AGRISCIENTIA, 2019, VOL. 36 (2): 67-80

COMUNICACIÓN

Gestión de datos espacio-temporales de imágenes satelitales

Castillo Moine, M. A. y Balzarini, M. G.

RESUMEN

El manejo de datos de largas series temporales del índice de vegetación de diferencia normalizada (NDVI) en territorios extensos demanda un uso eficiente del recurso computacional. En este trabajo se discuten e ilustran estrategias para la construcción y procesamiento estadístico de bases de datos masivos espacio-temporales provenientes de imágenes satelitales. Se detalla la implementación de un protocolo de manejo de datos en el software R, con implementación de cómputos paralelizada. Los resultados muestran que el concepto *dividir-aplicar-combinar* resultó adecuado para filtrar y clasificar largas series de tiempo de NDVI distribuidas territorialmente a escala regional.

Palabras clave: SIG; MODIS; series de tiempo de imágenes satelitales; dividir-aplicar-combinar; protocolo de gestión de datos; procesamiento en paralelo.

Castillo Moine, M. A. and Balzarini, M. G., 2019. Spatio-temporal data management of satellite imagery. Agriscientia 36 (2): 67-80

SUMMARY

The management of long time series data of Normalized Difference Vegetation Index (NDVI) over large territories demands efficient use of computational resources. This paper discusses and illustrates strategies for the construction and statistical processing of massive spatio-temporal databases from satellite images. The implementation of a data management protocol in the R software is detailed, with implementation of parallel computations. The results show that the concept *divide-apply-combine* was adequate to filter and classify long time series of NDVI territorially distributed at a regional scale.

Keywords: GIS; MODIS; satellite image time series; divide-apply-combine; data management protocol; parallel processing.

Fecha de recepción: 05/02/2018; fecha de aceptación: 25/09/2019.

Castillo Moine, M. A.: Universidad de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Investigaciones Fisiológicas y Ecológicas Vinculadas a la Agricultura (IFEVA), Facultad de Agronomía. Av. San Martín 4453, (C1417DSE), Buenos Aires, Argentina. Balzarini, M. G.: Universidad Nacional de Córdoba, Consejo Nacional de Investigaciones Científicas y Técnicas, Cátedra de Estadística y Biometría, Facultad de Ciencias Agropecuarias. Ing. Agr. Félix Aldo Marrone 746, Ciudad Universitaria, Córdoba, Argentina. Correspondencia a: mcastillomoine@agro.uba.ar

INTRODUCCIÓN

Existen numerosas técnicas de medición y recolección de datos provenientes de variables espacialmente distribuidas, es decir, asociadas a una ubicación espacial. Del mismo modo, se pueden medir y recolectar datos de variables temporalmente distribuidas. La realización de variables asociadas a ubicaciones espaciales a lo largo del tiempo genera un tercer tipo de dato, el dato espacio-temporal. Dado que para cada unidad de análisis pueden registrarse una o más variables, pueden quedar definidos vectores multivariados para el espacio muestral de datos espacio-temporales. El almacenamiento y procesamiento de estas bases de datos requiere especial atención computacional. En las últimas décadas, el incremento de tecnologías para la medición de variables biofísicas, la reducción de costos de la industria de los sensores, y el incremento de la capacidad de cómputo con procesamiento secuencial y en paralelo, han favorecido el desarrollo y consolidación de bases de datos masivos de naturaleza espacio-temporal para estudios ambientales.

Los datos provenientes de sensores remotos revisten particular interés. Montados en plataformas satelitales permiten tener registros históricos de estimaciones de variables biofísicas para todo el planeta; su volumen asciende al orden de los petabytes, y su heterogeneidad es alta ya que abarcan una gran cantidad de variables. Nuevas tecnologías en sistemas de información geográfica (SIG) han permitido mejorar la sistematización de las bases de datos espaciales y espacio-temporales. La integración de los datos cartográficos y los provenientes de sensores remotos contribuyen al estudio de procesos territoriales de gran escala, como los cambios de coberturas del suelo a bajo costo. Una característica importante a tener en cuenta para integrar bases de datos espacio-temporales en un SIG es la diversidad de formatos en la que podría presentarse la información. Los formatos más usuales incluyen las geometrías línea, punto y polígono (vectorial) y matricial (*raster*).

Diferentes entornos de escritorio permiten conformar bases de datos de distintos formatos, asociadas a coberturas del suelo. Los cambios de uso de suelo responden a muchos factores, y una de las principales consecuencias directas en el territorio es el cambio de las coberturas. Por esto, los cambios de cobertura de la superficie terrestre son evaluados a través de la respuesta espectral de las mismas, esto es, la reflectancia o absorbancia medidas a diferentes longitudes de onda, según corresponda. Entre las herramientas de software para la construcción y manejo de estas bases de datos se destacan PostGIS (OSGEO Development Team, 2018), GRASS (GRASS Development Team, 2018), QGIS (QGIS Development Team, 2018) y R (R Core Team, 2018), siendo este último el más recomendado para implementar análisis estadísticos simples y avanzados. La mayoría de los programas SIG de escritorio de código abierto son lo suficientemente versátiles como para gestionar en un mismo proyecto numerosas capas de datos, en diferentes geometrías y formatos de almacenamiento. Sin embargo, la eficiencia de gestión de las capas de información aún está lejos de ser óptima (principalmente porque una gran cantidad de datos es volcada en memoria RAM), lo que plantea desafíos a la hora de planificar el armado de un SIG. Los análisis estadísticos para datos espaciales, temporales y espacio-temporales están ligados a lenguajes de programación como C++, R y Python. Los dos últimos son lenguajes de código abierto (*open source*) que integran desarrollos de C++, resultando en una mayor accesibilidad para el usuario. En particular, R ha tomado especial importancia en la comunidad estadística. GDAL (GDAL/OGR Contributors, 2018) es una librería desarrollada para soportar datos en formato *raster* (155 formatos) y vectorial (95 formatos), que puede ser usada desde R de manera directa mediante *rgdal* (Bivand, Keitt y Rowlingson, 2014). Existen

numerosos desarrollos estadísticos en R para datos espaciales, temporales y espacio-temporales, pero una limitante intrínseca del lenguaje se impone: los procesos ocurren todos en memoria. R utiliza en las operaciones matemáticas de la mayoría de las funciones, formato matriz o vector (opcionalmente *data frames* que son convertidos internamente a matriz o vector), y el formato preferido para la presentación de resultados es la lista. Existen desarrollos de *Memory Mapped Files* (Adler, Gläser, Nenadic, Oehlschlägel y Zucchini, 2018; Kane, Emerson, Haverty y Determan, 2018; Ryan, 2018), *Shared Memory* (Kane et al., 2018), y *Map and Reduce Memory* (Henry y Wickham, 2018) y *dividir-aplicar-combinar* (Wickham, 2011), que convierten a R en un entorno apto para la ejecución de algoritmos de procesamiento de grandes volúmenes de datos en máquinas locales, optimizando el uso de recursos de hardware, principalmente memoria RAM. La computación en paralelo que puede implementarse en R es una práctica frecuente para el procesamiento de grandes volúmenes de datos. En este trabajo, se discuten estrategias para la construcción de bases de datos de series de tiempo de imágenes satelitales con el software R y se ilustran las ventajas del marco de trabajo *dividir-aplicar-combinar* para el preprocesamiento de los datos.

MATERIALES Y MÉTODOS

Datos de ilustración

Se conformó una base de datos de series de tiempo de imágenes satelitales (STIS) para un área de Córdoba, Argentina, que se extiende aproximadamente entre los 30° y 33° de latitud sur y los 63° y 66° de longitud oeste. El área presenta diversidad topográfica y climática, incluyendo dos regiones naturales: la del Espinal y la del Chaco Seco (Luti *et al.*, 1979) y diversidad de suelos representada en nueve órdenes (Capitanelli, 1979; Jarsún, 1981; Gorgas y Tassile, 2006). De norte a sur se extiende un conjunto de cordones montañosos; estructuralmente se trata de bloques tectónicos elevados, separados por valles elevados. Estos cordones montañosos funcionan como trampa adiabática para los vientos húmedos provenientes del sureste, favoreciendo las precipitaciones sobre las laderas orientales o en los valles entre ambos cordones. Por consiguiente, las cuencas de montaña son de especial importancia. El extremo noroccidental del área está dominado por una extensa salina. El clima es semiárido de régimen estacional monzónico con déficit hídrico, y cambia de sureste a noroeste de más húmedo y fresco a más cálido y seco, con presencia de microclimas. Las precipitaciones y temperaturas son las principales variables meteorológicas formadoras del paisaje.

Se utilizaron datos provenientes de los sensores MODIS (*Moderate Resolution Imaging Spectoradiometer*) (plataforma *Terra*), a resolución espacial de 250, 500 y 1000 m cada uno a dos días, en 36 bandas espectrales entre los 0,4 y 1,4 µm (nivel 3). Con resolución 250 m se obtuvieron capas de información de las bandas relacionadas a la vegetación, con 500 m de las bandas relacionadas a otras propiedades de las coberturas terrestres y con 1000 m de las bandas de temperatura de superficie. La cobertura temporal abarca el período comprendido entre los años 2000 a 2018. Una característica destacable es la sistematización de los datos MODIS en productos cuyas características pueden ser consultadas en MODIS *Products Table (*LP DAAC - *Land Processes Distributed Active Archive Center,* s. f.*)*. Cada producto es una colección de datos de una o más bandas espectrales y sus respectivos atributos de calidad en un único archivo, para un área y período de tiempo fijos. El nivel de los productos indica el grado de preprocesamiento. Las capas de información representando cada variable espacialmente distribuida, pueden almacenarse combinadas en un único archivo o en archivos individuales.

Se recolectaron datos de los productos MOD13Q1 (Solano, Didan, Jacobson y Huete, 2015) MOD09Q1 (Vermote y Ray, 2015), MOD11A2 (Wan, Hook y Hulley, 2015), MOD09A1 (Vermote y Ray, 2015), colección 6. MOD13Q1 provee los valores de NDVI (Normalized Difference Vegetation Index) y EVI (Enhanced Vegetation Index) a nivel de píxel, a resolución espacial de 250 m cada 16 días junto a los valores de reflectancia de las bandas 1, 2, 3 y 7 y medidas de calidad de observación. El NDVI se define como

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

donde _NIR_ (near infrared) es la reflectancia en el infrarrojo cercano (banda 2), y _Red_ es la reflectancia en el rojo (banda 1). El _EVI_ se define como

$$EVI = 2.5\,\frac{NIR - Red}{NIR + C_1 \times RED - C_2 \times Blue + L}$$

donde *L* es un factor de ajuste para el efecto de fondo de la canopia, y c1 y c2 coeficientes de la resistencia de los aerosoles atmosféricos, y *Blue* la

banda 3 (azul). El dato se construye con el máximo valor registrado con máxima calidad en el período de 16 días. Los valores fuera de rango (para el mínimo y máximo valor posible de cada banda) y con presencia de nubes se consideran de baja calidad. Luego, es elegido el píxel de máximo valor (Solano *et al*., 2015) empirical MOD09Q1*,* provee reflectancias de superficie para las bandas 1 y 2 a 250 m de resolución. Cada píxel contiene la mejor observación posible para un período de 8 días (Vermote y Ray, 2015). Con este producto se construye el producto MOD13Q1. Con el producto MOD11A1 se obtuvieron los valores promedio de temperatura de superficie (*Land Surface Temperature*, LST) para un período de 8 días, con una resolución de 1 km. El valor de cada píxel se construye como el promedio simple de los valores registrados para el producto MOD11A1 (Wan, 2013). Finalmente, se compila MOD09A1, el cual provee los valores de reflectancia de superficie para las bandas 1 a 7 a una resolución de 500 m. Cada píxel contiene la mejor observación posible para el período de 8 días, teniendo en cuenta la máxima cobertura de observación, ángulo de observación cenital bajo, ausencia de nubes o aerosoles atmosféricos (Vermote y Ray, 2015). Para el tratamiento de datos masivos se usaron los paquetes *bigmemory* (Kane, Emerson, y Weston, 2013)*, foerach* (Microsoft y Weston, 2017) *y doParalell* (Calaway, Microsoft Corporation, Weston, & Tenenbaum, 2018) de R. *Bigmemory* permite el uso y creación de grandes matrices (llamadas *big.matrix*) que no caben en memoria RAM, gracias a la implementación de *descriptor pointers*, que son descriptores de puntos de acceso a los datos, los cuales se encuentran almacenados en disco (*shared memory*). Si bien la principal limitante pasa a ser la velocidad del disco duro, los métodos de indexación permiten usar una *big.matrix* como cualquier otra matriz en R, cargando en memoria RAM pequeños sectores de manera secuencial (*memory-mapped files*). Solo algunos métodos tratarán de convertir la *big.matrix* a matriz por completo y darán error de memoria insuficiente. Sin embargo, cualquier algoritmo de base matricial puede potencialmente reprogramarse usando internamente objetos de clase *big.matrix*. Algunos paquetes asociados a *bigmemory* ofrecen herramientas eficientes para operaciones matemáticas. Todas las operaciones pueden paralelizarse usando los paquetes *foreach* y *doParalell*. Para una paralelización eficiente, es aconsejable pensar cómo la matriz puede dividirse en partes, aplicar el algoritmo en cada parte, y recombinar los resultados en un formato adecuado (*dividir-aplicar-combinar*).

Otro paquete de R que se destaca por su versatilidad y eficiencia, para la manipulación de grandes volúmenes de datos espaciales, es el paquete *raster* (Hijmans, 2018). Al igual que *bigmemory*, utiliza *descriptor pointers*, con la salvedad de que soporta todos los drivers admitidos por *rgdal*, evitando en etapas tempranas de procesamiento la duplicación de datos. Para la gestión de series de tiempo de imágenes satelitales en formatos aceptados por el paquete *raster*, puede usarse *rts* (Naimi, 2018), que combina clases para datos espaciales y temporales valiéndose de la especificación de un índice temporal. En este trabajo se definió un índice espacial (número de celda) y uno temporal (fecha de adquisición de la imagen definido como el primer día del período de 8 días para MOD09Q1) que se conservó durante todo el proceso.

Protocolo para la gestión de bases de datos espacio-temporales

Paso 1. Definición de la base de datos

Conocer cuál es el origen de los datos, su formato, volumen de almacenamiento, cantidad de capas de información. Para ello es recomendable:

a. Obtener los archivos conteniendo las capas de entrada y organizarlos en un directorio de trabajo (estructura jerárquica de archivos).
b. Verificar la integridad de los metadatos asociados a cada capa y archivo obtenidos.
c. Verificar la integridad de la base de datos, corregir faltantes si los hubiera, y generar un índice espacio-temporal.
d. Definir el sistema de proyección más adecuado (geográfico o cartográfico) y fijar un Sistema de Coordenadas de Referencia (SRC) en función de los marcos de referencia o estándares oficiales para la región, si los hubiere.

Paso 2. Definición de los algoritmos de procesamiento

Tener presente que cada paso de procesamiento implica una transformación de los datos y que ocurre en memoria RAM (algunos lenguajes de programación tienen herramientas para predecir la demanda de RAM de algoritmos determinados). Se recomienda:

a. De cada archivo, definir cuáles son las variables (capas) sobre las que se trabajarán si estos son multijerárquicos.
b. Definir qué cómputos se realizarán con cada variable.

c. Establecer un flujo de trabajo para cada variable.
d. Definir los *inputs* y los *outputs* de cada paso y verificar la concordancia de formato de las capas entre el *output* de un paso y el *input* del siguiente.
e. Determinar el paso en el que todas las capas deberían tener el mismo SRC y reproyectar todas desde ese paso en adelante. Este punto debería ser tal que se tenga que realizar la menor cantidad de reproyecciones posibles. Verificar con algunas capas la calidad espacial (al menos geométrica) de las reproyecciones.
f. Generalizar el flujo de trabajo de una capa *a todas* las demás que sufrirán el mismo procesamiento.

Paso 3. Definición del formato de resultados.

Definir el formato de presentación de resultados, los cuales pueden reportarse desde tablas y gráficos como hasta en un SIG en nuevas capas. En todos los casos el dominio temporal y la extensión espacial deben ser especificados.

RESULTADOS Y DISCUSIÓN

Paso 1. Definición de la base de datos

Dependiendo de la geometría y tipo de datos que conformarán la base de datos y del software con el que se quiera realizar el procesamiento, dependerá el formato de almacenamiento en R. Un formato moderno que admite todas las geometrías es GeoJSON (Butler *et al.*, 2016). Unificar el formato de almacenamiento y el SRC implica tareas de extracción de datos y reproyección de cada capa de datos. En este trabajo se usó un SRC de acuerdo con los marcos de referencia usados en Argentina (POSGAR, 2007). La reproyección implica duplicar el archivo de datos por completo, ya sea en memoria RAM o en una copia física. Si se opta por copias físicas, se debe pensar en los formatos de almacenamiento soportados por las diferentes librerías del software. El formato del dato es también importante, ya que siempre que se pueda es recomendable guardarlos en un tipo de dato que ocupe menos espacio (Wickham, 2015). Por ejemplo, el tipo de dato entero INT4S del paquete *raster* permite números enteros positivos y negativos usando 4 bytes por número, y el tipo FLT8S numérico flotante de doble precisión 8 bytes por número. En el paquete *bigmemory* los tipos de datos pueden ser flotante de doble precisión, entero, corto o catacter (8, 4, 2, y 1 bytes, respectivamente). Por ejemplo un valor de NDVI de 0,6549 (flotante) puede ser representado como 654 × 0,001 (entero, redondeando al tercer decimal).

Los productos MODIS son distribuidos como *.hdf*, un formato de compresión altamente eficiente (Folk, Heber, Koziol, Pourmal y Robinson, 2011). Dado que este formato no puede ser accedido desde *rgdal*, en este trabajo se extrajeron los datos de cada imagen para el área de estudio y se almacenaron como *GTiff*, un formato que permite operar desde prácticamente cualquier entorno, incluyendo *rgdal*.

Los datos MODIS para el período 18/02/2000 al 18/02/2018, totalizaron 828 archivos en formato *.hdf* para el *tile* h12v12 de la grilla MODIS de los productos MOD11A2, MOD09Q1 y MOD09A1, y 417 para MOD13Q1. Todos fueron descargados desde la aplicación *Earthdata Search (*«NASA», s. f.). Se usaron gestores de descargas paralelas. Paquetes de R como *MODIStsp* (Busetto y Ranghetti, 2016) ofrecen descarga de datos MODIS. Una vez descargados, se extrajeron los datos para el área de estudio definida por la extensión con la utilidad *gdalwarp* (GDAL/OGR Contributors, 2018). Los *flags -multi* y *-wo* aceleraron el proceso permitiendo el primero optimizar la lectura de cada imagen por partes, y el segundo habilitando el procesamiento en paralelo mediante el argumento *NUM_THREADS=val/ALL_CPUS*. Los datos fueron reproyectados al SRC planar EPSG 5345 Argentina Faja 3 (POSGAR 2007) de acuerdo a la ubicación de la región. El método de extracción utilizado fue *vecino más cercano*. Se conservaron los atributos de metadatos. Cada banda extraída fue almacenada en imágenes en formato *GTiff* unibanda, y organizadas en directorios individuales –un directorio por banda por producto–. La decisión de reproyección en este paso estuvo ligada a la aplicación de algoritmos no mostrados en esta ilustración. Para el caso de la ilustración, se puede mantener el SRC de origen hasta el paso 3. Para la gestión eficiente de datos *raster* en R se utilizó la función *stack*, que crea un apilado de imágenes *raster* con la misma resolución, generando un *descriptor pointer* a cada archivo. Un apilado de imágenes *raster* es equivalente a una serie de tiempo de imágenes. Se eliminaron los datos fuera de rango para cada variable y los que no hubieran sido generados con calidad óptima, y se predijeron los valores faltantes por interpolación lineal en el dominio del tiempo. Para operar con la banda de calidad se programaron operaciones lógicas, creando máscaras con la función *mask* del paquete *raster*. Los *datasets* resultantes fueron almacenados como *stacks* en un único archivo.

Si se sabe que la base de datos será actualizada regularmente es recomendable guardar archivos separados, y crear nuevos *descriptor pointers* a los archivos con cada actualización de la base de datos.

Paralelamente se almacenaron capas vectoriales de información auxiliar provenientes de los grupos de trabajo asociados a IDERA (Infraestructura de Datos Espaciales de la República Argentina) (IDERA, s. f.), proporcionadas a través de su portal de datos y servicios WMS y WFS. Estas capas fueron: Suelos de la República Argentina 1:500000 y 1:1000000 (INTA, 2018), mapa de unidades de vegetación de la Argentina (UVA) (Oyarzabal *et al*., 2018), mapa de sistemas de usos de tierras en Argentina presentando 25 usos (MAyDS, 2018) y un mapa de ecorregiones (Morello, Matteucci, Rodríguez y Silva, 2012). Del Instituto Geográfico Nacional se descargaron las capas de información para aguas continentales (Puchet y IGN, 2018), límites políticos administrativos (Puchet y IGN, 2017) y mapa de coberturas del suelo histórico (IGN, 2013), en formato vectorial. En formato *raster* se compilaron los datos del producto *MDE-Ar* (Modelo Digital de Elevaciones de la Argentina) (IGN, 2018), derivado de la misión SRTM (*Shuttle Radar Topography Mission*). La resolución de MDE-Ar es de 30 m, y su calidad ha sido verificada y mejorada (filtrado de datos extremos, relleno de vacíos, enmascarados y límites). Con el *MDE-Ar* se obtuvieron las variables topográficas orientación, pendiente, índice de posición topográfica, índice de irregularidad del terreno. De IDECOR se obtuvo el mapa de cobertura del suelo nivel 2 de la provincia de Córdoba (García *et al*., 2018) que fue recategorizado a once categorías más generales. El detalle de la recategorización fue documentado en los metadatos de la capa. A partir de esta capa se generó una máscara para áreas urbanas. Todas las capas de información auxiliar fueron reproyectadas al SRC EPSG 5345 Argentina Faja 3 POSGAR 2007. Si bien la reproyección puede posponerse a la instancia de resultados, haberla realizado en esta etapa permitió realizar controles visuales de calidad de la geometría. Es imprescindible que todas las capas tengan una alta calidad geométrica en todas las etapas de procesamiento. Mínimamente debe usarse un sistema de proyección adecuado para la región de estudio y la localización de los píxeles de las imágenes satelitales, capas vectoriales auxiliares y de los productos obtenidos debe coincidir lo más exactamente posible con los datos usados de referencia. Como datos de referencia pueden usarse puntos georreferenciados a campo u otros productos georreferenciados de calidad conocida. Si esta condición no se cumple los métodos de geoproceso que involucran diferentes capas fallan y la propagación de errores espaciales puede ser alta (Congalton, 2015). Los mapas de tres capas auxiliares de información se muestran en la Figura 1.

Para las etapas posteriores al preprocesamiento se tuvo en cuenta el formato en que aceptan los datos cada algoritmo utilizado. Dado que la mayoría de los algoritmos usa formatos matriciales, se iteró a través de las capas de los *stacks* del paso anterior (recordar que cada capa de un *stack* contiene los datos espaciales para una fecha) para construir matrices adecuadas. Si bien la mayoría de los paquetes de R que trabajan con imágenes ofrecen métodos automáticos de conversión entre imágenes y matrices, una precaución debe tomarse: las imágenes digitales (entre ellas los *raster*) son consideradas matrices, dado que cada celda se corresponde con un número de fila y columna en el lienzo de dibujo, y contiene el valor registrado para una variable. La conversión directa de *raster* a matriz da como resultado una matriz con un número de columnas ≥ 1 (tantas como columnas tenga el lienzo de dibujo). En consecuencia, para todos los procesamientos, R considera que cada columna de la matriz resultante es una variable diferente. Para evitar esta interpretación errónea, la imagen *raster* debe ser vectorizada. Cada vector tiene longitud igual al número de celdas del *raster*, y será ubicado rellenando una columna de la matriz. En los pasos subsiguientes se operará con la matriz, donde cada fila representa un píxel y cada columna una fecha de adquisición. Si se mantiene el orden de las filas y de las columnas invariante, es posible reconstruir cada *raster* o *stack* original, o construir nuevos *raster* y *stacks* con los resultados de los métodos. Todas las matrices fueron construidas usando el formato *big.matrix*, conservando una única vez el índice espacial tomando como molde una imagen de extensión y resolución espacial adecuada. Las matrices resultantes contienen en cada fila una serie de tiempo para un píxel, y en cada columna las observaciones para una fecha dada. La ventaja de este formato radica en la indexación más rápida de las series de tiempo respecto a un *stack*. El método de indexación es el mismo que para las matrices clásicas, usando *[i, j]*, donde *i* representa las filas y *j* las columnas.

Si el destino de las capas *raster* espacio-temporales es realizar análisis univariado, la resolución espacial y temporal de cada STIS no tiene importancia; solo es necesario que las series de tiempo estén completas y tengan igual longitud. Si el destino de los *raster* temporales es realizar análisis multivariados, se suman a esos

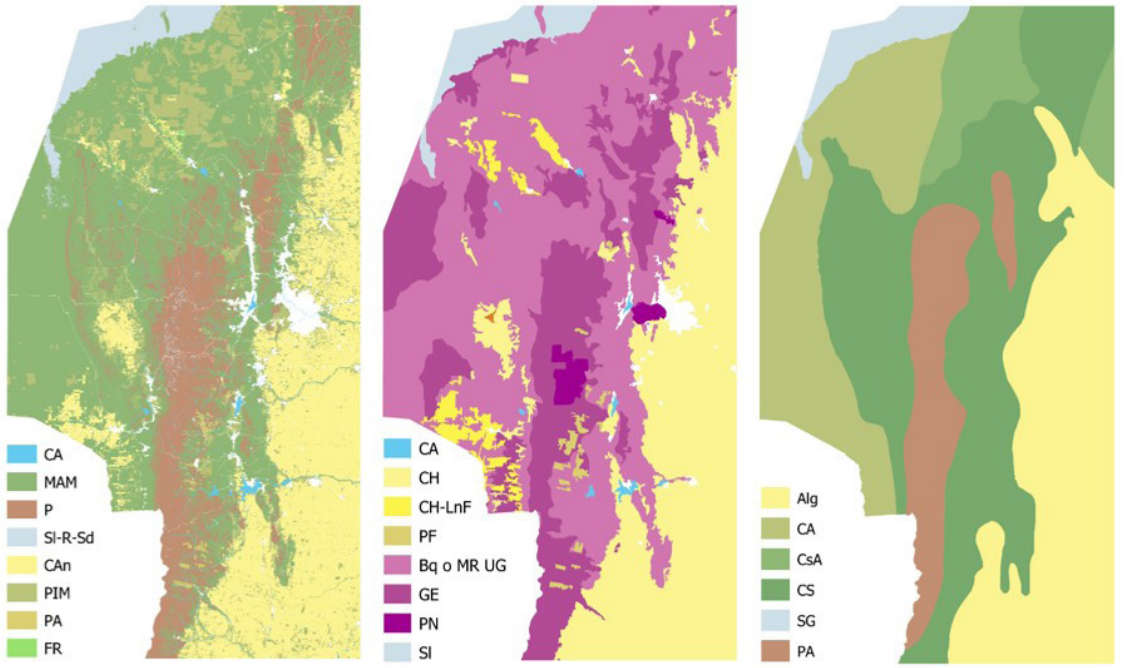

Figura 1. Mapas de capas de información auxiliar. (a) Mapa de cobertura del suelo Nivel 2 de la provincia de Córdoba (IDECOR, 2018). Referencias: CA: cuerpos de agua; MAM: monte arbustos y matorrales; P: pastizales; Sl-R-Sd salinas, roca o suelo desnudo; Can: cultivos anuales; PIM: pasturas implantadas manejadas; PA: plantaciones arbóreas; FR: frutales bajo riego. (b) Mapa de sistemas de usos de tierras en Argentina (MAyDS, 2018). Referencias: CA: cuerpos de agua; CH: cultivos de herbáceas; CH-LnF: cultivo de herbáceas o leñosas no forestales; PF: plantaciones forestales; Bq o MR UG: bosque o matorrales con uso ganadero; GE: ganadería extensiva; PN: parques nacionales; Sl: salares o salina (c) UVA: mapa de unidades de la vegetación de la Argentina (Oyarzábal *et al.*, 2018). Referencias: Alg: algarrobal; CA: Chaco Árido; CsA: Chaco Semiárido; CS: Chaco Serrano; SG: Salinas Grandes; PA: pastizales de altura. En blanco: áreas enmascaradas.

requerimientos que: para todas las variables, el dato espacial en formato *raster* tenga la misma resolución espacial, los centroides de cada celda entre capas coincidan, y que todos los *raster* tengan la misma extensión. Se ajustó la extensión y resolución de cada capa al producto de menor resolución. Dado que las bandas 1 y 2 son las de menor resolución con 250 m, las demás bandas fueron reescaladas a 250 m usando interpolación espacial por vecino más cercano. Se calculó la variable NDVI a 250 m de resolución cada 8 días usando las bandas 1 y 2 del producto MOD09Q1. También se acumularon las precipitaciones para el período comprendido entre dos archivos MODIS consecutivos, esto es cada 8 y 16 días. Los archivos correspondientes a la serie de tiempo de imágenes satelitales para la variable NDVI a 8 días y 250 m de resolución espacial. Los archivos de datos necesarios para construir la variable NDVI que se usó como ilustración en este trabajo, junto a los *script* de R necesarios, se encuentran disponibles como material suplementario. Se usó como formato de salida de las STIS de este paso el formato *RasterStack*.

Paso 2. Definición de los algoritmos de procesamiento

Con fines de procesamiento se operó con formatos *matriz*, *big.matrix* y *RasterStack*. Para interoperar entre matrices y raster es importante que la dimensión de las capas de datos sea consistente entre los datos de entrada y de salida. Revisar las salidas de cada algoritmo de R y guardar solo el resultado necesario es una práctica recomendable. Exportar los resultados a disco permite liberar memoria RAM. De esta manera, grandes volúmenes de datos, del orden de los cientos de gigabytes, pueden ser procesados en hardware promedio de escritorio.

Para el análisis de series de tiempo el filtrado y la eliminación de estacionalidad es una técnica de preprocesamiento común que mejora las clasificaciones de STIS (Kim *et al.*, 2014, Cao *et al.*, 2018). Para el conjunto de datos de ilustración se usó el filtro de suavizado Savitzky-Golay (Savitzky y Golay, 1964), que elimina parte de variabilidad residual de alta frecuencia y corrige para *outliers* conservando la forma de las series de tiempo. Se

usó la función *sgolayfilt* del paquete *signal* (Signal Developers, 2013) parametrizada con longitud de filtro de tamaño 9 y orden 3, con un factor de escala de 1. Para estos preprocesamientos se transformaron los *RasterStack* a *big.matrix*, se hicieron constantes las filas de la matriz correspondientes a píxeles enmascarados y se procesaron las series de tiempo con un algoritmo en paralelo basado en el concepto de *dividir-aplicar-combinar*. Las áreas enmascaradas se mantienen inalteradas en los pasos siguientes. La *big.matrix* se particionó en partes de 50 000 pixeles (filas) y tantas columnas como longitud de la serie de tiempo (828). Se desarrolló la función *prepare_time_series_stack* que procesa cada serie de cada parte en un núcleo diferente y finalmente reordena las partes en una nueva *big.matrix*. En la Figura 2 se muestra una serie de tiempo de NDVI, antes y después del filtrado. Para ilustrar otra implementación del concepto *dividir-aplicar-combinar* se aplicó un procesamiento de clasificación a las series de tiempo sin depurar de cada píxel. Se realizó una clasificación no supervisada sobre un conjunto de entrenamiento reducido con un algoritmo de k-*means* usando distancia euclídea (Hartigan y Wong, 1979) , un enfoque robusto para agrupar datos ordenados secuencialmente (Paparrizos y Gravano, 2015). Para este fin, se seleccionó aleatoriamente una muestra de entrenamiento de 2502 series de tiempo de NDVI; el tamaño muestral fue calculado de acuerdo a la propuesta de Cochran (Cochran, 1977) según se discute en Olofsson *et al.* (2014) para el muestreo aleatorio. Para elegir la cantidad de grupos se usó el método del codo sobre la muestra. El método consiste en analizar cuánta variabilidad es explicada por el agrupamiento conforme se aumenta el número de grupos. Se selecciona el número *C* de grupos de significancia biológica conocida –si es posible– por encima del cual añadir más grupos no explica mejor la variabilidad total observada. Se agruparon las series de tiempo filtradas en $C = 7$ grupos con el modelo k-*means* ajustado se predijeron los resultados del agrupamiento de cada parte de la *big.matrix* aplicando la misma idea en particiones para 50 000 píxeles cada una. El resultado de la clasificación se exportó como mapa, y se acondicionó junto a capas de información auxiliar en un SIG para visualizar la distribución y relación espacial de las mismas. El gráfico del método del codo para este ejemplo se muestra en la Figura 3a; se observó que pocos grupos tenían poca significancia biológica y que en más de 7 grupos el agrupamiento no explicaba mejor la estructura de los datos.

Si bien el flujo de trabajo presentado no es el único posible, la estrategia *dividir-aplicar-combinar* usada permitió procesar STIS con dos funciones de R desarrolladas *ad hoc*. La función *prepare_time_series_stack* permitió filtrar las series de tiempo y acondicionarlas para el procesamiento mientras que la función *predict_km_bigmem* permitió realizar el agrupamiento de los píxeles con el algoritmo k-*means* según el patrón espacio-temporal subyacente. Ambas funciones se encuentran en el *script* del material suplementario. La función *prepare_time_series_stack* crea una *big.matrix* de dimensión igual al *RasterStack* de entrada, en la que se almacenarán las series de tiempo originales, y realiza lo propio para almacenar las series de tiempo filtradas. Luego, de manera iterativa lee cada capa del *RasterStack*, lo vectoriza y guarda en una columna de la matriz (iterar a través de las capas del *RasterStack* es más rápido que iterar a través de las series de tiempo). De este modo las series de tiempo de los pixeles se almacenan en las filas de la *big.matrix*. Para ejecutar tareas en paralelo se generó un índice espacial que se usó para dividir la matriz en partes; cada parte es cargada a la memoria RAM y procesada usando computación en paralelo. La función *foreach* a través de su operador binario *%dopar%* es la que permitió procesar en paralelo

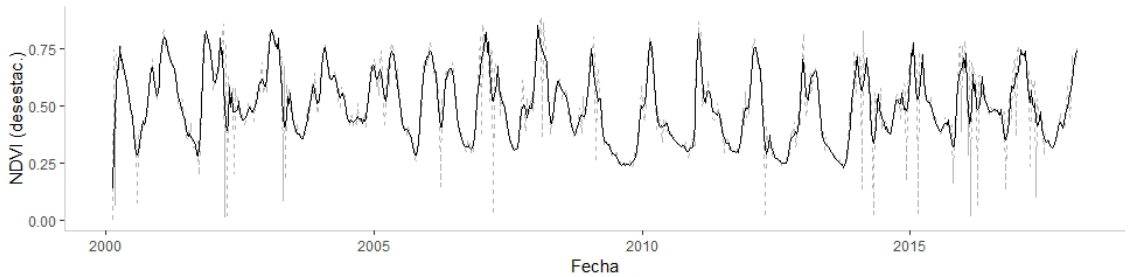

Figura 2. Preprocesamiento de una serie de tiempo de NDVI para un píxel. En línea punteada la serie de tiempo original, en línea sólida la serie de tiempo procesada con el filtro de Savitzky y Golay.

cada parte de la matriz. Las partes de la matriz con series de tiempo filtradas son grabadas en la *big.matrix* creada para tal fin. Luego se crea una copia permanente de la *big.matrix*, se cierran las sesiones de procesamiento en paralelo y se genera el resultado. En la etapa de procesamiento, las STIS filtradas fueron muestreadas de manera aleatoria y la muestra fue clasificada con la función *k-means*, y el modelo obtenido fue usado para predecir el grupo de pertenencia del resto de las STIS con la función *predict_km_bigmem*. Así, la implementación de las dos funciones creadas (filtrado y clasificación) ilustra la utilidad del concepto *dividir-aplicar-combinar*.

Paso 3. Definición de formatos de resultados

El resultado de la clasificación es un vector que indica el grupo de pertenencia de la serie de tiempo de cada píxel. En el ejemplo, contiene las etiquetas para siete clases cuya distribución espacial se muestra en la Figura 3b en un mapa, generado en formatos vectorial y *raster*. Se eligieron ambos formatos teniendo en cuenta las operaciones de muestreo posteriores en R. Las capas de información auxiliar y de los resultados fueron ordenadas y compiladas en un SIG que se encuentra disponible bajo requerimiento a los autores. En este trabajo se usó el entorno QGIS porque se puede integrar con R –exportando los resultados en algún formato compatible, mediante el paquete *RQGIS* (Muenchow, Schratz y Brenning, 2017) o mediante el R *Processing Scripts Module* (Ghetta y QGIS DevelopmentTeam, 2018)– y por su potencial para la visualización y edición de las capas de resultado, más sencilla que en R. No existe una recomendación única en cuanto a la geometría en la que deben gestionarse las capas de salida que resultan de los análisis. Se recortaron las capas auxiliares al área de interés. La reproyección, recorte y gestión de las capas vectoriales puede realizarse de diferentes maneras. En los *script* de ilustración del material suplementario se usó la herencia de atributos al crear capas nuevas a partir de una existente en R. Operaciones complejas de SIG, como el recorte vectorial, no son eficientes en R pero pueden realizarse cuando el software ha sido elegido como entorno de trabajo (Muenchow *et al.*, 2017), en los últimos pasos del código de ilustración se realiza esta operación. El mapa resultante de la clasificación espacio-temporal se muestran en la Figura 3b. Las áreas en blanco representan zonas enmascaradas que no han sido analizadas.

Dado que el mapa sintetiza la historia del área de estudio, no puede compararse de manera directa con las clasificaciones logradas en las capas de información auxiliar de la Figura 1, que representan información para un momento específico de tiempo o características más constantes como el mapa de las UVA. Asimismo el algoritmo de clasificación y el tipo de centroide de grupo usados para clasificar las series de tiempo pueden generar resultados diferentes. No obstante en el mapa obtenido a través de la clasificación de las STIS las siete clases logradas copian aproximadamente la forma de los límites de las principales unidades de vegetación (Salinas Grandes, Chaco, Pastizal de Altura y Algarrobal) y de los usos del suelo (agricultura/ganadería). Se evaluó el solapamiento espacial entre los grupos del mapa obtenido y el mapa de unidades de vegetación. Para esta comparación se usó el paquete *sabre* (Nowosad y Stepinski, 2018) que provee los índices de asociación entre regionalizaciones independientes *mapcurves* (Hargrove, Hoffman y Hessburg, 2006) y *V-measure* (Rosenberg y Hirschberg, 2007; Nowosad y Stepinski, 2018). Valores más altos de *MapCurves* y de *V-measure* indican más coincidencia entre las regiones del mapa comparado y el mapa de referencia; los valores de ambos índices oscilan entre 0 y 1. Los índices *Mapcurves* y *V-measure* arrojaron valores de 0.44 y 0.31 respectivamente.

El mapa obtenido con las STIS muestra fragmentación dentro de cada una de las unidades de vegetación ya que no solo indica el tipo de unidad o vegetación predominante si no también su evolución temporal en los últimos 18 años del área de estudio. La distribución espacial de los grupos obtenidos muestra patrones más heterogéneos en el noroeste del área, donde la variabilidad temporal ha sido mayor. Dado que el mapa surge de un método de clasificación no supervisado, los nombres de los grupos obtenidos pueden traducirse a nuevas clases que tengan un mayor significado para su interpretación. Las clasificaciones de las STIS logradas con otro número de grupos dieron lugar a interpretaciones diferentes, por ejemplo se observó que con seis grupos no se pudo discriminar pastizales de cultivos porque son grupos de firmas espectrales similares. Aun así las clasificaciones no supervisadas son valiosas ya que permiten explorar y describir de manera preliminar algunas características de la base de datos espacio-temporal. Las clases en el mapa de la Figura 3b son: Salinas y Cuerpos de Agua (grupo 5), los Pastizales Naturales al centro del área (grupo 2), y la zona agrícola (grupos 1 y 6). El resto del área está comprendida por las categorías 3-4-7. Esta zona presenta más parches, indicando

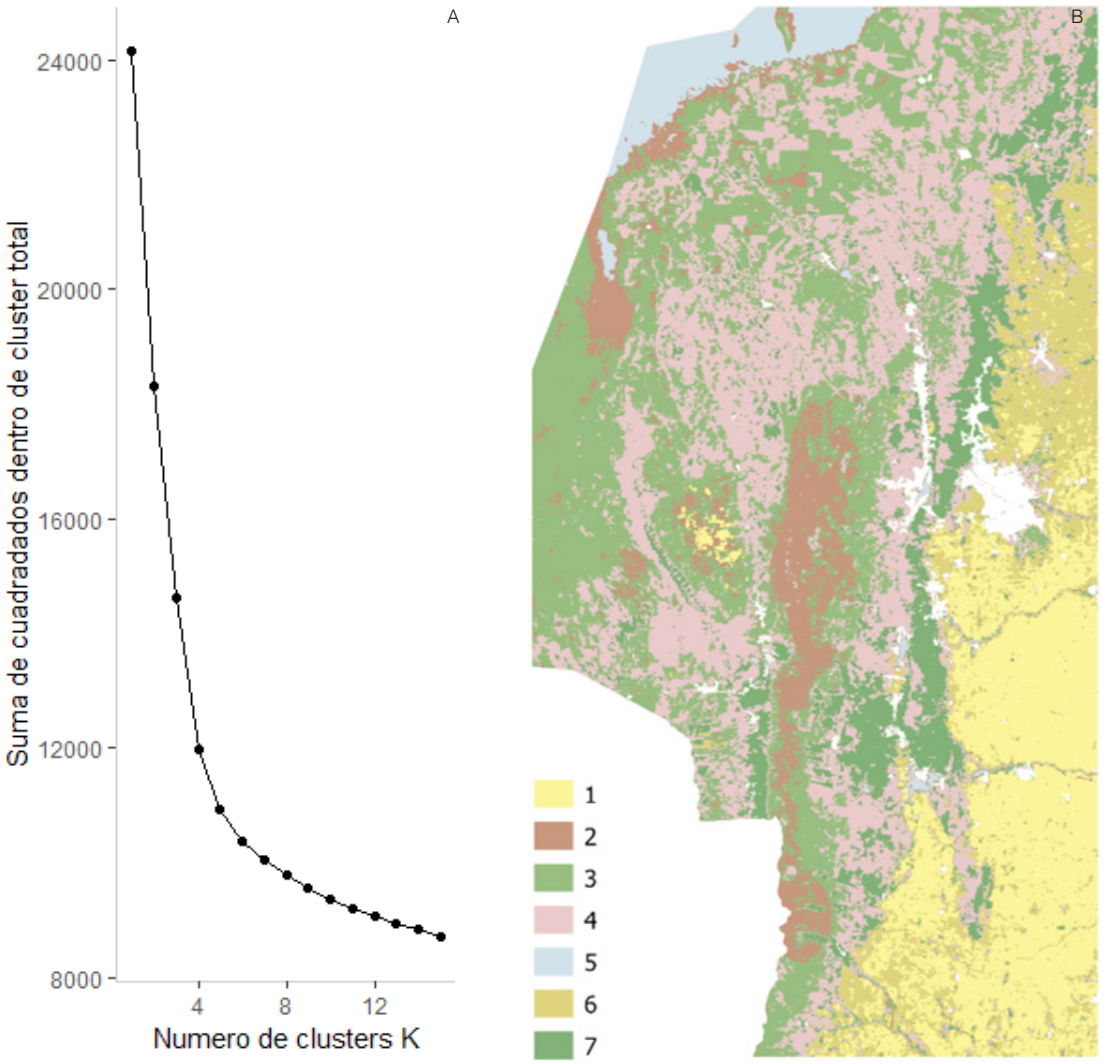

Figura 3. A. Gráfico del criterio del codo para la selección del número óptimo de grupos obtenido sobre una muestra aleatoria de series de tiempo. El gráfico sugiere entre 5 y 8 grupos; conociendo la variabilidad del área se eligieron 7 grupos. **B**. Mapa obtenido para la clasificación con el algoritmo k-*means*de las series de tiempo de NDVI depuradas usando 7 grupos.

que es donde ha habido historias muy diferentes entre unidades contiguas del paisaje en los últimos 18 años. El grupo 7 pertenece al Chaco Serrano, pero está solo representado mayoritariamente en las laderas orientales de las Sierras Chicas, donde las limitantes climáticas respecto a los mismos bosques del resto de la región son menores. En la Figura 4 se muestran en negro las series de tiempo medianas de todos los grupos. El área gris alrededor de cada serie de tiempo mediana representa el intervalo intercuartílico conteniendo el 50 % de los datos registrados para cada grupo. Las STIS de NDVI son marcadamente estacionales. Como se espera dadas las características climáticas y de vegetación en cada área, se observa mayor media de las series del grupo 7 que corresponde a los bosques en las laderas orientales de las Sierras Chicas, y similar patrón de variabilidad estacional para todas las clases excepto para el grupo 5.

Los resultados muestran que el marco de trabajo *dividir-aplicar-combinar* permitió implementar en un entorno de escritorio la gestión de una base de datos de STIS de series de tiempo de NDVI de 6 GB en R usando los paquetes *raster*, *bigmemory*, *foreach* y *doParallel*. Es posible implementar los procedimientos propuestos en computadoras personales, para áreas y períodos de tiempo más o menos extensos: el ejemplo de

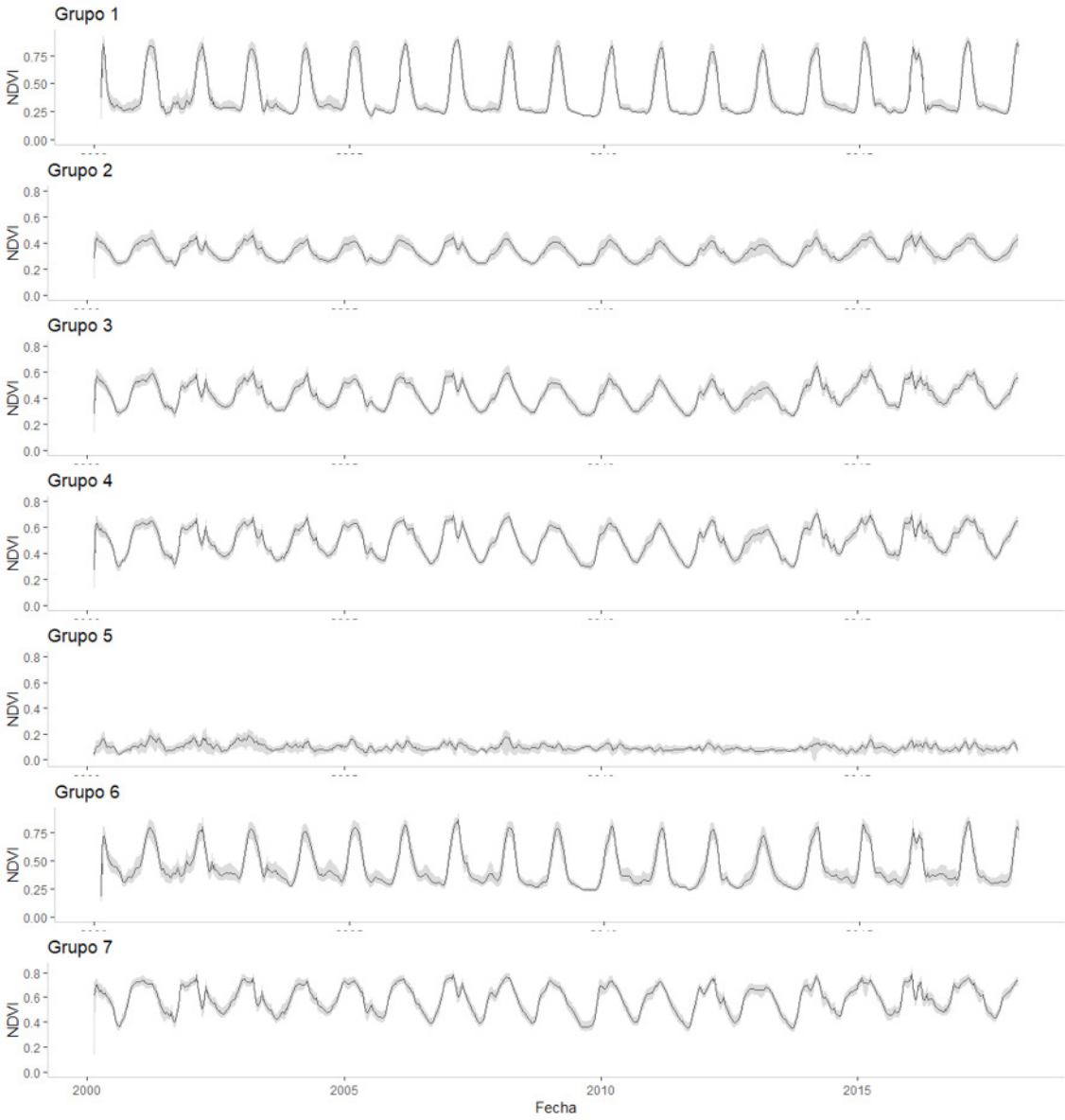

Figura 4. Series de tiempo para los grupos obtenidos clasificando las series de tiempo depuradas. Las líneas negras representan la serie de tiempo mediana para cada grupo, en tanto que las áreas sombreadas representan el 50 % de las series de tiempo restantes comprendidas entre los cuartiles 0,25 y 0,75. Se observa similar patrón de variabilidad estacional para 5 de los 7 grupos.

ilustración completo el procesamiento completo demoró aproximadamente 15 horas, en una computadora con procesador Intel Core i5, 8va generación y 16 GB de memoria RAM. El marco de trabajo presentado puede aplicarse de manera directa en servidores de mayor envergadura y en distintos sistemas operativos para bases de datos sustancialmente más grandes.

CONCLUSION

Este trabajo presenta un protocolo de gestión de datos espacio-temporales de imágenes satelitales y detalla su implementación en el software R y entornos de escritorio. El concepto *dividir-aplicar-combinar* resultó adecuado para su implementación en el contexto de largas series de tiempo de NDVI

distribuidas territorialmente a escala regional. La integración de entornos estadísticos como R y de entornos SIG como QGIS amplía las posibilidades de gestión y procesamiento de este tipo de bases de datos.

BIBLIOGRAFÍA

Adler, D., Gläser, C., Nenadic, J., Oehlschlägel, J. y Zucchini, W. (2018). ff: Memory-efficient Storage of Large Data on Disk and Fast Access Functions (versión 2.2-14) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Bivand, R., Keitt, T. y Rowlingson, B. (2014). rgdal: Bindings for the Geospatial Data Abstraction Library(versión 0.9-1) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Busetto, L. y Ranghetti, L. (2016). MODIStsp: An R package for automatic preprocessing of MODIS Land Products time series. *Computers & Geosciences*, 97, 40-48. https://doi.org/http://dx.doi.org/10.1016/j.cageo.2016.08.020

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S. y Schaub, T. (2016). The GeoJSON Format. En *Internet Engineering Task Force (IETF)*. https://doi.org/10.17487/RFC7946

Calaway, R., Microsoft Corporation, Weston, S. y Tenenbaum, D. (2018). Foreach Parallel Adaptor for the «parallel» Package (versión 1.4.7) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Cao, R., Chen, Y., Shen, M., Chen, J., Zhou, J., Wang, C. y Yang, W. (2018). A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter. *Remote Sensing of Environment*, 217, 244-257. https://doi.org/10.1016/J.RSE.2018.08.022

Capitanelli, R. G. (1979). Geomorfología. En R. A. Miatello, M. E. Roque y J. B. Vázquez (Eds.), *Geografía física de la provincia de Córdoba* (213-296). Córdoba, Argentina: Boldt.

Cochran, W. G. (1977). *Sampling Techniques(3ªed.)*. Nueva York, Estados Unidos: Wiley.

Congalton, R. G. (2015). Assessing positional and thematic accuracies of maps generated from remotely sensed data. En Prasad Thenkabail (Ed.), *Remote Sensing Handbook-Three Volume Set (pp. 617-636)*. CRC Press. https://doi.org/10.1201/b19294

Developers, S. (2013). signal: Signal processing (versión 0.7-6) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Folk, M., Heber, G., Koziol, Q., Pourmal, E. y Robinson, D. (2011). An overview of the HDF5 technology suite and its applications. *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases - AD '11*. https://doi.org/10.1145/1966895.1966900

García, C., Piumetto, M., Teich, I., Morales, H., Kindgard, A., Fuentes, M.… Ravelo, A. (2018). *Mapas de cobertura del suelo de la provincia de Córdoba 2017/2018 - niveles 1 a 3*. Recuperado de https://idecor-ws.cba.gov.ar/geoserver/idecor/wms

GDAL/OGR Contributors (2018). GDAL - Geospatial Data Abstraction Library (versión 2.1) [Software de cómputo]. Recuperado de https://gdal.org

Ghetta, M. y QGIS Development Team. (2018). R Processing Scripts Module. QGIS Geographic Information System (versión 1) [Software de cómputo]. Recuperado de http://www.qgis.org/

Gorgas, J. A. y Tassile J. L. (2006). Recursos Naturales de la Provincia de Córdoba. Los Suelos. Nivel de Reconocimiento 1:500000. Córdoba, Argentina: Agencia Córdoba Ambiente S.E. e Instituto Nacional de Tecnología Agropecuaria Estación Experimental Agropecuaria Manfredi.

GRASS Development Team. (2018). GRASS (versión 7.6) [Software de cómputo]. Recuperado de grass.osgeo.org/

Hargrove, W. W., Hoffman, F. M. y Hessburg, P. F. (2006). Mapcurves: a quantitative method for comparing categorical maps. *Journal of Geographical Systems*, 8 (2), 187. https://doi.org/10.1007/s10109-006-0025-x

Hartigan, J. A. y Wong, M. A. (1979). Algorithm AS 136: a k-means clustering algorithm. En *Applied Statistics*. https://doi.org/10.2307/2346830

Henry, L. y Wickham, H. (2018). purrr: Functional programming tools (versión 0.2-2) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Hijmans, R. J. (2018). Geographic Data Analysis and Modeling (versión 2.8-19) [Software de cómputo]. Viena, Ausitra: The Comprehensive R Archive Network: Contributed Packages.

IDERA.(s. f.). Recuperado de https://idera.gob.ar

IGN (2013). *Coberturas del Suelo* [Archivo de datos]. Recuperado de https://ign.gob.ar

IGN. (2018). *MDE-Ar: Modelo Digital de Elevaciones de la Argentina* [Archivo de datos]. Recuperado de https://ign.gob.ar

Instituto Nacional de Tecnología Agropecuaria (INTA) (2018). *Suelos de la República Argentina 1:500000 y 1:1000000* [Archivo de datos]. Recuperado de http://geointa.inta.gov.ar/geoserver/web/

Jarsún, B. (1981). Génesis, taxonomía y aptitud de los suelos. Departamentos Marcos Juárez y Unión (Provincia de Córdoba). Disertación doctoral no publicada, Universidad Nacional de Córdoba, Córdoba, Argentina.

Kane, M. J., Emerson, J. W., Haverty, P. J. y Determan, C. (2018). bigmemory: Manage Massive Matrices with Shared Memory and Memory-Mapped Files(versión 4.5.33) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Kane, M. J., Emerson, J. y Weston, S. (2013). Scalable Strategies for Computing with Massive Data. *Journal of Statistical Software55(14), 1-19*. doi: 10.18637/jss.v055.i14

Kim, S.-R., Prasad, A. K., El-Askary, H., Lee, W.-K., Kwak, D.-A., Lee, S.-H. y Kafatos, M. (2014). Application of the Savitzky-Golay Filter to Land Cover Classification Using Temporal MODIS Vegetation Indices. *Photogrammetric Engineering y Remote Sensing, 80(7), 675-685*. doi: 10.14358/PERS.80.7.675

LP DAAC - Land Processes Distributed Active Archive Center (s. f.). *MODIS Products Table*.Recuperado de lpdaac.usgs.gov

Luti, R., de Solís, M. B., Galera, F., de Ferreira, N. M., Berzal, M., Nores, M. y Roque, M. (1979). Vegetación. En J. Vázquez, R. Miatello, y M. Roque (Eds.), *Geografía Física de la Provincia de Córdoba (pp. 268-297)*. Buenos Aires: Boldt.

MAyDS (2018). Sistemas de Usos de Tierras en Argentina [Archivo de datos]. Recuperado de http://geo2.ambiente.gob.ar/geoserver/wfs

Microsoft y Weston, S. (2017). foreach: Provides Foreach Looping Construct for R(versión 1.4.7) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Morello, J., Matteucci, S. D., Rodríguez, A. F. y Silva, M. E. (2012). Ecorregiones GEPAMA. Ecorregiones y complejos ecosistémicos argentinos [Archivo de datos]. Recuperado de http://geo2.ambiente.gob.ar/geoserver/wms?SERVICE=WMS&

Muenchow, J., Schratz, P. y Brenning, A. (2017). RQGIS: Integrating R with QGIS. *The R Journal, 9(2), 409-428*. Recuperado de https://cran.r-project.org/package=RQGIS

Naimi, B. (2018). rts: Raster Time Series Analysis(versión 1.0-47) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

National Aeronautics and Space Administration (NASA) (s. f.). Recuperado de https://search.earthdata.nasa.gov

Nowosad, J. y Stepinski, T. F. (2018). Spatial association between regionalizations using the information-theoretical V-measure. *International Journal of Geographical Information Science, 32(12), 2386-2401*. doi: 10.1080/13658816.2018.1511794

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E. y Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment, 148, 42-57*. doi 10.1016/j.rse.2014.02.015

OSGEO Development Team (2018). PostgreSQL/PostGIS (versión 3.0) [Software de Cómputo]. Recuperado de https://postgis.net/

Oyarzabal, M., Clavijo, J., Oakley, L., Biganzoli, F., Tognetti, P., Barberis, I. y Len, R. J. C. (2018). *Unidades de vegetación de la Argentina. Ecología Austral, 28(1).* doi: 10.25260/EA.18.28.1.0.399

Paparrizos, J. y Gravano, L. (2015). E55--k-shape: efficient and accurate clustering of time series. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 1855-1870*. doi: 10.1145/2723372.2737793

Puchet, A. y IGN. (2017). Límites políticos administrativos [Archivo de datos]. Recuperado de https://ign.gob.ar

Puchet, A. e Instituto Geográfico Nacional (IGN) (2018). Aguas continentales [Archivo de datos]. Recuperado de https://ign.gob.ar

QGIS Development Team (2018). QGIS (versión 2.6) [Software de cómputo]. Zúrich, Suiza: Qgis. Recuperado de https://www.qgis.org

R Core Team (2018). R: A language and environment for statistical computing (versión 3.6.1) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages. Recuperado de https://cran.r-project.org/

Rosenberg, A. y Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. *Proceedings of the 2007 Joint Conference on Emprical Methods in Natural Language Processing and Computational Natural Language Learning*. doi: 10.7916/D80V8N84

Ryan, J. A. (2018). mmap: Map Pages of Memory (versión 0.6-17) [Software de cómputo]. Viena, Austria: The Comprehensive R Archive Network: Contributed Packages.

Savitzky, A. y Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8), 1627-1639. https://doi.org/10.1021/ac60214a047

Solano, R., Didan, K., Jacobson, A. y Huete, A. (2015). MODIS Vegetation Index User's Guide (MOD13 Series). Vegetation Index and Phenology Lab, The University of Arizona. doi: 10.5067/MODIS/MOD13Q1.006

Vermote, J. C. y Ray, J. P. (2015). MODIS Surface Reflectance User's Guide collection 6. MODIS Land

Surface Reflectance Science Computing Facility. doi: 10.5067/MODIS/MOD09A1.006.

Wan, Z. (2013). MODIS Land Surface Temperature Products User's Guide. doi: 10.5067/MODIS/MOD11B3.006

Wan, Z., Hook, S. y Hulley, G. (2015). MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. [Archivo de datos]. doi: 10.5067/MODIS/MOD11A2.006

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1), 1-29. doi: 10.18637/jss.v040.i01

Wickham, H. (2015). Advanced R. doi: 10.1007/978-1-4842-2077-1