Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires

Facultad de Ciencias Económicas

Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Clasificación de la productividad de áreas sembradas mediante K -Medias -Un enfoque a partir del NDVI

AUTOR: PATRICIA GIRIMONTE

DICIEMBRE 2019

Resumen

La actividad agrícola es una de las actividades económicas más importantes de nuestro país, por lo cual es, y ha sido de importancia el desarrollo de conocimiento que permita diseñar estrategias para maximizar la productividad de los cultivos, reduciendo los costos y el impacto ambiental.

Una herramienta utilizada con este objeto es la teledetección satelital la cual ha sido ampliamente estudiada en distintas partes del mundo. Su aplicación comienza en la década del 70 y se ha mantenido en constante crecimiento. En los últimos años el avance tecnológico en los sensores satelitales ha permitido obtener información más precisa, y las mejoras computacionales tanto a nivel de hardware como de software permite hoy analizar el gran volumen de datos que la teledetección siempre ha generado, desde una perspectiva diferente.

Con la información obtenida de las imágenes satelitales, se han definido diferentes índices relacionados con la productividad de los cultivos, siendo uno de los más utilizados el NDVI (Índice de Vegetación de Diferencia Normalizada). Este índice es derivado de información captada por sensores remotos y está asociado a la fracción de la radiación solar absorbida por las plantas.

Uno de las aplicaciones del Big Data en las ciencias agronómicas es la que utiliza la información obtenida mediante teledetección satelital, la cual a partir de la política de Datos Abiertos iniciadas por la NASA (National Aeronautics and Space Administration) y la ESA (European Spatial Aeronautics) en los años 2013 y 2017 respectivamente, está disponible más allá de las áreas gubernamentales, representando una oportunidad para generar conocimiento, tanto en el ámbito académico, como en el sector público y privado.

En este trabajo se seleccionará una zona de la región pampean, dada la densidad de las áreas sembradas en la región, dentro de las imágenes satelitales disponibles del satélite Landsat 8 de la NASA. A partir de las imágenes obtenidas se calculará el NDVI para cada píxel, con el objeto de clasificar el área seleccionada mediante el método de aprendizaje automático de clasificación no supervisado de K-medias.

Palabras clave: NDVI, Big Data, teledetección, K-medias, productividad

Estructura

Introducción

La actividad agropecuaria, de las que forman parte las producciones agrícolas tiene un rol preponderante en el engranaje de la economía argentina, siendo una actividad competitiva, generadora de empleo tanto en el ámbito estatal como privado, aportando al PBI y con una importante contribución tributaria.

Con los avances tecnológicos y la capacidad del productor de adaptarse e innovar, la producción agraria ha tenido un crecimiento sostenido, tanto por la expansión de la frontera agrícola como de la mejora en los rendimientos por hectárea. El censo agropecuario 2018 realizado por INDEC (Instituto nacional de Estadísticas y Censos) ha relevado que un 21 % de las hectáreas de nuestro país corresponden a superficie implantada con algún cultivo, correspondiendo estas mayoritariamente a la región pampeana. (INDEC,2019).

En las últimas dos décadas ha habido una evolución de las principales producciones agrícolas dándose los principales crecimientos en la producción de soja, maíz y trigo, Si bien la soja ha tenido una caída en los últimos dos años, continúa siendo la que mayor rentabilidad representó para el sector (Bosa de Cereales, 2019), y la de mayor contribución tributaria. (FADA, 2019).

En la región pampeana norte, uno de los departamentos de mayor producción de soja, es Rio Cuarto en la provincia de Córdoba, la cual ha representado el 27% del total del país en la campaña 2018/19 (Bolsa de Cereales de Córdoba, 2019).

La incorporación de la información que brinda la teledetección satelital en las ciencias agronómicas tiene su origen en la década del 70, donde también se definieron los primeros índices relacionados con la productividad. Uno de los más utilizados es el Índice de Vegetación de Diferencia Normalizada (NDVI) también llamado Índice Verde (Rouse, et al, 1973). Este índice es derivado de información captada por sensores remotos, y está asociado a la fracción de la radiación solar absorbida por las plantas. Por este motivo, existe una fuerte relación del índice verde con algunas características de la vegetación como por ejemplo la biomasa, el índice de área foliar o la productividad, entre otras. El conocimiento y seguimiento temporal de la cobertura vegetal, la información de pronósticos evolutivos de la vegetación, los pronósticos meteorológicos y la zonificación de eventos destacables

resultan fundamentales en el proceso de toma de decisiones para el manejo adecuado y sustentable de los agroecosistemas y los recursos naturales.

En los últimos años los avances tecnológicos en los sensores satelitales han permitido obtener información más precisa, y las mejoras computacionales tanto a nivel de hardware como de software permite hoy realizar el análisis del gran volumen de datos que la teledetección siempre ha generado, desde una perspectiva diferente.

Por otro lado, la política de Datos Abiertos iniciadas en el año 2013 por la NASA y en el año 2017 por la ESA, permite la disponibilidad de esta información no solo a organismos gubernamentales o militares. Esta apertura ha posibilitado a los distintos países poner a disposición de sus ciudadanos información georreferenciada relacionada con otros indicadores económicos. En nuestro país algunos de los organismos públicos que utilizan información satelital aplicada a la actividad agropecuaria para obtener tanto un mapeo de la cobertura del suelo como índices de productividad, entre otros usos, son el INTA (Instituto nacional de tecnología Agropecuaria), la Bolsa de Cereales de Buenos Aires, de Córdoba, de Santa Fe, y el Ministerio de Agricultura, Ganadería y Pesca.

El último censo agropecuario llevado a cabo por INDEC fue realizado mediante dispositivos de captura móvil, lo que permitió incorporar por primera vez un módulo geográfico, que ayudó al censista tanto a ubicarse en campo como a mejorar la visual de la cobertura del suelo (INDEC, 2019).

Las imágenes satelitales más usadas por los organismos gubernamentales son las del Landsat 8 de la NASA, Sentinel 2 (Satelite) de la ESA, y complementariamente el SPOT (Satélite Para la Observación de la Tierra) desarrollado por el CNES (Centro Nacional de Estudios Espaciales francés) en colaboración con Bélgica y Suecia. Las imágenes del SPOT pueden encontrarse o ser requeridas en nuestro país en el portal de la CONAE.

Dentro del ámbito académico la disponibilidad de esta información permite plantear diferentes líneas de investigación, entre las que podrían mencionarse las relacionadas con la productividad en las actividades agropecuarias, la tributación agropecuaria en base a la productividad, la sustentabilidad del suelo, contaminación ambiental, entre otras.

Dentro del ámbito privado existen también diferentes propuestas que le permiten al productor agrónomo gestionar y monitorear su campo de manera remota a través de

imágenes satelitales provistas por la NASA, ESA, CONAE y en algunas iniciativas más costosas por drones. Algunas de estas propuestas digitalizan los diferentes puntos del campo, identifican y realizan un mapeo de la variabilidad del terreno, para así poder elegir la mejor estrategia para maximizar el rinde, reducir los costos y el impacto ambiental. Estas aplicaciones son utilizadas también por las compañías aseguradoras de la actividad agropecuaria.

En este contexto el objetivo del presente trabajo es clasificar mediante el método no supervisado de K-medias un área sembrada a partir del NDVI obtenido mediante teledetección satelital de la región de Rio Cuarto, provincia de Córdoba en Argentina.
El enfoque del estudio es cuantitativo, y exploratorio descriptivo, correspondiendo a un diseño transversal. Para cumplimentar el objetivo planteado se seleccionó a partir de las imágenes satelitales disponibles del Landsat 8 en el catálogo de la Comisión Nacional de Actividades Espaciales (CONAE) y de la United States Geological Survey (USGS) del período enero-marzo de 2019 una imagen que contuviese la zona de interés. Estas imágenes satelitales son de acceso libre, previo registro y aceptación de los términos y condiciones de uso, en las páginas de la CONAE y USGS.
Se consideró este período por ser la época de mayor vigorosidad de la planta de soja, dado que, en nuestro país, y en general en el hemisferio sur la siembra oscila entre el mes de septiembre y el mes de diciembre, siendo la cosecha entre los meses de marzo y mayo.
Para cumplimentar con el objetivo planteado se obtuvo una imagen satelital que contuviese la región de Río Cuarto, a partir de esta imagen obtenida se seleccionó un área del departamento de la que se contaba con información de la cobertura del suelo (IDECOR, 2019), se calculó el NDVI para cada pixel del área seleccionada y luego se realizó la clasificación mediante K-medias (Hastie, T 2001).

1. El manejo de grandes volúmenes de datos en las ciencias agronómicas obtenidos mediante teledetección a partir de la política de datos abiertos de las agencias espaciales

La teledetección satelital ha producido siempre un gran volumen de datos, pero es a partir de las políticas de datos abiertos llevadas a cabo por la NASA y la ESA que la posibilidad de contar con esta información se abre a todos los organismos públicos, privados y a los

ciudadanos en general. Por otro lado, los avances a nivel de hardware y software, las nuevas formas de almacenar y procesar los datos hacen posible el procesamiento y análisis de esta información. La combinación de estos factores representa una oportunidad y un desafío en el ámbito de las ciencias agronómicas tanto para obtener, a partir de los índices relacionados con la productividad, un mapeo y mejor conocimiento de los suelos cultivados, como para predecir su productividad.

En este apartado se parte de la definición de Big Data y como ha impactado en el desarrollo de las ciencias agronómicas, particularmente a partir de las políticas de datos abiertos de las principales agencias espaciales.

Luego se describe la aplicación de la teledetección en agronomía, en particular el uso del NDVI como un indicador de la productividad.

Finalmente se hace referencia a la oportunidad que ha significado la política de Datos Abiertos llevada a cabo por la NASA y la ESA para mejorar la toma de decisiones relacionadas con la productividad en agronomía, y se ejemplifica con casos tanto del ámbito público como privado en nuestro país.

1.1 Definición de Big Data y su aplicación en las ciencias agronómicas

Existen diferentes usos y definiciones para el término Big Data, e infinidad de publicaciones que intentan su descripción desde distintos puntos de vista. Diferentes estudios mencionan que el término fue utilizado por primera vez en una publicación de investigadores de la NASA en 1997 (Cox & Ellsworth, 1997), en el resumen los autores mencionan “Los objetos de Big Data son solo eso, objetos de datos individuales (o conjuntos) que son demasiado grandes para ser procesados por algoritmos estándar y el software en el hardware que se tiene disponible”.

En el año 2001, Lanyn, en una publicación para Meta Group en referencia a la “cantidad de

datos” que empezaban a surgir en las transacciones de los e-commerce, utililiza las 3 Vs, del Big Data “Volumen, Velocidad y Variedad” (Laney, 2001). Pero es en el año 2012 donde la consultora Gartner (que había adquirido Meta Group) publica en su estudio anual Hype Cycle for Big Data (Gartner, 2012), la popularizada definición de Big Data basada en las tres Vs, “"un gran volumen, velocidad o variedad de información que demanda formas

costeables e innovadoras de procesamiento de información que permitan ideas extendidas, toma de decisiones y automatización del proceso".

A estas primeras 3 Vs del Big Data, algunos autores han agregado otras dos, veracidad y valor. Podemos resumirlas, citando a distintos autores, como:

Volumen: a partir del año 2012 el crecimiento de los datos es exponencial esperando para el año 2020 aproximadamente 7 zettabytes solo en USA (Gantz, 2012). En cada instante se están generando distintos tipos de datos, estructurados y no estructurados a partir de una mayor pluralidad de fuentes, incluidos los clics en Internet, las transacciones móviles, el contenido generado por el usuario y las redes sociales, así como el contenido generado intencionalmente a través de redes de sensores o transacciones comerciales como consultas de ventas y transacciones de compra (Lehrer, Wieneke, & Otros, 2018). A partir de la teledetección satelital se genera un gran volumen de datos, si bien en la actualidad no podemos hablar de "gran volumen" a partir solo de una imagen determinada de una región, pero si de un ensamble de imágenes, y de una serie de ellas en el tiempo.

Velocidad: Para muchas aplicaciones, la velocidad de creación de datos es aún más importante que el volumen. La información en tiempo real o casi en tiempo real hace posible que una empresa sea mucho más ágil que sus competidores (Brynjolfsson & McAfee, 2012).

Variedad: los datos son de diferente tipo, estructurados y también no estructurados como texto, datos de sensores, audio, video, secuencias de clics, archivos de registro y más. (Muley, 2018)

Veracidad: las fuentes de datos deberían ser confiables, para quienes toman las decisiones en las empresas menciona Muley en su publicación del año 2018 (Muley, 2018) "La confianza juega un mayor rol en la ciencia de datos y, por lo tanto, la integridad de los datos es una dimensión central".

Valor: Muley también menciona en el mismo artículo que los datos se tratan como el "nuevo petróleo" dado que los conocimientos a los que se puede llegar a partir de ellos son muy valiosos, por eso algunas instituciones modernas ven a sus datos como un repositorio de activos. (Muley, 2018).

Big Data tiene aplicación en todas las áreas del conocimiento, y hay infinidad de investigaciones al respecto no solo en el ámbito académico sino también en los organismos públicos y en las empresas privadas.

En el caso particular de las ciencias agronómicas, según la publicación de la revista de la NewAg International (NewAg International, 2017), "la agricultura está enfrentando una revolución con la integración de herramientas y sistemas de decisiones potenciados por Big Data. El gran volumen de datos, entre otras aplicaciones, se está utilizando para aumentar la eficiencia y al mismo tiempo disminuir el impacto sobre el medioambiente. La capacidad computacional moderna ha permitido aumentar la capacidad de recolectar, intercambiar, procesar y sintetizar datos de una forma tal que está impactando en todo el ámbito agrícola, a nivel de maquinarias, optimización de semillas, fertilizantes e insumos, riego y gestión de de parcelas. Para poder obtener valor del Big Data, la información que brindan los datos debe ser procesada y analizada a tiempo y sus resultados deben estar disponibles para tomar decisiones en las operaciones agrícolas. Los productos basados en métodos de machine learning serán exponencialmente mejores en la medida que más usuarios se unan a su uso."

Uno de los usos del Big Data en agronomía es la que utiliza imágenes satelitales provistas principalmente por los satélites Landsat 8, Sentinel y SPOT. Estas imágenes satelitales contienen información que permite construir índices relacionados con la productividad de los cultivos, por lo cual resultan de utilidad, entre otros usos, para clasificar la cobertura del suelo, analizar sufrimiento hídrico, predecir productividad y detectar anomalías si las hubiera.

1.2 NDVI, un índice obtenido a partir de teledetección satelital asociado a la productividad agrícola

1.2.1 La teledetección satelital y su aplicación a las ciencias agronómicas

La teledetección satelital es la adquisición de radiación electromagnética a distancia a través de sensores localizados en plataformas móviles, sin que exista contacto material con el objeto observado, y la trasformación de los datos obtenidos mediante técnicas de interpretación y reconocimiento de superficies (Sobrino, 2000).

La teledetección en general, y en particular la satelital, se fundamenta en el hecho de que todo objeto o cubierta de suelo absorbe, transmite y refleja el flujo de luz que incide en él, de forma que la proporción reflejada dependerá de la naturaleza del objeto o de la cubierta iluminada; esta radiación reflejada y captada por un dispositivo sensible a su registro digital (sensor) será la respuesta espectral propia de dicho objeto o cubierta. Este es el principio por medio del cual, a partir de los colores, formas y texturas, nuestra visión opera para discriminar diferencias en la interpretación de fotografías en blanco y negro, y principalmente en color. El desafío ese encuentra en aquello que cambia pero que no podemos distinguir a simple vista. El ojo humano es incapaz de ver longitudes de onda por debajo del rojo, en el infrarrojo, y son justo esas longitudes en las que las partes verdes de las plantas vigorosas reflejan mejor la luz.

Una vez obtenida la información por intermedio de sensores montados en plataformas, la porción de energía electromagnética es digitalizada y convertida en imágenes. La teledetección comprende el tratamiento de esa información mediante técnicas desarrolladas para la obtención de productos que podrán ser analizadas de acuerdo a las distintas perspectivas de aplicación.

La clasificación de escenas satelitales obtenidas mediante teledetección consiste en transformar una imagen pancromática, multi o hiper espectral en una imagen compuesta por clases temáticas que luego serán asociadas a coberturas del suelo de interés de acuerdo al objetivo de estudio (CONAE, 2018).

La utilidad de la teledetección en las ciencias agronómicas ha sido ampliamente estudiada en distintas partes del mundo (Schomwandt, 2015). Su aplicación comienza en la década del 70 y se ha mantenido en constante crecimiento. Como ya se ha mencionado, los avances tecnológicos en los sensores satelitales que permiten obtener información más precisa, los avances a nivel de hardware y software, las nuevas formas de almacenar y procesar los datos que se generan, sumado a las políticas de datos abiertos llevadas a cabo por la NASA a partir del año 2013 y por la ESA en el año 2017, ha permitido que no solo las áreas gubernamentales o militares de los países tengan acceso a la información, sino también el público en general, lo que ha significado nuevas oportunidades para desarrollar conocimiento aplicado a las ciencias agronómicas tanto dentro del ámbito académico, como en el empresarial.

1.2.2 Definición de NDVI y su relación con la productividad

La utilización de datos espectrales para evaluar parámetros de vegetación se basa en la reflectancia diferencial de los tejidos fotosintéticos en la porción rojo e infrarrojo del espectro electromagnético (Rouse et .al., 1973).

La fenología de la vegetación se define como los ciclos recurrentes de las actividades biológicas estacionales y su relación con condiciones ambientales y meteorológicas como temperatura, luz, humedad y tipo de suelo. Los distintos tipos de vegetación responden de manera diferente a dichas condiciones. Debido a la absorción de la clorofila, las hojas verdes reflejan muy poca luz correspondiente al rojo, mientras que muestran una alta reflectancia en la zona del infrarrojo cercano.

El NDVI (Normalized Difference Vegetation Index) es un índice que deriva del cociente entre la reflectancia del rojo y el infrarrojo cercano: Ir – R/ Ir+R (1) donde Ir es la reflectancia correspondiente al infrarrojo cercano y R la reflectancia correspondiente al rojo del espectro electromagnético (Rouse et .al., 1973). Los valores del índice oscilan entre -1 y 1. Los valores negativos están relacionados con cuerpos de agua y superficies degradadas por acción del fuego, mientras que valores positivos más bajos (cercanos a 0) corresponden a vegetación senescente o de baja cobertura. Los valores positivos altos (cercanos a +1) representan alto contenido de biomasa fotosintética. Se han encontrado fuertes relaciones entre el NDVI y algunas características funcionales y estructurales de la vegetación como biomasa, índice de área foliar, cobertura y productividad primaria neta, es por esto que es uno de los índices más utilizados como indicador de la productividad. En el caso particular de las investigaciones relacionadas con la productividad de los cultivos, se basan principalmente en el NDVI como un indicador de la productividad (Deering 1978), existiendo numerosas publicaciones (Tucker et al., 1980; Quarmby et al., 1993;Doraiswamy y Cook, 1995; Boken y Shaykewich, 2002; Mkhabela et al., 2005; Moriondo et al., 2007), las más recientes han incorporado en su análisis métodos de machine learning (Mellor et.al ,2012; Joshil Raj K, et.al ,2015; Zheng et.al ,2015; Long, L.S., et.al ,2017; Kala, S. et.al ,2018).

1.3 Datos Abiertos para mejorar la toma de decisiones relacionadas con la productividad en agronomía

Los conceptos de Big Data y Datos Abiertos están estrechamente relacionados en el sentido que la "apertura" de datos ha generado un gran volumen de datos. Los datos abiertos no solo son generados por los organismos gubernamentales, sino también por empresas privadas que deciden abrir su información.

Para que un dato sea considerado abierto, tiene que tener disponibilidad y acceso, debe poder ser reutilizado pudiendo incluso ser cruzado con la información suministrada por otros datos, y tiene que garantizar la participación universal, sin discriminación con las áreas de actuación, personas o grupos.

En octubre de 2015, durante la cumbre de la Alianza para el Gobierno Abierto (AGA) que se realizó en México, fue presentada oficialmente la "Carta Internacional de Datos abiertos". Esta carta es una iniciativa multilateral y colaborativa, que ha sido apoyada por gobiernos, entre los que se encuentra nuestro país, organizaciones de la sociedad civil, sector privado y expertos en la materia (Open Data Charter). Esta postula que los "Datos abiertos son datos digitales que son puestos a disposición con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar.

Establece seis principios básicos y amplios, que pueden ser de gran utilidad, tanto para aquellos gobiernos que ya establecieron una política de datos abiertos como para aquellos que aún no han comenzado. Los 6 principios para que un dato sea considerado abierto son: 1) abiertos por defecto, 2) oportunos y exhaustivos, 3) accesibles y utilizables, 4) comparables e interopeorables, 5) con el objeto de mejorar la gobernanza y la participación ciudadana y 6) para el desarrollo incluyente y la innovación.

Nuestro país, firmante de este acuerdo promulgó en el año 2016 la Ley de Acceso a la Información Pública, que establece que la información en poder del Estado debe ser accesible para todas las personas y estar disponible en formatos electrónicos abiertos para facilitar su circulación y redistribución, y mediante el decreto 117/2016 el gobierno nacional impulsó el "Plan de Apertura de Datos". Con este fin puso a disposición de las ciudades y provincias una serie de pasos a seguir para que puedan comenzar una política de datos abiertos. (Presidencia de la Nación-Ministerio de Modernización, s.f.). En el mismo portal

se postula que “Datos Abiertos es una iniciativa global, ligada a las políticas de Gobierno Abierto. Se trata de un medio que posibilita un mejor conocimiento del funcionamiento del gobierno, el fortalecimiento del rendimiento de cuentas y la mejora de la vida en ciudadanía”.

Respecto de la apertura de datos de los organismos públicos hay que tener en cuenta que “No todos los datos públicos son o pueden ser abiertos.” Los datos públicos son aquellos que los organismos del Estado generan y/o administran para el cumplimiento de sus misiones y funciones. El acceso a la información pública es un derecho reconocido en el país por la Ley 27275/16, que contempla diferentes principios para fomentar la publicidad de toda la información poseída y generada por el sector público.
No todos los datos públicos pueden o deben publicarse (algunos están protegidos por legislación específica que lo prohíbe o regula) y un dato publicado no es abierto si no cumple con las condiciones de la sección anterior.
Existen normativas que protegen determinados datos públicos en casos particulares. Estas situaciones que en el marco de las políticas de acceso se toman como excepciones, al implementar una iniciativa de datos abiertos debemos contemplarlas.
Las normativas nacionales al respecto son, Ley de protección de datos personales 23.526/00, Ley del Sistema Estadístico 17.622/68, Ley de Procedimiento Fiscal 11.683 y Ley de Propiedad Intelectual 11.723/33.” (Presidencia de la Nación-Ministerio de Modernización, s.f.)

La apertura de los datos llevada a cabo por la NASA se encuadra dentro de las políticas de datos abiertos llevada a cabo por el gobierno de los Estados Unidos la cual en parte se encuentra plasmada en diferentes documentos publicados por la Casa Blanca (White house 2014).

En nuestro país en el portal de Datos Abiertos del Ministerio de Agricultura Ganadería y Pesca, del INTA, la Bolsa de cereales de Córdoba, la Bolsa de cereales de Buenos Aires, por citar algunos ejemplos, es posible obtener informes, series numéricas relacionada con la actividad agropecuaria e información georreferenciada.

Esta información disponible, en el caso particular de las ciencias agronómicas representan una oportunidad y un desafío de innovación. Como se mencionó anteriormente, contar con la información satelital de un área de interés en el tiempo, más la información histórica de otras fuentes, permite entre otras aplicaciones analizar la variabilidad del terreno, para así

poder elegir la mejor estrategia para maximizar el rinde, reducir los costos y el impacto ambiental

2. Metodología y datos utilizados para clasificar la cobertura del suelo a partir del NDVI

En este apartado se describen los pasos seguidos para cumplimentar los objetivos planteados, relativos a la obtención de los datos y su posterior procesamiento en el software libre R (R Core Team,2019)

En el apartado 2.1, se justifica la elección de la zona en el departamento de Río Cuarto en la provincia de Córdoba y el día seleccionado. Seguidamente en el apartado 2.2 se describen los pasos seguidos para la obtención de la imagen satelital dentro del catálogo de imágenes disponibles del Landsat 8 de la NASA que contuviese la zona y el período del año de interés, y finalmente en el apartado 2.3 se describe el procesamiento de la imagen obtenida y posterior cálculo del NDVI.

2.1 Determinación de la zona y período de análisis

La soja es el principal cultivo de Argentina no sólo por la producción, sino por la superficie ocupada. Nuestro país es el tercer productor de soja después de Estados Unidos y Brasil, siendo las provincias de Buenos Aires, Córdoba y Santa Fe las de mayor producción interna.

La producción de soja en la provincia de Córdoba en la campaña 2018/19 alcanzó las 14.970.100 toneladas, representando el 27% del total del país, que de acuerdo con el Ministerio de Agricultura, Ganadería y Pesca se produjeron más de 55 millones de toneladas. Esto sucedió principalmente por el importante aumento ocurrido en el rendimiento, ya que la superficie se mantuvo prácticamente igual al año pasado (solo un 1% más). El rendimiento promedio ponderado fue de 37,6 qq/ha, siendo de 5 quintales superior a la media nacional (Bolsa de Cereales de Córdoba, 2019).

La provincia de Córdoba provee información detallada de la cobertura de su suelo (IDECOR, 2019) con mapas de coberturas obtenidos mediante trabajo en campo y teledetección satelital.

A partir de esta información de la cobertura de suelo, y la disponible en la Bolsa de Cereales de la misma provincia pudo determinarse un área con siembra mayoritariamente de soja 18

km al norte aproximadamente de Adela María en el departamento de Río Cuarto. ((Bolsa de Cereales de Córdoba, 2019) .

Para determinar el período, se tuvo en cuenta que en Argentina el cultivo de la soja en secano comienza entre los meses de septiembre y diciembre (siembra tardía, generalmente para la soja de segunda, luego del trigo), siendo entonces el período febrero-marzo el de mayor vigorosidad de la planta.

2.2 Obtención de la imagen satelital

Existen distintos satélites que permiten obtener imágenes terrestres, entre los más utilizados en la actualidad para los análisis de cobertura digital se encuentra el Landsat 8, con dos sensores, uno de ellos, Oli (Operational Land Imager) permite obtener mediciones en el infrarrojo cercano (CONAE, 2016). El Landsat 8 es un satélite de observación terrestre estadounidense lanzado el 11 de febrero de 2013. Es el octavo y último a la fecha del proyecto Landsat operado por la NASA y el Servicio Geológico de los Estados Unidos (United States Geological Survey, USGS) desde 1972. (NASA, 2015). El Landsat 8 tiene una revista cada 16 días, una resolución de 30m x30 m, y barre un área de aproximadamente 190 km de ancho y 180 km de alto (USGS, 2015)

Para la obtención de la imagen satelital, se buscó en el catálogo de la CONAE en forma interactiva una imagen del Landsat 8 del período enero-marzo de 2019 que contuviese al departamento de Río Cuarto. Para acceder a este catálogo es necesario un registro previo y aceptación de los términos y condiciones.

A partir de esta búsqueda interactiva se pudo determinar que una imagen con menos del 20% de nubes (para poder asegurar visibilidad de la zona) para el período y región de interés correspondía al día 12 de marzo de 2019.

Para acceder a las 11 bandas del satélite Landsat 8, una de las formas es a partir de la página de United States Geological Survey (USGS, s.f.), previo registro y aceptación de los términos y condiciones de uso. Luego de realizado el registro, se puede seleccionar la imagen de interés poniendo directamente el nombre de la región, en este caso Río Cuarto (es posible hacerlo en español), luego se selecciona el rango de fechas de interés, el satélite, en este caso se seleccionaron las imágenes del Landsat 8 con sus dos sensores Oli (Operational Land Imager) y Tiers (Thermal infrared sensor) y se pidió a lo sumo un 20% de nubes.

Dentro de las opciones disponibles están las imágenes “LandsatLook”, estos son archivos optimizados para la interpretación visual y no se recomienda su utilización para procesamiento; mientras que “Level 1 GeoTiff data Product” contiene imágenes georreferenciadas y corregidas en el terreno, estas son las adecuadas para un posterior procesamiento y análisis (CONAE, 2018). Antes de descargar la imagen, que comprimida en formato “tar” tiene un tamaño aproximado de 1.65 GB, puede visualizarse para corroborar el porcentaje de nubes y que corresponda a la región buscada,

Una vez descargada la imagen al descomprimirla se obtienen 11 archivos, 9 de formato TIFF (Tagged Image File Format), y 2 correspondientes a los metadatos que contiene información de fecha, hora, latitud y longitud de la zona. El nombre del archivo comprimido y luego la primera parte del nombre de cada una de las bandas corresponde a la plataforma, el sensor, path-row, y el año y día juliano, de adquisición. Descomprimidas, cada una de las bandas tienen un tamaño aproximado de 113 MB salvo la banda 8 que tiene un tamaño de 453 MB. El formato Tiff permite el procesamiento de las imágenes satelitales en diferentes softwares específicos y otros no específicos como Python y R.

2.3 Procesamiento de la imagen satelital y cálculo del NDVI

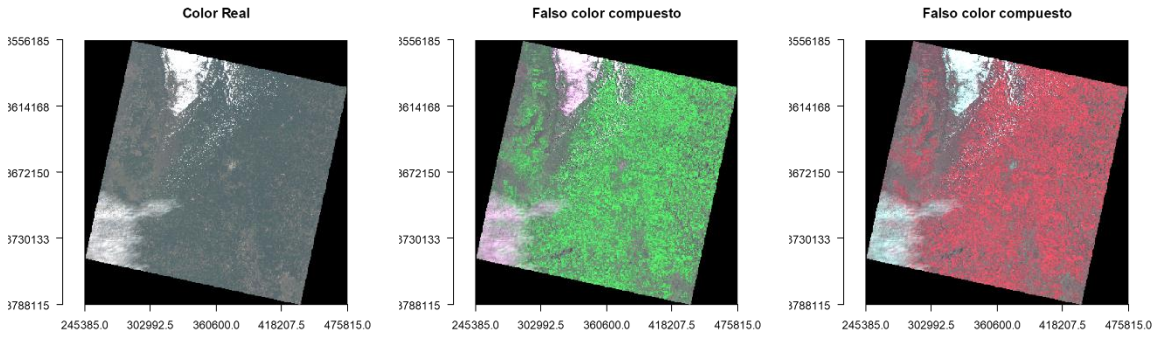
Una vez obtenidas las 11 bandas de la imagen se realizó un primer procesamiento en SOPI (Software de procesamiento de imágenes) desarrollado por la CONAE.
Este primer análisis descriptivo permitió observar la región de interés, y decidir el apilado de bandas adecuado para visualizar la imagen de acuerdo a nuestro interés de clasificar la cobertura del suelo.
La interpretación visual es una parte fundamental de la teledetección, dado que nos permite observar características biofísicas y fenológicas de los diversos usos y coberturas del suelo. La reflectancia como se mencionó anteriormente es la energía que refleja una cobertura del total de energía que recibe del Sol, toma valores entre 0 y 1, pero suele expresarse también como porcentaje, por ejemplo, una reflectancia del 20% para un cultivo, significa que refleja el 20% de la energía solar incidente. La reflectancia espectral es la medida para una longitud de onda determinada, nos permite definir la respuesta espectral de una cobertura. En una imagen, cuanto más alto (e intenso) es el valor del píxel de una determinada cobertura, mayor reflectancia presentará en determinada zona del espectro (CONAE, 2018).

A partir de la interpretación de las firmas espectrales y de la disponibilidad de bandas de las misiones satelitales, es posible observar la respuesta espectral de los usos y coberturas y relacionarlas con variables biofísicas. Existen diferentes combinaciones de bandas y su utilización dependerá del fenómeno en estudio. Por ejemplo, para visualizar vegetación se utiliza una combinación de bandas que incluyan la del rojo e infrarrojo cercano.

El posterior procesamiento de las imágenes se realizó en R 3.6.1 con las librerías raster (Hijmans, 2019), dplyr (Wickham et.al. ,2019), sp (Pebesma, E.J., R.S. Bivand, 2005) y rgdal (Bivand, et.al, 2019), con un script realizado para tal fin (Hijmans R., 2019) (Ver Apéndice).

Para visualizar las imágenes se realizaron 3 apilados de bandas distintos (**Figura 1)**.

Figura 1



Bandas 432 Bandas 453 Bandas 543

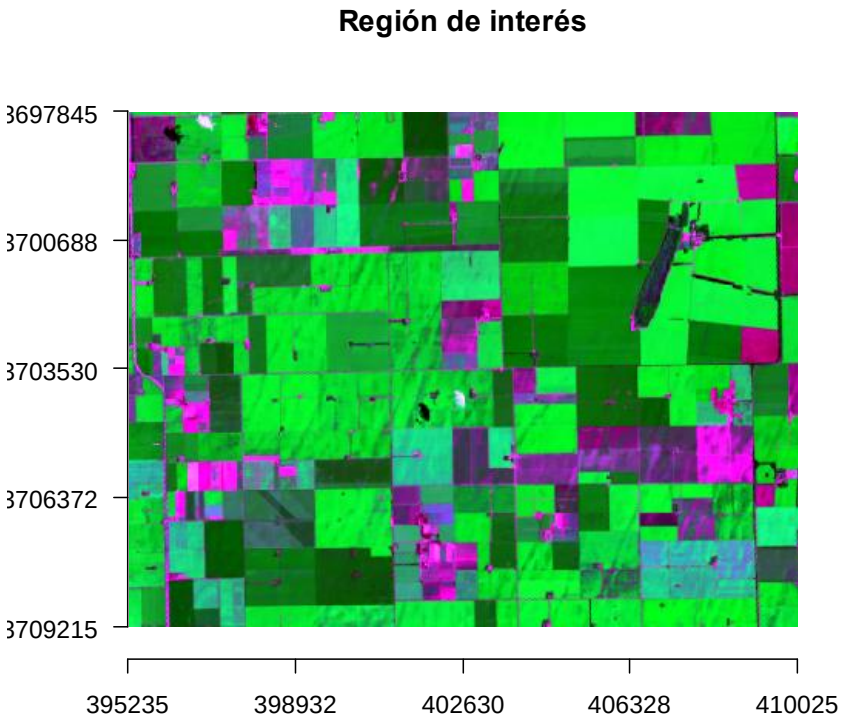
Como nuestro objeto es clasificar la cobertura del suelo a partir de NDVI, se continuó el análisis con el apilado de las bandas 4 (infrarrojo cercano),5 (infrarrojo de onda corta) y 3 (rojo), dado que permite tener una apreciación visual de las áreas cultivadas.

Se definió un polígono aproximadamente 18 km al norte de Adela María en el departamento de Río Cuarto, determinando su posición de acuerdo con la longitud y latitud en Google Maps.

Para poder extraer de la imagen satelital este polígono se lo definió como un “objeto espacial” con las mismas coordenadas geográficas que la imagen obtenida del Landsat 8 (Apéndice, Figuras 1 y 2).

La **Figura 2** corresponde a la región seleccionada. La imagen pixelada obtenida permite observar distintas coberturas del suelo. Las zonas con verde más claro deberían corresponderse con las zonas donde los cultivos presentan mayor vigorosidad.

Figura 2



Luego para la región seleccionada se obtuvo para cada píxel el valor de NDVI operando entre las bandas 5 y 4 de acuerdo a la fórmula (1) del apartado 1.2.

3. Clasificación y análisis de la cobertura del suelo a partir del NDVI

Distintas investigaciones han establecido que existe una relación entre el NDVI y el nivel de productividad, si bien no es el único determinante. Clasificar a partir del NDVI y analizar los valores obtenidos es de importancia para evaluar la cobertura del suelo, y detectar, si las hubiera, zonas donde el crecimiento del cultivo podría presentar problemas, dado que se obtendría un valor de NDVI inferior al esperado.

En este apartado, en 3.1 se comienza realizando un análisis descriptivo del NDVI, seguidamente en 3.2 se aplica el método de Kmedias para clasificar el NDVI, considerando

de 3 a 20 clusters. Finalmente, en 3.3 se describe brevemente en que consiste el método de K medias, su implementación en R y se presentan los resultados obtenidos más relevantes.

3.1 Análisis descriptivo del NDVI

Para el área seleccionada, de acuerdo a la resolución del Landsat 8 el número de pixeles es de 186847; como nos interesa un mapeo del suelo, a partir del NDVI, se trabajó con el total de pixeles.

La **Tabla 1**, muestra las medidas resumen para el NDVI calculado.

Tabla 1: Medidas descriptivas de NDVI

Mínimo	Máximo	Media	Mediana	Rango	Desvío	MAD
0.096	0.650	0.624	0.539	0.554	0.102	0.093

Las Figuras 3 y 4 corresponden al histograma y boxplot del NDVI. respectivamente. A partir del análisis descriptivo realizado puede observarse una distribución asimétrica a izquierda, con presencia de valores atípicos.

Figura 3

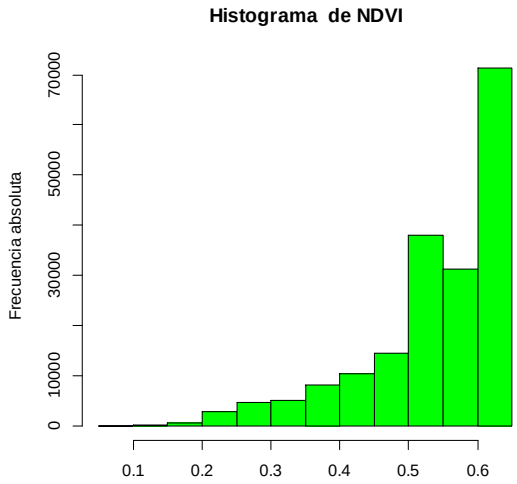
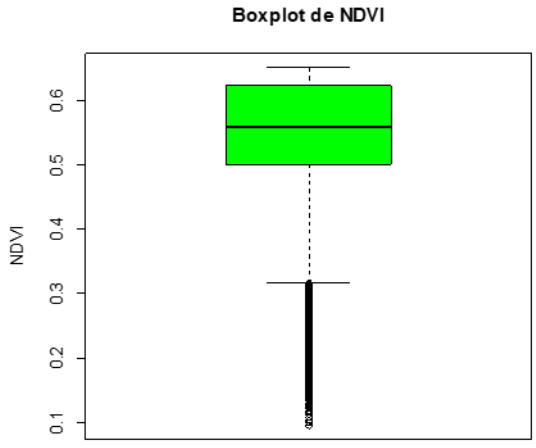


Figura 4

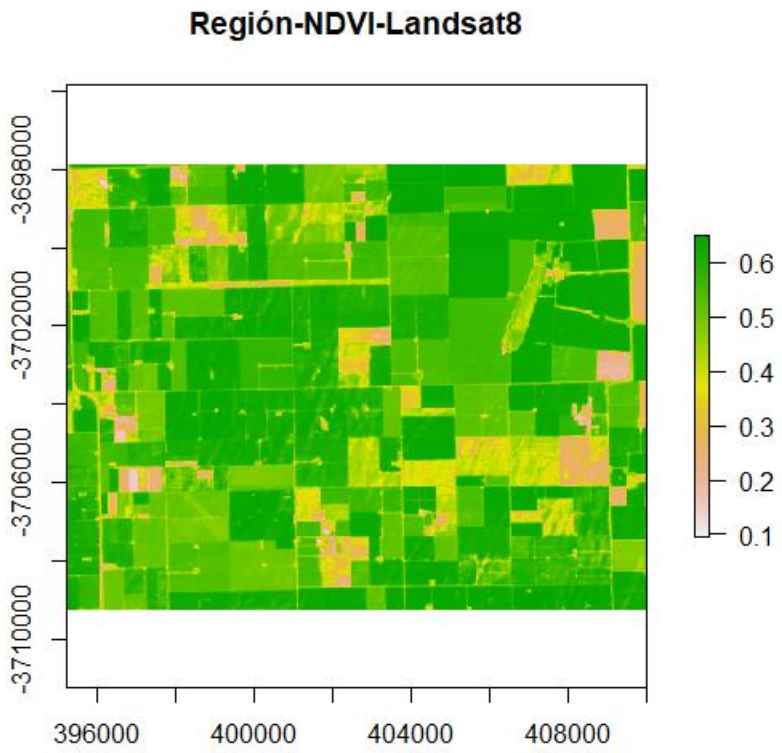


Los valores de NDVI se encuentran en un rango entre 0.1 y 0.65, siendo la mayoría de los mayores a 0.5 indicando un nivel alto de vigorosidad, esto es de esperar dado que

corresponde a una zona con cultivo principalmente de soja, y al período elegido de mayor vigorosidad de la planta.
De acuerdo con la clasificación de NDVI los valores más bajos corresponderían a zonas con vegetación senescente o de baja cobertura (Tabla 1, Anexo).

El **Figura 5** corresponde al gráfico por default de R, que realiza un mapeo de la zona a partir del valor de NDVI de cada píxel.

Figura 5



Para poder realizar este mapeo se combina la información de la capa raster con los valores calculados de NDVI, con la capa vectorial con los datos georreferenciados del área de estudio. Las coordenadas de la capa vectorial son las que permiten identificar la región.
En este gráfico puede observarse como ya se mencionó que el NDVI toma valores entre 0.10 y 0.65. El color preponderante en la región es el verde en distintas tonalidades, cuanto más intenso, es mayor el valor del índice.

Como se mencionó anteriormente, los que valores positivos más bajos podrían corresponder a vegetación senescente o de baja cobertura y los más altos representan un alto contenido de biomasa fotosintética.

3.2 Clasificación del NDVI mediante el algoritmo de K- medias

Como hemos mencionado diversas investigaciones relacionadas con la productividad de los cultivos, se basan principalmente en el NDVI como un indicador de la productividad.
El machine learning, aprendizaje automático o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras "aprendan". En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de la estadística inferencial, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático incorpora las preocupaciones de la complejidad computacional de los problemas. De acuerdo al abordaje de los problemas se lo clasifica en aprendizaje supervisado y no supervizado (Hastie et. al, 2001). En el aprendizaje supervisado, el objetivo es predecir el valor del output en función de distintas variables medidas; mientras que, en el aprendizaje no supervisado, no existe un output y el objetivo es describir las asociaciones y patrones entre un conjunto de medidas de entrada. Dentro de los métodos de aprendizaje no supervisado que permiten resolver problemas de clasificación, y que se aplicará en este trabajo, se encuentra el de K medias (k-means) (Hastie et. al, 2001).

3.2.1 Método de K-medias

Uno de los métodos de análisis de cluster (cluster analysis) más utilizados en la práctica es el método de K-medias (K-means).
El término cluster análisis (usado por primera vez por Tryon, 1939) se refiere a diferentes métodos para agrupar objetos de tipo similar en categorías. Es una técnica de análisis exploratorio que intenta ordenar diferentes objetos en grupos de manera que los objetos dentro de cada grupo están más estrechamente relacionados entre sí que con los que están en diferente grupo. En los últimos años estos métodos forman parte de los llamados métodos de aprendizaje automático de clasificación no supervisada.

El análisis de cluster se usa para encontrar estructuras en los datos sin proveer una explicación ni interpretación, solo "descubre" estructura en los datos, es luego el analista quien debe interpretar los posibles motivos de esa estructura.
Para poder realizar esta segmentación o agrupamiento es necesario establecer una medida de semejanza entre dos objetos (o de desemejanza).

El algoritmo de las K-medias es uno de los más populares iterativos descendentes métodos de cluster, y es usado cuando todas las variables son del tipo cuantitativo, y se considera como medida de desemejanza la distancia euclídea.
Luego para un total de N puntos a asignar a K cluster ($K \leq N$), el criterio es asignar las N observaciones a los K clusters de modo que dentro de cada cluster el promedio de las diferencias de cada observación a la media del cluster, definido por los puntos del cluster, sea mínima.
Es decir, el procedimiento tiene como objetivo minimizar la suma de cuadrados dentro de clusters, lo que equivale a asignar cada elemento al cluster en el que la distancia a la media del cluster es mínima. Como input se requiere el número k de clusters y la matriz de datos.
El algoritmo consiste en los siguientes pasos:
1. Particionar los objetos en k clusters iniciales.
2. Reasignar cada ítem de la lista de a uno por vez al cluster de cuyo centroide (media) se encuentre más cerca. Recalcular el centroide del cluster que recibe el ítem y del cluster que pierde el ítem.
3. Repetir el paso 2 hasta que no haya más reasignaciones.

En este trabajo, para cada píxel obtuvimos el valor de NDVI, por lo cual los datos son unidimensionales, luego la distancia euclídea es el módulo de la diferencia.

3.2.2 Implementación en R

Para clasificar la zona seleccionada mediante el algoritmo de K-medias, se trabajó con el total de los pixeles. Se utilizó la función kmeans de R (R Core Team, 2019) considerando un máximo de 500 interacciones, y el algoritmo de Lloyd (Lloyd, 1982). Esta función informa como una "medida de la calidad" de la clasificación obtenida el valor porcentual del cociente entre la suma de cuadrados entre los cluster (SSBC) y la suma de cuadrados total

(SSTC), este valor podría considerarse con el porcentaje de variabilidad explicada por la clasificación.

Para determinar el número de clusters, se aplicó el método heurístico del "codo", para esto se aplicó la función kmeans considerando de 1 a 20 clusters, para cada una de las clasificaciones se consideró la suma de cuadrados "intra-clusters" (SSWC) para determinar gráficamente cuando comienza a mantenerse constante.

De acuerdo con el resultado obtenido se analizaron los resultados de las clasificaciones obtenidas mediante el método de K-Medias considerando de 5 a 20 clusters. Para cada una de estas clasificaciones se realizó además del mapeo de la región clasificada, el gráfico de los boxplots en paralelo del NDVI en cada uno de los clusters obtenidos. Para complementar los resultados se realizó además un análisis descriptivo, de NDVI dentro de cada cluster.

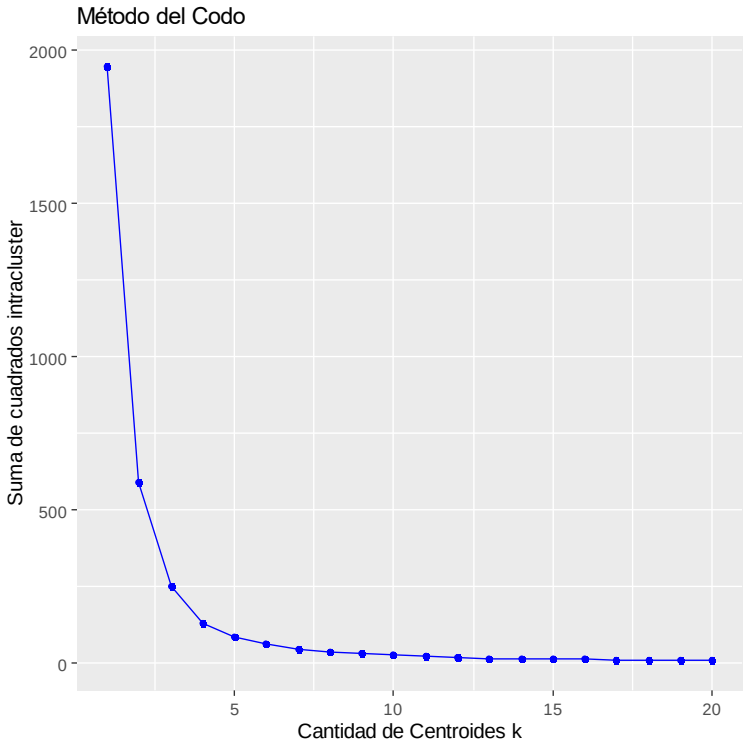
Para el caso de datos unidimensionales existe un método alternartivo, basado en programación dinámica propuesto por Wang y Song (Wang, 2011). Esta propuesta está implementada en R. Se realizó la clasificación mediante este algoritmo obteniéndose resultados similares.

Dada la presencia de valores atípicos, algunos valores muy bajos respecto de la mayoría, se aplicó el algoritmo de clustering CLARA (Clustering Large Applications) (Kaufman and Rousseeuw,1990) los resultados obtenidos también fueron similares a los obtenidos con el algoritmo kmeans, incluso considerando 20 clusters los resultados obtenidos fueron mejores con k-means en el sentido de la separación de los grupos.

3.3 Resultados obtenidos a partir del procesamiento y análisis de los datos.

Para determinar el número de clusters a considerar se aplicó el método heurístico "del codo". Con los resultados obtenidos mediante el método de K-medias considerando de 1 a 15 clusters, se construyó la **Figura 5,** que muestra la suma de cuadrados intra-cluster, en función del número de clusters. Puede observarse que a partir de 5 clusters (donde se forma el "codo") no hay una alta variación en los resultados de la suma de cuadrados intra-cluster

Figura 5



A partir de los resultados obtenidos y teniendo en cuenta la clasificación de NDVI (Anexo, Tabla 1), se compararon los resultados obtenidos considerando 5, 7, y 9 clusters. Los gráficos 6 7 y 8, muestran los resultados obtenidos correspondientes al mapeo de la región considerada y los boxplots en paralelo.

Puede observarse en los boxplots de la **Figura 6**, 3 clusters con valores de NDVI mayor a 0.5, un cluster con valores menores que 0.3, y dos cluster con valores entre 0.3 y 0.5, que de acuerdo a la clasificación de NDVI, sería conveniente separar. Para esta clasificación el valor porcentual del cociente entre la suma de cuadrados entre los cluster (SSBC) y la suma de cuadrados total (SSTC) es 95.80%.

Puede observarse en los boxplots de la **Figura 7**, 3 clusters con valores de NDVI mayor a 0.5, un cluster con valores menores a 0.3, un cluster con valores entre 0.3 y 0.4 y un cluster con algunos valores menores que 0.4 y otros mayores que de acuerdo a la clasificación de NDVI, sería conveniente separar. Para esta clasificación el valor porcentual del cociente entre la suma de cuadrados entre los cluster (SSBC) y la suma de cuadrados total (SSTC) es 97.68%.

Figura 6: Clasificación con 5 clusters

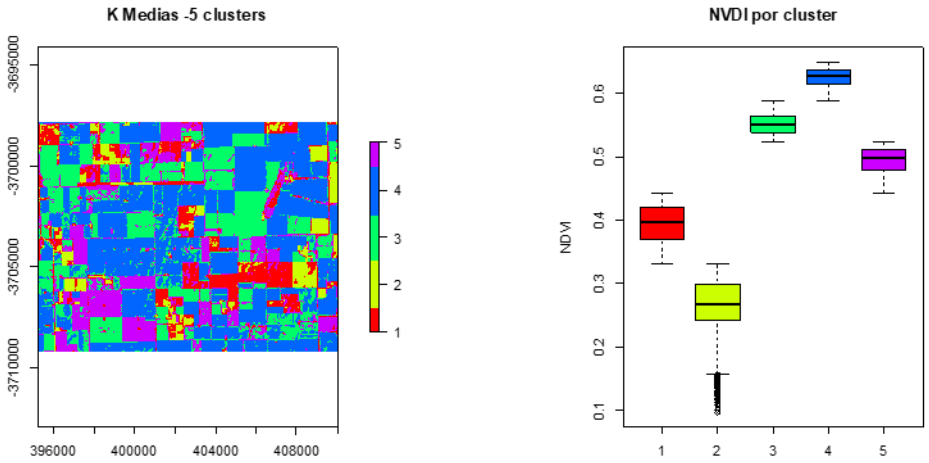
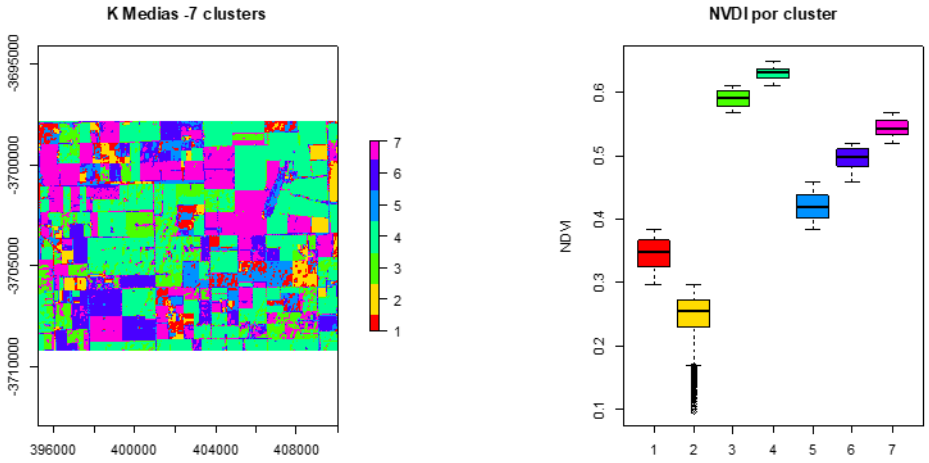


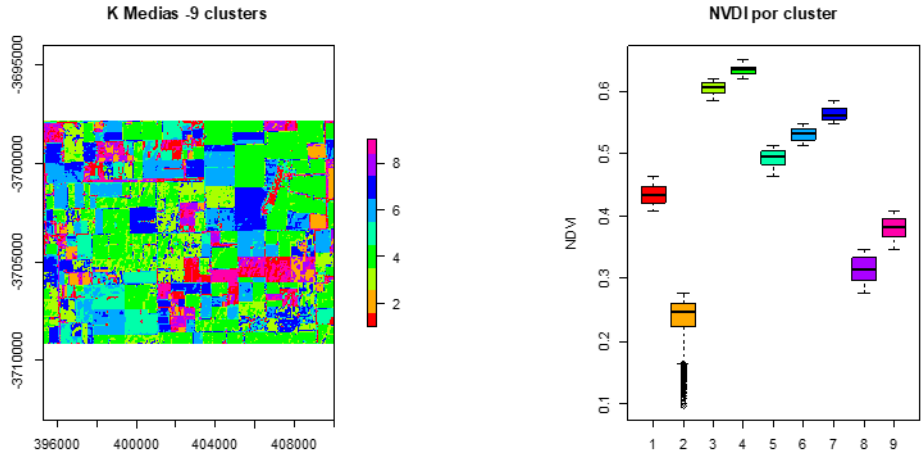
Gráfico 7: Clasificación con 7 clusters



Puede observarse en los boxplots del Gráfico 8 que no hay ningún cluster con valores menores y mayores de 0.4. Para esta clasificación el valor porcentual del cociente entre la suma de cuadrados entre los cluster (SSBC) y la suma de cuadrados total (SSTC) es 98.53%.

En las 3 clasificaciones realizadas, no se logra separar un cluster con valores menores y mayores a 0.2. Para lograrlo hay que considerar 20 clusters, pero se pierde la separación lograda con 9 clusters con punto de corte 0.4. Esta dificultad se debe a la presencia de valores atípicos, algunos valores muy bajos respecto del resto.

Figura 8: Clasificación con 9 clusters



Teniendo en cuenta que se tiene conocimiento de la cobertura estimada del suelo (IDECOR 2018), se considera adecuada para la región analizada considerar 9 cluster. Las **tablas 1 y 2** del Apéndice muestran respectivamente los datos de la clasificación obtenida considerando 9 clusters y el análisis descriptivo de NDVI para cada uno de estos.

A partir del análisis realizado, para la región de interés, donde tenemos un conocimiento previo de que se trata de zonas con plantaciones de soja mayoritariamente, hemos obtenido como era esperable valores de NDVI mayores a 0.5 mayoritariamente.

Conclusión

Este trabajo se ha estructurado de la siguiente manera. Luego de la introducción, en el apartado 1, se han desarrollado los conceptos y su aplicación a la agronomía de Big Data, NDVI obtenido mediante teledetección satelital y datos abiertos. Estos tres conceptos se articulan entre sí como ha sido mencionado, dado que una de las aplicaciones del Big Data en agronomía es a partir de la información que se obtiene por teledetección satelital, mediante la cual es posible calcular el índice NDVI relacionado con la productividad de las áreas cultivada; y es a partir de las políticas de datos abiertos llevadas a cabo por las principales agencias espaciales, que esa información está disponible para el público en

general, representando un desafío y oportunidad de generar conocimiento en las ciencias agronómicas tanto en el ámbito académico con en el sector público y privado en particular.

En el segundo apartado se ha desarrollado la metodología para cumplimentar el objetivo planteado en este trabajo de obtener una clasificación de áreas sembradas a partir del NDVI. Con este objeto se ha seleccionado como zona de interés a analizar el departamento de Río Cuarto en la provincia de Córdoba por ser una de las zonas de la región pampeana con mayor área cultivada, principalmente de soja. Se ha elegido para el análisis la cobertura del suelo con cultivo de soja dado que junto con el maíz y el trigo representan el mayor porcentaje de la producción de granos del país. La información disponible en el portal de la Bolsa de Cereales de Córdoba y los mapas interactivos desarrollados por IDECOR para la misma provincia, ha permitido establecer dentro del departamento plantaciones agrícola con soja. Como la siembra de soja se realiza principalmente entre septiembre y diciembre, una de las etapas de mayor vigorosidad de la planta debería darse entre los meses de enero y marzo, es por esto por lo que se ha seleccionado dentro del catálogo disponible de las imágenes del Landsat 8 de la NASA la imagen del 12 de marzo de 2019. Esta imagen ha sido seleccionada por contener a la región de interés, encontrarse en el período de mayor vigorosidad del cultivo y tener un porcentaje de nubes inferior al 20%, lo que ha permitido una correcta visualización de esta. Con la imagen obtenida se ha procesado y extraído la información de sus capas vectoriales, lo que ha permitido georreferenciarla, y las capas raster con la información de las bandas que ha permitido calcular para cada pixel de la región seleccionada el valor de NDVI.

Finalmente, en el apartado 3, a partir de los valores de NDVI obtenidos se ha realizado un análisis descriptivo que ha posibilitado observar que los valores de NDVI en la región de interés se han encontrado entre 0.10 y 0.65; y se ha obtenido la clasificación de la región y su mapeo mediante el método de clasificación de K medias considerando 9 clusters.

De acuerdo con el objetivo planteado se ha podido clasificar con un alto valor de performance la zona sembrada seleccionada a partir del NDVI.

Por lo mencionado anteriormente este trabajo ha permitido tomar conocimiento de las bases de datos satelitales disponibles, y de las herramientas que permiten su procesamiento. Dentro de los softwares que permiten el procesamiento de las imágenes satelitales y su posterior análisis se ha elegido trabajar con el software libre R. El código que se ha utilizado tanto

para el procesamiento como análisis posterior ha sido incorporado al Apéndice de este trabajo como parte de la transferencia de este.

Por otro lado, ha posibilitado reflexionar sobre el acceso a los datos abiertos y la privacidad de los datos utilizados. Si bien en este trabajo no se han involucrado personas, si, la posesión de personas individuales o sociedades. La disponibilidad de los datos a tal fin debería estar anonimizada de manera que, si bien a partir de la imagen es posible identificar una región geográficamente, que esta identificación no pueda ser cruzada con otra información, como por ejemplo la nomenclatura catastral, que permita la identificación de las personas o sociedades propietarias de la tierra.

Ha permitido pensar en futuras líneas de trabajo de investigación, que involucren desde lo metodológico, la adquisición de una serie de imágenes de una misma región con el objeto de plantear modelos para estimar la productividad y detectar anomalías. Por otro lado, como se ha sido mencionado, la actividad agropecuaria es una de las que mayor porcentaje aporta a la tributación, por lo cual, en relación con la política tributaria, otra línea sería analizar un escenario posible que permita establecer un valor de "renta presunta" a la actividad agrícola proporcional a la productividad, estimada mediante métodos de machine learning a partir del NDVI.

Si bien ha sido estudiado que el NDVI guarda una fuerte relación con la productividad, no es el único indicador, por lo cual otras variables, consideradas de acuerdo con la zona geográfica, el tipo de suelo, de cultivo, deberían ser contempladas en el análisis y predicción de la productividad.

Referencias bibliográficas

Bivand R., Keitt T.,Rowlingson R. (2019). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-4. https://CRAN.R-project.org/package=rgdal

Boken, V. K. y Shaykewich, C. F. (2002). Improving an operational wheat yield model using phenological phase-based Normalized Difference Vegetation Index. *International Journal of Remote Sensing*, 23, 4155−4168.

Bolsa de Cereales de Córdoba. (2019). Obtenido de http://www.bccba.com.ar

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

CONAE, (2016). Índices Espectrales derivados de imágenes satelitales Landsat 8 Sensor OLI. Guía de Usuario.

CONAE, (2018). Material del Curso SoPI: Introducción a la Teledetección

Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Cox, M., & Ellsworth, D. (1997). Managing Big Data for Scientific Visualization.

Cristianini, N., & Shawe-Taylor, J. (2000). Frontmatter. In *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (pp. I-Iv). Cambridge: Cambridge University Press.

Deering, D. W. (1978). Rangeland reflectance characteristics measured by aircraft and spacecraft sensors. Ph.D. Dissertation, Texas A & M University, College Station, TX, 338 pp

Doraiswamy, P. C. y Cook, P. W. (1995). Spring wheat yield assessment using NOAA AVHRR data. Canadian Journal of Remote Sensing, 21, 43−51.

FADA. (Septiembre de 2019). Obtenido de http://fundacionfada.org/informes/indice-fada-septiembre-2019-564/

Gantz, R. (2012). The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyse the future 2007, 1-16.

González R., Woods R. (2007). Digital Image Processing . Ed. Prentice Hall, 3th Ed

Hastie, T., Tibshirani, R., Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hijmans, R (2019). raster: Geographic Data Analysis and Modeling. R package version 3.0-2. https://CRAN.R-project.org/package=raster

Hijmans, R (2019). https://rspatial.org/

IDECOR ,2019. Obtenido de: https://idecor.cba.gov.ar

INTA, 2017. (s.f.). Obtenido de https://inta.gob.ar/sites/default/files/inta_informe_estadistico_del_mercado_de_soja.pdf

Joshil Raj K., SivaSathya S., (2014) SVM and Random Forest Classification of Satellite Image with NDVI as an Additional Attribute to the Dataset. In: Pant M., Deep K., Nagar A., Bansal J. (eds) Proceedings of the Third International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol 258. Springer, New Delhi

Kala, Sumi, Singh, Magan, Dutta, S,Singh, Narendra , Dwivedi, S. (2018).Applicattion of support vector machines for fodder crop assessment. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. IV-5. 415-420. 10.5194/isprs-annals-IV-5-415-2018.

Lloyd, S. P. (1957, 1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in *IEEE Transactions on Information Theory*, 28, 128–137.

Long, L.S., Solana, C., Canters, F., Chen, L., & Kervyn, M. (2017). Testing random forest classification for identification 1 and aging of lava flows from a single Landsat 8 image 2.

Mellor, Andrew & Haywood, Andrew & Jones, Simon & Wilkes, Phil. (2012). Forest Classification using Random forests with multisource remote sensing and ancillary GIS data.

Merg, C., Petri, D., Bodoira, F., Nini, M., Fernández, M., Schmidt, F., Montalva, R., Guzmán, L., Rodríguez, K., Blanco, F., & Selzer, F. (2011). Mapas digitales regionales de lluvias, índice estandarizado de precipitación e índice verde. Revista Pilquen, Sección Agronomía, 13(11), 1-11.

Ministerio de Agricultura, Ganaderia y Pesca. Presidencia de la Nación. (2019). Obtenido de https://geoadmin.agroindustria.gob.ar/geonetwork/srv/spa/catalog.search#/home

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S. y Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology, 151, 385-393Moriondo et al., 2007

Open Data Charter. (s.f.). Obtenido de https://opendatacharter.net/principles-es/

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

Presidencia de la Nación-Ministerio de Modernización. (s.f.). Obtenido de Paquete de Apertura de Datos de la República Argentina: https://datosgobar.github.io/paquete-apertura-datos/guia-subnacionales/#1-que-son-los-datos-abiertos

Quarmby, N. A., Milnes, M., Hindle, T. L. y Silleos, N. (1993). The use of multitemporal NDVI measurements from AVHRR data for crop yield estimation and prediction. International Journal of Remote Sensing, 14, 199−210. DOI: 10.1080/01431169308904332.

Raschka, Sebastian (2015). Python Machine Learning, Packt Open Source. ISBN 978-1-78355-513-0

Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

Rouse, J.W., Jr., R.H. Haas, J.A. Schell, and D.W. Deering. (1973). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. Prog. Rep. RSC 1978-1, Remote Sensing Center, Texas A&M Univ., College Station, 93p. (NTIS No. E73-106393).

Schomwandt, David. (2015). Teledetección aplicada a las Ciencias Agronómicas y recursos naturales. http://www.siia.gob.ar/joomla_files/images/mapas/ManualSensores.pdf

Sobrino, J. A., Raissouni, N., Kerr, Y., Olioso, A., López-García, M. J.,Belaid, A., El Kharraz, M. H., Cuenca, J., Dempere, L., (2000).Teledetección. Sobrino, J. A. (Ed.), Servicio de Publicaciones, Universidad de Valencia (ISBN 84-370-4220-8), Valencia (España).

Tucker, C. J., Holben, B. N., Elgin, J. H. y McMurtrey, J. E. (1980). Relationships of spectral data to grain yield variation. Photogrammetric Engineering and Remote Sensing, 46, 657−666.

U.S. Department of the Interior, U.S. Geological Survey. (2015). LANDSAT 8 (L8) Data users handbook (pdf). p. 106.

USGS. (s.f.). Obtenido de https://earthexplorer.usgs.gov/

White House. (2014). Big data: Seizing opportunities, preserving values. CFR.org. Retrieved from http://www.cfr.org/technology-and-science/white-house-big-data---seizing-opportunitiespreserving-values/p32916

Wickham H., François R., Henry L. ,MüllerK.(2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

Zheng, Baojuan, Myint, Soe,Thenkabail, Prasad,Aggarwal, Rimjhim.(2015). A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. International Journal of Applied Earth Observation and Geoinformation. 34. 103–112. 10.1016/j.jag.2014.07.002.

Anexos

Tabla 1:Clasificación de NDVI:

Los valores de NDVI corresponden a los propuestos por Merg et al. (2011)

Clasificación	Valor
Nubes y agua (NA)	< 0.01
Suelo sin vegetación (SV)	0.01 - 0.1
Vegetación ligera (VL)	0.1 - 0.2
Vegetación mediana (VM)	0.2 - 0.4
Vegetación alta (VA)	> 0.4

Apéndices

Figura 1: Identificación en Google Maps de la zona de interés.

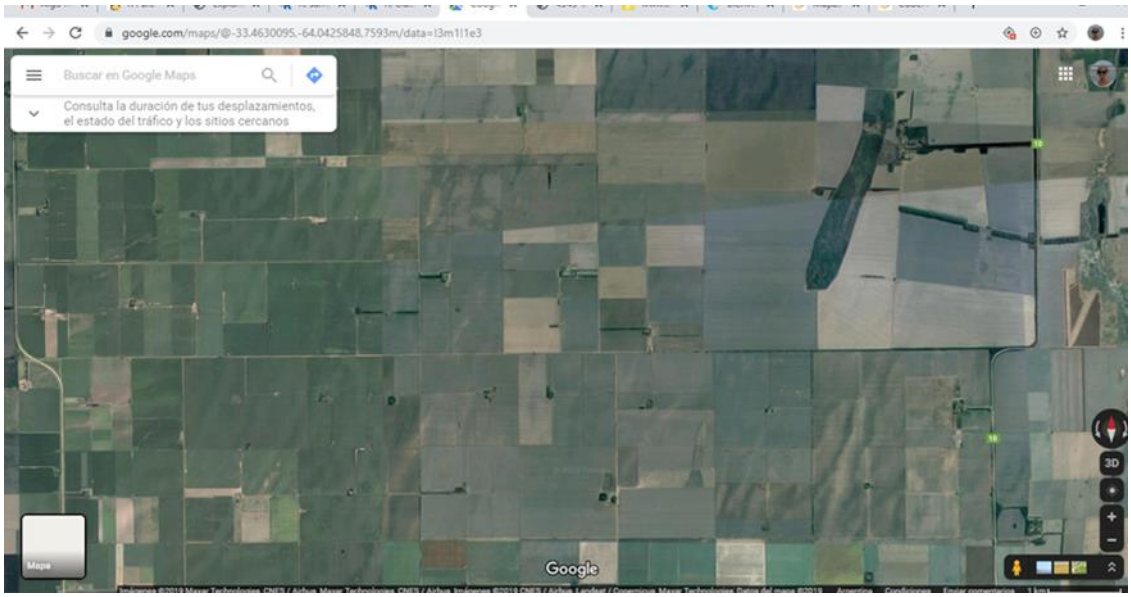
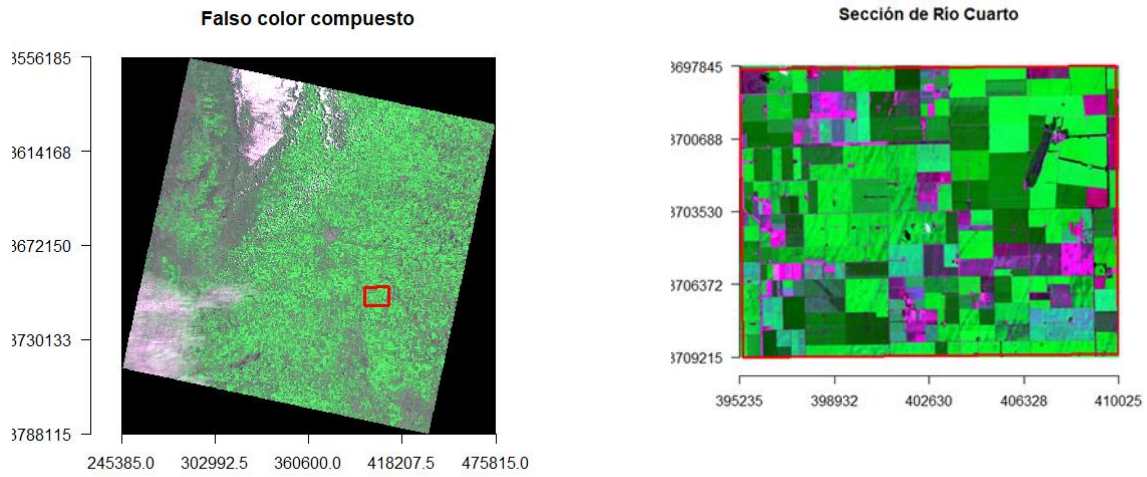


Figura 2: Selección de la zona de interés:



Resultados de la clasificación con 9 clusters

La tabla 1 muestra los datos de cada uno de los clusters

Tabla 1: Datos del cluster

Cluster	Tamaño	Centroide	SSWC
1	11265	0,434	2,913
2	6343	0,238	6,262
3	25802	0,605	2,430
4	53472	0,633	2,375
5	18773	0,492	3,221
6	29280	0,531	2,979
7	25392	0.563	2,710
8	6538	0,314	2,853
9	9982	0,380	2,882

SSWC: es la suma de cuadrados intra-cluster

La Tabla 2 muestra el análisis descriptivo de NDVI en cada cluster ordenados en forma ascendente.

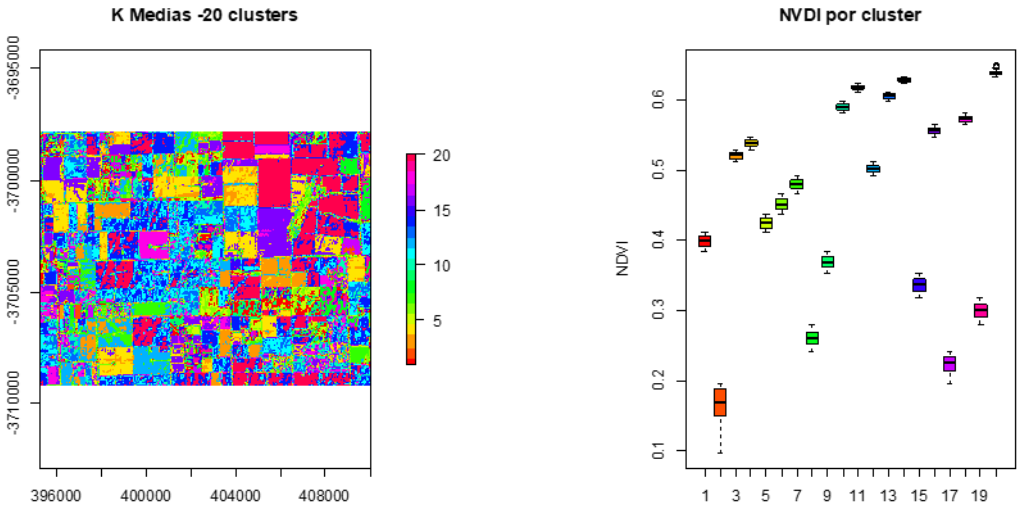
MAD: es la desviación absoluta mediana

CV : el el coeficiente de variación porcentual.

Tabla 2: Análisis descriptivo de NDVI

Cluster	Mínimo	Máximo	Media	Mediana	Desvio	MAD	CV
2	0,096	0,276	0,238	0,247	0,031	0,025	13%
8	0,276	0,347	0,314	0,316	0,021	0,027	7%
9	0,347	0,407	0,380	0,382	0,017	0,021	4%
1	0,407	0,463	0,434	0,433	0,016	0,020	4%
5	0,463	0,512	0,492	0,494	0,013	0,015	3%
6	0,512	0,547	0,531	0,532	0,010	0,012	2%
7	0,547	0,584	0,561	0,561	0,010	0,012	2%
3	0,584	0,619	0,605	0,607	0,010	0,011	2%
4	0,619	0,650	0,633	0,633	0,007	0,007	1%

Figura 3: Clasificación K medias con 20 clusters



Script en R.

#Esta sintaxis fue escrita para el procesamiento y análisis de imágenes Landsat 8 pero dado que
#se trabajó con archivos con extensión tif, puede utilizarse para imágenes Sentinel, revisando la definición de cada banda

#Referencia

```
#Spatial Data Science with R
#https://rspatial.org/
#Created by the GFC for the Innovation Lab for Collaborative Research on Sustainable Intensification
#License: CC BY-SA 4.0 --- Source code on github
#© Copyright 2016-2019, Robert J. Hijmans


# Se cargan las librerías necesarias para el procesamiento de los datos,
# si no fueron instaladas hay que instalarlas previamente
install.packages("raster")
install.packages("sp")
install.packages("dplyr")
install.packages("rgdal")

library(raster)
library(sp)
library(dplyr)
library(rgdal)

#El directorio está fijado en la carpeta que contiene la carpeta "Datos"
#Cada una de las bandas fue guardada en la carpeta Datos,
#y se le cambió el nombre a cada archivo.tif con el nombre de la banda
# Se leen los raster con las bandas necesarias para graficar a color real y falso color
#Cada una de las bandas de la imagen "descargada" del Landsat 8 fue renombrada como Bandai i=1,...11
# Para realizar el análisis se le asignó un nombre a las bandas de interés
#Solo se trabajó con 3 bandas, el trabajo con las restantes bandas es similar

# Blue
b2 <- raster('Datos/Banda2.tif')
# Green
b3 <- raster('Datos/Banda3.tif')
# Red
b4 <- raster('Datos/Banda4.tif')
# Near Infrared (NIR)
b5 <-raster('Datos/Banda5.tif')

#Chequeo  de algunos datos de las bandas
crs(b2)
ncell(b2)
dim(b2)
res(b2)
crs(b3)
ncell(b3)

#Para poder visualizar es necesario hacer un apilado adecuado de bandas
#stack es para apilar las bandas
#Color real
```

```
landsatRGB <- stack(b4, b3, b2)
plotRGB(landsatRGB, axes = TRUE, stretch = "lin", main = "Color Real")

#Falso color compuesto visualizando verde

landsat453<-stack(b4, b5, b3)
x11() #Para no cerrar el gráfico anterior
plotRGB(landsat453, axes = TRUE, stretch = "lin", main = "Falso color compuesto")

#Falso color compuesto visualizando rojo

landsat543<-stack(b5, b4, b3)
x11()
plotRGB(landsat543, axes = TRUE, stretch = "lin", main = "Falso color compuesto")


#A partir de ahora se trabajó con  la combinación de bandas 453
#Chequeamos el tipo de archivo

showDefault(landsat453)

#Extraemos del archivo "RasterStack" landsat 453 la región de interes
#Previamente seleccionamos la región de interés
#Las coordenadas de Google Maps no son las mismas, armanos el poligono y luego cambiamos el crs
#Se arman  las 4 esquinas del polígono armado en sentido antihorario, en este caso se seleccionó un rectángulo

loncba <- c(-33.416156, -33.416156,-33.517468, -33.517468)#coordenadas correspondiemtes a la longitu
latcba<- c(-63.968872,-64.126873,-64.126873,-63.968872)#coordenadas correspondientes a la latitud

verticescba<-cbind(latcba,loncba)
#Para armar un archivo de tipo vectorial, es necesario asignarle un crs, en este caso el de Google maps dado que
#es de donde se tomó la información
policba<- spPolygons(verticescba, crs= CRS('+init=epsg:4326'))

#Chequeamos que el tipo de objeto construido sea "SpatialPolygons"
showDefault(policba)

#Asignamos nombre al crs del archivo raster

crslandsat<-crs(landsat453)

#Cambiamos el crs para que tenga el mismo del archivo de tipo raster

policbalandsatcba<-spTransform(policba, crslandsat)
```

```
#Chequeamos el cambio
crs(policbalandsatcba)

#Vemos que tipo de objeto hemos creado
showDefault(policbalandsatcba)

#Graficamos el polígono superpuesto sobre la imagen

plot(policbalandsatcba,border='red', col='transparent', lwd=3, add=TRUE)

#Nos interesa la información de ese rectángulo
#Primero extraemos el rectángulo con la sentencia crop del paquete raster
#El poligono que extraemos es policbalandsatcba

Region<-crop(landsat453, policbalandsatcba)

#Vemos de que tipo de objeto se trata
showDefault(Region)
head(Region)
data.class(Region)
#Es un objeto de tipo RasterBrick
#La información que extrajimos es el valor de cada banda
#Podemos guardar este objeto como archivo convirtiendolo en un data frame
write.csv2(as.data.frame(Region),"Region453.csv")

#Graficamos la región seleccionada
x11()
plotRGB(Region, axes = TRUE, stretch = "lin", main = "")
title( "Región de interés")

#Para cada pixel de la región de interés calculamos el NDVI, también se puede calcular
#el NDVI para toda la imagen y luego extraer la región de interés

names(Region)
#En Landsat 8 (chequear el número de banda si
#Banda 5 es NIR
#Banda 4 es Red
ndvi <- (Region[['Banda5']] - Region[['Banda4']]) / (Region[['Banda5']] +
Region[['Banda4']])
data.class(ndvi)
#Es de tipo RasterLayer
#Chequeamos dimensión, resolucin y crs
dim(ndvi)
res(ndvi)
crs(ndvi)
ncell(ndvi)
ncell(Region) # Los objetos ndvi y Region tienen que tener el mismo número de pixeles
```

```
#Realizamos el gráfico de la región de interés "mapeada" de acuerdo al NDVI

x11()
plot(ndvi, main = 'Región-NDVI-Landsat8')

#########################################################
#Análisis descriptivo de los valores obtenidos de NDVI#
#########################################################
#Para trabajar con los valores de NDVI tenemos que extraer sus valores
#Usamos getValues
ndvir <- getValues(ndvi)
#es un vector con los valores de NDVI de cada pixel
#Chequemos la longitud
length(ndvir)
#Ibservamos los primeros valores
str(ndvir)
#Obtenemos medidas resumen de posición y de dispersión
summary(ndvir)
sd(ndvir)
mad(ndvir)
# Boxplot e Histograma
x11()
boxplot(ndvir, col="green", main="Boxplot de NDVI", ylab="NDVI")
x11()
hist(ndvir, col="green",xlab=" ",ylab="Frecuencia absoluta",main="Histograma  de NDVI")
################################
#Clasificación no supervisada #
################################

#Kmeans

#Para trabajar con los valores de NDVI tenemos que extraerlos sus valores
#Si ya no fue hecho antes
#ndvir <- getValues(ndvi)
#es un vector con los valores de NDVI de cada pixel

#Hacemos kmeans
#Trabajamos con todos los pixeles
#Para todo el subconjunto
#Fijamos una semilla, para generar los mismos centroides
#Hacemos el gráfico para aplicar el "método del codo" y determinar el número de clusters

set.seed(255)
wcss <- vector()
for(i in 1:20){
  wcss[i] <- sum(kmeans(na.omit(ndvir), centers = i, iter.max = 500, algorithm="Lloyd")
$withinss)
}
```

#Una vez calculados los valores de WCSS en función de la cantidad de centroides k, vamos a graficar los resultados:

```
library(ggplot2)
ggplot() + geom_point(aes(x = 1:20, y = wcss), color = 'blue') +
  geom_line(aes(x = 1:20, y = wcss), color = 'blue') +
  ggtitle("Método del Codo") +
  xlab('Cantidad de Centroides k') +
  ylab('Suma de cuadrados intracluster')

#De acuerdo a este métdo sería adecuado trabaja con un mínimo de 5 clusters
#Se consideran entonces

#5 Clusters
set.seed(255)
kndvicluster5<- kmeans(na.omit(ndvir), centers = 5, iter.max = 500, algorithm="Lloyd")
# Se genera un objeto de tipo kmeans
data.class(kndvicluster5)

# Con setValues "fijamos" los valores de ndvi en cada cluster
kndvir5 <- setValues(ndvi, kndvicluster5$cluster)
#Es un objeto RasterLayer
data.class(kndvir5)
#Otra forma es:
#kndvir5 <- raster(ndvi)
#values(kndvir5) <- kndvicluster5$cluster
#kndvir5
#Hacemos el gráfico  por default
x11()
plot(kndvir5, main = 'K Medias-  clusters' )

#Podemos hacer el gráfico y boxplt en paralelo para cada cluster

x11()
par(mfrow=c(1,2))
plot(kndvir5, main ="K Medias -5 clusters",col=rainbow(5) )
boxplot(ndvir~as.factor(kndvicluster5$cluster),col=rainbow(5),
xlab="",ylab="NDVI",main="NVDI por cluster")

#Análisis descriptivo de NDVI por cluster

tapply(ndvir,as.factor(kndvicluster5$cluster),summary)
tapply(ndvir,as.factor(kndvicluster5$cluster),sd)
tapply(ndvir,as.factor(kndvicluster5$cluster),mad)


#7 Clusters
set.seed(255)
kndvicluster7<- kmeans(na.omit(ndvir), centers = 7, iter.max = 500, algorithm="Lloyd")
```

```
# Se genera un objeto de tipo kmeans
data.class(kndvicluster7)

# Con setValues "fijamos" los valores de ndvi en cada cluster
kndvir7 <- setValues(ndvi, kndvicluster7$cluster)
#Es un objeto RasterLayer
data.class(kndvir7)
#Otra forma es:
#kndvir7 <- raster(ndvi)
#values(kndvir7) <- kndvicluster7$cluster
#kndvir7
#Hacemos el gráfico  por default
x11()
plot(kndvir7, main = 'K Medias- 7 clusters' )



#Podemos hacer el gráfico y boxplot en paralelo para cada cluster

x11()
par(mfrow=c(1,2))
plot(kndvir7, main ="K Medias -7 clusters",col=rainbow(7) )
boxplot(ndvir~as.factor(kndvicluster7$cluster),col=rainbow(7),
xlab="",ylab="NDVI",main="NVDI por cluster")

#Análisis descriptivo de NDVI por cluster

tapply(ndvir,as.factor(kndvicluster7$cluster),summary)
tapply(ndvir,as.factor(kndvicluster7$cluster),sd)
tapply(ndvir,as.factor(kndvicluster7$cluster),mad)

#9 Clusters
set.seed(255)
kndvicluster9<- kmeans(na.omit(ndvir), centers = 9, iter.max = 500, algorithm="Lloyd")
# Se genera un objeto de tipo kmeans
data.class(kndvicluster9)

# Con setValues "fijamos" los valores de ndvi en cada cluster
kndvir9 <- setValues(ndvi, kndvicluster9$cluster)
#Es un objeto RasterLayer
data.class(kndvir9)
#Otra forma es:
#kndvir9 <- raster(ndvi)
#values(kndvir9) <- kndvicluster9$cluster
#kndvir9
#Hacemos el gráfico  por default
x11()
plot(kndvir9, main = 'K Medias- 9 clusters' )
```

```
#Podemos hacer el gráfico y boxplot en paralelo para cada cluster

x11()
par(mfrow=c(1,2))
plot(kndvir9, main ="K Medias -9 clusters",col=rainbow(9) )
boxplot(ndvir~as.factor(kndvicluster9$cluster),col=rainbow(9),
xlab="",ylab="NDVI",main="NVDI por cluster")

#Análisis descriptivo de NDVI por cluster

tapply(ndvir,as.factor(kndvicluster9$cluster),summary)
tapply(ndvir,as.factor(kndvicluster9$cluster),sd)
tapply(ndvir,as.factor(kndvicluster9$cluster),mad)
#11 Clusters
set.seed(255)
kndvicluster11<- kmeans(na.omit(ndvir), centers = 11, iter.max = 500, algorithm="Lloyd")
# Se genera un objeto de tipo kmeans
data.class(kndvicluster11)

# Con setValues "fijamos" los valores de ndvi en cada cluster
kndvir11 <- setValues(ndvi, kndvicluster11$cluster)
#Es un objeto RasterLayer
data.class(kndvir11)
#Otra forma es:
#kndvir11 <- raster(ndvi)
#values(kndvir11) <- kndvicluster11$cluster
#kndvir11
#Hacemos el gráfico  por default
x11()
plot(kndvir11, main = 'K Medias- 11 clusters' )

#Podemos hacer el gráfico y boxplot en paralelo para cada cluster

x11()
par(mfrow=c(1,2))
plot(kndvir11, main ="K Medias -11 clusters",col=rainbow(11) )
boxplot(ndvir~as.factor(kndvicluster11$cluster),col=rainbow(11),
xlab="",ylab="NDVI",main="NVDI por cluster")

#Análisis descriptivo de NDVI por cluster

tapply(ndvir,as.factor(kndvicluster11$cluster),summary)
tapply(ndvir,as.factor(kndvicluster11$cluster),sd)
tapply(ndvir,as.factor(kndvicluster11$cluster),mad)

#20 Clusters
set.seed(255)
kndvicluster20<- kmeans(na.omit(ndvir), centers = 20, iter.max = 500, algorithm="Lloyd")
```

```
# Se genera un objeto de tipo kmeans
data.class(kndvicluster20)

# Con setValues "fijamos" los valores de ndvi en cada cluster
kndvir20 <- setValues(ndvi, kndvicluster20$cluster)
#Es un objeto RasterLayer
data.class(kndvir20)
#Otra forma es:
#kndvir20 <- raster(ndvi)
#values(kndvir20) <- kndvicluster20$cluster
#kndvir20
#Hacemos el gráfico  por default
x11()
plot(kndvir20, main = 'K Medias- 20 clusters' )

#Podemos hacer el gráfico y boxplot en paralelo para cada cluster

x11()
par(mfrow=c(1,2))
plot(kndvir20, main ="K Medias -20 clusters",col=rainbow(20) )
boxplot(ndvir~as.factor(kndvicluster20$cluster),col=rainbow(20),
xlab="",ylab="NDVI",main="NVDI por cluster")

#Análisis descriptivo de NDVI por cluster

tapply(ndvir,as.factor(kndvicluster20$cluster),summary)
tapply(ndvir,as.factor(kndvicluster20$cluster),sd)
tapply(ndvir,as.factor(kndvicluster20$cluster),mad)


#####################################################
#Una forma de resumir la información obtenida es: #
#####################################################

#3 clusters
nc3<-length(kndvicluster3$centers)
n3<-kndvicluster3$size
cent3<-kndvicluster3$centers
perf3<-kndvicluster3$betweenss/kndvicluster3$totss
s3<-tapply(ndvir,as.factor(kndvicluster3$cluster),summary)
sd3<-tapply(ndvir,as.factor(kndvicluster3$cluster),sd)
mad3<-tapply(ndvir,as.factor(kndvicluster3$cluster),mad)
lista3<-list( nc3,n3, cent3, perf3,s3,sd3,mad3)
names(lista3) <-c("Número de Clusters", "Tamaño del cluster","Centroides",
"Performance", "Medidas Resumen","Desvío","MAD")

#5 clusters
nc5<-length(kndvicluster5$centers)
n5<-kndvicluster5$size
```

```
cent5<-kndvicluster5$centers
perf5<-kndvicluster5$betweenss/kndvicluster5$totss
s5<-tapply(ndvir,as.factor(kndvicluster5$cluster),summary)
sd5<-tapply(ndvir,as.factor(kndvicluster5$cluster),sd)
mad5<-tapply(ndvir,as.factor(kndvicluster5$cluster),mad)
lista5<-list( nc5,n5,cent5, perf5,s5,sd5,mad5)
names(lista5) <-c("Número de Clusters","Tamaño del cluster", "Centroides",
"Performance", "Medidas Resumen","Desvío","MAD")

nc7<-length(kndvicluster7$centers)
n7<-kndvicluster7$size
cent7<-kndvicluster7$centers
perf7<-kndvicluster7$betweenss/kndvicluster7$totss
s7<-tapply(ndvir,as.factor(kndvicluster7$cluster),summary)
sd7<-tapply(ndvir,as.factor(kndvicluster7$cluster),sd)
mad7<-tapply(ndvir,as.factor(kndvicluster7$cluster),mad)
lista7<-list( nc7,n7,cent7, perf7,s7,sd7,mad7)
names(lista7) <-c("Número de Clusters","Tamaño del cluster", "Centroides",
"Performance", "Medidas Resumen","Desvío","MAD")

nc9<-length(kndvicluster9$centers)
n9<-kndvicluster9$size
cent9<-kndvicluster9$centers
perf9<-kndvicluster9$betweenss/kndvicluster9$totss
s9<-tapply(ndvir,as.factor(kndvicluster9$cluster),summary)
sd9<-tapply(ndvir,as.factor(kndvicluster9$cluster),sd)
mad9<-tapply(ndvir,as.factor(kndvicluster9$cluster),mad)
lista9<-list( nc9,n9,cent9, perf9,s9,sd9,mad9)
names(lista9) <-c("Número de Clusters", "Tamaño del cluster","Centroides",
"Performance", "Medidas Resumen","Desvío","MAD")

nc11<-length(kndvicluster11$centers)
n11<-kndvicluster11$size
cent11<-kndvicluster11$centers
perf11<-kndvicluster11$betweenss/kndvicluster11$totss
s11<-tapply(ndvir,as.factor(kndvicluster11$cluster),summary)
sd11<-tapply(ndvir,as.factor(kndvicluster11$cluster),sd)
mad11<-tapply(ndvir,as.factor(kndvicluster11$cluster),mad)
lista11<-list( nc11,n11,cent11, perf11,s11,sd11,mad11)
names(lista11) <-c("Número de Clusters","Tamaño del cluster", "Centroides",
"Performance", "Medidas Resumen","Desvío","MAD")

nc20<-length(kndvicluster20$centers)
n20<-kndvicluster20$size
cent20<-kndvicluster20$centers
perf20<-kndvicluster20$betweenss/kndvicluster20$totss
s20<-tapply(ndvir,as.factor(kndvicluster20$cluster),summary)
sd20<-tapply(ndvir,as.factor(kndvicluster20$cluster),sd)
mad20<-tapply(ndvir,as.factor(kndvicluster20$cluster),mad)
```

```
lista20<-list( nc20,n20,cent20, perf20,s20,sd20,mad20)
names(lista20) <-c("Número de Clusters","Tamaño del cluster", "Centroides", "Performance", "Medidas Resumen","Desvío","MAD")

c(lista3,lista5,lista7,lista9,lista11,lista20)

#########################################################
Anexo
#########################################################
#Clasificación con CLARA
library(cluster)
#Se realiza solo la clasificación, no el mapeo de la región
set.seed(255)
c9<-clara(na.omit(ndvir),k=9,metric ="euclidean")
c9
set.seed(255)
c20<-clara(na.omit(ndvir),k=20,metric ="euclidean")
c20
boxplot(ndvir~as.factor(c9$clustering),col=rainbow(20),
xlab="",ylab="NDVI",main="NVDI por cluster")
boxplot(ndvir~as.factor(c20$clustering),col=rainbow(9),
xlab="",ylab="NDVI",main="NVDI por cluster")
```