento en los caudales de los ríos Paraguay, Paraná y Uruguay, dado el emplazamiento de su alta cuenca. [23]

rutas, puentes y caminos. Este tipo de obras modifican la composición y el funcionamiento de los humedales.

Otros problemas de conservación tienen que ver con la disminución de la calidad del agua debido a la recepción de efluentes domiciliarios, agropecuarios e industriales, la invasión de especies exóticas de flora y fauna que pueden desplazar a la biota local, e incluso el impacto de obras y actividades que se desarrollan fuera de esta región, tales como las numerosas represas ubicadas en el tramo superior del río Paraná.

Como se explica en [7], hemos alterado los ecosistemas más veloz y extensivamente en los últimos cincuenta años que en cualquier otro período de tiempo comparable en la historia de la humanidad. Estos cambios son el resultado del crecimiento exponencial de la población, nuestro consumo de recursos y nuestra creciente capacidad tecnológica para alterar los ecosistemas y el ambiente terrestre.

La actividad humana frecuentemente modifica el régimen de distrurbios mediante su iniciación o supresión. El endicamiento de los ríos o arroyos puede eliminar inundaciones que arrastran sedimentos de los canales, resultando en enormes modificaciones en las redes alimenticias de los mismos y su capacidad para soportar la vida de los peces. Las represas hidroeléctricas pueden eliminar inundaciones asociadas a las lluvias y regular el flujo basándose en la demanda de electricidad, lo que causa una discrepancia entre el tiempo en el que ocurren los disturbios y el régimen al que los organismos están adaptados. Por su sensibilidad y su efecto en otros mecanismos interactivos de control, los cambios en los regímenes de disturbios alteran la estructura y función de los ecosistemas.

4. TELEDETECCIÓN

4.1. Definición y usos

Para superar las dificultades de acceso y cobertura de grandes áreas, la teledetección es muy utilizada al momento de estimar variables biogeofísicas. Es una herramienta clave para mejorar nuestras capacidades de conocimiento, monitoreo e inventario; y nos provee una forma de acceder a diversas escalas temporales y espaciales [18].

Pero ¿De qué hablamos cuando hablamos de teledetección? Existen diversas definiciones, pero podemos decir que se trata de la estimación de propiedades de los objetos que se encuentran en la superficie de la Tierra a través de datos adquiridos de manera remota [6].

Cuando la radiación electromagnética alcanza la superficie terrestre y sus objetos, esta puede ser transmitida, absorbida o reflejada. La magnitud de cada uno de estos procesos depende de las propiedades de los mismos.

Por ejemplo, mediante sensores remotos podemos medir la cantidad de radiación solar reflejada en función de la longitud de onda, llamada reflectancia espectral. Esta es la magnitud más importante de la teledetección óptica, y es usada regularmente en modelos para obtener variables biogeofísicas a partir de datos ópticos [23]. En resumen: nos interesa medir la interacción entre la radiación y los objetos de la superficie para así inferir propiedades de los mismos.

Si bien existen diferentes tipos de sensores remotos, nos vamos a interesar en este trabajo en los sensores ópticos. Los mismos utilizan al sol como fuente de energía, y miden la porción de la energía emitida por esta fuente que los alcanza luego de interactuar con los elementos de la superficie terrestre y con la atmósfera. Las imágenes multiespectrales son producidas por sensores ópticos que poseen detectores que miden en rangos distintos de longitudes de onda, y la reflectancia es registrada independientemente en cada uno de ellos, correspondiendo cada uno a una banda de la imagen.

En [7] se explica que la teledetección nos provee un conjunto de herramientas para determinar la estructura y algunos aspectos de la función de paisajes heterogéneos. La heterogeneidad espacial de interés ocurre, en su mayoría, en escalas espaciales que no pueden ser observadas desde un único punto en tierra. Los desarrollos recientes en teledetección han transformado nuestra habilidad para analizar la heterogeneidad espacial dado que nos facilitan técnicas que nos permiten visualizar, de una sola vez, ecosistemas a lo largo y ancho de un área grande, para poder comprenderlos como un todo.

4.2. Índices de vegetación

El objetivo de trabajar con un índice es la reducción de múltiples bandas de datos en una sola, condensando la información que consideremos más importante según el estudio que estemos llevando a cabo.

En el caso de estudios relacionados con la cubierta vegetal (como es el nuestro), se utilizan los índices de vegetación o verdes. Su objetivo principal es formular una medida sintética precisa sobre las variaciones espacio-temporales que ocurren en los ecosistemas dado que asumen la existencia de una relación con variables biogeofísicas de la superficie.

Estos índices son calculados a partir de la combinación de bandas espectrales de las

imágenes satelitales. Se basan en el comportamiento espectral típico de la vegetación en las regiones del espectro electromagnético correspondientes al rojo (R) (aproximadamente 0,550-0,70 μm) y el infrarrojo cercano (IRc, aproximadamente 0,730-1 μm)

4.2.1. NDVI

Debido a la actividad de los pigmentos fotosintéticos de las plantas, casi toda la radiación en el infrarrojo cercano se refleja dependiendo del índice de área foliar (LAI por sus siglas en inglés), la distribución angular de las hojas y su anatomía y morfología. Lo contrario ocurre para la energía en la parte visible del espectro, que se absorbe en su mayoría en las porciones del azul (470 nm) y del rojo (670 nm) [9].

El contraste entre las respuestas del rojo y el infrarrojo cercano (llamado comúnmente "borde rojo") constituye una medida de la cantidad de vegetación. Las características de reflectividad también varían según el estado de la biomasa verde en pie, donde la máxima diferencia entre R e IRc corresponde al estadio de mayor densidad o vigor de la vegetación y el mínimo contraste a áreas de muy poca vegetación o en estado senescente.

El índice de vegetación más ampliamente usado es el Índice de Vegetación de Diferencia Normalizada (en la literatura lo encontramos con sus siglas en inglés, NDVI, que usaremos de aquí en adelante), que se calcula de la siguiente manera:

$$NDVI = \frac{\rho_{IRc} - \rho_R}{\rho_{IRc} + \rho_R}$$

donde ρ_{IRc} es la reflectancia en la longitud de onda del infrarrojo cercano y ρ_R la correspondiente al rojo.

A partir de esta fórmula podemos ver que este índice toma valores entre -1 y 1. El mismo es muy utilizado debido a que se relaciona directamente con el estado de la biomasa verde en pie (a mayor cantidad de biomasa verde en pie, mayores valores de NDVI) [23]. Además de eso, realza cambios pequeños en las respuestas espectrales de las coberturas, y describe adecuadamente el comportamiento de la vegetación en ambientes donde la cobertura vegetal es alta y hay baja proporción de suelo desnudo.

4.3. MODIS

El espectrorradiómetro de imágenes de resolución moderada (MODIS por sus siglas en inglés) es un instrumento a bordo de los satélites Terra (lanzado en 1999) y Aqua (lanzado en 2002) de la NASA. En conjunto capturan nuestro planeta cada 1-2 días en 36 bandas espectrales. A partir de este instrumento existen varios productos que capturan características de la atmósfera, la superficie terrestre, el océano, etc. En esta tesis utilizaremos el producto MYD13Q1 de la colección Aqua-MODIS.

Debido a su resolución espacial media (250m en el caso del producto que utilizaremos), en una sola imagen podemos cubrir toda nuestra región de estudio. Su resolución temporal es de 16 días, ya que cada imagen provista es el producto de un algoritmo que compone las imágenes obtenidas en ese período de tiempo. Cada píxel de la misma minimiza la presencia de nubes, aerosoles, contaminación y ángulo de visión, obteniendo un valor razonable de NDVI para este período. Además de eso, se multiplica por un factor de 10000 y se descartan aquellos con valores inferiores a -2000 o superiores a 10000. De esta manera se maximiza la calidad del producto para asegurar la utilidad de cada imagen provista [9].

Para esta tesis utilizaremos 416 imágenes. La primera es de julio de 2002 y la última del mismo mes de 2019.

5. TRABAJOS PREVIOS

En esta tesis nos basaremos fundamentalmente en [19] y en [20]. El objetivo de estos *papers* es el análisis de la dinámica fluvial a lo largo del tiempo en el Paraná medio. Para esto, utilizan un enfoque que integra capas relacionadas a la elevación del terreno, redes de drenaje, unidades geomórficas y patrones de NDVI.

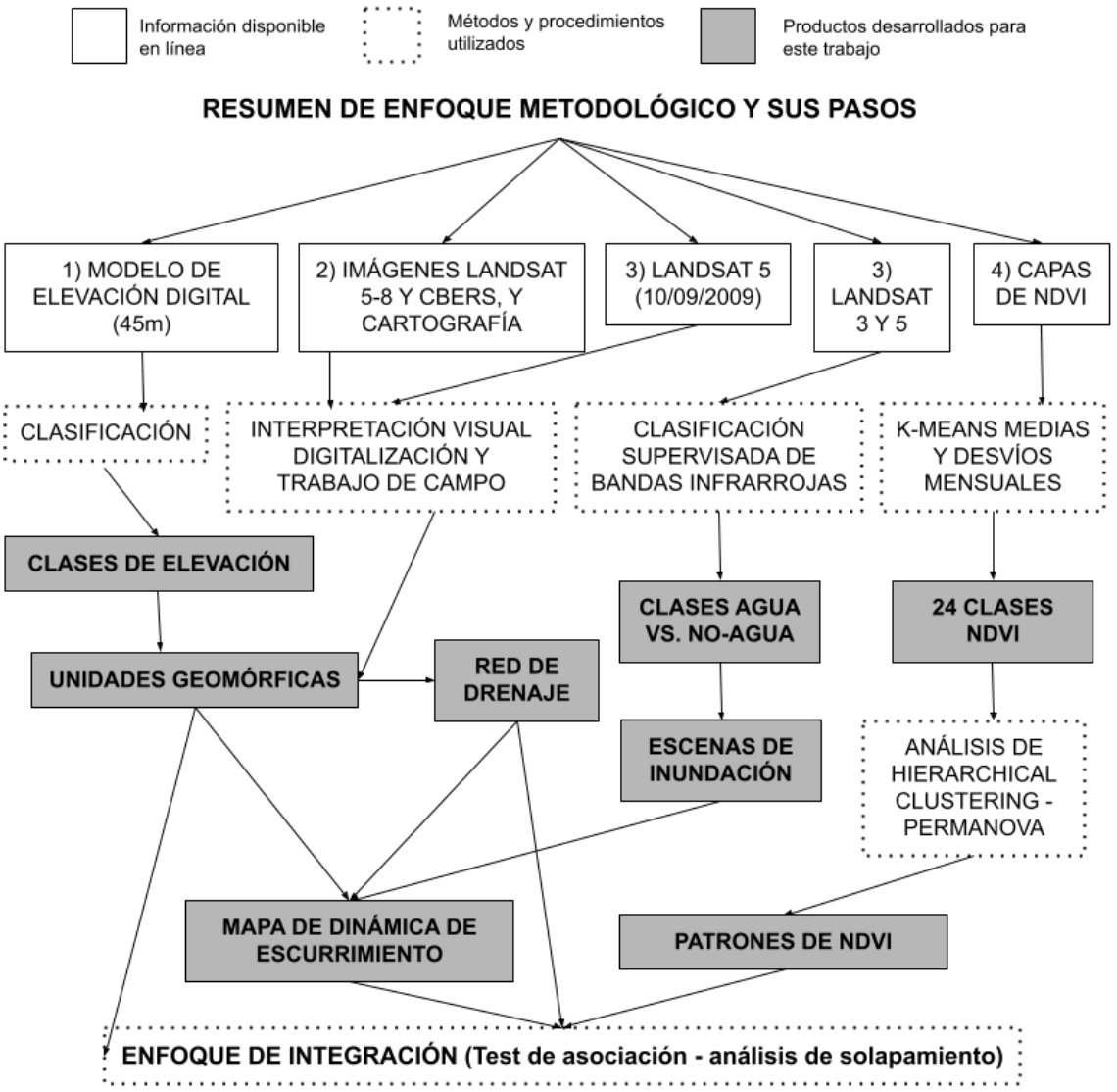


Fig. 5.1: Enfoque de [20] para analizar dinámicas de escurrimiento superficial y su relación con los patrones de NDVI (traducción propia del original en inglés).

En [20] se explica que, al no existir mapas topográficos del área, se emplearon modelos digitales de elevación provistos por el Instituto Geográfico Nacional. El mismo se clasificó en clases de elevación de 2 metros. La red de drenaje se delineó utilizando una imagen

Landsat 5 TM en un momento de aguas bajas (septiembre 2009) para incluir cursos de agua permanentes y cuerpos de agua de tamaño considerable. Para estudiar las dinámicas de escurrimiento superficial, se utilizó un set de quince imágenes Landsat que representaban diferentes niveles hidrométricos para las principales inundaciones de los últimos cuarenta años. Cada una se clasificó de manera supervisada en dos categorías (agua o no agua) y las escenas clasificadas se ordenaron de manera ascendente según el nivel de agua para entender de qué manera se expande el agua al ingresar en la planicie de inundación. A partir de eso se elaboraron dos mapas: uno para condiciones de inundación y uno para condiciones previas al desborde del río.

Como se ve en [20], confeccionar un mapa de dinámicas de escurrimiento superficial requiere interpretación visual y trabajo de campo. El objetivo de establecer vínculos entre la dinámicas de escurrimiento superficial y patrones de NDVI tiene que ver con elaborar indicadores que estén disponibles fácilmente y nos permitan interpretar estas dinámicas.

En esta tesis nos focalizamos en el la elaboración de patrones que proponen los *papers* mencionados, cambiando el área de estudio: en lugar del Paraná Medio nos centraremos en el Sitio Ramsar.

Para esto, nos propusimos utilizar imágenes del lugar tomadas en diferentes momentos, como explicamos en 4.3, y emplear técnicas de reducción de la dimensionalidad y posteriormente de *clustering*. Nuestro objetivo es refinar las técnicas propuestas en los trabajos previos (esto lo explicaremos más adelante) y comparar técnicas de reducción de dimensionalidad.

Más en concreto, elaboramos dos mapas de patrones de NDVI a través de una serie de 416 imágenes Aqua-MODIS (MYD13Q1) del sitio Ramsar:

1. para el primero comenzamos tomando las medias y desvíos estándar mensuales de NDVI para reducir su dimensionalidad (tal como en los trabajos anteriores) y

2. para el segundo comenzamos haciendo lo propio con análisis de componentes principales (PCA).

Posteriormente utilizamos técnicas de *clustering* para encontrar patrones de NDVI.

6. ELABORACIÓN DE NUESTROS MAPAS

6.1. ¿Qué es *clusterizar*?

En el área de *machine learning* o aprendizaje automático el problema de *clustering* se refiere al de encontrar grupos, o *clusters*, de puntos de datos en un espacio multidimensional [5]. Algunos autores también definen este problema como aquel que trata de encontrar alguna estructura en datos que no están etiquetados [12]. Por este motivo es que se trata de un asunto de aprendizaje no supervisado (*unsupervised learning*).

Para ser más precisos: supongamos que tenemos un *set* de datos $\{x_1, \ldots, x_N\}$ correspondientes a N observaciones de una variable aleatoria de dimensión D. El objetivo es particionar este conjunto en K grupos (para algún K a determinar) buscando similitud entre los puntos que quedan dentro de una clase pero que sean diferentes a aquellos que quedan fuera de la misma. La forma en la que definamos la medida de similitud, la distancia entre los *clusters*, la manera de agruparlos, la estructura que asumimos en los datos, etcétera, nos dan distintos algoritmos de *clustering*.

6.2. Nuestro *dataset*

En el Capítulo 4 decíamos que teníamos 416 imágenes. Las mismas son de 449×421 píxels cada una, y cada uno de esos píxels que corresponda al sitio Ramsar contiene el NDVI (reescalado entre -2000 y 10000) observado por MODIS en la región de 250m$\times$250m. Cada una de las mismas corresponde a una fecha (en verdad es una composición de imágenes de dieciséis días, pero a lo largo de esta tesis nos referiremos a esta composición como si correspondiera a la imagen de una fecha). Una vez que "enmascaramos" nuestras imágenes para tener solamente puntos correspondientes al sitio Ramsar y las "aplanamos" (dado que los algoritmos de *clustering* no tienen en cuenta la espacialidad) tenemos que nuestro *set* de datos está compuesto por 45260 puntos de dimensión 416.

Un primer enfoque podría haber sido tratar de "clusterizar" estos datos directamente. Sin embargo, tendríamos que lidiar con la "maldición de la dimensionalidad" descripta por Richard Bellman. La misma consiste en el hecho de que, a medida que la dimensión en la que trabajamos se incrementa, ciertas propiedades deseables del espacio se pierden. Por ejemplo, el concepto de distancia entre dos puntos se vuelve menos preciso dado que el volumen de una esfera en dimensiones altas está concentrado en el borde. Además de eso, el volumen del espacio crece tan rápido que los datos disponibles pueden fácilmente convertirse en *sparse* (mayoritariamente 0) para nuestro espacio, que pasa a estar prácticamente vacío [5]. Por último, en el caso de que exista algún tipo de estructura subyacente en los datos (por ejemplo, periodicidad o alguna regularidad en el tiempo) la estaríamos perdiendo.

6.3. Reducción de dimensionalidad

6.3.1. Medias y desvíos mensuales

En [19] y [20] la técnica de reducción de dimensionalidad consiste en aprovechar el conocimiento disciplinar de las autoras con respecto a la ecología del paisaje: puesto que

el NDVI es un índice verde que representa el área foliar, debería presentar un carácter estacional. Es decir, las imágenes que representan el NDVI de enero de 2003 no deberían ser muy diferentes a las de enero de otros años, y así sucesivamente con otros meses.

Por eso es que las autoras redujeron la dimensión temporal de 416 a 24 tomando las medias y las desviaciones estándar para cada mes. Por ejemplo, la primera imagen ahora correspondería al promedio de todas las imágenes del *set* anterior correspondientes a enero, la segunda a febrero, etc. La decimotercera ahora corresponde al desvío estándar de todas las imágenes correspondientes a enero, la decimocuarta a febrero, etc.

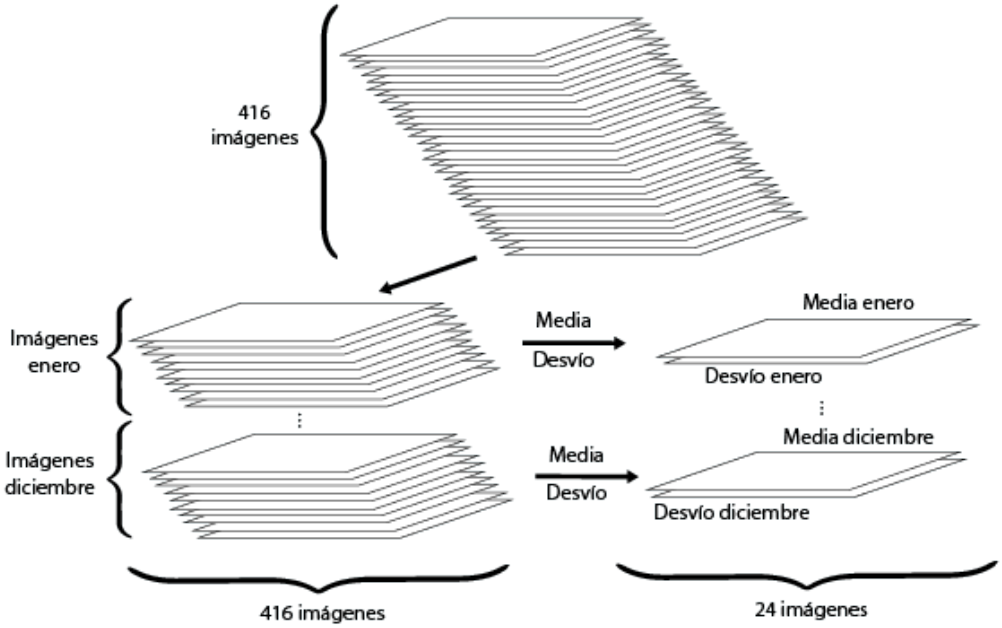


Fig. 6.1: Reducción de dimensionalidad usando medias y desvíos mensuales.

Un potencial problema de este enfoque es que se basa, justamente, en la asunción de periodicidad y, por lo tanto, pueden llegar a no captarse ciertos eventos extremos, como inundaciones o sequías extraordinarias.

Además de eso, consideramos que este método no se trata de una reducción de la dimensión sino de creación de variables nuevas a través de medidas resumen que no garantizan explicar la varianza de nuestros datos.

6.3.2. PCA

Otro enfoque posible es una reducción de dimensionalidad más clásica usando Análisis de Componentes Principales (PCA por sus siglas en inglés). Este método tiene dos grandes interpretaciones (que resultan ser equivalentes). La primera no es probabilística y define PCA como la proyección ortogonal de nuestros datos en un espacio de dimensión menor de forma tal que la varianza de de los datos proyectados se maximice. La segunda tiene un enfoque generativo: modelamos nuestros datos como

$$x_i = W h_i + \mu + \epsilon \tag{6.1}$$

donde h_i son variables latentes, μ representa un corrimiento del 0 y ϵ es "ruido" que sigue una distribución gaussiana con media 0 y matriz de covarianza $\Psi = \sigma^2 \mathbf{I}$. De esta manera,

si las h_i están dadas, tenemos la siguiente interpretación probabilística:

$$p(x_i|h_i) = \mathcal{N}(Wh_i + \mu, \Psi) \tag{6.2}$$

Para un modelo probabilístico completo necesitamos alguna distribución para las variables latentes. Si asumimos $h_i \sim \mathcal{N}(0, \mathbf{I})$ nos queda una distribución gaussiana para nuestros datos:

$$p(x) = \mathcal{N}(\mu, WW^T + \Psi) \tag{6.3}$$

Asumimos homocedasticidad en ϵ dado que lleva a un modelo más simple que asumir que Ψ simplemente es una matriz diagonal y porque creemos que tiene sentido dado que todos los píxels fueron medidos con la misma herramienta y sus valores se encuentran dentro del mismo rango y escala.

La biblioteca libre *scikit-learn* de Python implementa la reducción de dimensionalidad utilizando la descomposición en valores singulares (SVD) de los datos para proyectarlos a un espacio de dimensión menor. Además de eso, dada una descomposición, implementa un *score* del modelo, dado por la verosimilitud o *likelihood* de los datos observados dado el modelo. Esto nos da un criterio para decidir con cuántos componentes principales nos quedamos. Evidentemente a medida que vamos incrementando la dimensión la verosimilitud sube, y es máxima cuando llegamos a 416, nuestra cantidad original de imágenes. Sin embargo, buscamos justamente reducir la dimensionalidad y simplificar nuestro modelo, por lo que es necesario encontrar el equilibrio entre su simpleza y su explicabilidad.

El paquete de Python *kneed* implementa el algoritmo *Kneedle* que apunta a buscar este balance. Lo que hace es hallar la "rodilla" de una función, definiéndola como su punto de máxima curvatura (esto se puede calcular con la derivada segunda para funciones suaves o utilizando interpolación o bien métodos geométricos en el caso de funciones no continuas o *datasets* discretos) [24].

Lo que hicimos, entonces, fue una reducción de dimensionalidad usando PCA para diferentes D (cantidad total de componentes principales). Graficamos el *score* del modelo en función de este parámetro y le buscamos la *rodilla* a este gráfico.

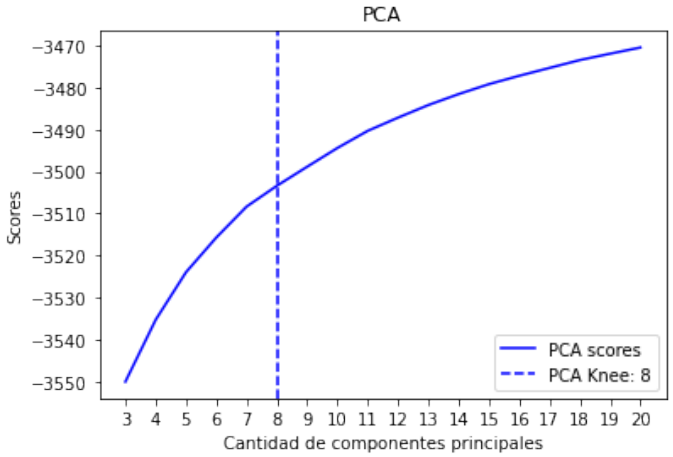


Fig. 6.2: Logaritmo de la verosimilitud en función de la cantidad de componentes principales.

Como se ve, *kneed* encontró la *rodilla* de este gráfico en ocho componentes principales.

6.3.3. Contribuciones

Los siguientes gráficos nos pueden ayudar a entender ciertas ventajas del método de análisis de componentes principales por sobre el de las medias y desvíos. Cuando realizamos esto último, todas las fechas tienen el mismo grado de importancia. Es decir, todas contribuyen en igual medida al *dataset* nuevo con dimensión menor. Supongamos que nuestro *dataset* original se llama $\texttt{imágenes} \in \mathbb{R}^{416\times\#\text{píxels}}$.

Sabemos que nuestra secuencia de fechas es de la siguiente manera:

$$\texttt{fechas} = (\text{04-07-2002}, \text{20-07-2002}, \text{05-08-2002}, \ldots, \text{19-07-2020})$$

Por lo tanto, la componente perteneciente a la media del mes de julio será

$$1 \cdot \texttt{imágenes}[0] + 1 \cdot \texttt{imágenes}[1] + 0 \cdot \texttt{imágenes}[2] + \cdots + 1 \cdot \texttt{imágenes}[415] \tag{6.4}$$

Si vemos los coeficientes en la suma anterior ($[1, 1, 0, \ldots, 1]$) estos representan la secuencia de contribuciones de los miembros de `imágenes` a la séptima componente de nuestro *dataset* nuevo con dimensión menor (la correspondiente a julio). Podemos normalizar el vector para expresar estas contribuciones en forma de porcentaje, pero en cualquier caso se vería como en el gráfico 6.3.

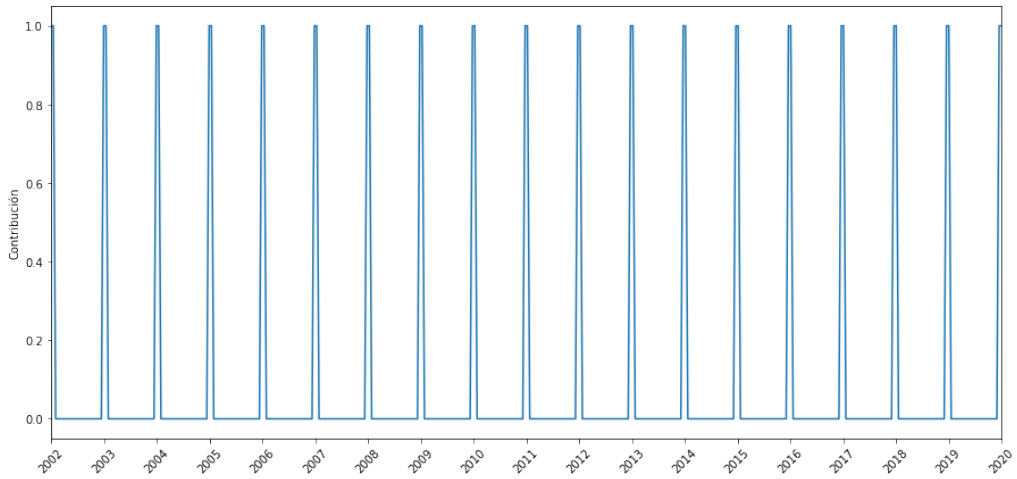
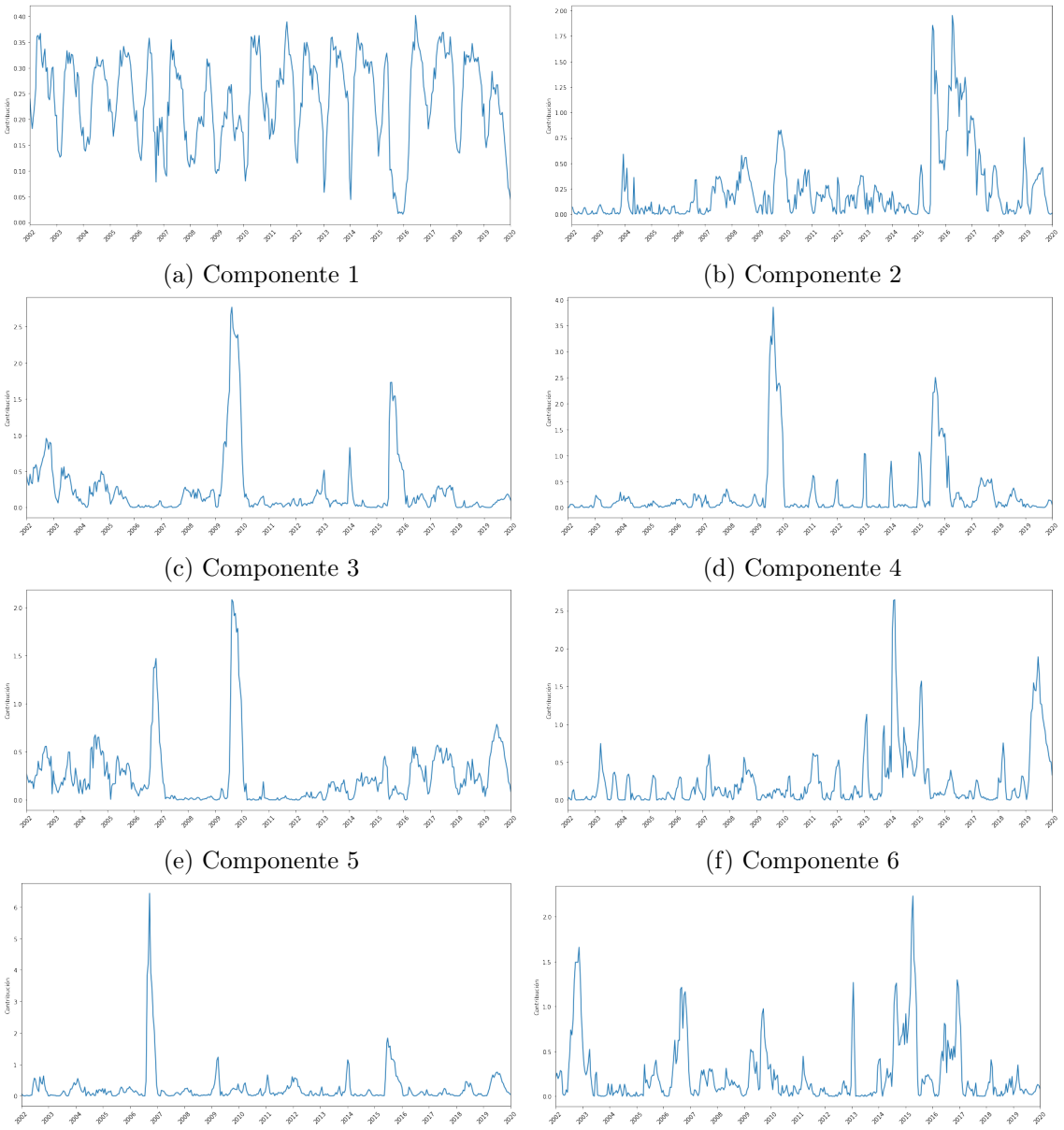


Fig. 6.3: Contribuciones de cada imagen a la séptima componente del *dataset* reducido

De manera similar podemos ver que los gráficos para todas las componentes serán como este, solo que trasladados un poco horizontalmente.

En el caso de PCA, las contribuciones de las diferentes fechas a las componentes principales es dispar, como podemos ver en los gráficos de la figura 6.4.

Al observar los gráficos vemos que, si bien en la primera componente principal (que representa la que mayor varianza explica en los datos) le asigna a las diferentes fechas una importancia prácticamente periódica (como en el caso anterior), en las otras componentes principales la contribución no es homogénea. De esta manera, tenemos menos chances de "perdernos" eventos extremos que con las medias y desvíos mensuales.



(a) Componente 1

(b) Componente 2

(c) Componente 3

(d) Componente 4

(e) Componente 5

(f) Componente 6

(g) Componente 7

(h) Componente 8

Fig. 6.4: Componentes 1 a 8 resultantes de PCA

6.4. Clustering

6.4.1. *Gaussian Mixture Model* y algoritmo E-M

El Modelo de mezcla de gaussianas o *Gaussian Mixture Model* (GMM de acá en adelante) es un modelo probabilístico que consiste en la superposición de componentes gaussianas. La distribución correspondiente podemos escribirla así:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \tag{6.5}$$

donde $\mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$ representa la densidad de la distribución normal con media y covarianza μ_k y $\mathbf{\Sigma}_k$ en la $k-$ésima componente y además $0 \leq \pi_k \leq 1$ y $\sum_{k=1}^{K} \pi_k = 1$ para que la densidad sea no negativa e integre 1.

Bishop en [5] introduce la variable aleatoria K-dimensional $\mathbf{z}$ en la cual $z_k \in \{0, 1\}$ pero solo una de las z_k es 1 y todas las otras son 0. Especificamos su distribución en términos de π_k de esta manera: $p(z_k = 1) = \pi_k$. Así, podemos "aislar" cada componente normal de $\mathbf{x}$:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \tag{6.6}$$

Introducimos también la probabilidad de $\mathbf{z}$ dada $\mathbf{x}$ y la calculamos utilizando el teorema de Bayes:

$$\gamma\left(z_k\right) \equiv p\left(z_k = 1 \mid \mathbf{x}\right) = \frac{p\left(z_k = 1\right) p\left(\mathbf{x} \mid z_k = 1\right)}{\sum_{j=1}^{K} p\left(z_j = 1\right) p\left(\mathbf{x} \mid z_j = 1\right)} = \frac{\pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)} \tag{6.7}$$

Ahora, si tenemos un *set* de datos de dimensión D (en nuestro caso será 8 o 24 según el método utilizado para la reducción de la dimensionalidad) $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ vamos a tener una variable latente $\mathbf{z}$ para cada $\mathbf{x}$. Podemos, por lo tanto, representar nuestros datos y estas variables latentes en dos matrices $\mathbf{X}$ (de $N \times D$) y $\mathbf{Z}$ (de $N \times K$) respectivamente, donde en cada una de ellas ($\mathbf{x}_i^T$ o $\mathbf{z}_i^T$ según corresponda) es la fila i-ésima.

Si asumimos que nuestros datos son realizaciones i.i.d. (independientes e idénticamente distribuidas) de la mezcla de gaussianas de la que hablamos, podemos escribir su *log-likelihood* (el logaritmo de la verosimilitud) así:

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right) \right\} \tag{6.8}$$

para intentar maximizarla y así encontrar los parámetros ($\boldsymbol{\mu}_k$, $\mathbf{\Sigma}_k$ y π_k) de nuestra distribución.

Si derivamos respecto de las medias $\boldsymbol{\mu}_k$ e igualamos a 0, nos queda

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)}}_{\gamma(z_k)} \mathbf{\Sigma}_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) \tag{6.9}$$

Multiplicamos por $\mathbf{\Sigma}_k^{-1}$ (asumimos que $\mathbf{\Sigma}_k$ no es singular) y obtenemos

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n \tag{6.10}$$

donde definimos

$$N_k = \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \tag{6.11}$$

Lo que acá vemos es que la media $\boldsymbol{\mu}_k$ para la componente gaussiana k-ésima se obtiene como un promedio ponderado de todos los puntos en nuestro *set* de datos, donde el peso que le damos al punto $\mathbf{x}_n$ está dado por la probabilidad *a posteriori* $\gamma\left(z_{nk}\right)$ de la componente k por haber generado $\mathbf{x}_n$.

Si ahora derivamos (6.8) con respecto a $\mathbf{\Sigma}_k$ e igualamos a 0 nos queda

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}} \tag{6.12}$$

Ahora tenemos que maximizar (6.8) también con respecto a los coeficientes π_k, teniendo en cuenta que los mismos deben sumar 1. Esto podemos hacerlo usando multiplicadores de Lagrange, maximizando

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{6.13}$$

lo cual nos lleva a

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)} + \lambda \tag{6.14}$$

Acá podemos multiplicar por π_k y sumar sobre k para usar la restricción de la maximización y nos queda $\lambda = -N$ y $\pi_k = \frac{N_k}{N}$, de forma tal que el coeficiente para la componente k está dado por la responsabilidad promedio de la misma por explicar los puntos de nuestros datos.

Ahora bien, estos resultados no constituyen una solución explícita para nuestro problema, dado que los mismos dependen de $\gamma(z_{nk})$ que depende a su vez de los parámetros que buscamos a través de (6.7). Sin embargo, sí sugieren un esquema iterativo: el algoritmo de ***Expectation-Maximization*** o E-M para mezcla de Gaussianas (el mismo es más general y puede utilizarse para otras distribuciones). Los pasos son los siguientes:

1. Inicializamos nuestros parámetros $\boldsymbol{\mu}_k$, $\mathbf{\Sigma}_k$ y π_k y evaluamos la *log-likelihood*.

2. Paso ***Expectation***: evaluamos las responsabilidades usando los valores actuales de los parámetros:
$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)} \tag{6.15}$$

3. Paso ***Maximization***: Reestimamos los parámetros usando las responsabilidades del paso anterior:
$$\begin{aligned} \boldsymbol{\mu}_k^{\text{nueva}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n \\ \mathbf{\Sigma}_k^{\text{nueva}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{nueva}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{nueva}}\right)^{\mathrm{T}} \\ \pi_k^{\text{nueva}} &= \frac{N_k}{N} \end{aligned} \tag{6.16}$$
donde $N_k = \sum_{n=1}^{N} \gamma\left(z_{nk}\right)$

4. Chequeamos convergencia evaluando la *log-likelihood* como en (6.8) (en general se decide frenar cuando el cambio en esta magnitud es menor que un cierto ϵ). En caso de que no se satisfaga el criterio de convergencia, se retorna al paso 2.

Es necesario tener en cuenta que que la *log-likelihood* puede tener varios máximos locales, y no hay garantías de que este algoritmo encuentre el mayor de ellos. Por eso es que más tarde vamos a usar una heurística para quedarnos con alguno suficientemente bueno.

Para llevar a cabo el *clustering* empleamos de nuevo la biblioteca *scikit-learn* que implementa el algoritmo anterior para estimar los parámetros de una mezcla de gaussianas en nuestro *set* de datos.

Las autoras en [19] utilizan el conocido método K-means para realizar el *clustering*. Sin embargo, tal como se puede leer en [5] y en otros lugares, este último método es un caso particular de un GMM en el que todas las $\mathbf{\Sigma}_k$ son matrices del tipo $\boldsymbol{\sigma}_k \cdot \mathbf{I}$ (múltiplos de la identidad). En otras palabras, solo puede encontrar *clusters* "circulares", por lo que decidimos utilizar GMM que nos provee mayor flexibilidad.

El método *fit* obtiene los parámetros del algoritmo E-M, mientras que *predict* utiliza el modelo ajustado para, con el mismo, asignarle a cada punto de nuestro *set* de datos un *cluster*. ¿De qué manera lo hace? Una vez que tenemos $\boldsymbol{\mu}_k$, $\mathbf{\Sigma}_k$ y π_k podemos evaluar las responsabilidades $\gamma(z_n k)$ como en (6.7). Recordemos que esta magnitud es la probabilidad *a posteriori* de z_k una vez observada $\mathbf{x}_n$. Bishop explica que también podemos interpretarla como la responsabilidad que tiene la componente k en explicar la observación $\mathbf{x}_n$. Por lo tanto, a cada una de estas últimas le asigna la componente que maximiza $\gamma(z_{nk})$.

6.4.2. Criterio de Información Bayesiana (BIC)

La pregunta que por ahora no abordamos es ¿Cómo encontramos K, la cantidad de *clusters*? Existen muchos criterios para la selección de este número. Algo que nos gustaría de este criterio es que, tal como venimos diciendo en este capítulo, logre un balance entre simpleza y explicabilidad. Si tomásemos cada punto como un *cluster*, tendríamos un modelo que se ajusta perfectamente a nuestros datos, pero sería un modelo extremadamente complejo que no nos permite buscar una estructura subyacente.

En [5] se introduce el Criterio de Información Bayesiana (BIC por sus siglas en inglés). Consideremos un *set* de datos $\mathcal{D}$ y un conjunto de modelos $\mathcal{M}_i$. Para cada modelo definimos una función de verosimilitud $p(\mathcal{D}|\theta_i, \mathcal{M}_i)$. Si introducimos una probabilidad *a priori* $p(\theta_i|\mathcal{M}_i)$, nos interesa calcular la evidencia del modelo $p(\mathcal{D}|\mathcal{M}_i)$ para los distintos modelos posibles que tenemos. Omitimos condicionar en $\mathcal{M}_i$ para mantener limpia la notación. La evidencia del modelo está dada por

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{6.17}$$

Ahora consideremos la expansión de Taylor del logaritmo de una función f alrededor de un punto estacionario $\mathbf{z}_0$. Como su gradiente se anula en ese punto tenemos (ignorando el resto y tomando $\mathbf{A} = -\mathcal{H}\left[\ln f(z)\right]|_{\mathbf{z}=\mathbf{z}_0}$)

$$\begin{aligned} \ln f(\mathbf{z}) &\simeq \ln f\left(\mathbf{z}_0\right) - \frac{1}{2}\left(\mathbf{z}-\mathbf{z}_0\right)^{\mathrm{T}} \mathbf{A}\left(\mathbf{z}-\mathbf{z}_0\right) \\ \Rightarrow f(\mathbf{z}) &\simeq f\left(\mathbf{z}_0\right) e^{-\frac{1}{2}\left(\mathbf{z}-\mathbf{z}_0\right)^{\mathrm{T}} \mathbf{A}\left(\mathbf{z}-\mathbf{z}_0\right)} \\ \Rightarrow \int f(\mathbf{z}) \mathrm{d}\mathbf{z} &\simeq f\left(\mathbf{z}_0\right) \int e^{-\frac{1}{2}\left(\mathbf{z}-\mathbf{z}_0\right)^{\mathrm{T}} \mathbf{A}\left(\mathbf{z}-\mathbf{z}_0\right)} \mathrm{d}\mathbf{z} \\ &= f\left(\mathbf{z}_0\right) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \end{aligned} \tag{6.18}$$

donde usamos en el último paso que el integrando se parece mucho a la densidad de una distribución normal (que integra 1) y por lo tanto la normalizamos.

Si ahora tomamos $f(\theta)$ como el integrando en (6.17) nos queda

$$\ln p(\mathcal{D}) \simeq \ln p\left(\mathcal{D} \mid \hat{\boldsymbol{\theta}}\right) + \ln p\left(\hat{\boldsymbol{\theta}}\right) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \tag{6.19}$$

donde $\hat{\boldsymbol{\theta}}$ es el valor de $\boldsymbol{\theta}$ en la moda de la distribución *a posteriori* y, si asumimos que $p(\theta) = \mathcal{N}(\theta|\mathbf{m}, \mathbf{V}_0)$,

$$\begin{aligned}\mathbf{A} &= -\mathcal{H}\left[\ln p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)\right]\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= -\mathcal{H}\left[\ln p\left(\mathcal{D} \mid \boldsymbol{\theta}\right)\right]\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \mathcal{H}\left[\ln p\left(\boldsymbol{\theta}\right)\right]\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \mathbf{H} + {\mathbf{V}_0}^{-1}\end{aligned} \tag{6.20}$$

Si asumimos que nuestra cantidad de puntos es grande, el último término es despreciable en relación al primero. De esta forma, podemos escribir (6.19) así:

$$\ln p(\mathcal{D}) \simeq \ln p\left(\mathcal{D} \mid \hat{\boldsymbol{\theta}}\right) - \frac{1}{2}\left(\hat{\boldsymbol{\theta}} - \mathbf{m}\right)\mathbf{V}_0^{-1}\left(\hat{\boldsymbol{\theta}} - \mathbf{m}\right) - \frac{1}{2}\ln|\mathbf{H}| + \text{constante} \tag{6.21}$$

y acá también podemos despreciar el segundo término del lado derecho si N es grande en relación al primer término.

Como asumimos datos i.i.d., $\mathbf{H} = -\mathcal{H}\left[\ln p\left(\mathcal{D} \mid \boldsymbol{\theta}\right)\right]\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ es una suma de términos, uno por cada punto. Así que podemos escribir

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{H}_n = N\widehat{\mathbf{H}} \tag{6.22}$$

donde $\mathbf{H}_n$ es la contribución del dato n-ésimo y $\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n$. Ahora, usando propiedades del determinante,

$$\ln|\mathbf{H}| = \ln|N\widehat{\mathbf{H}}| = \ln\left(N^M|\widehat{\mathbf{H}}|\right) = M\ln N + \ln|\widehat{\mathbf{H}}| \tag{6.23}$$

donde M es la dimensión de θ, y estamos asumiendo que $\widehat{\mathbf{H}}$ tiene rango máximo M. Finalmente, en esta última igualdad podemos despreciar $\ln|\widehat{\mathbf{H}}|$ debido a que es $O(1)$ comparado con $\ln N$.

Combinando todo esto, llegamos a que

$$\ln p\left(\mathcal{D}\right) \simeq \ln p(\mathcal{D}|\hat{\theta}) - \frac{1}{2}M\ln N \tag{6.24}$$

donde omitimos las constantes aditivas. Al lado derecho de (6.24) se lo llama Criterio de Información Bayesiana o BIC por sus siglas en inglés. El mismo penaliza la complejidad del modelo mientras que premia su explicabilidad, tal como queríamos.

La biblioteca *scikit-learn* implementa BIC como método de una instancia de un modelo de mezcla de gaussianas. Es decir, dado un modelo ya entrenado puede calcular el BIC para los datos que uno provee. Esta biblioteca multiplica por -2 la ecuación (6.24), por lo tanto un modelo será "mejor" cuanto menor sea el BIC, y de aquí en más llamaremos BIC a esta versión negativizada para mantener la consistencia.

6.4.3. *Clusterizamos*

Recordemos que por ahora tenemos dos *datasets* a partir de nuestras imágenes: en uno de ellos cada píxel $\in \mathbb{R}^{24}$ y contiene las medias y desvíos mensuales de NDVI. En el otro, cada píxel $\in \mathbb{R}^8$ y contiene las componentes que explican mayor varianza.

Para elegir un *clustering* de nuestras imágenes, iremos iterando $K_1, \ldots, K_N$ que representan la cantidad de *clusters* que le diremos al algoritmo de E-M que busque.

1. Dado un K_n, corremos el algoritmo de E-M veinte veces con semillas de aleatorización distintas para su inicialización (las semillas son elegidas y fijadas por nosotros para tener replicabilidad). Para cada corrida i-ésima nos quedamos con su BIC_i^n. Llamaremos $\overline{\mathrm{BIC}}^n \coloneqq \frac{1}{20}\Sigma_{i=1}^{20}\mathrm{BIC}_i^n$
2. Tomamos como $\mathrm{K} = \mathrm{argmín}_n\, K_n$
3. Dado entonces ese K, volvemos a correr el algoritmo veinte veces con esas semillas y nos quedamos con aquella que haya producido el mejor BIC.

Esto lo hacemos debido al grado de aleatoriedad que tiene la inicialización de los parámetros en los algoritmos de E-M. Queremos tener varios resultados y promediarlos para asegurarnos que un BIC alto o bajo no sea producto de haber comenzado con parámetros especialmente buenos o malos.

Como se ve en los gráficos 6.5 y 6.6, nos quedamos con 50 *clusters* para el caso de la imagen a la que le redujimos la dimensionalidad usando PCA (al *clustering* resultante lo llamaremos *clustering 1*, figura 6.7) y 22 para aquella en la que calculamos las medias y desvíos mensuales (*clustering 2*, figura 6.8).

6.5. Elección de un mapa *ground-truth*

Para poder hacer un análisis de nuestros *clusterings* resultantes, es necesario contar con algún mapa que represente la "verdad del terreno" o, en el lenguaje del aprendizaje automático, la *ground-truth*.

Con este fin decidimos emplear un mapa de unidades de humedal.

¿Qué son las unidades de humedal? Tal como plantean Kandus y Minotti en [17], las unidades de humedal son elementos del paisaje, predecibles en cuanto a su tipología, dada por el contexto hidrogeomórfico del paisaje donde se encuentran. A las mismas les corresponde la definición de humedal del comienzo de la tesis: permiten la acumulación permanente o temporaria de agua somera, y presentan rasgos distintivos asociados a criterios diagnósticos (régimen hidrológico, biota y suelo o sustrato).

Las mismas autoras, en [16], que es la propuesta que ellas hacen para un Inventario Nacional de Humedales, proponen cuatro niveles de análisis de acuerdo a la escala de la que estemos hablando. De menor a mayor granularidad, los niveles son las regiones, los sistemas de paisajes, las unidades de paisajes y por último las unidades de humedal. Es decir que estas últimas vienen a ser la unidad mínima de análisis de los humedales en términos espaciales.

En el mapa de unidades de humedal del Sitio Ramsar (figura 6.9) hay unas 29 unidades distintas. Nos pareció, sin embargo, que estas unidades constituían un nivel muy granular de análisis. Por lo tanto, las agrupamos en cinco categorías según el paisaje preponderante: cursos de agua (categoría 0), bosques (categoría 1), espiras (categoría 2), lagunas (categoría 3) y media loma (categoría 4). Estas se pueden ver en la figura 6.10.

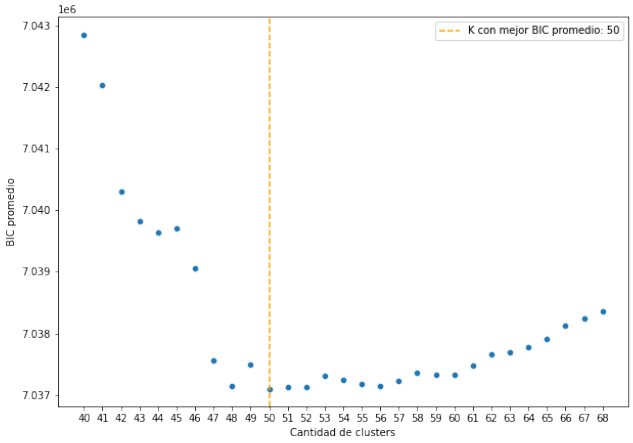


Fig. 6.5: Dimensionalidad reducida con PCA. Cantidad de *clusters* vs. BIC promedio.

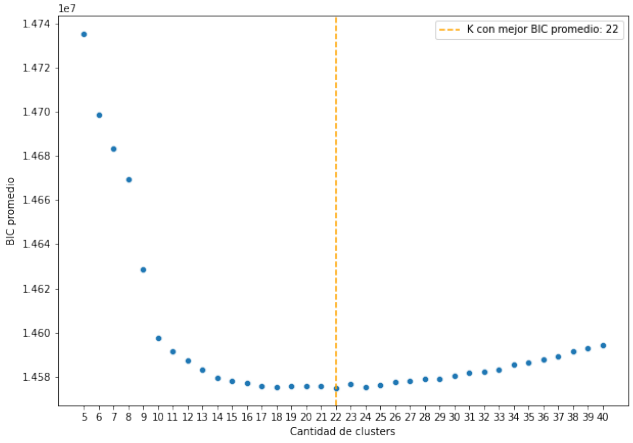


Fig. 6.6: Dimensionalidad reducida con medias y desvíos mensuales. Cantidad de *clusters* vs. BIC promedio.

Por otra parte, el mapa de unidades de humedal tiene una resolución alta (10m), por lo que una comparación directa con nuestros mapas de resolución media (250m) nos llevaría a confusiones. Pensemos que en cada píxel de nuestros mapas entran 625 píxels del mapa de unidades de humedal.

Decidimos entonces crear un "mapa de la verdad" que tuviese la misma resolución que nuestros *clustering* y que a la vez represente la realidad del terreno. Para esto, imaginemos que sobre el mapa de cobertura simplificado superponemos otro (por ahora "transparente") pero con una resolución más gruesa. En cada píxel "grande" caben (como dijimos en el párrafo anterior) 625 de los "chicos". Algunos serán de ríos, otros de bosques, etc. y en diferentes proporciones.

Entonces, con esta información armamos un mapa con cinco bandas (es decir, cada píxel es un vector de $\mathbb{R}^5$): cada una de ellas tiene la proporción de cada clase en ese píxel "grande". En la figura 6.11 podemos ver un ejemplo de esto: el píxel resultante es

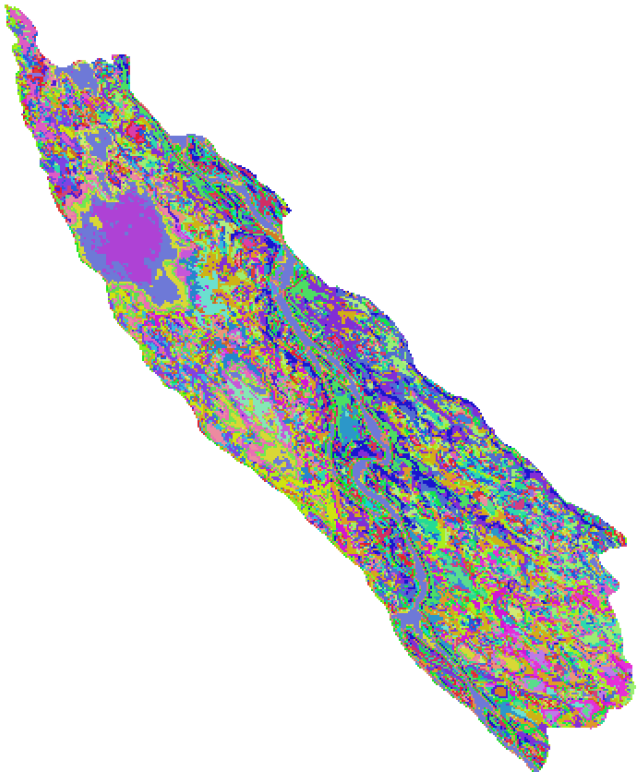


Fig. 6.7: Clustering 1

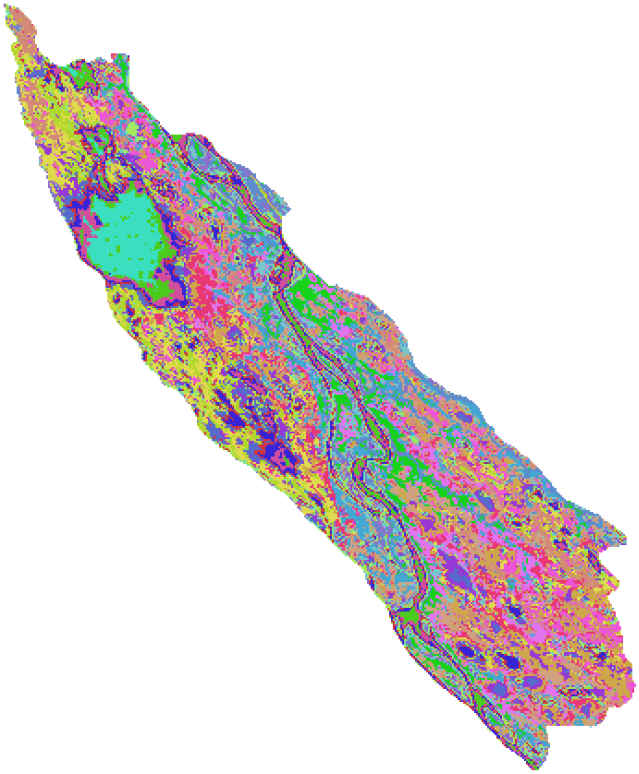


Fig. 6.8: Clustering 2

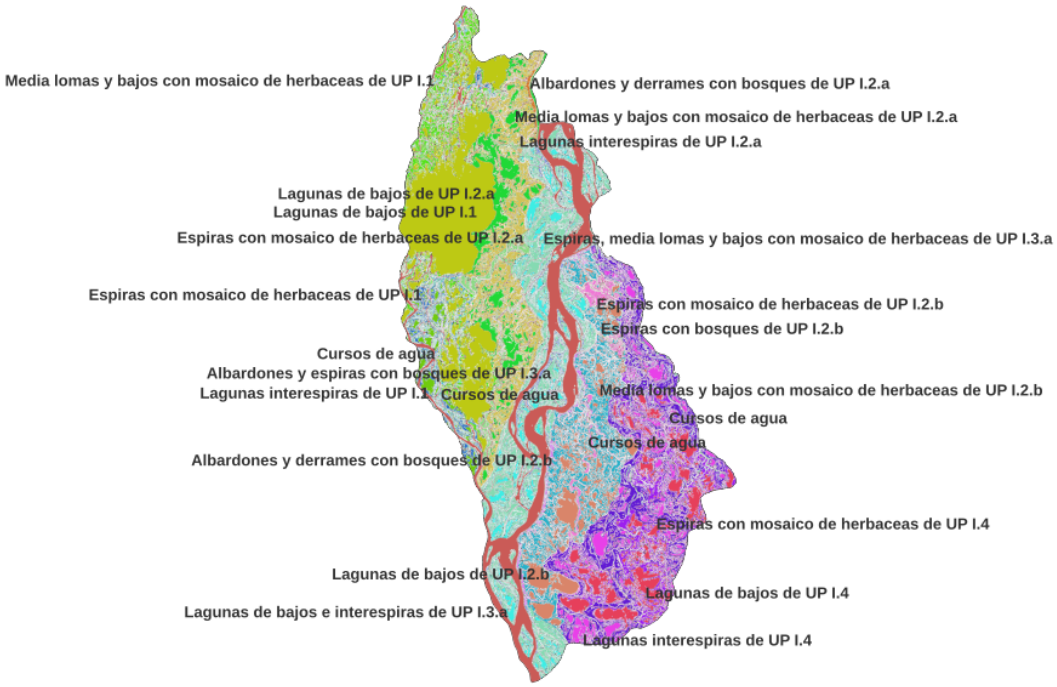


Fig. 6.9: Unidades de humedal Sitio Ramsar.

$(0{,}3472; 0; 0; 0{,}4032; 0{,}2496)$ porque la matriz de la derecha tiene 34,72 % de la clase 0 (cursos de agua), 0 % de de las clases 1 y 2 (bosques y espiras), 40,32 % de la clase 3 (lagunas) y 24,96 % de la clase 4 (media loma). Un píxel "puro" de cursos de agua hubiera sido el $(1; 0; 0; 0; 0)$, y su matriz toda de color violeta [1].

Ahora bien, cada uno de estos nuevos píxels, al ser multibanda es difícil de comparar con los mapas generados por nuestros *clustering*. No solamente por su dimensión sino porque los píxels de estos últimos tienen números enteros (la clase a la que pertenecen) y estos están en $\mathbb{R}^5$. Nosotros nos preguntamos ¿Cuáles son los píxels "parecidos" entre sí en este mapa multibanda? Los que tengan proporciones similares en cada componente. Entonces corrimos un *clustering* utilizando el modelo de mezcla de gaussianas, igual al que empleamos para elaborar nuestros mapas.

El resultado del mismo es un mapa de 26 clases al que le llamaremos *mapa GT* (por *ground-truth*), y lo podemos ver en la figura 6.12.

[1] En realidad este procedimiento y la figura 6.11 son ilustrativos. Para ser fieles a los hechos, partimos del mapa de cobertura que vemos en la figura 6.10 que es un vector en Qgis y fue hecho a partir de un mapa de tipo *raster* con una resolución de 10m × 10m. Este vector fue *rasterizado* con una resolución de 250m × 250m produciendo píxels mixtos, por lo cual se *rasteriza* realizando un mapa multibanda con las proporciones de cada tipo de píxel en cada banda.

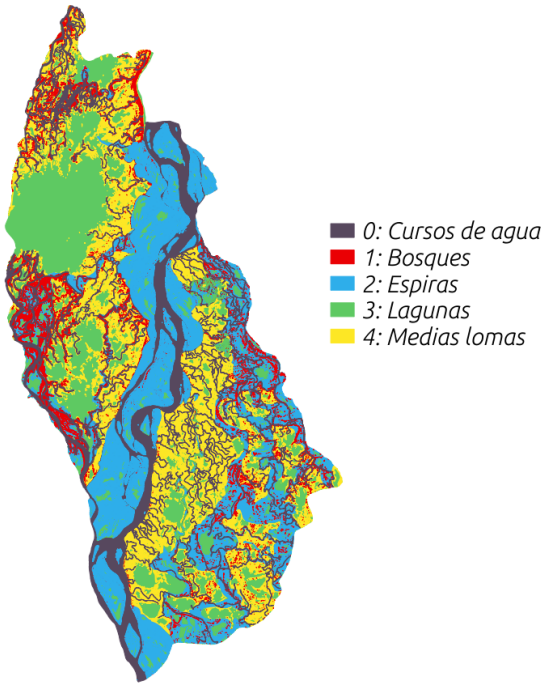


Fig. 6.10: Unidades de humedal simplificadas.

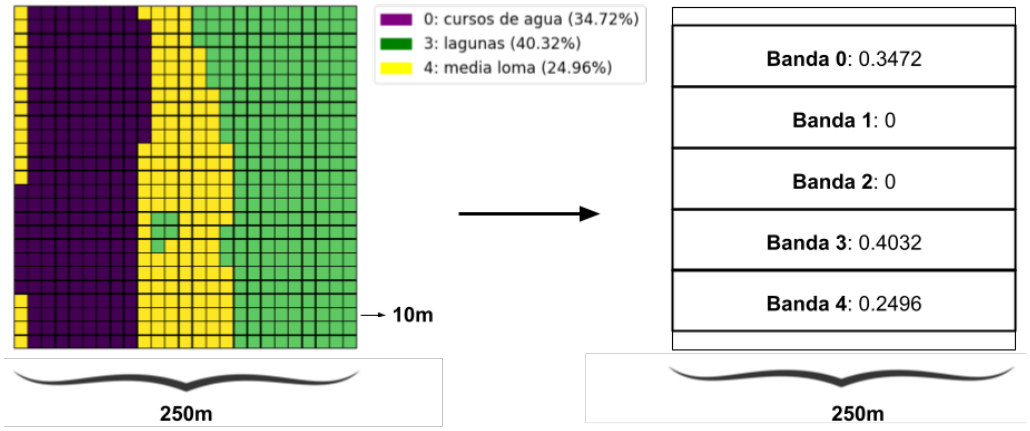


Fig. 6.11: Cómo construimos los píxels multibanda. A la derecha, 625 píxels del mapa de cobertura simplificado que cabe en un píxel de 250m de lado. En el píxel resultante, a la izquierda, tenemos en cada banda la proporción correspondiente a esa clase en la matriz de píxels de la derecha.

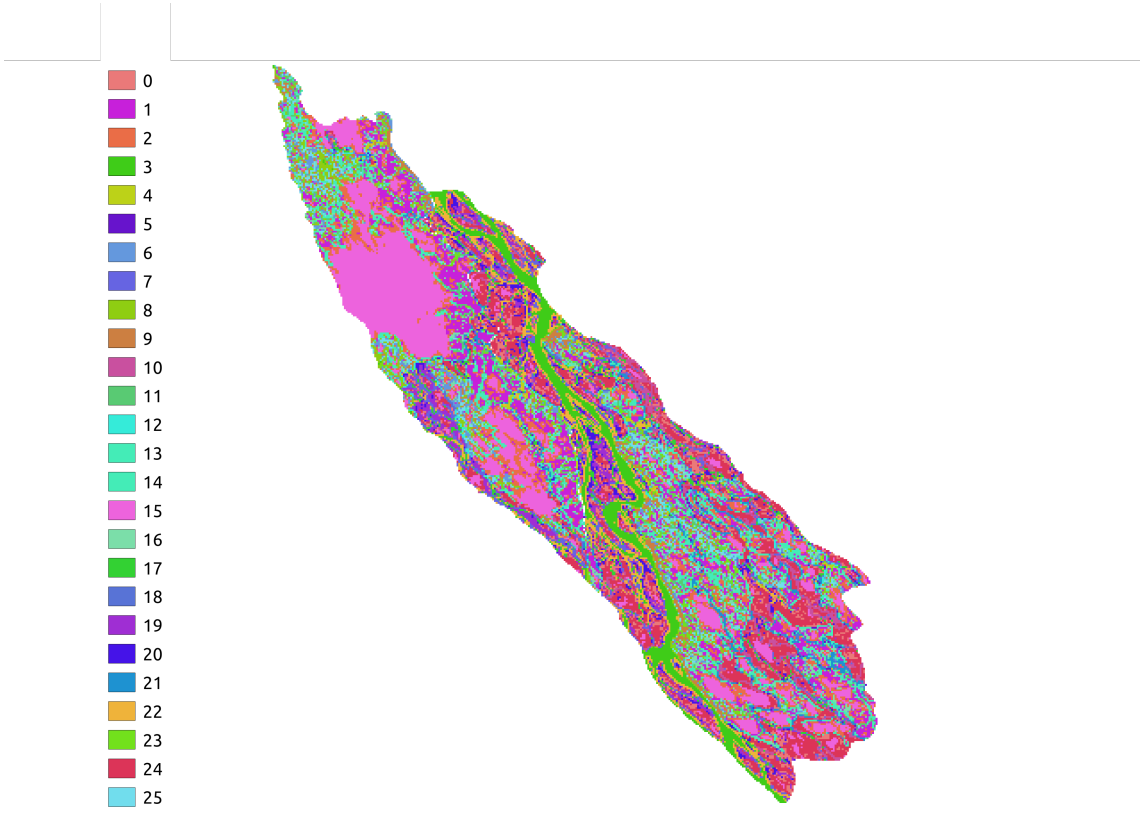


Fig. 6.12: Mapa GT.

7. COMPARANDO MAPAS

7.1. Definiciones y preguntas

En esta sección vamos a llamar "mapa" a un *clustering* o clasificación de manera indistinta, es decir, será un $M \in \mathbb{N}_0^n$, $M = [m_0, \ldots, m_{n-1}]$.

Y llamaremos "clases" a sus *clusters* o clases, independientemente de si les ponemos una etiqueta ("lagunas", "espiras", etc.) o no. Es decir

$$\texttt{clases}(M) = \{m_0, \ldots, m_{n-1}\} \tag{7.1}$$

Por otro lado, llamaremos $C_j(M)$ (usaremos C_j cuando sea evidente el M por el contexto para simplificar la notación) a la clase de M donde todos sus elementos (sus píxels) son iguales a j:

$$C_j(M) = \{i \in [0, \ldots, n-1] \mid m_i = j\} \tag{7.2}$$

Cuando comparamos dos mapas (llamémoslos $M = [m_0, \ldots, m_{n-1}]$ y $M' = [m'_0, \ldots, m'_{n-1}]$, con $k = \#\texttt{clases}(M)$ y $k' = \#\texttt{clases}(M')$), algunas preguntas pertinentes que nos podríamos hacer son:

- ¿Cómo se reparten los píxels de $C_j(M)$ a lo largo de las clases de M'?
- ¿Están en su mayoría representados por alguna de ellas?

Por eso decidimos armar un conjunto de funciones que nos permiten responder estas preguntas.

La primera toma un mapa M y un número de clase j del mismo (recordemos que su clase será $C_j(M)$), además de un M'; y nos responde cuál es el mejor representante de $C_j(M)$ en M'. La manera más directa de hacerlo es tomar todos los píxels de $C_j(M)$; contar, para cada $C_i(M')$, cuántos píxels de $C_j(M)$ pertenecen a $C_i(M')$; y buscar dónde se realiza el máximo de ese vector. O sea:

$$\begin{aligned} X^j &= [x_0^j, \ldots, x_{k'-1}^j] \\ x_i^j &= \#\{k \in [0, \ldots, n-1 \mid m_k = j \wedge m'_k = i\} \end{aligned} \tag{7.3}$$

y devolver $\operatorname{argmáx}_i X_i^j$.

Algo que nos puede ocurrir, sin embargo, es que muchos píxels de $C_j(M)$ se encuentren en una clase determinada de M' (por ejemplo, $C_k(M')$) pero que esto se deba a que $C_k(M')$ cubre un porcentaje muy grande de M'. Por eso es que decidimos "escalar" cada uno de esos números por el tamaño de cada clase.

Entonces podemos definir, finalmente,

$$\begin{aligned} X^j &= [x_0^j, \ldots, x_{k'-1}^j] \\ x_i^j &= \frac{\#\{l \in [0, \ldots, k'-1 \mid m_l = j \wedge m'_l = i\}}{\#C_i(M')} \\ \texttt{mejor\_representante\_para}_{M,M'}(j) &= \operatorname*{argmáx}_i X_i^j \end{aligned} \tag{7.4}$$

Entonces ahora tenemos una función que a cada una de las clases del mapa M le asigna su mejor representante en el mapa M'.

Esto nos permite definir una función que toma dos mapas (M y M') y nos devuelve una función

$$\texttt{mejor\_asignación}_{M,M'} : \texttt{clases}(M) \to \texttt{clases}(M') \tag{7.5}$$

que nos permite "mapear" cada una de las clases de M a las clases de M' de acuerdo a su mejor representante.

La otra función que nos interesa introducir es una consecuencia directa de la anterior: tomamos el M, y a cada uno de sus píxels le asignamos, según la clase a la que pertenezca (su valor) su mejor representante en M':

$$\begin{aligned} f_{M'}(M) = f_{M'}([m_0, \ldots, m_{n-1}]) = \\ \big[\texttt{mejor\_representante\_para}_{M,M'}(m_0), \ldots, \\ \texttt{mejor\_representante\_para}_{M,M'}(m_{n-1})\big] \end{aligned} \tag{7.6}$$

Es decir, tenemos un nuevo mapa al que podemos llamar $f_{M'}(M)$.

7.2. Aplicación a nuestros mapas

Pensemos ahora cómo podemos aplicar estas funciones al estudio de nuestros *clusterings* y sus comparaciones con el *mapa GT*.

Como explicamos antes, si utilizamos la función $\texttt{mejor\_asignación}_{clustering,GT}$, la misma nos da un "mapeo" de cada *cluster* del *clustering* a la clase que mejor lo representa en el mapa GT. Es decir que, en cierta manera, nos permite "etiquetar" cada uno de nuestros *clusters* y transformar nuestro *clustering* en una clasificación a través de $f_{GT}(clustering)$, permitiéndonos decir, por ejemplo "los *clusters* 3, 2 y 15 representan las lagunas, el 17 representa las espiras". Diremos entonces que $clasificación = f_{GT}(clustering)$.

Una vez que tenemos esto podemos pensar en razonar esto para el otro lado (es decir, empezando desde nuestro mapa GT) pero teniendo en cuenta la clasificación y preguntarnos qué representa $f_{clasificación}(GT)$. Por cómo definimos las funciones anteriormente, $\texttt{mejor\_asignación}_{GT,clasificación}$ nos da, para cada clase del mapa GT, su mejor representante dentro de $\texttt{clases}(clasificación)$.

Pensemos que

$$\begin{aligned} \mathrm{Im}(f_{M'}(M)) \subseteq \texttt{clases}(M') \implies \\ \mathrm{Im}(f_{GT}(clustering)) \subseteq \texttt{clases}(GT) \implies \\ \mathrm{Im}(f_{clasificación}(GT)) \subseteq \texttt{clases}(clasificación) \\ = \texttt{clases}(f_{GT}(clustering)) \subseteq \texttt{clases}(GT) \end{aligned} \tag{7.7}$$

Dado que $\texttt{mejor\_asignación}_{clustering,GT}$ es una función cuyo dominio es $\texttt{clases}(GT)$ y su codominio es $\texttt{clases}(clasificación)$ (que a su vez es $\texttt{clases}(GT)$), lo que puede interpretarse es que $\texttt{mejor\_asignación}_{GT,clasificación}$ nos da un "mapeo" de $\texttt{clases}(GT)$ a $\texttt{clases}(GT)$. De esta manera, podemos ver qué clases del mapa GT nuestro método de *clustering* "junta".

Lo que podemos pensar es que nuestros mapas 1 y 2 contienen clases espectrales, mientras que el mapa GT contiene clases de información. Las del primer tipo tienen que ver con el mundo de los datos numéricos de NDVI, es lo que generan nuestros algoritmos

de *clustering*. Las del segundo tipo tienen una interpretación, una semántica correspondiente a las unidades de humedal y al paisaje predominante en el área que esas clases ocupan. Las funciones construidas nos permiten hacer, de alguna manera, una asociación entre ambos tipos de clases: $\texttt{mejor\_asignación}_{clustering,GT}$ (la "ida") nos permite saber qué clases espectrales corresponden a la misma clase de información y, viceversa, $\texttt{mejor\_asignación}_{GT,clasificación}$ (la "vuelta") nos dice qué clases de información se confunden espectralmente. Por ejemplo, las lagunas vegetadas o con agua libre pertenecen a la misma clase de información aunque espectralmente pertenezcan a *clusters* diferentes. A su vez, los cuerpos de agua lóticos y lénticos (ríos y lagunas) son objetos de clases de información muy diferentes pero espectralmente pueden confundirse en un mismo *cluster*.

7.3. Herramientas para el análisis de los resultados

En esta sección queremos analizar nuestros métodos de *clustering* desde la perspectiva que contamos en la sección anterior.

7.3.1. Descripciones de las clases

A cada una de las clases del *mapa GT* les asignamos una descripción, que se corresponde con los tipos de unidades de humedal que se ven en la Figura 8.5. Esto lo hicimos para las clases que ocuparan más de un 2 % de los píxels del mapa, dado que tratar con clases tan minoritarias no aporta al análisis. Cuando una clase tiene en su descripción algún tipo de unidad de humedal en negritas, es porque la misma ocupa más del 50 % de los píxels de esa clase. Asimismo, los signos = y > muestran el orden de predominancia de tipos de unidades de humedal en cada clase. Por ejemplo, la 14 está compuesta en su mayoría por media loma (más del 50 %) seguida en menor medida por lagunas y cursos, que representan una proporción similar en la clase.

7.3.2. Firmas temporales

Un elemento muy útil a la hora de analizar las clases es el concepto de firma temporal promedio. La firma temporal promedio de una clase intenta capturar el comportamiento de la misma a lo largo del tiempo. Recordemos que tenemos una serie de 416 imágenes, separadas cada 16 días. Entonces, dada una clase j:

1. tomamos de nuestras imágenes originales la primera (correspondiente al 4 de julio de 2002),
2. de esa imagen tomamos todos los píxels correspondientes a la clase j,
3. calculamos el NDVI promedio y la desviación estándar. Ese será el primer punto de la firma temporal promedio para la clase j.
4. Repetimos los pasos 1 a 3 para todas las fechas.

Así, vamos construyendo el gráfico de una función que representa la evolución a lo largo del tiempo del NDVI promedio (y desvío) para cada clase. En lo que sigue veremos varios ejemplos aplicados a nuestros mapas.

Clase	Descripción
0	**Espira** > Laguna
1	**Media loma** > laguna
2	**Laguna** > Media loma
3	**Curso**
6	**Media loma** > laguna = bosque
7	Espira > Curso > bosque
8	Media loma = curso = laguna
9	**Media loma** > bosque
10	**Espira** > Laguna
14	**Media loma** > laguna = curso
15	**Laguna**
16	**Media loma** > laguna
17	**Espira** > bosque > curso
19	**Espira** > bosque > laguna
20	**Espira** > bosque
22	**Curso** > espira > bosque
24	**Espira** > Laguna

Tab. 7.1: Descripciones de las clases del mapa GT

7.4. Resultados

En las tablas siguientes podemos ver lo que mencionábamos al final de la sección anterior: cada tabla representa $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa)}$ para el mapa 1 (el *clustering* con GMM que comenzó con reducción de dimensionalidad utilizando PCA) y el mapa 2 (el *clustering* con GMM que comenzó con reducción de dimensionalidad utilizando las medias y desvíos mensuales). Para poder ver los mapas de manera más detallada e interactiva, sugerimos descargar los mapas del repositorio publicado (ver apéndice) para visualizarlos con el *software* Qgis.

Algo que podemos ver en ambas tablas 7.2 y 7.3 es que en las descripciones de $\mathrm{Im}(f_{clasificación}(GT))$ (es decir en la última columna) no hay descripciones repetidas. Si las hubiera, sería una deficiencia del *clustering* el hecho de no juntarlas. O al menos deberíamos preguntarnos por qué no lo hace, dado que tienen características similares.

Otro aspecto a remarcar es que en muchos casos (para el mapa 1: 0, 1, 2, 3, 6, 8, 9, 14, 15, 16 y 21; para el mapa 2: 1, 2, 3, 8, 9, 10, 12, 15 y 22) $\texttt{mejor\_representante\_para}_{GT,f_{GT}(mapa)}(j) = j$, lo cual significa que nuestros *clusterings* no están mezclando clases. Pensemos cómo sería un *clustering* perfecto: sería uno que clasifique todos los píxels en exactamente la misma clase a la que pertenecen en el mapa GT. Eso nos daría una $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa)}$ igual a la identidad.

Por otro lado, en los casos en los cuales $\texttt{mejor\_representante\_para}_{GT,f_{GT}(mapa)}(j) = i$ con $i \neq j$ se puede ver que las clases i y j tienen descripciones similares e, incluso, en muchos casos idénticas.

Para hacer un análisis un poco más profundo debemos utilizar el conocimiento del régimen de disturbios del área de estudio que, tal como mencionamos en 3.2.3 y 3.2.2, consisten en las inundaciones regulares y extraordinarias; así como también las firmas temporales promedio.

clase	Descripción	mejor_representante_para$_{\mathbf{GT,f_{GT}(mapa1)}}$(clase)	Descripción
0 24	**Espira** > Laguna **Espira** > Laguna	0	**Espira** > Laguna
1	**Media loma** > laguna	1	**Media loma** > laguna
2	**Laguna** > Media loma	2	**Laguna** > Media loma
3	**Curso**	3	**Curso**
10	**Espira** > Laguna	5	
6 12 25	**Media loma** > laguna = bosque Espira > media loma = bosque = curso **Espira** > Laguna	6	**Media loma** > laguna = bosque
8	Media loma = curso = laguna	8	Media loma = curso = laguna
9 19 20	**Media loma** > bosque **Espira** > bosque > laguna **Espira** > bosque	9	**Media loma** > bosque
14	**Media loma** > laguna = curso	14	**Media loma** > laguna = curso
15	**Laguna**	15	**Laguna**
16	**Media loma** > laguna	16	**Media loma** > laguna
17	**Espira** > bosque > curso	20	**Espira** > bosque
21 7 22	Curso > espira > bosque Espira > Curso > bosque **Curso** > espira > bosque	21	Curso > espira > bosque

Tab. 7.2: mejor_asignación$_{GT,f_{GT}(mapa1)}$

clase	Descripción	mejor_representante_para$_{\mathbf{GT,f_{GT}(mapa2)}}$(clase)	Descripción
1 16	**Media loma** > laguna **Media loma** > laguna	1	**Media loma** > laguna
2	**Laguna** > Media loma	2	**Laguna** > Media loma
3	**Curso**	3	**Curso**
8 14	Media loma = curso = laguna **Media loma** > laguna = curso	8	Media loma = curso = laguna
6 9 19 20	**Media loma** > laguna = bosque **Media loma** > bosque **Espira** > bosque > laguna **Espira** > bosque	9	**Media loma** > bosque
10 11 21	**Espira** > Laguna **Media loma** > curso Curso > espira > bosque	10	**Espira** > Laguna
12 17 25	Espira > media loma = bosque = curso **Espira** > bosque > curso **Espira** > Laguna	12	Espira > media loma = bosque = curso
15	**Laguna**	15	**Laguna**
0 24	**Espira** > Laguna **Espira** > Laguna	20	**Espira** > bosque
7 22	Espira > Curso > bosque **Curso** > espira > bosque	22	**Curso** > espira > bosque

Tab. 7.3: mejor_asignación$_{GT,f_{GT}(mapa2)}$

Las mismas nos muestran, para cada clase del mapa GT, la manera en la cual su estructura varía fenológicamente[1] y cómo está sujeta a las inundaciones y secas. Su grado de oscilación nos marca la variación estructural de la cubierta vegetal a lo largo del tiempo. Analicemos esto más detalladamente.

La clase 3 (Figura 7.1), que representa los cursos de agua, queda sola en ambas asignaciones (es decir, son asignadas a sí mismas). Es distintiva por tener una desviación estándar muy baja, oscilaciones muy pequeñas y mantenerse constantemente cerca del 0 y con valores negativos. Ninguna otra clase se comporta de esta manera. Además de eso, debido a que representa cursos de agua permanentes, su firma temporal prácticamente no se ve afectada por las inundaciones. Sí vemos el efecto de la sequía en 2008, durante la cual la misma alcanza valores negativos. El pico que se ve entre 2006 y 2007 puede deberse a

[1] La fenología es el estudio de eventos biológicos periódicos; en la práctica aplicado frecuentemente a fenómenos de por sí periódicos, tales como los patrones de crecimiento, desarrollo y reproducción en un organismo a lo largo de su vida en relación a las estaciones. [2]

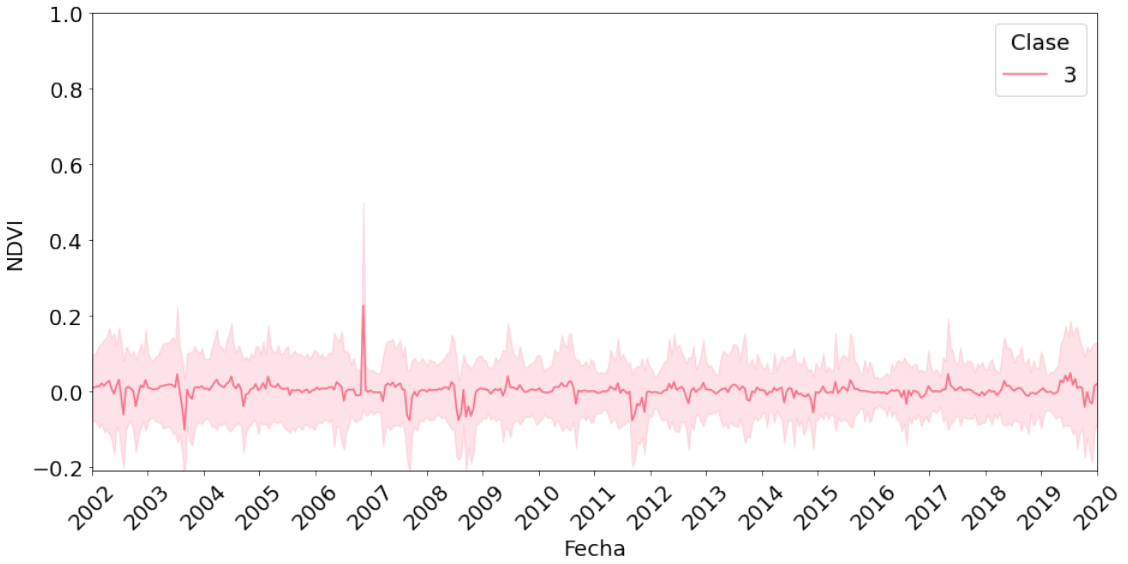


Fig. 7.1: Firma temporal de la clase 3.

errores de medición del instrumento o bien puede haber ocurrido que durante una época de bajante haya quedado expuesta mucha vegetación por estar en un suelo que suele estar cubierto de agua y por lo tanto no se llegó a secar.

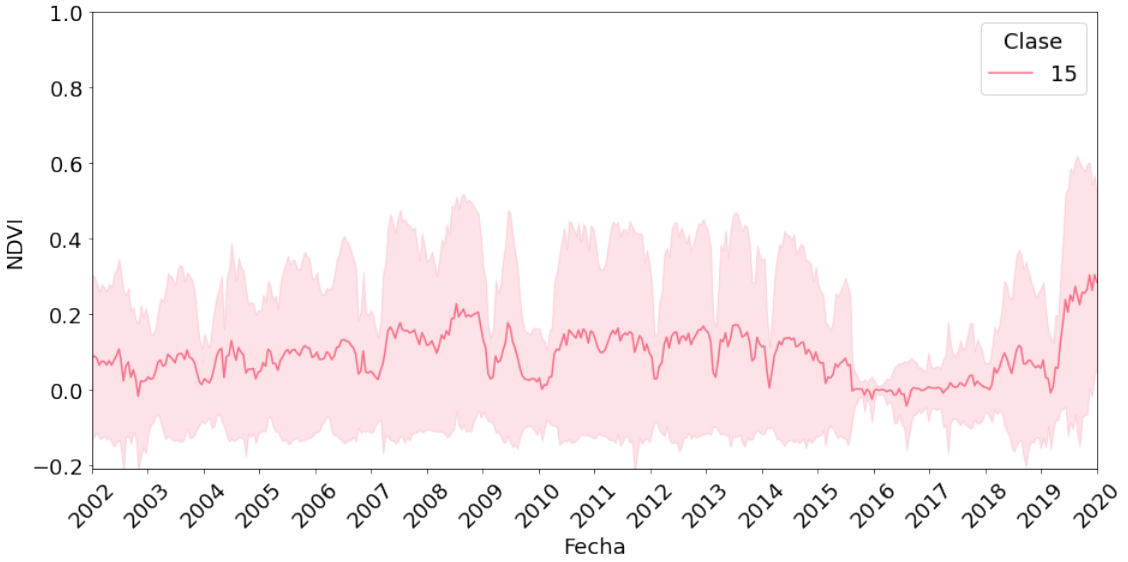


Fig. 7.2: Firma temporal de la clase 15.

La clase 15 (Figura 7.2) también es asignada a sí misma en ambos mapeos y representa las lagunas. En este caso se ve, a pesar de que la firma temporal promedio se mantiene cerca del 0, mucho más desvío estándar. Si recordamos que nuestros gráficos representan la firma temporal promedio de la clase, es natural pensar que un alto desvío representa la heterogeneidad espacial de la clase, es decir, lo diversa que es internamente. Podemos ver que el gráfico presenta picos y valles marcados e irregulares (aunque mucho menos

que las clases que veremos después) entre 2007 y 2010, años con disturbios frecuentes que provocan que los procesos de sucesión no lleguen a terminarse. En los cinco años subsiguientes presenta un patrón regular hasta que en 2016 la firma se desploma hacia su mínimo debido a las inundaciones, y no solo eso sino que también su desvío se achica dado que estos disturbios homogeneizan toda la región ocupada por la clase, llenándola de agua. Mientras que otras clases tardan solamente un año en recuperar sus niveles de NDVI, es decir, sus niveles de vegetación, esta clase lo hace tres años después, llegando en 2020 a niveles incluso superiores a los previos.

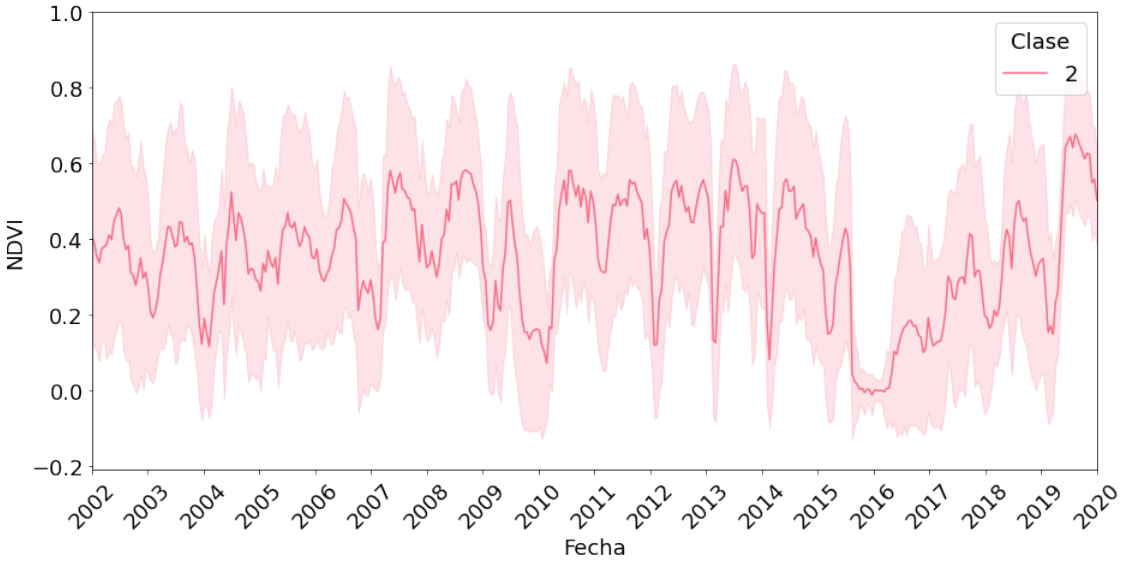


Fig. 7.3: Firma temporal de la clase 2.

La clase 2 (Figura 7.3), que también tiene la identidad como asignación en ambos mapas, tiene oscilaciones mucho más marcadas, es decir, mucha más variabilidad fenológica. Esto se debe a que representa parte del interior y los bordes de las lagunas, estos últimos compuestos por medias lomas que contienen plantas herbáceas. Por eso mismo es que se ve más afectada por las inundaciones, su NDVI promedio es más alto (alrededor de 0,3) y también lo es su desviación estándar salvo durante las inundaciones de 2016, tal como ocurre en la clase anterior, con la que también comparte su régimen irregular entre 2007 y 2010 y su lenta recuperación después de estos disturbios hasta llegar a niveles de NDVI superiores a los picos de años anteriores.

No es el caso, por ejemplo, de las clases 12, 17 y 25 (Figura 7.4), formadas fundamentalmente por espiras que se encuentran en zonas aledañas a los cursos de agua (no solo el Paraná sino también otros secundarios como el Paranacito). A través de $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa2)}$ las tres pasan a formar parte de una única clase. Tienen en común un NDVI promedio cercano a 0,6, oscilaciones grandes pero regulares, un desvío estándar bajo y una recuperación rápida después de las inundaciones (se ven marcados los valles en 2007, 2010 y 2016), lo cual puede deberse en parte a la topografía (suelen ser zonas más elevadas, sobre todo aquellas que tienen bosques como las clases 12 y 17). Algo a destacar es que la firma temporal de la clase 25, aunque es muy similar al resto, se encuentra consistentemente por debajo de las firmas de las otras clases. Esta clase contiene lagunas además de espiras, lo cual puede explicar este comportamiento. A través de $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa1)}$ la clase 6 va a

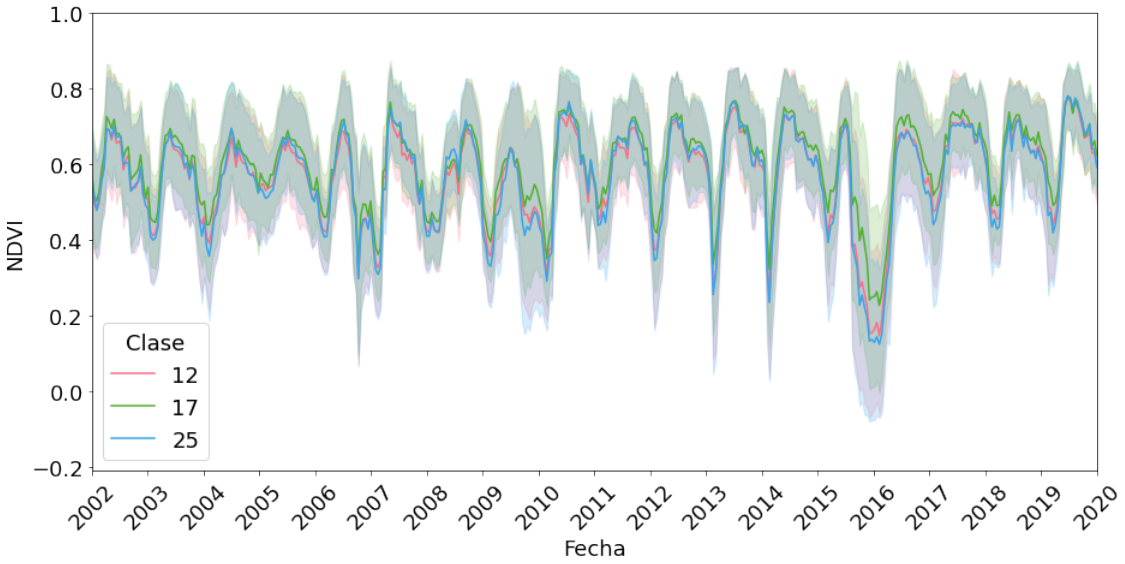


Fig. 7.4: Firmas temporales de las clases 12, 17 y 25.

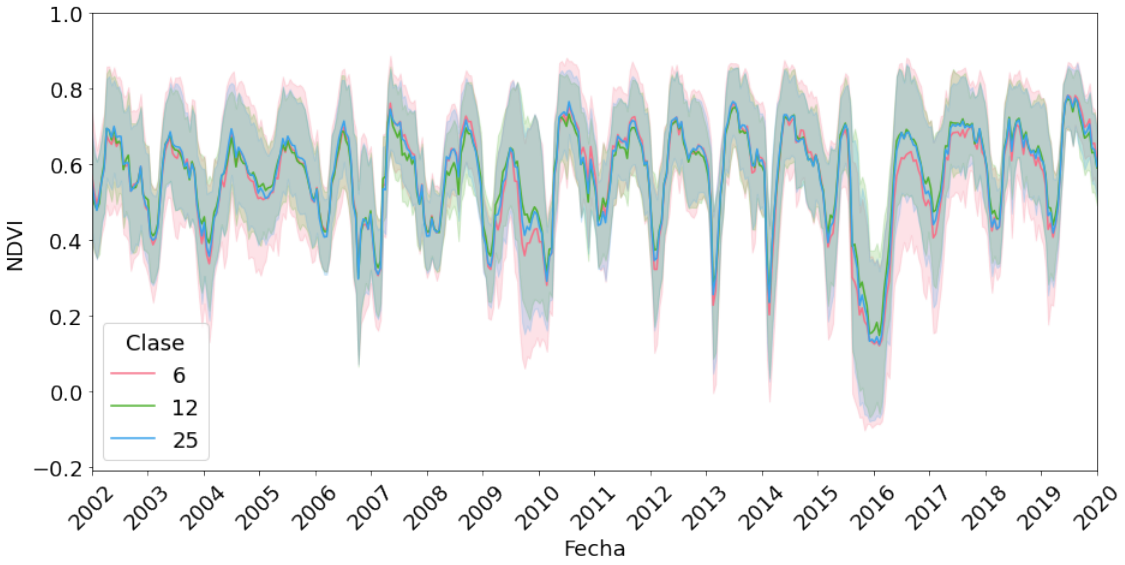


Fig. 7.5: Firmas temporales de las clases 6, 12 y 25.

parar al grupo de clases que veremos a continuación, que tiene varias similitudes (Figura 7.5).

Las clases 9, 19 y 20 (Figuras 7.6 y 7.7), que también son "fusionadas" a través de ambas asignaciones), comparten con el grupo anterior una alta oscilación, picos y valles marcados y bajo desvío estándar; lo que nos indica poca variación estructural al interior de cada una de las clases y una variabilidad importante a nivel fenológico. A diferencia del grupo de clases anterior, el NDVI es más alto y la recuperación es dispar entre las clases después de las inundaciones de 2016. Por otra parte, en este grupo las firmas están más separadas que en el anterior, donde se juntan en aguas bajas. Esto nos indica que las clases de este último, durante el estiaje, tienen niveles de vegetación similares pero durante las crecientes hay clases más inundadas que otras. De todos modos, la separación es menor.

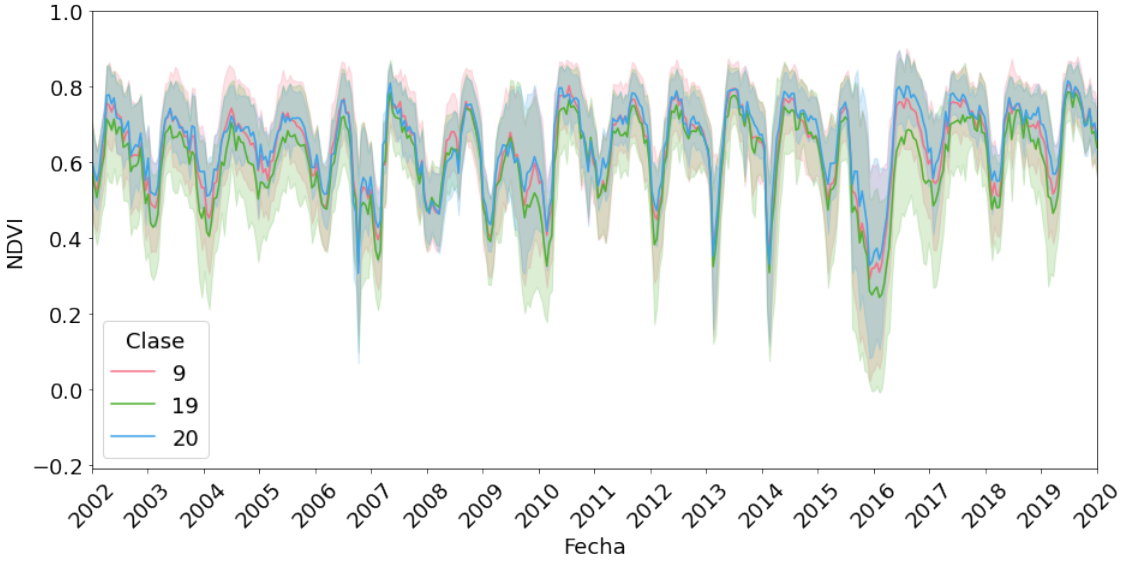


Fig. 7.6: Firmas temporales de las clases 9, 19 y 20.

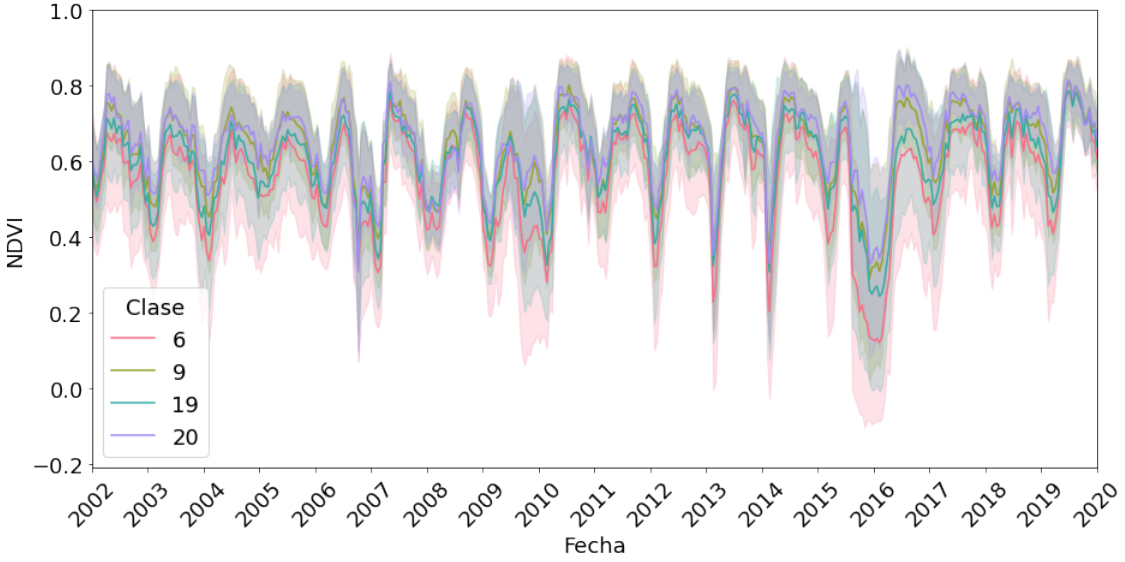


Fig. 7.7: Firmas temporales de las clases 6, 9, 19 y 20.

En estos grupos, por el contrario, las señales están separadas tanto en aguas bajas como en aguas altas. Si vamos al mapa GT, se pueden ver los albardones del margen derecho del Paraná, los bosques del Parque Nacional Pre Delta, los derrames del Paraná y las islas del cauce principal.

Un comportamiento parecido encontramos en las clases 0, 20 y 24 (Figuras 7.8 y 7.9), con la diferencia de que este conjunto tiene más presencia de lagunas. La clase 20 aparece

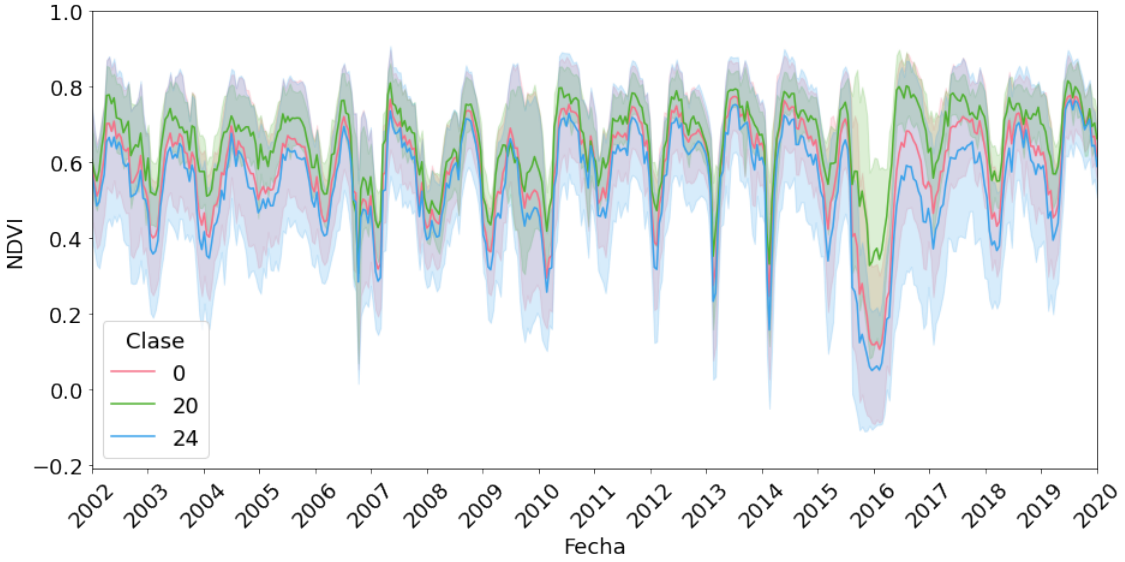


Fig. 7.8: Firmas temporales de las clases 0, 20 y 24.

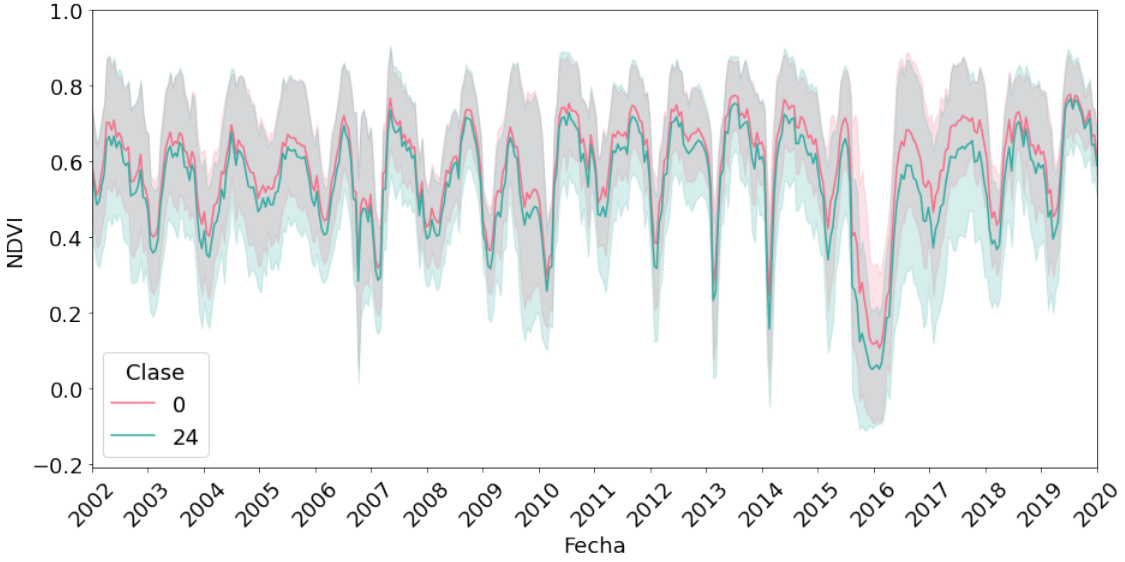


Fig. 7.9: Firmas temporales de las clases 0 y 24.

en el conjunto de clases anterior y en este debido a que, como se ve en la Tabla 7.3,

$$\begin{aligned} \texttt{mejor\_representante\_para}_{\mathbf{GT},\mathbf{f_{GT}(mapa2)}}(\mathbf{20}) &= 9 \quad , \\ \texttt{mejor\_representante\_para}_{\mathbf{GT},\mathbf{f_{GT}(mapa2)}}(\mathbf{0}) &= 20 \quad \text{y} \\ \texttt{mejor\_representante\_para}_{\mathbf{GT},\mathbf{f_{GT}(mapa2)}}(\mathbf{24}) &= 20 \end{aligned}$$

Las mismas están formadas fundamentalmente por espiras, y la 20 es la única que tiene bosques, por eso no es sorpresa que su firma temporal se mantenga consistentemente por encima de las demás y presente una recuperación mucho más inmediata luego de las inundaciones. En estas clases vemos los derrames hacia la laguna de Victoria y el área activa del sistema Paraná-Coronda. Se observa una mezcla de áreas altas y bajas que explican la

separación de las señales de NDVI en casi todo momento.

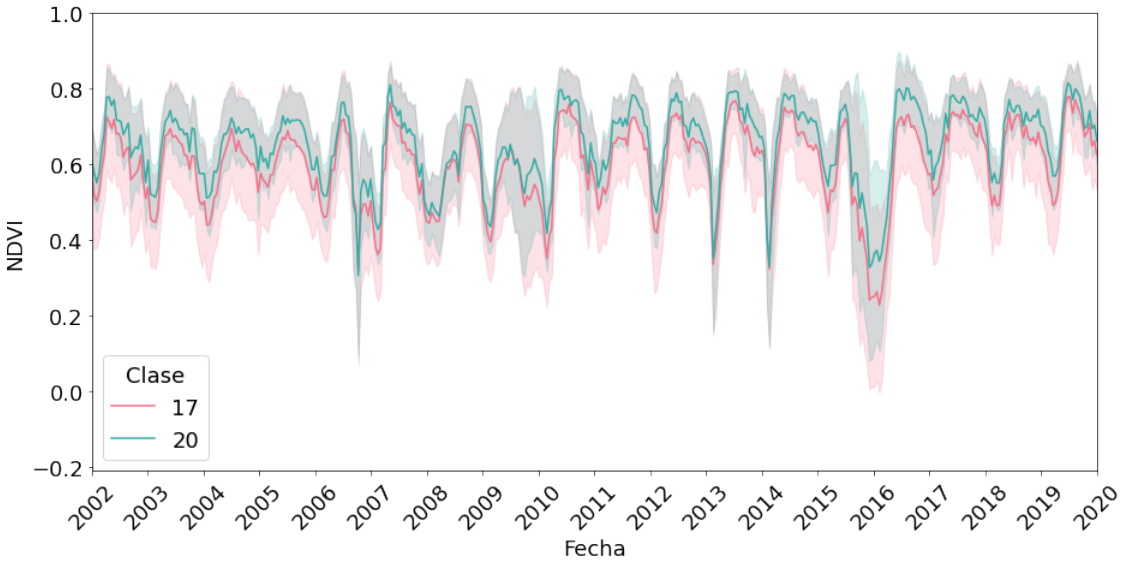


Fig. 7.10: Firma temporal de las clases 17 y 20.

La clase 20 también es particular en $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa1)}$ dado que

$\texttt{mejor\_representante\_para}_{\mathbf{GT},\mathbf{f_{GT}(mapa1)}}(\mathbf{20}) = 9$ (al igual que para el mapa 2) y
$\texttt{mejor\_representante\_para}_{\mathbf{GT},\mathbf{f_{GT}(mapa1)}}(\mathbf{17}) = 20$

Este grupo de clases es muy similar al grupo 9, 19, 20 porque se ven compactas (poco desvío estándar), patrones regulares y recuperación veloz luego de disturbios. Representan las zonas próximas a los cursos de agua, generalmente elevadas topográficamente, lo cual puede explicar su alta resiliencia y NDVI promedio elevado. La firma temporal de la clase 20 se encuentra por encima de la 17 porque esta última clase toma bordes de ríos, por lo que puede tener mayor presencia de agua.

Otro grupo que se diferencia mucho de los anteriores es el de las clases 7 y 22 (Figuras 7.12 y 7.11), ambas compuestas por espiras, cursos de agua y bosques. Su NDVI promedio se encuentra alrededor de 0,4, su oscilación es baja y su desvío estándar es alto. No se encuentra una baja tan abrupta en su firma temporal durante los disturbios como en los otros grupos de clases y la recuperación es rápida. Podemos ver que la firma de la clase 7 se mantiene consistentemente por encima de la firma de la clase 22, lo cual tiene sentido dado que en la segunda clase predominan los ríos mientras que en la segunda no. La excepción a esto son las las crecientes (o los mínimos de nuestro gráfico), donde se junta la señal. Esto puede explicarse por el hecho de que el agua homogeneiza los terrenos. También puede estar ocurriendo que en los períodos de aguas bajas (cuando hay más vegetación al descubierto) la clase 7 tenga más cubierta vegetal. A través de $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa1)}$ estas clases también se juntan, solo que en este último mapeo también la 21 entra en este grupo. No creemos que sea una clase muy relevante debido al porcentaje de píxels bajo que ocupa (1,75 % del total) pero su firma de NDVI promedio se encuentra por encima de las otras dos (alrededor de 0,6) salvo durante el 2016 donde se ve mucho más afectada por las inundaciones, llegando a tocar el 0.

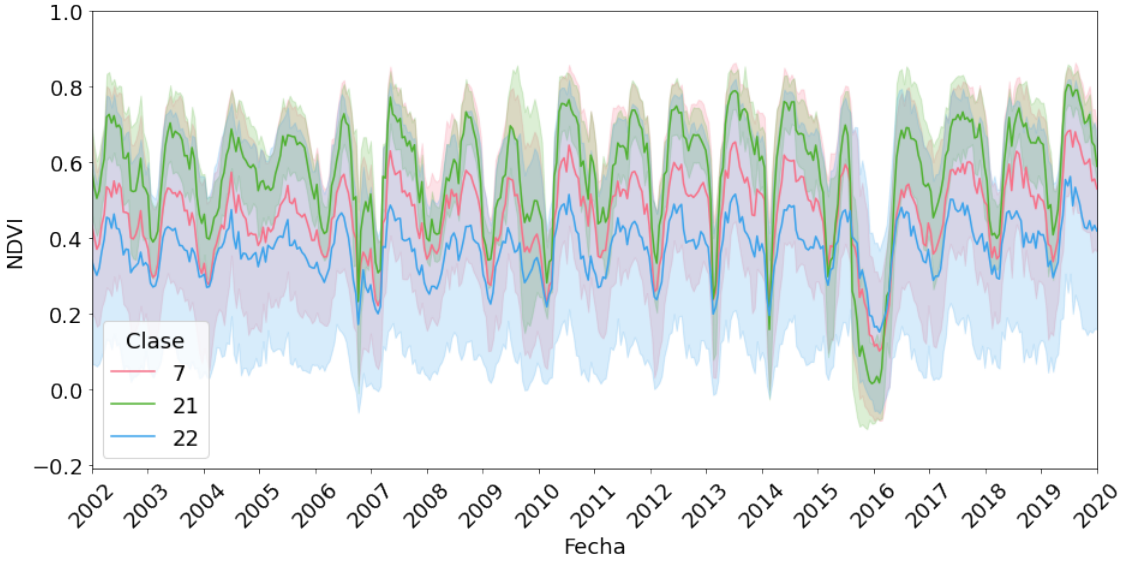


Fig. 7.11: Firmas temporales de las clases 7, 21 y 22.

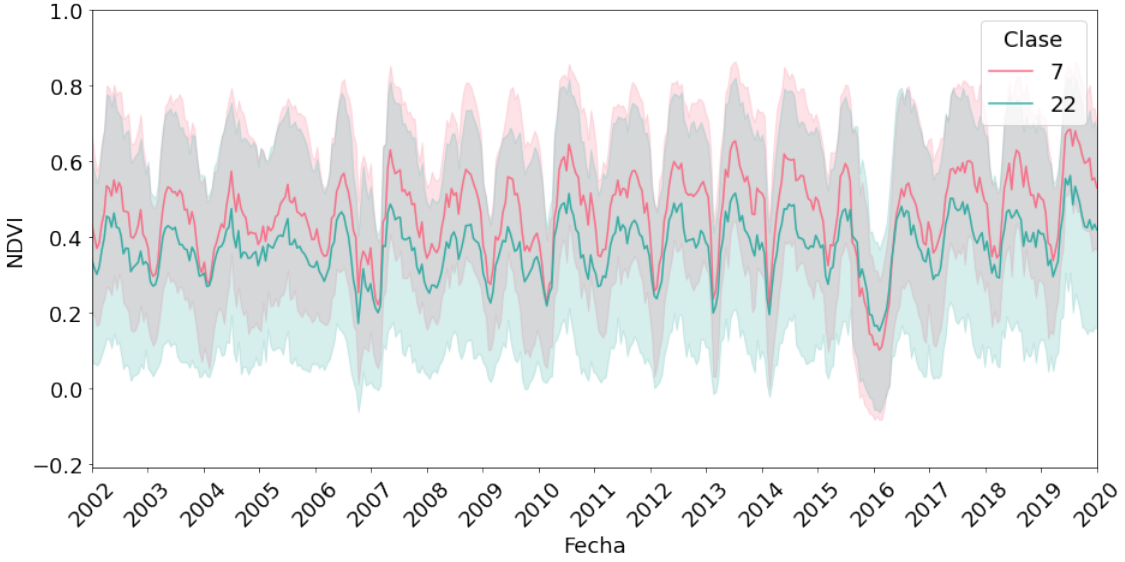


Fig. 7.12: Firmas temporales de las clases 7 y 22.

Esta última clase forma parte de otro grupo a través de $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa2)}$): 10, 11, 21. Las tres firmas son muy parecidas, con poco desvío estándar y picos y valles marcados. Los primeros tienen un patrón bastante regular (es decir que en períodos de aguas bajas la vegetación suele llegar siempre a los mismos niveles, lo cual nos indica una alta resiliencia) mientras que los segundos son más caóticos, mostrándonos niveles dispares de afectación a lo largo de las diferentes inundaciones. Tal como ocurre en otras firmas temporales, durante las mismas las señales se juntan por la presencia de agua.

Las clases 1 y 16 (que quedan separadas bajo $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa1)}$ pero juntas bajo $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa2)}$) son prácticamente idénticas, están compuestas mayormente por medias lomas y (de forma secundaria) por lagunas. Sus patrones son bastante más irregulares y su recuperación después de los disturbios es lenta.

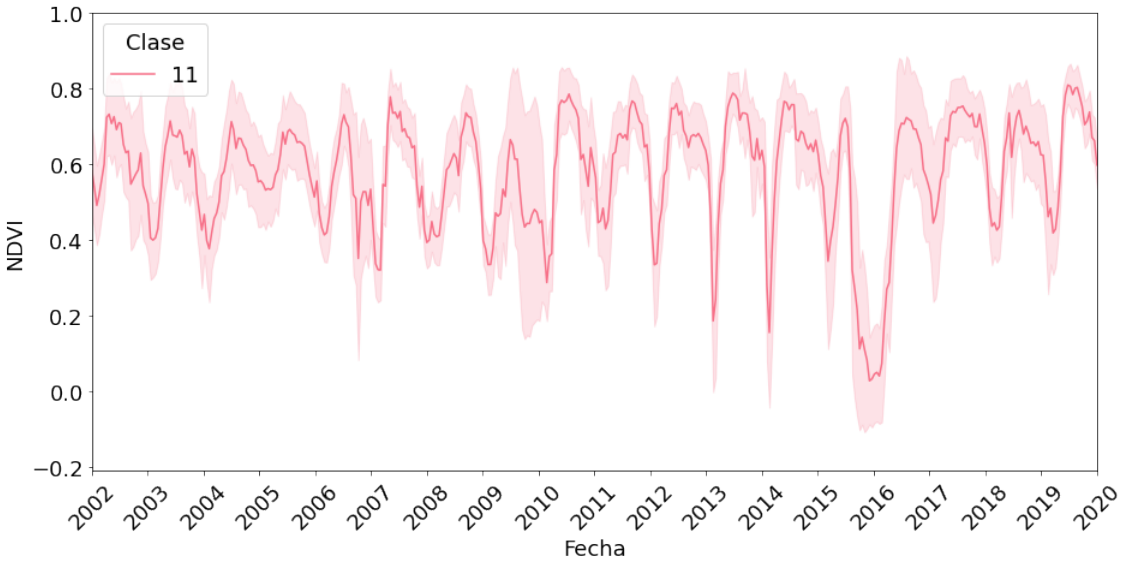


Fig. 7.13: Firmas temporales de la clase 11.

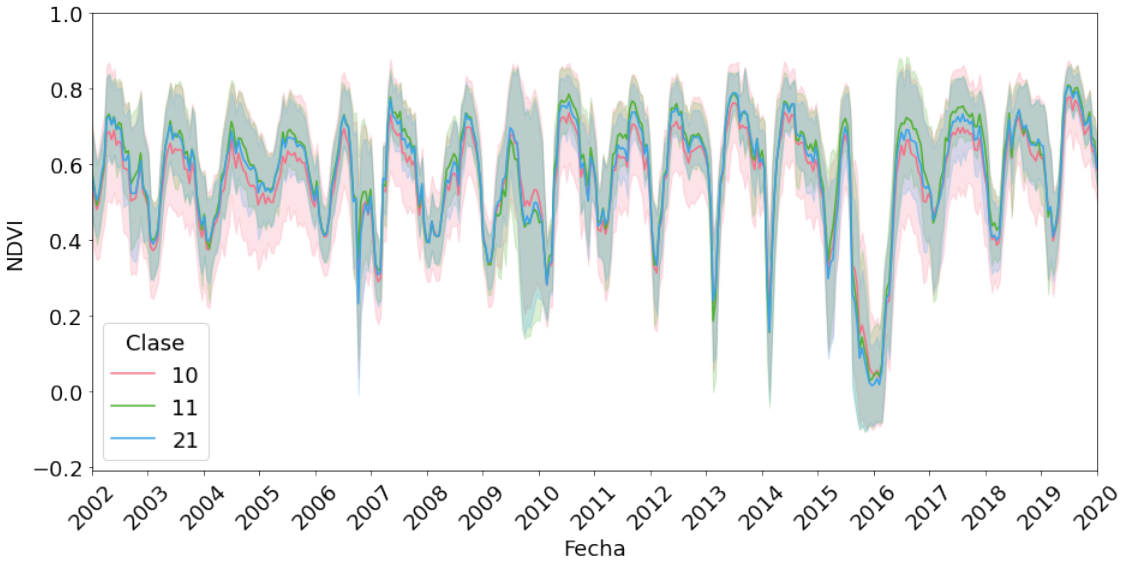


Fig. 7.14: Firmas temporales de las clases 10, 11 y 21.

Lo mismo pasa para las clases 8 y 14 (que quedan juntas a través de $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa2)}$ pero solas en $\texttt{mejor\_asignación}_{GT,f_{GT}(mapa1)}$), similares en sus composiciones al grupo anterior y a su patrón de recuperación luego de las inundaciones. Presentan, sin embargo, un NDVI promedio más bajo (ronda el 0,4) debido a su mayor presencia de agua. Están compuestas por el área activa del Paraná y sus cursos tributarios, además de bordes de ríos y lagunas.

En resumen: en una gran parte de las clases ambos mapas "filtran" de la misma manera (es decir, las clases que quedan solas en un mapa también lo están en el otro y los grupos de clases). Las diferencias se dan en los casos de las clases 6 y 17 (ambas clases que contienen zonas altas como media lomas y bosques, que van a parar a grupos distintos pero ambos contienen firmas muy parecidas), 1 y 16 (el mapa 1 las separa pero el 2 las junta a pesar

de que son casi idénticas), 8 y 14 (sus firmas son bastante distintas pero ocurre lo mismo que con el par de clases anterior).

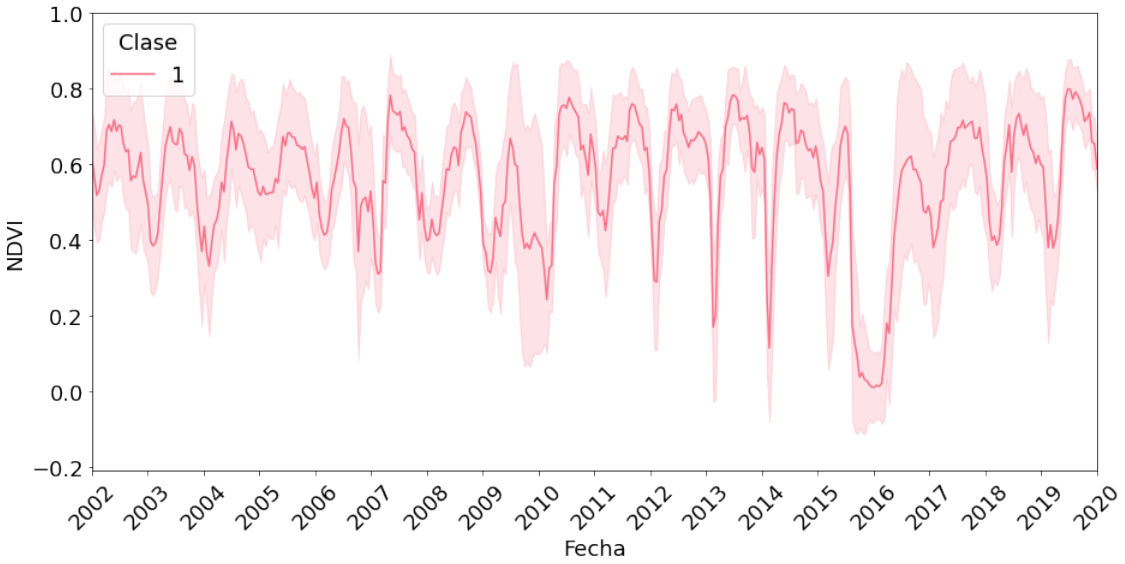


Fig. 7.15: Firmas temporales de la clase 1.

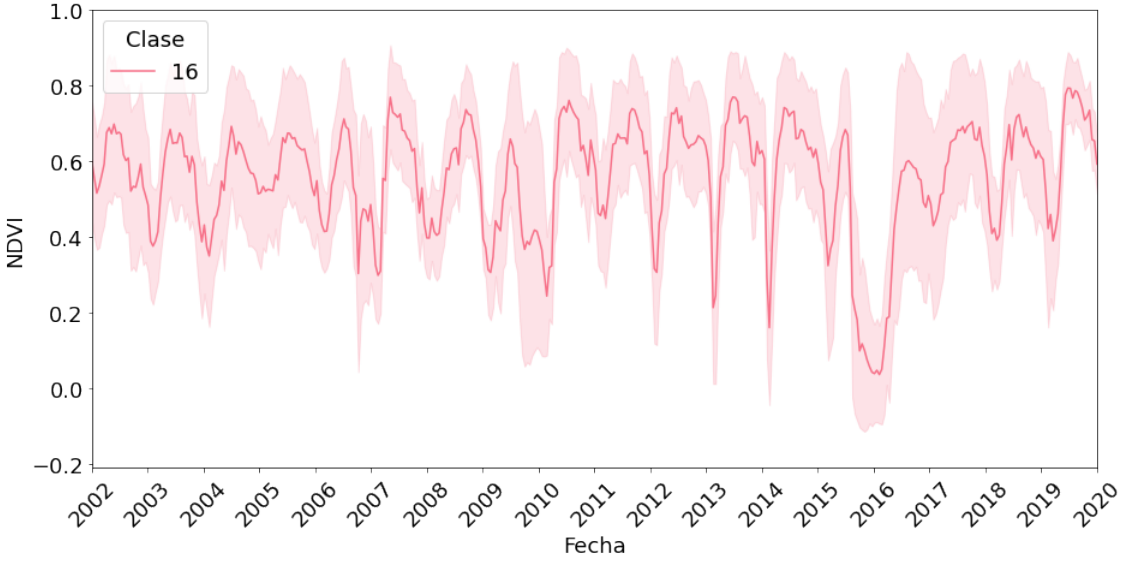


Fig. 7.16: Firmas temporales de la clase 16.

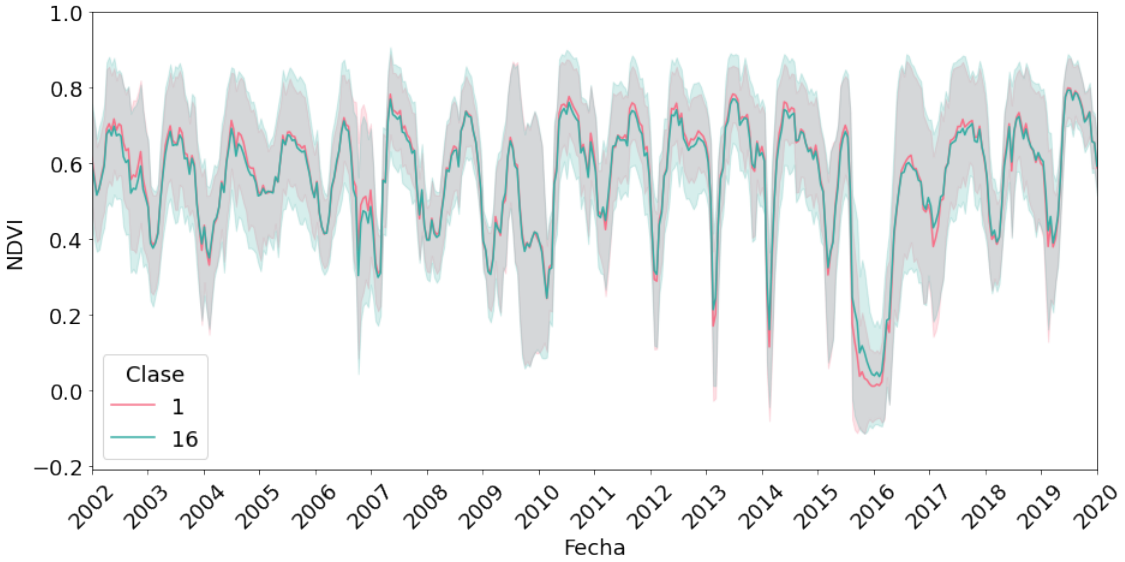


Fig. 7.17: Firmas temporales de las clases 1 y 16.

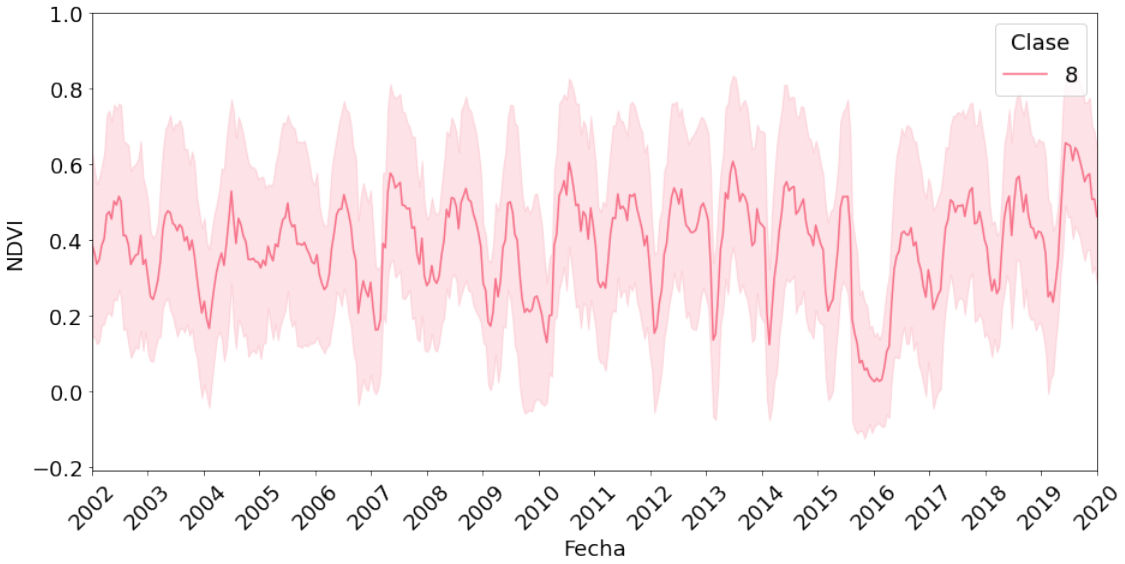


Fig. 7.18: Firmas temporales de la clase 8.

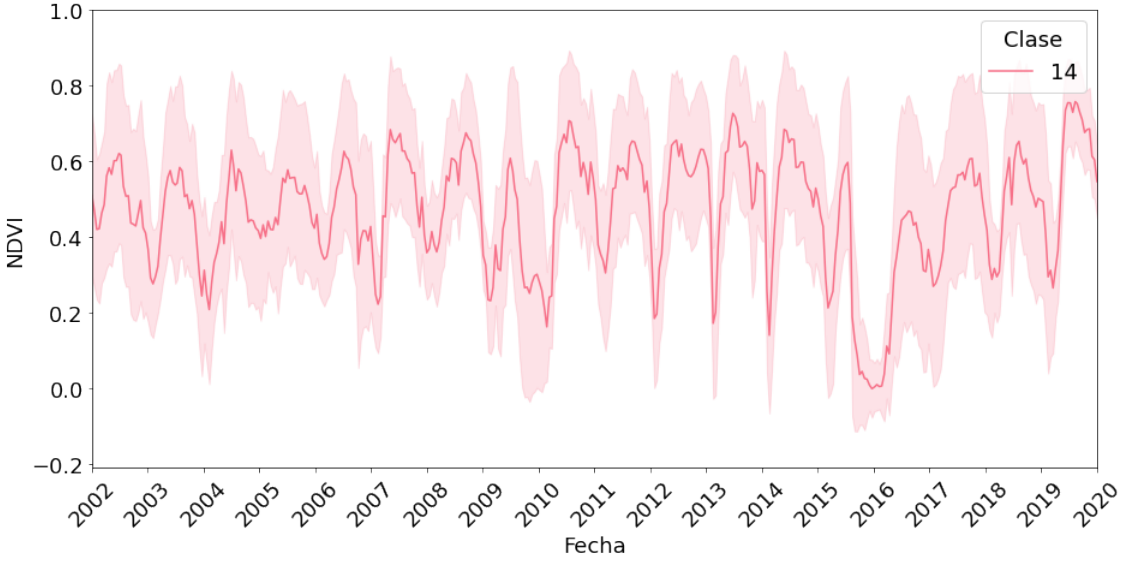


Fig. 7.19: Firmas temporales de la clase 8.

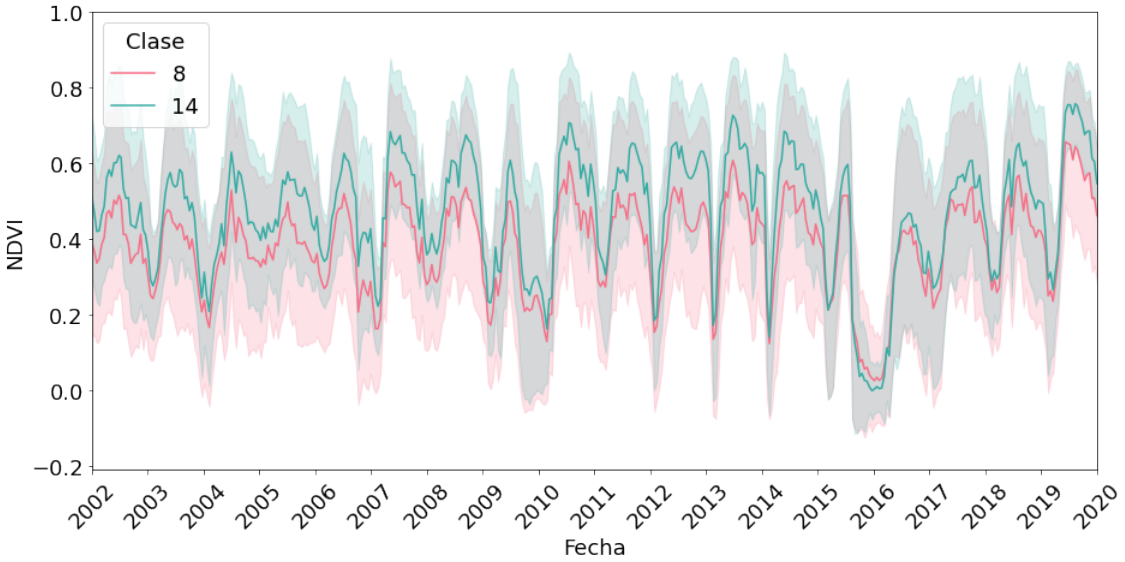


Fig. 7.20: Firmas temporales de las clases 8 y 14.

8. AGGLOMERATIVE

8.1. *Agglomerative clustering*

En [19] y [20] como parte del enfoque integrador que mencionábamos en el capítulo 5 se realiza un posprocesamiento del *clustering* obtenido a partir de *K-means* (no se realiza un trabajo como el del capítulo 7). Por eso decidimos, como en los capítulos anteriores, partir de los mapas obtenidos en el capítulo 6 y replicar (mejorando) los métodos utilizados en los *papers* consultados.

Las autoras toman cada *cluster* como si fuese un punto bidimensional, asignándole su NDVI promedio y su desvío a lo largo del tiempo. A partir de eso corren un algoritmo de *agglomerative clustering* que consiste en:

Algoritmo 1: Algoritmo de *agglomerative clustering* original

```
1 def agglomerative(clustering):
2 |   Comenzar considerando cada punto como un cluster
3 |   while #clusters (clustering) > 1 do
4 |   |   Elegir los dos clusters más cercanos (para alguna métrica) y unirlos.
5 |   end
```

Nos parece, no obstante, que esto tiene una serie de desventajas. Por ejemplo, no trabaja con las distancias originales entre los puntos sino que los "simplifica" y trabaja con las distancias entre los promedios y desvíos entre clases. Una vez que uno "mergea" dos clases habría que recomputar las medias y desvíos, y no nos queda claro por los trabajos mencionados anteriormente qué es lo que representan las clases resultantes una vez que son "fusionadas".

Además de eso, ninguna de las rutinas que encontramos en bibliotecas conocidas de Python corren *agglomerative clustering* tomando como dato de entrada un *clustering* para iterar el ciclo *while* del algoritmo anterior.

Por eso fue que decidimos programar nuestra propia rutina de *agglomerative clustering*, para hacer un posprocesamiento de los *clusterings* 1 y 2 obtenidos en el capítulo 6, que consiste justamente en eso y se puede ver en el algoritmo 2.

Para la línea 6 utilizamos el criterio de Ward: la minimización de la varianza *intra-cluster*. A diferencia de otros métodos (como la distancia máxima o promedio entre los puntos, que fueron probados pero no arrojaron buenos resultados), lo que nos dice el criterio de Ward es que el costo $\Delta(A, B)$ de *mergear* dos *clusters* A y B es cuánto aumentará la varianza del *cluster* resultante $A \cup B$ si combináramos A con B:

$$\Delta(A,B) = \sum_{i \in A \cup B} \|\vec{x_i} - \vec{m}_{A\cup B}\|^2 - \sum_{i \in A} \|\vec{x_i} - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x_i} - \vec{m}_B\|^2 = \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2 \tag{8.1}$$

donde $\vec{m}_j$ es el centro del *cluster* j, n_j es la cantidad de puntos en él y $\|\cdot\|$ es la distancia euclídea.

La biblioteca de Python *Scipy* calcula $\Delta(C_i, C_j)$ en nuestro algoritmo. Así, va *mergeando* los *clusters* más cercanos hasta que todos quedan unidos en una misma gran clase.

Algoritmo 2: Agglomerative clustering

```
1  def agglomerative(clustering_original):
2      historial ← ∅
3      clustering ← clustering_original
4      clusters_remanentes ← #clusters (clustering)
5      while clusters_remanentes >1 do
6          clusters_a_unir, distancia ← clusters_más_cercanos (clustering)
7          merge (clustering, clusters_a_unir)
8          registrar (historial, clusters_a_unir, distancia)
9          clusters_remanentes ← #clusters (clustering)
10     end
11     return historial
```

Obviamente, tenemos tantas iteraciones como *clusters*, ya que en cada una tenemos un *cluster* menos.

Para ser más específicos, el resultado directo del algoritmo de *agglomerative clustering* es una matriz $A \in \mathbb{R}^{(n-1)\times 4}$ donde n es la cantidad de clases en el *clustering* original, es decir que cada fila j representa una iteración. Para cada j, en A_{j1} y A_{j2} tenemos las clases que se van a *fusionar*, A_{j4} es la distancia entre ellas y A_{j3} es la cantidad de píxels en el nuevo *cluster*.

El motivo por el cual decidimos que el *output* de este algoritmo fuera una matriz de estas características tiene que ver con que la biblioteca *scipy* requiere este formato para elaborar un dendrograma. Esto último consiste en un gráfico que nos muestra los sucesivos pasos del algoritmo y cómo las clases se van fusionando. La longitud de los "brazos" que unen dos clases representa la distancia entre ellas. En su documentación *scipy* además establece, al graficar el dendrograma, que en cada iteración j las dos clases que se unan lo harán para formar la clase $n + j$. Por eso es que veremos números altos en los nombres de las clases. Podemos ver ambos dendrogramas en las figuras 8.1 y 8.2.

El desafío tiene que ver con elegir con qué iteración nos quedamos. Y para esto es preciso poder comparar de alguna forma el *mapa GT* con nuestras diferentes iteraciones de *agglomerative clustering*.

8.1.1. Elección de una iteración de *agglomerative clustering*

Existen distintas medidas para comparar *clusterings*. El más simple de ellos se llama índice de Rand (Rand Index en inglés), y se define de la siguiente manera: dadas dos particiones X e Y de un conjunto S, si decimos que

- a es el número de elementos de S que pertenecen al mismo subconjunto tanto en X como en Y (para nuestro caso, son píxels que pertenecen a la misma clase en los dos mapas a comparar),
- b es el número de elementos de S que pertenecen a subconjuntos distintos tanto en X como en Y (píxels que en tanto en un mapa como en el otro pertenecen a clases distintas),
- c es el número de elementos de S que pertenecen a subconjuntos distintos en X pero

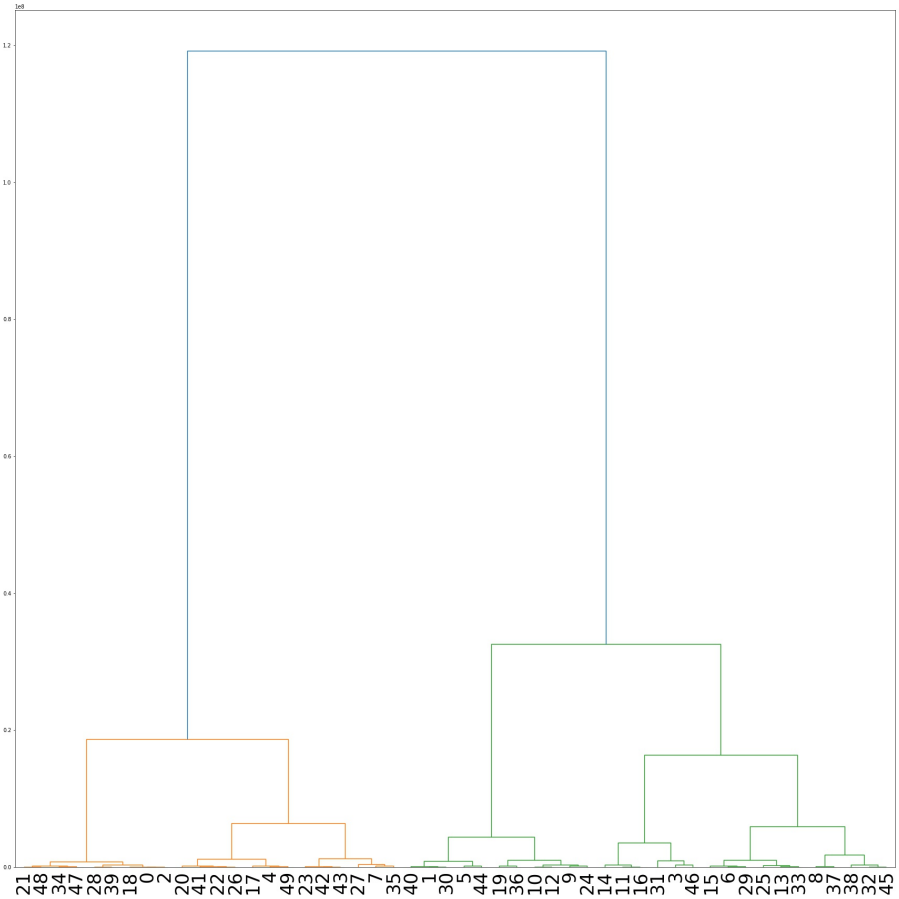


Fig. 8.1: Dendrograma del *clustering* 1.

en Y pertenecen al mismo subconjunto (píxels que en el mapa GT pertenecen a la misma clase pero a clases distintas en el *clustering*) y

- d es el número de elementos de S que pertenecen al mismo subconjunto en X pero en Y pertenecen a subconjuntos diferentes (el caso inverso del anterior);

entonces el índice de Rand según [12] es

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{\#S}{2}} \tag{8.2}$$

donde el denominador de la derecha representa la cantidad de formas posibles de tomar elementos de a dos en el conjunto S. Por su definición, el RI se encuentra entre 0 y 1.

El problema con este índice es que, dados dos *clusterings* aleatorios, el mismo no es 0 como uno esperaría. Es más: es cercano a 1 incluso si difieren significativamente. Por eso

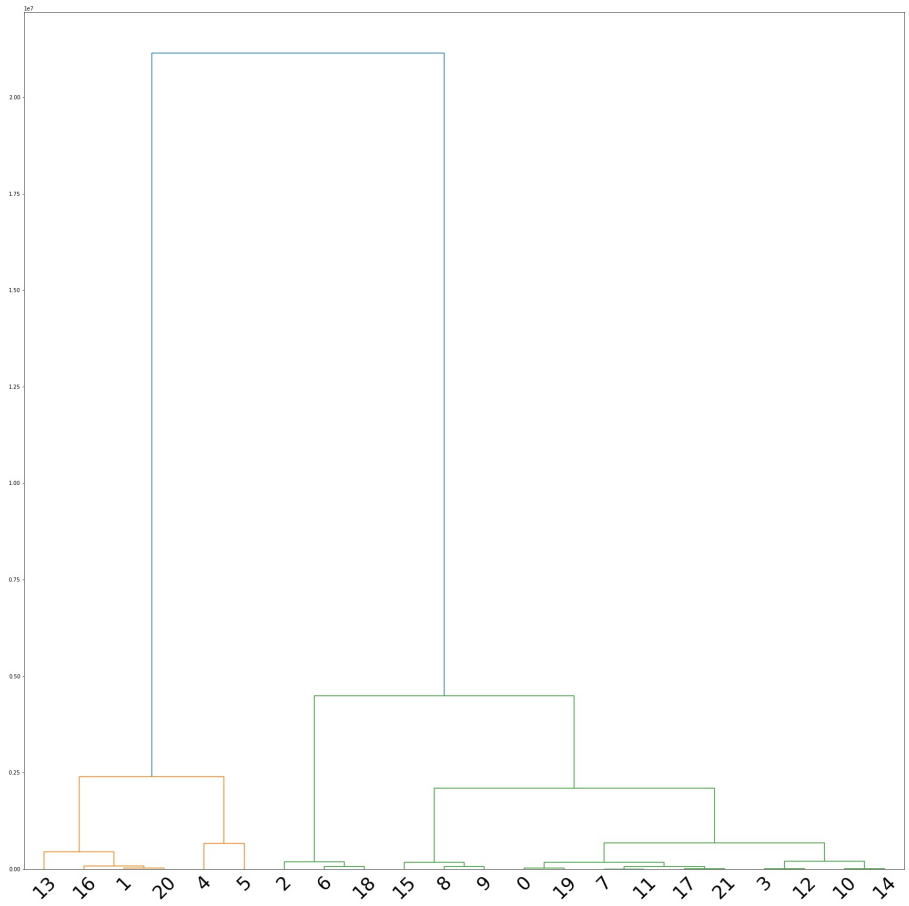


Fig. 8.2: Dendrograma del *clustering* 2.

es que existe el índice de Rand ajustado (Adjusted Rand Index en inglés), que como su nombre lo indica tiene una "corrección por azar". El mismo establece una base usando el índice de Rand esperado en un modelo aleatorio. Es decir, el índice de Rand ajustado lo podemos calcular así:

$$ARI = \frac{RI - \mathbb{E}[RI]}{\text{máx}\, RI - \mathbb{E}[RI]} \tag{8.3}$$

La biblioteca *scikit learn* lo calcula de la siguiente manera:

$$2 \cdot \frac{a \cdot b - c \cdot d}{(a+c) \cdot (c+b) + (a+d) \cdot (d+b)} \tag{8.4}$$

De esta forma, el ARI toma valores entre -1 y 1, es 0 para *clusterings* aleatorios y 1 para *clusterings* que coinciden perfectamente. Es preciso notar, además, que las clases de

ambos *clusterings* para este índice son intercambiables (es decir, no importan los nombres de las clases) y además es un índice simétrico (es decir $ARI(X, Y) = ARI(Y, X)$).

Lo que decidimos hacer para encontrar nuestra iteración del *agglomerative clustering* fue buscar aquella iteración que maximizara el ARI al compararla con el *mapa GT*. El resultado de esto es la iteración 34 del algoritmo para aquel mapa cuya dimensionalidad redujimos con PCA, y la 10 para el mapa en el cual hicimos lo propio con las medias y desvíos mensuales.

Algo que notamos en los dos *clusterings* que generamos utilizando GMM es que tienen muchas clases minoritarias, es decir, compuestas por muy pocos píxels. Esto nos dificulta el análisis debido a que son clases para las cuales es muy complicado pensar qué es lo que representan. Supusimos que esto disminuiría a medida que el algoritmo de *agglomerative clustering* va avanzando, debido a que este va fusionando clases entre sí. Esto es lo que efectivamente ocurre, y de hecho ninguna de nuestras iteraciones elegidas tiene clases que ocupen menos del 2 % de los píxels.

Con todo esto ya podemos comenzar a analizar. Para ser más concisos con los nombres de los mapas, en este capítulo final llamaremos "*mapa a*" al mapa que comenzó con una reducción de dimensionalidad mediante PCA, siguió con un *clustering* mediante GMM y terminó con la iteración 34 de *agglomerative*; y llamaremos "*mapa b*" a aquel que comenzó con una reducción de dimensionalidad empleando las medias y desvíos mensuales, siguió con un *clustering* mediante GMM y terminó con la iteración 10 de *agglomerative*. Podemos ver ambos en las figuras 8.3 y 8.4.

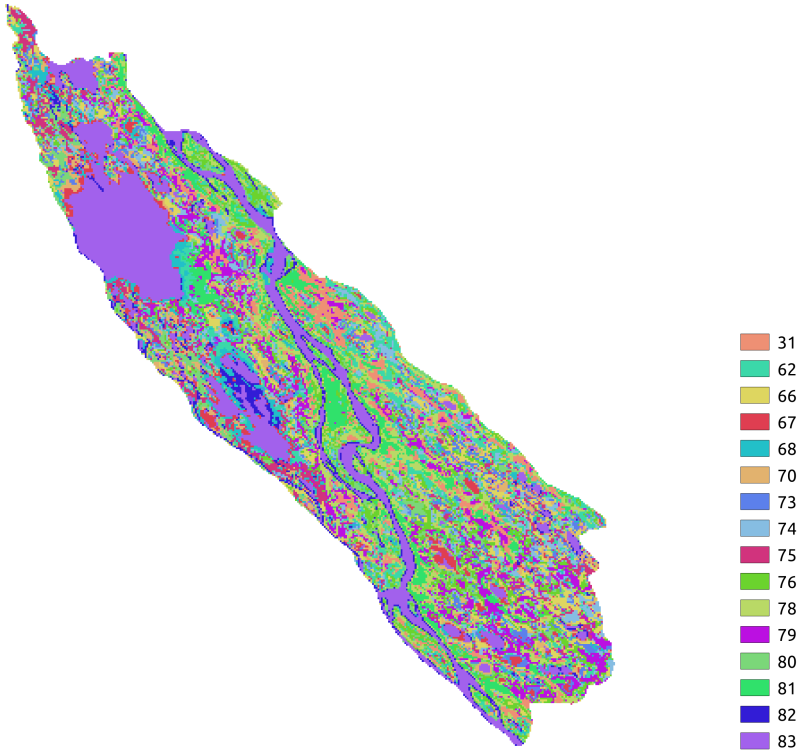


Fig. 8.3: *mapa a.*

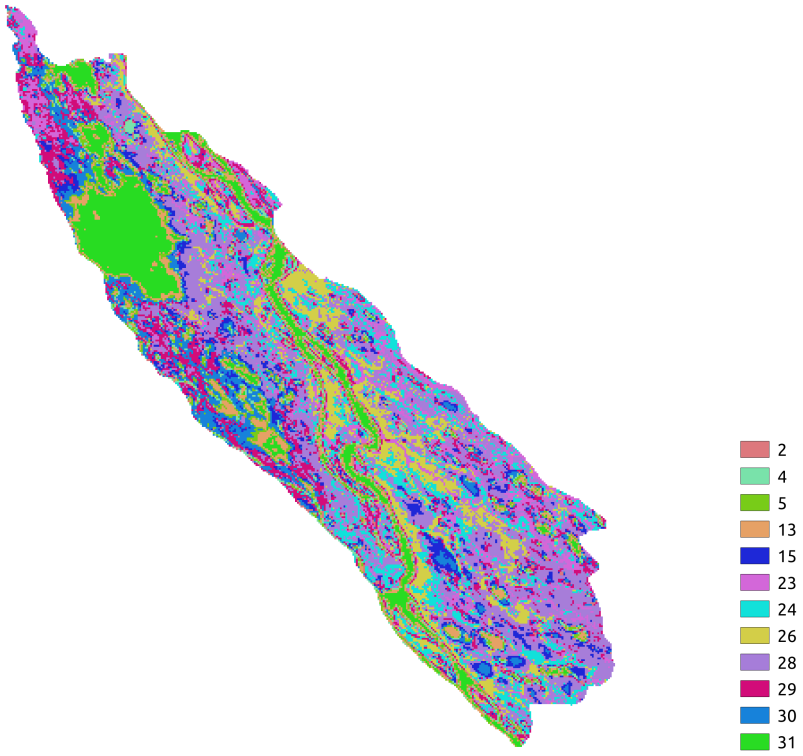


Fig. 8.4: *mapa b.*

8.2. Análisis de nuestros mapas

Cuando elaboramos nuestro *mapa GT* nos preguntamos cuál era la composición de cada una de estas nuevas clases: ¿Cuál es su relación con las unidades de humedales? ¿Son clases puras en relación a estas últimas? ¿Son mixtas? ¿En qué grado? Para respondernos esta pregunta elaboramos la matriz que se ve en la figura 8.5, en la que cada fila representa el promedio de cada clase generada. Por ejemplo, la cuarta fila nos dice que la clase 3 (comenzamos contando desde 0) consiste completamente de cursos de agua, por eso vemos que es $(100; 0; 0; 0; 0)$.

Otra pregunta que nos hicimos es dónde estaba localizado cada tipo de unidad de humedal. Por ejemplo: ¿Las lagunas están solo en una clase? ¿O en varias? Esto lo respondimos con la matriz de la figura 8.6.

Este tipo de diagramas tiene una relación directa con las matrices de confusión, muy utilizadas para analizar resultados de algoritmos de clasificación. La diferencia en este caso es que, en estas últimas, las entradas representan cantidades absolutas; mientras que nosotros elegimos valores porcentuales.

Para todos los "mapas de calor" (o *heatmaps*) utilizamos las bibliotecas libres Pandas y Seaborn de Python. Para leer cada uno de ellos nos paramos en una fila y nos podemos preguntar cómo está compuesta la clase correspondiente a la misma. Además de eso, en la leyenda de cada fila y columna podemos ver qué porcentaje del mapa correspondiente

Fig. 8.5: Grado de pertenencia promedio de cada clase del *mapa GT* a cada tipo de unidad de humedal.

ocupa la clase.

8.2.1. Análisis por tipo de humedal

Lo que haremos ahora es analizar cada tipo de humedal, y su pertenencia a las diferentes clases del *mapa GT*. A su vez, discutiremos la inclusión de las mismas en nuestros mapas a y b.

Para esto, hicimos *heatmaps* que comparan los mapas a y b con el *mapa GT* en ambos sentidos. Excluimos de los mismos las clases del *mapa GT* cuyos píxels ocupan menos del 2 % del total, dado que complejizan su lectura y pensamos que no aportan demasiado al análisis. Por ese motivo, es posible que las filas no sumen exactamente 100.

Cursos de agua

Empecemos con los cursos de agua. Algo que podemos observar es que los mismos se encuentran repartidos entre dos clases del *mapa GT*. La mayoritaria (clase 3) está casi enteramente contenida en la clase 83 del *mapa a*. En el caso del *mapa b* está repartida entre tres clases, una que representa los cursos en sí (31) y otras dos que contienen una variedad de clases (29 y 15). La segunda clase del *mapa GT* en orden de importancia para los cuerpos de agua (la 22) representa los bordes de los mismos, y la vemos bastante más repartida entre diferentes clases en los mapas realizados por nosotros. Algunas de ellas las vemos justamente en los bordes de los cursos de agua y otras no.

Lagunas

Sigamos con las lagunas. Pasa algo muy parecido en este caso que en el anterior: las mismas están repartidas (en gran parte) en tres clases del *mapa GT*.

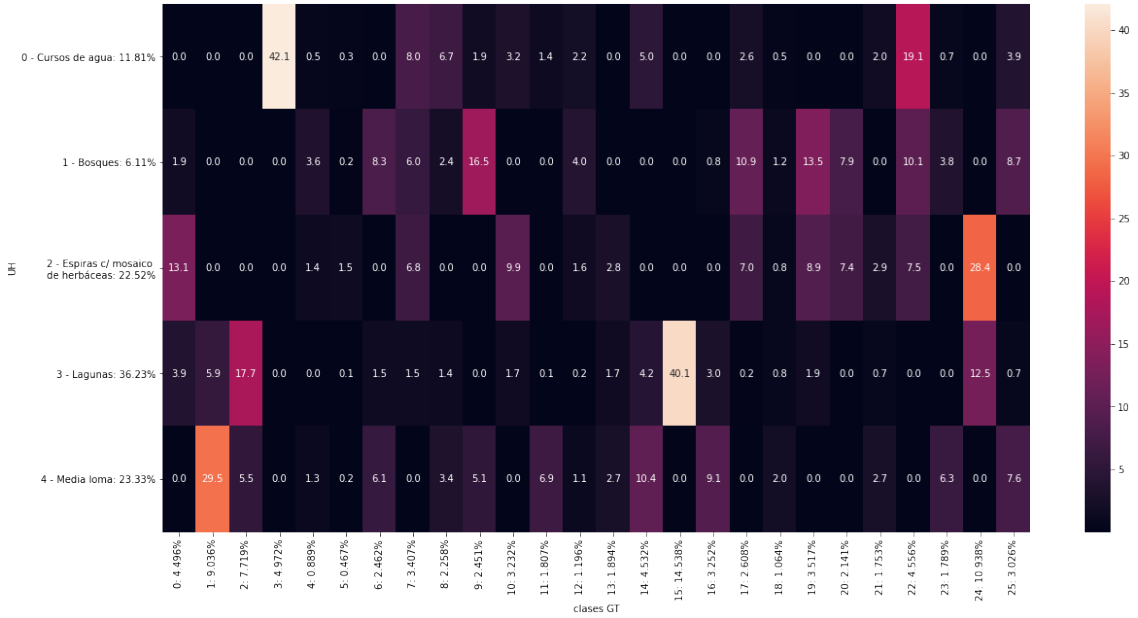


Fig. 8.6: Grado de pertenencia promedio de cada tipo de unidad de humedal a cada clase del *mapa GT*.

Acá podemos notar una diferencia entre el *mapa GT* y los mapas elaborados por nosotros: estos últimos no distinguen los ecosistemas lóticos (ríos) de los lénticos (lagunas). Si miramos el *mapa b*, vemos que las clases 13 y 31 están compuestas por los bordes y el interior (respectivamente) tanto de ríos como de lagunas. La primera clase en orden de importancia del *mapa GT* para las lagunas (15) está mayoritariamente contenida en la clase 83 (como en el caso anterior), que consiste de ríos y lagunas. La segunda clase del *mapa GT* en orden de importancia para las lagunas (2) está más repartida a lo largo de los *clusters* de nuestros mapas, que de todos modos están contenidas en su mayoría en las clases 2 y 15 (las clases del *mapa GT* que representan las lagunas).

Aún así, es interesante mencionar que las dos clases en las que están las lagunas están mayoritariamente compuestas en su mayoría, justamente, por lagunas. Esto nos habla bien de nuestro *mapa GT*, dado que significa que en este caso no "mezcla" clases.

Bosques

Como vemos en el *heatmap* 8.6 los bosques se encuentran bastante repartidos entre las clases de nuestro *mapa GT*. En su mayoría están en las clases 9, 19, 17 y 22 de este último. Si miramos el otro *heatmap*, el 8.5, vemos que en todas ellas los bosques tienen un porcentaje significativo pero no necesariamente es la clase preponderante.

Los píxels de estas clases los vemos distribuidos entre varias clases de ambos mapas. Esto puede tener que ver con el hecho de que los bosques ocupan un porcentaje muy pequeño del mapa de unidades de humedal (6 %), por lo que los píxels del *mapa GT* que contienen bosques están mezclados con otros tipos de humedales.

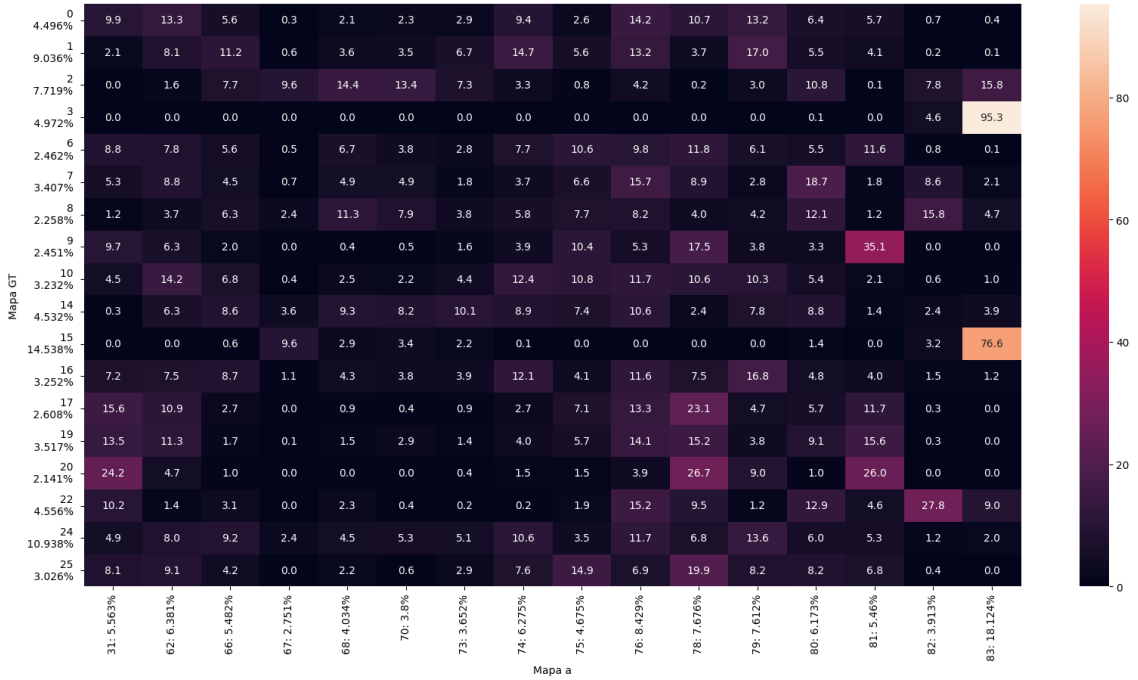


Fig. 8.7: Grado de pertenencia de cada clase del *mapa GT* a cada clase del *mapa a*.

Espiras con mosaico de herbáceas

La mayor parte de las espiras con mosaico de herbáceas las encontramos en las clases 24 y 0 de nuestro *mapa GT*. Ambas clases se encuentran mayoritariamente en los *clusters* 93 y 86 del *mapa a*, con una diferencia grande de porcentaje entre ellas.

Ambas las vemos preponderantemente en las clases 24 y 28 del *mapa b*, y muy distribuidas entre varias clases en el *mapa a*.

Pensando al revés, podemos ver que las clases mencionadas recién tienen un porcentaje importante de su composición en las clases 24 y 0 del *mapa GT*.

Por último, al igual que vemos en varios otros casos, las clases 24 y 0 del *mapa GT* están compuestas en gran medida por espiras con mosaico de herbáceas, lo cual una vez más nos dice que este mapa no está mezclando clases.

Media loma

Esta clase de humedal la podemos ver predominantemente en las clases 1 y 14 del *mapa GT*. Las mismas (como en los casos anteriores) están compuestas en su mayoría por media loma, aunque también (en menor medida) por espiras.

Las clases mencionadas del *mapa GT* las encontramos distribuidas entre muchas clases del *mapa a* pero se destacan las 79, 74 y 75; que son clases que a su vez tienen a las primeras como mayoritarias en su composición, además de la clase 24.

Con respecto al *mapa b*, las clases 1 y 14 del *mapa GT* se encuentran repartidas entre las clases 28, 30 y 29 (en orden de importancia). Esto y la presencia de la clase 24 para el *mapa a* nos habla de que este mapa mezcla espiras y medias lomas.

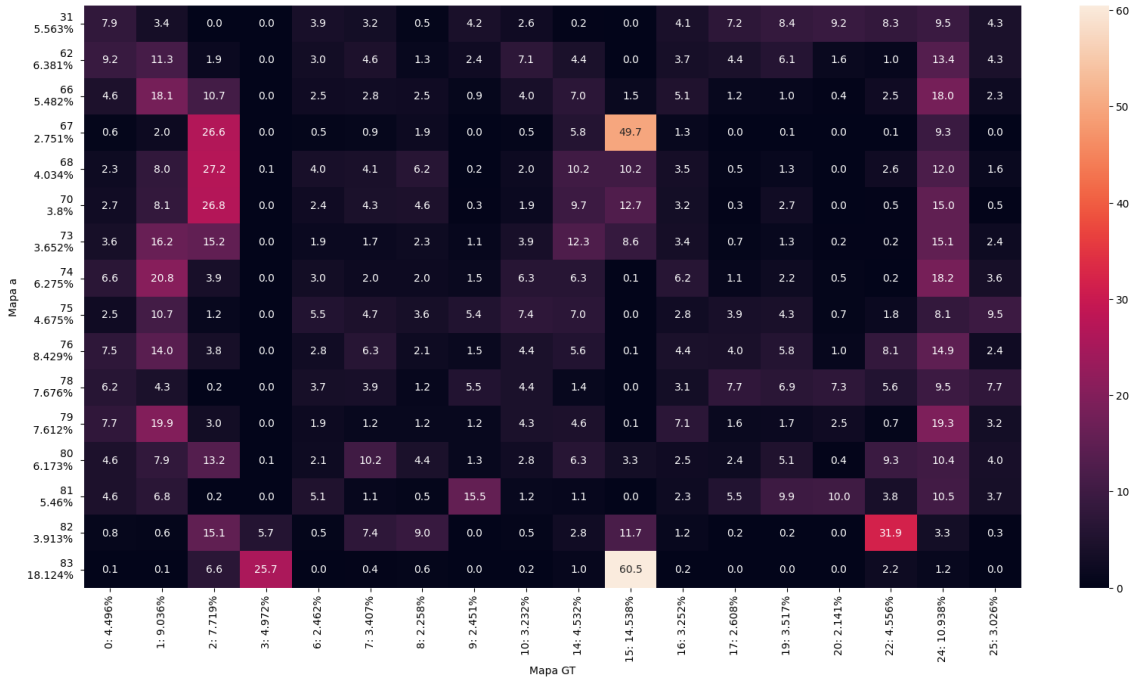


Fig. 8.8: Grado de pertenencia de cada clase del *mapa a* a cada clase del *mapa GT*.

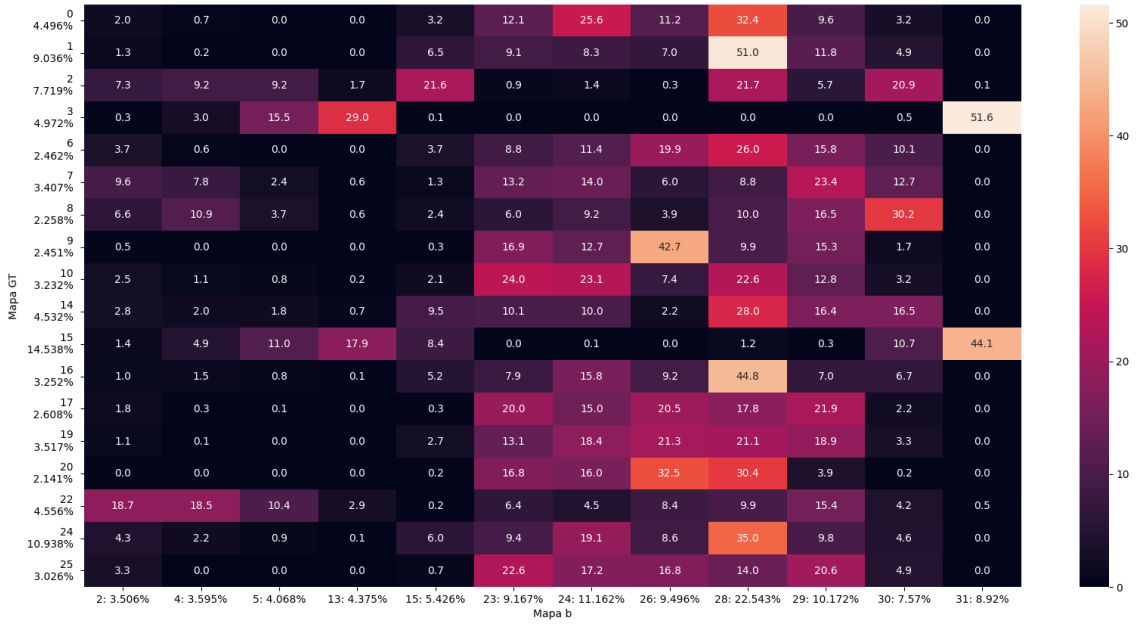


Fig. 8.9: Grado de pertenencia de cada clase del *mapa GT* a cada clase del *mapa b*.

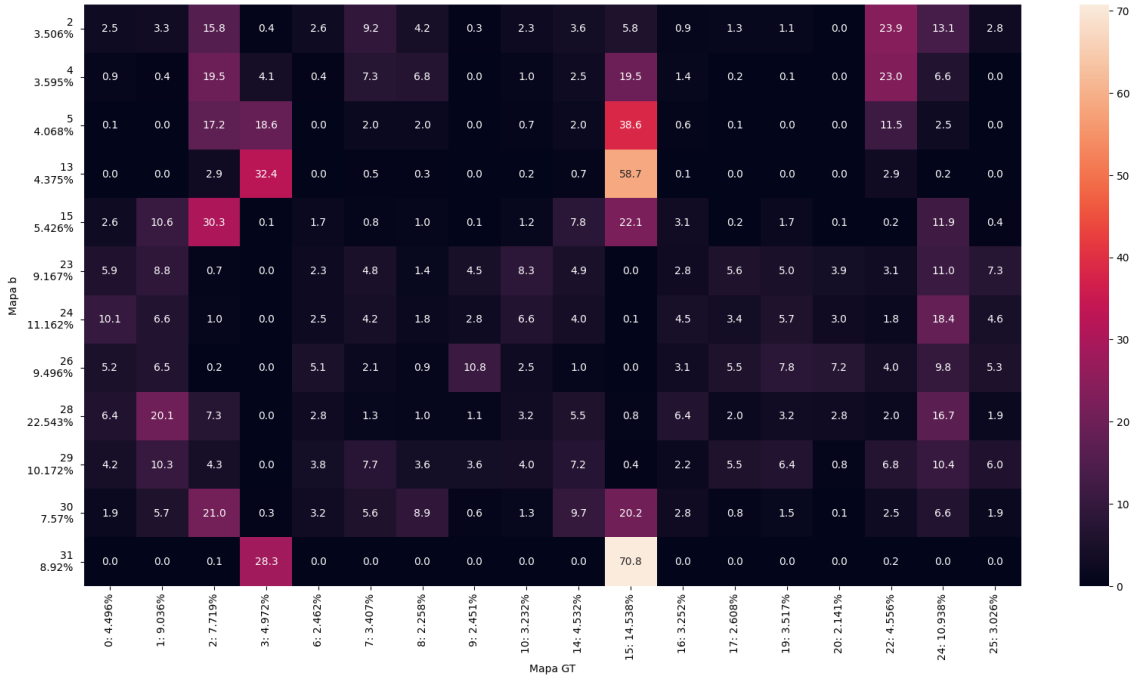


Fig. 8.10: Grado de pertenencia de cada clase del *mapa b* a cada clase del *mapa GT*.

9. CONCLUSIONES

En este trabajo exploramos dos formas de reducir la dimensionalidad en series de tiempo de imágenes satelitales. Nos basamos en trabajos previos realizados en la región del Paraná medio para replicar la tarea de clasificar unidades de paisaje utilizando un índice verde pero en una región más acotada, aunque con muchas similitudes.

Decidimos, además, emplear métodos de reducción de la dimensión y de *clustering* que creemos más pertinentes para el problema en cuestión. Los dos métodos utilizados arrojan resultados similares en términos de la preservación de la información contenida en la serie de tiempo de las imágenes: ninguno de los dos *clusterings* distingue entre ecosistemas lóticos y lénticos, ambos distinguen los bosques y zonas altas (las últimas en inundarse), aunque estas unidades de humedal se encuentren mezcladas con otras clases debido a que los bosques son minoritarios y no llegan a ocupar un píxel entero. Además de eso, exhiben gran riqueza en su representación de las lagunas, debido a que ambos distinguen los bordes y el interior de las mismas. Lo que los diferencia es que los primeros no están siempre cubiertos de agua (dependen de las inundaciones y secas), exhibiendo una vegetación bastante diferente. Por último, las espiras y las medias lomas en general se mezclan, dependiendo por lo general de la topografía, variable que queda pendiente integrar en el análisis para futuros trabajos.

Lo que es importante mencionar es que el método de reducción de la dimensionalidad por análisis de componentes principales (PCA) nos permitió llegar a resultados similares con una dimensión mucho menor y asegurándonos de que las componentes que conservamos fueran, efectivamente, las que más varianza explicaran en nuestros datos.

Nos encontramos en el camino con desafíos de diferencias de escala que nos obligaron a crear nuestro propio mapa de verdad del terreno, haciendo que nuestro análisis sea más indirecto. Este, tal como se explica en [15] es un problema constante en el estudio de los ecosistemas mediante teledetección en general, y de los humedales en particular.

Programamos además un algoritmo de *clustering* jerárquico y lo utilizamos como método de posprocesamiento de otro algoritmo de *clustering* basado en un modelo de mezcla de gaussianas. En lo que respecta a esto, existen otros métodos para definir un criterio de parada a la hora de decidir la cantidad final de *clusters*. La utilización del índice de Rand ajustado (ARI) es una entre muchas que podrían ser exploradas. En el capítulo correspondiente utilizamos varios mapas de calor a la vez, dado que estábamos comparando tres mapas al mismo tiempo (mapa GT, mapa de unidades de humedal, mapa 1 o 2 según el caso). Resta explorar la existencia de otras herramientas que nos permitan realizar análisis multidimensionales.

Una forma de abordar el análisis comparativo de dos clasificaciones consiste en "normalizar" su cantidad de clases con un mapa *ground-truth*. A través de las biyecciones que creamos entre las clases de este último mapa, pero que están "filtradas" por la lente de nuestros *clusterings*, uno puede "hacerles las mismas preguntas" a ambas clasificaciones.

Observando las firmas temporales de los promedios y desviaciones estándar de cada clase podemos ver que nuestros *clusterings* mezclan clases de nuestro mapa *ground-truth*. Pero no lo hacen de cualquier manera sino que juntan clases que tienen un comportamiento similar a lo largo del tiempo (su fenología) aunque sean estructuralmente distintas. En este sentido, es posible que las unidades geomórficas (ver [20]) capturen mejor la dimensión temporal y la evolución de las dinámicas de escurrimiento que las unidades de humedal.

Elegimos esta forma de comparación porque el gran tamaño de las primeras (con respecto a la extensión del sitio Ramsar) hacía que toda nuestra área de estudio quedara cubierta, en mayor parte, por una unidad geomórfica, dificultando el análisis. Queda pendiente la realización de un análisis cuantitativo de comparación entre los distintos grupos de firmas temporales que nuestros algoritmos de *clustering* juntan o separan.

Los productos de MODIS, con su resolución media y su algoritmo de composición de imágenes, nos dan *datasets* ideales para introducirnos en la teledetección. Sin embargo, es posible que se puedan tener resultados más concluyentes con imágenes de resolución espacial más alta, como Landsat. En ese caso no sería necesario construir nuestro mapa con la verdad del terreno porque las resoluciones coincidirían, pero sumaríamos desafíos de infraestructura, dado que la cantidad de píxels sería mucho mayor (recordemos que también tenemos la dimensión temporal), y deberíamos correr nuestros modelos en la nube. Otro desafío sería el proceso de preprocesamiento y selección de imágenes, dado que las mismas tienen nubes y (si seleccionáramos aquellas que están "limpias") no estarían obtenidas en intervalos regulares de tiempo. Otra opción, si hiciéramos esto, sería elegir una imagen de la serie que consideremos pertinente y usar allí los algoritmos de *clustering*. Usando este método, que encontramos en varios trabajos, perderíamos la temporalidad.

Por último, queda pendiente realizar un análisis que incluya de manera cuantitativa la evolución a lo largo del tiempo de los niveles hidrométricos del Paraná, dado que los mismos determinan en gran parte la fenología en nuestra área de estudio.

Apéndice

A. SOFTWARE LIBRE, DATOS ABIERTOS Y CÓDIGO ABIERTO

Nos parece que es una buena oportunidad para resaltar algunos beneficios de la investigación científica a través del uso del Software Libre y Open Source (FOSS por sus siglas en inglés) [25]. Un programa es software libre si los usuarios tienen las cuatro libertades esenciales:

- La libertad de ejecutar el programa como se desee, con cualquier propósito.
- La libertad de estudiar cómo funciona el programa, y cambiarlo para que haga lo que se desee. El acceso al código fuente es una condición necesaria para ello.
- La libertad de redistribuir copias para ayudar a otros.
- La libertad de distribuir copias de sus versiones modificadas a terceros. Esto le permite ofrecer a toda la comunidad la oportunidad de beneficiarse de las modificaciones. El acceso al código fuente es una condición necesaria para ello.

Todo el código que escribimos para esta tesis (bajo licencia GNU-General Public Licence) está en un repositorio en *Github*, al que cualquiera puede acceder[1] para ver su contenido, "clonarlo" (bajarlo en su computadora), hacer un *fork* (copiarlo y modificarlo sin afectar el original), ejecutarlo y difundirlo. Nos ocupamos de que nuestro código sea comprensible y bien organizado, y de que los resultados sean reproducibles.

Utilizamos FOSS a lo largo de toda la tesis[2]. La misma está organizada alrededor de un paquete de *Python* (desarrollado por nosotros) al que llamamos `clasificacion_humedales`, que consiste en varios *jupyter notebooks* ejecutables que se valen de distintos métodos y clases. Para evitar problemas de dependencias e incompatibilidades de versiones utilizamos *Docker*, que permite crear un *container* (una especie de máquina virtual simplificada) donde se puede correr el proyecto previa instalación automática de todo lo necesario. Esto hace que su ejecución (y posterior desinstalación completa si se desea) sea sencilla e idéntica tanto para nosotros como para cualquier otro usuario, además de independiente de la computadora en la que esté trabajando. Las bibliotecas utilizadas son todas de FOSS (se pueden ver en `requirements.txt`), así como también el programa que empleamos para visualizar los mapas generados (*Qgis*).

Las únicas excepciones son las herramientas *Overleaf* y *Google Docs*, que utilizamos en la web (de ahí su conveniencia) pero solo para la escritura de esta tesis y pueden ser reemplazadas por cualquier usuario por un compilador local de LaTeXy procesadores de texto, respectivamente. Además de eso, existen diversas formas de acceder a los productos de MODIS. Nosotros usamos la plataforma *AppEEARS* [1] de la cual no conseguimos la licencia ni el código, por lo cual no creemos que sea FOSS. De todos modos, en la página de MODIS se presentan otras formas de descargar las imágenes y en el repositorio incluimos un archivo estandarizado del *request* para poder replicarlo a través de herramientas libres.

[1] Ver https://github.com/maianumerosky/tesis

[2] Existen diferencias entre el *free software* y *open source*, además de una gran variedad de licencias entre los programas que utilizamos. No todos cumplen estrictamente las cuatro libertades del *software* libre pero todas se clasifican dentro del FOSS a los efectos de esta tesis.

Algunas ventajas de hacer público el código en los trabajos académicos se pueden encontrar en [13]. Por ejemplo, permite que cualquier persona pueda acceder al código y reproducir los resultados, lo cual no ocurre en todas las producciones científicas y nos parece importante estimular. Cada vez más revistas exigen la publicación del código producido como un requisito para publicar artículos. La independencia de *software* privativo, además de librarnos de tener que pagar licencias para usar los programas, hace que la producción científica sea más transparente. El acceso al código permite ver la implementación de los diferentes algoritmos y bibliotecas que se están utilizando, si hay cambios en la misma a lo largo de las versiones del programa y solicitar su corrección en caso de que encontremos algún error. Además de eso, nos parece muy positiva la comunidad que surge alrededor de los proyectos de *software* libre, que permite consultar dudas, responderlas, contribuir e incluso armar comunidades y conferencias.

Por último, con respecto a las imágenes MODIS que utilizamos para este trabajo, si bien la entidad que las produce, LP DAAC (Land Processes Distributed Active Archive Center, parte del Sistema de Datos de Observación Terrestre y Sistema de Información de la NASA) no publica el código de los algoritmos que procesan las imágenes crudas, los productos son accesibles de manera gratuita y ofrecen un Documento de Bases Teóricas del Algoritmo (Algorithm Theoretical Basis Document o ATBD) que encontramos en [9]. Los mismos, sin embargo, son insuficientes a los efectos de la transparencia dado que el lenguaje coloquial puede traducirse a código de muchas formas, lo cual puede introducir ambigüedades en su revisión.

BIBLIOGRAFíA

[1] AppEEARS Team. *Application for Extracting and Exploring Analysis Ready Samples*. Ver. 2.41. USGS/Earth Resources Observation y Science (EROS) Center, Sioux Falls (South Dakota), USA. URL: https://lpdaacsvc.cr.usgs.gov/appeears (visitado 16-05-2020).

[2] Michael Begon, Colin R. Townsend y John L. Harper. *Ecology: from individuals to ecosystems*. 4th ed. Malden, MA: Blackwell Pub, 2006. 738 págs. ISBN: 9781405111171.

[3] Laura Benzaquén, ed. *Inventario de los humedales de Argentina: sistemas de paisajes de humedales del corredor fluvial Paraná-Paraguay*. 1a edición. Buenos Aires: Secretaría de Ambiente y Desarrollo Sustentable de la Nación, 2013. ISBN: 978-987-29340-0-2.

[4] Laura Benzaquén y Andrea E. Izquierdo, eds. *Regiones de Humedales de La Argentina*. tex.lccn: GB628.31 .R44 2017. Buenos Aires, Argentina: G.I.E.H. ; UNSAM 3iA, Instituto de Investigación e Ingeniería Ambiental ; Wetlands International : Ministerio de Ambiente y Desarrollo Sustentable, Presidencia de la Nación, 2017.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. tex.lccn: Q327 .B52 2006. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[6] James B. Campbell y Randolph H. Wynne. *Introduction to Remote Sensing*. 5th ed. tex.lccn: G70.4 .C23 2011. New York: Guilford Press, 2011. ISBN: 978-1-60918-176-5.

[7] F. Stuart Chapin, P. A. Matson y Peter Morrison Vitousek. *Principles of terrestrial ecosystem ecology*. 2nd ed. New York: Springer, 2011. 529 págs. ISBN: 9781441995032 9781441995025 9781441995049.

[8] *Convención relativa a los humedales de importancia internacional especialmente como hábitat de aves acuáticas*. 1971. URL: https://www.ramsar.org/sites/default/files/documents/library/current_convention_s.pdf.

[9] Didan, Kamel. *MYD13Q1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V006*. Type: dataset. 2015. DOI: 10.5067/MODIS/MYD13Q1.006. URL: https://lpdaac.usgs.gov/products/myd13q1v006/ (visitado 09-11-2021).

[10] Rolando García. "Interdisciplinariedad y sistemas complejos". En: *Revista Latinoamericana de Metodología de las Ciencias Sociales* 1.1 (2011), págs. 66-101. ISSN: 1853-7863. URL: https://www.relmecs.fahce.unlp.edu.ar/article/view/v01n01a04 (visitado 18-09-2022).

[11] Beatriz Giacosa. *Plan de Manejo del Sitio RamsarDelta del Paraná*. Ed. por David Balderrama, Marta Andelman y Mateo Matarasso. Wetlands International, 2019. ISBN: 978-987-29811-8-1. URL: https://lac.wetlands.org/publicacion/plan-de-manejo-del-sitio-ramsar-delta-del-parana/.

[12] Laura Igual y Santi Seguí. *Introduction to Data Science*. New York, NY: Springer Berlin Heidelberg, 2017. ISBN: 978-3-319-50016-4.

[13] Darrel C. Ince, Leslie Hatton y John Graham-Cumming. "The case for open computer programs". En: *Nature* 482.7386 (feb. de 2012). Number: 7386 Publisher: Nature Publishing Group, págs. 485-488. ISSN: 1476-4687. DOI: 10.1038/nature10836. URL: https://www.nature.com/articles/nature10836 (visitado 05-03-2022).

[14] Ursula Jaramillo Villa, Jimena Cortés-Duque y Carlos Flórez-Ayala. *Colombia anfibia un país de humedales*. 2016. ISBN: 9789588889481 9789588889818.

[15] Patricia Kandus. "Ecosistemas de humedal. Importancia y expresión espacial en Argentina." Coloquio. Coloquio. Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 4 de sep. de 2021. URL: https://youtu.be/eRimbVDJ3wQ (visitado 01-05-2022).

[16] Patricia Kandus y Priscila Minotti. *Propuesta de un marco conceptual y lineamientos metodológicos para el Inventario Nacional de Humedales. Informe final. Documento Rector del Inventario Nacional de Humedales DI-2018-3-APN-SSPYOAD#MAD*. 2018.

[17] Patricia Kandus y Priscilla Minotti. "Conceptos y enfoques metodológicos para un inventario de humedales a escala nacional: el paisaje como organizador." En: *Revista de la Asociación Argentina de Ecología de Paisajes* 9 (dic. de 2019), págs. 84-89.

[18] Patricia Kandus y col. "Remote Sensing of Wetlands in South America: Status and Challenges". En: *International Journal of Remote Sensing* 39.4 (feb. de 2018), págs. 993-1016. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431161.2017.1395971.

[19] Zuleica Marchetti y col. "NDVI Patterns as Indicator of Morphodynamic Activity in the Middle Paraná River Floodplain". En: *Geomorphology* 253 (ene. de 2016), págs. 146-158. ISSN: 0169555X. DOI: 10.1016/j.geomorph.2015.10.003.

[20] Zuleica Marchetti y col. "Vegetation and Hydrogeomorphic Features of a Large Lowland River: NDVI Patterns Summarizing Fluvial Dynamics and Supporting Interpretations of Ecological Patterns". En: *Earth Surface Processes and Landforms* 45.3 (mar. de 2020), págs. 694-706. ISSN: 0197-9337, 1096-9837. DOI: 10.1002/esp.4766.

[21] Subsecretaría de Recursos Hídricos de la Nación e Instituto Nacional del Agua. *Catálogo y Visualizador de información hidrológica - SIyAH - INA*. URL: https://alerta.ina.gob.ar/pub/mapa (visitado 21-06-2022).

[22] Conference of the Parties. *The Ramsar Strategic Plan 2016-2024*. 2015. URL: https://www.ramsar.org/sites/default/files/documents/library/cop12_res02_strategic_plan_e_0.pdf (visitado 16-09-2022).

[23] Mercedes Salvia. "Aporte de la teledetección al estudio del funcionamiento del macrosistema Delta del Paraná: análisis de series de tiempo y eventos extremos". Tesis doct. Buenos Aires, Argentina: UBA, 2010.

[24] Ville Satopaa y col. "Finding a 'Kneedle' in a Haystack: Detecting Knee Points in System Behavior". En: *2011 31st International Conference on Distributed Computing Systems Workshops*. Minneapolis, MN, USA: IEEE, jun. de 2011, págs. 166-171. ISBN: 978-1-4577-0384-3. DOI: 10.1109/ICDCSW.2011.20.

[25] *¿Qué es el Software Libre? - Proyecto GNU - Free Software Foundation*. URL: https://www.gnu.org/philosophy/free-sw.es.html#f1 (visitado 05-03-2022).