UNIVERSIDAD DE CANTABRIA

Proyecto de Ciclo de Vida de los Datos

Análisis de la relación entre la temperatura, localización geográfica y rendimiento en la producción de *Zea mays* en Colombia

Enero de 2020

Alumno (1): David Montero Loaiza
Alumna (2): Ana González Guerra
Alumno (3): Javier Alonso del Saso
Alumna (4): Silvia Magdalena López Monzó

Índice

1. Punto de partida del proyecto

En el siguiente trabajo, presentaremos resultados obtenidos y analizados a partir de tres fuentes de datos en abierto. Estas fuentes serán los siguientes datasets:

- Producción del maíz en Colombia [1]
- Imágenes de temperatura a escala global [2]
- Polígonos de veredas de Colombia, donde un conjunto de veredas define un municipio: https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-nivel-de-referencia-de-veredas/

2. Descripción general del proyecto

2.1. Objetivos

El objetivo principal del proyecto es obtener la dependencia de la producción de Zea mays (maíz) con la temperatura de las diferentes áreas de producción localizadas en municipios de Colombia. Para comparar dicha producción, usaremos el cociente de toneladas producidas por hectárea, que definiremos como rendimiento. También compararemos la proporción del terreno sembrado que genera producto que definiremos como eficiencia de la tierra.

Todo ello para determinar la zona (o municipio) y temperatura de mayor producción de zea mays en Colombia.

2.2. Resultados esperados

Como primera aproximación, esperamos que a medida la temperatura aumente, también lo haga la producción, con una temperatura óptima asociada a una máxima producción de unos 25ºC. A temperaturas inferiores el rendimiento es menor llegando a una producción nula para temperaturas muy bajas.

2.3. Requisitos

A continuación, desarrollaremos las funcionalidad de cada uno de los datasets introducidos anteriormente.

- Dataset de la cadena productiva de maíz: en este csv podemos encontrar el rendimiento de la producción de Zea mays (maíz) en unidades de toneladas/hectárea por cada municipio de Colombia. Los datos se encuentran divididos por semestres, es decir, tenemos un dato por cada seis meses.

- Imágenes de la temperatura: podemos observar la temperatura diara en cada vereda de Colombia, donde un conjunto de veredas constituye un municipio. Para cada día tenemos una temperatura máxima y una temperatura mínima. Es importante darse cuenta de que como los datos de rendimiento están por semestres, será necesario sacar una temperatura media semestral. Por tanto, haremos la media de la temperatura máxima y mínima para cada día, y después realizaremos otra media para obtener dos temperaturas semetrales al año.

- Dataset con los polígonos de veredas: el gobierno colombiano no dispone libremente los datos georreferenciados de municipios del país, no obstante, dispone los de veredas, que se encuentran administrativamente en un nivel inferior y por los cuales se pueden obtener los municipios y concordar así con los datos de producción de maíz, que se encuentran por municipios.

2.4. Requerimientos técnicos

Para poder generar las fuentes de datos necesarias para el proyecto es necesario establecer el cultivo del maíz en las diferentes regiones de Colombia, para ello se deberá disponer lo siguiente:

- Contratación de personal cualificado para el cultivo (agricultores).

- Calendario de siembra para el cultivo adaptado a cada región de Colombia.

  Atendiendo a la bibliografía encontrada se ha realizado una agrupación de los diferentes municipios de Colombia por zonas: ZAE1 (caribe) , ZAE2 (llanos orientales), ZAE3 (región interandina), ZAE4 (región andina), ZAE5 (región del pacífico) y ZAE6 (región amazónica) [FAO, 2006], esto con el objetivo de reducir la variabilidad en las condiciones de siembra y cosecha.

  Planteándose el siguiente calendario para la siembra:

Calendario de siembra de *Zea mays* en Colombia		
Región	Período de siembra	Referencia
ZAE1	Abril; septiembre	
ZAE2	Marzo	
ZAE3	Marzo; septiembre	modificado de [FAO, 2006]
ZAE4; ZAE5; ZAE6	Febrero; septiembre	

Se ha considerado un periodo estándar antes de la recogida de la cosecha de 5 meses dado que en general el ciclo de vida del maíz abarca entre 150 y 300 días [Ospina Rojas and Duarte Pérez, 20

- Análisis químico del suelo en cada una de las zonas geográficas con el objetivo de determinar las necesidades nutricionales del mismo y por tanto los complementos necesarios para el desarrollo del cultivo del maíz. Estos complementos se tendrán en cuenta en la elección del abono apropiado para el campo de cultivo.

- Sistema de riego optimizado para aquellas regiones donde las precipitaciones son escasas.

- Contratación del servicio de estaciones meteorológicas con el gobierno de Colombia para disponer de los datos de temperatura de las regiones durante el transcurso del cultivo.

- Contratación del servicio de imágenes de satélite con el gobierno de Colombia con el objetivo de delimitar los municipios donde se pretende efectuar el análisis.

2.5. Descripción del problema

2.5.1. Interés

esto falta por arreglar

- Importancia y usos del cultivo del maiz.

- Colombia, condiciones ideales para el cultivo de Zea mays en general, teniendo en cuenta esto elegir la mejor zona si se dispone de recursos limitados, por ejemplo solo se puede abarcar terreno en uno de los municipios, para así asegurar la producción más efectiva a nivel geográfico, decidiendo esto en base a la condición temperatura de la zona, pues parece ser que en Colombia es un aspecto muy variable entre las diferentes regiones.

El problema a enfocar es encontrar el municipio de Colombia óptimo para la producción de Zea Mays, además de la temperatura asociada para dicha producción óptima. En definitiva, optimizar la producción de maíz en Colombia eligiendo la mejor región de producción.

(TEMPORAL) El cambio climático es un problema global que afecta a varios sectores de la industria. En específico, la alimentaria es una de las más relevantes y la que es más afectada

por cambios en la climatología. Comprender el estado actual de la producción agrícola teniendo en cuenta. Javi? qué querías decir aquí?

2.5.2. Cobertura (geográfica, temporal)

La cobertura temporal abarca el período 2015-2018, dividiendo los años por semestres. Concretamente para 2018 las medidas finalizan con el primer semestre.

En cuanto a la cobertura geográfica se centra en Colombia y en los municipios productores del cultivo elegido.

La limitación estaría centrada en el período estudiado para la producción, se ofrecen esos años y nada más.

2.5.3. Objetivos

Parafrasear los objetivos del principio.

2.5.4. Descripción de las fuentes de los datos

Los datos han sido extraídos de las siguientes bases de datos en abierto:

- Base de datos en abierto del gobierno de Colombia (https://www.datos.gov.co/). Concretamente los datos de producción han sido oobtenidos de la división de agricultura y desarrollo rural.

- The Land Processes Distributed Active Center (LP DACC), que es una base de datos en abierto englobada dentro del sistema de observación de la NASA (NASA Earth Observing System Data and Information System, EOSDIS) en la división USGS Earth Resources Observation and Science (EROS) (https://lpdaac.usgs.gov/).

- Geoportal del Departamento Aministrativo Nacional de Estadística de Colombia (DANE) (https://geoportal.dane.gov.co/)

Los formatos en los que se han recogido los datos han sido los siguientes:

- Producción de maíz en Colombia: formato csv (comma-separated values).
- Imágenes de temperatura: formato tiff (Tagged Image File Format) por ser imágenes georreferenciadas.
- Polígonos de veredas en Colombia: formato shp, es decir, formato de archivo informático (shapefile) que almacena las entidades geométricas de los objetos

3. Data Management Plan

A continuación, describimos el plan de gestión de los datos, en el que explicaremos cómo se han creado los datos, se detallará cómo se documentarán, quién podrá obtenerlos y cómo. Se especificará cómo serán almacenados y si serán comparados.

Para ello, seguiremos el plan orginizado por H2020 [Commission European, 2019] [https://ec.europa.eu/rese
funding-guide/cross-cutting-issues/open-access-data-management/data-management$_e$n.htm].$Estedocume$
$findable, accessible, interoperable, re-usable.Portanto, losdatoscreadospodránseridentificadosconmeta$

3.1. Data summary

- Propósito de la recogida / generación de datos y su relación con los objetivos del proyecto.

Encontrar una relación entre temperatura y rendimiento de la producción de maíz que permita decidir cuál es el mejor municipio de Colombia para alcanzar una mayor producción de maíz.

- ¿Qué tipos y formatos de datos generará o recogerá el proyecto?

- ¿Usaremos datos ya existentes? ¿cómo?

Los datos a usar no van a ser datos ya existentes, pues comenzaremos el proceso desde 0, desde el sembrado, pasando por la instalación de las estaciones metereológicas que nos permitirán controlar la temperatura hasta la toma de imágenes de satélite de los municipios de Colombia estudiados.

- ¿Cuál es el origen de los datos?

Los datos no tienen un único origen, provienen de tres fuentes diferentes: estaciones meteorológicas, análisis de la producción de maiz e imágenes de satélite de las regiones estudiadas.

poner por aquí que cada una de estas ramas de datos ha sido depositada en el repositorio correspondiente: gobierno de colombia, dane , etc.

- ¿Cuál es el tamaño esperado de los datos? no entiendo a qué se refiere con esto. - ¿Para quién podrían ser útiles estos datos?

Estos datos podrían ser útiles para pequeños agricultures interesados en comenzar a producir maíz, pues la relación temperatura-producción, les permitiría saber si una zona es óptima o no. Reduciendo con mucho el riesgo inicial de producción, no se arriesgan a sembrar en una zona no adaptada climatológicamente para la producción de maíz. Así en el plazo de un año podrían tener ganancias fijas que les permitiesen ampliar zona de cultivo en caso de ser necesario.

También sería posible saber para una nueva zona, no tiene por qué ser uno de los municipios dados, conociendo su temperatura media si es adecuada o no para la producción.

3.2. Fair Data

Hacer los datos accesibles, incluyendo metadatos

¿Se pueden encontrar los datos producidos y / o usados en el proyecto con los metadados, identificables y localizables por medio de un mecanismo de identificación estándar? (por ejemplo, identificadores persistentes y únicos como Digital Object Identifiers (DOIs)?

Los datos pueden ser encontrados en el repositorio de github(), junto con este informe en el que se describe su obtención, uso y análisis.

¿Qué convenciones de nombres se han seguido?

¿Se proporcionan palabras clave para optimizar la posibilidades de reutilización?

Las palanras claves proporcionadas son las siguientes: maíz, producción, rendimiento, Colombia, municipio, temperatura, dependencia.

¿Qué metadados se han creado? En caso de que no haya metadatos estándar en la disciplina, indicar el tipo de metadado y cómo se ha creado.

Se han incluido los siguientes metadatos: título, creador, palabras clave, descripción, editor, contribuidor, fecha, tipo, formato, identificador, fuente, idioma, cobertura y derechos.

Hacer los datos accesibles

¿Qué datos producidos y/o utilizados en el proyecto se presentarán abiertamente? Si algunos datasets no pueden ser compartidos (o tienen que serlo bajo ciertas restricciones) explica por qué, separando claramente si el motivo es legal o voluntario.

Se harán públicos los datos producidos y de los tres datasets utilizados, tan solo se harán públicos los correspondientes a las imágenes georreferenciadas y a los polígonos de veredas de Colombia, pero no el dataset de la producción de maíz debido a que recientemente el gobierno

Colombiano denegó el acceso al público y ahora es un dataset privado. Por tanto, para respetar la medida del gobierno Colombiano, el dataset no se hará público.

¿Cómo se harán accesible los datos?

Los datos serán accesibles mediante un repositorio github. La dirección de dicho repositorio se encuentra disponible como identificador en los metadatos.

Si existe alguna restricción de uso del repositorio, ¿qué tipo de acceso se necesita?

No existe ninguna restricción para el uso del respositorio, es de libre acceso.

¿Son los datos producidos interoperables? Esto es que permiten el intercambio y reutilización entre investigadores, instituciones, organizaciones, países, etc. (es decir, se adhieren a los estándares de formatos y en la medida de lo posible, se adaptan con aplicaciones de software disponibles (abiertas) y, en particular, facilitan las combinaciones con diferentes conjuntos de datos de diferentes orígenes)

Todos los datos serán almacenados en distintos documentos con formato csv, por lo tanto, será fácilmente legible por máquinas independientemente del país o institución que pretenda usarlos.

¿Qué tipo de metodología, estándares o vocabulario de metadata será usado para hacer que los datos sean interoperables?

Para describir los datos/metadatos del documento final, se han seguido los convenios del esquema Dublin Core. Estos metadatos han sido incluidos en inglés, para garantizar una mayor interoperabilidad.

¿Qué licencia tendrán los datos para permitir la reutilización más amplia posible?

Nos apoyaremos en Atribución/Reconocimiento 4.0 Licencia Pública Internacional. Por lo tanto, el usuario es libre tanto de compartir, copiar, redistribuir el material tanto como de adaptar, trasnformar y crear a partir de dicho material.[https://creativecommons.org/licenses/by/4.0/deed.es$_ES$]

¿Cuándo estarán disponibles los datos para su reutilización? Si se busca un embargo para dar tiempo a publicar o buscar patentes, especifique por qué y durante cuánto tiempo se aplicará, teniendo en cuenta que los datos de la investigación deben estar disponibles lo antes posible.

HELP

¿Durante cuánto tiempo estarán los datos disponibles para su reúso?

Debido a que los datos se encuentrán disponibles en GitHub, deberían estar permanentemente allí, y si desaparecen, serán por razones ajenas a nosotros.

¿Se describen los procesos de garantía de calidad de los datos?

JAVI ?????

4. Curación de Datos y ETL

Las tres fuentes de datos tuvieron que pasar por un proceso de curación y posteriormente por un proceso ETL para obtener una base de datos limpia y lista para realizar cualquier tipo de análisis sobre ella.

4.1. Preprocesamiento de imágenes

Las imágenes georreferenciadas de temperatura superficial, obtenidas a partir del producto MOD11A1, del sensor MODIS abordo del satélite de la NASA Terra EOS AM-1, fueron las primeras en preprocesarse, ya que era necesario adecuar la temporalidad de estos datos (diaria) a la temporalidad de los datos de producción (semestral).

Cada imagen de temperatura superficial está compuesta por 12 sub-imágenes, conocidas como bandas, de las cuales se utilizaron dos:

1. LST_Day_1km
2. LST_Night_1km

Estas bandas contienen la información de la temperatura diaria de la superficie terrestre de día y de noche repartidas en una grilla de 1 km. Ambas bandas fueron utilizadas como temperatura superficial mínima y temperatura superficial máxima y con ellas calcular la temperatura superficial media por cada valor de pixel de la grilla. De manera adicional, la temperatura, que

se encontraba en grados Kelvin, fue convertida a grados centigrados.

Teniendo una imagen de temperatura superficial por día, se calculó el promedio por cada pixel de la grilla se manera semestral desde el año 2015 hasta el primer semestre del año 2018 (fecha hasta la cual se encuentran disponibles los datos de producción de maíz). Por tal motivo se redujo la cantidad de imágenes diarias a 7 imágenes correspondientes a la temperatura superficial media semestral de los años correspondientes para la cobertura espacial de Colombia.

4.2. Preprocesamiento de veredas

Colombia, hasta el nivel de veredas, se encuentra administrativamente dividida en 3 niveles:

- Nivel superior: departamentos.
- Nivel intermedio: municipios.
- Nivel inferior: veredas.

El nivel en el que se encuentran los datos de producción de maíz es el de municipios, sin embargo, el dataset de municipios georreferenciados no se encuentra disponible en los datos abiertos ofrecidos por el Gobierno Colombiano, por tal razón, se han elegido los datos correspondientes a las veredas, que se encuentran a un nivel inferior y pueden ser transformados a un dataset georreferenciado de municipios en un simple paso.

El dataset georreferenciado de veredas contiene asociado una base de datos, en donde cada vereda se encuentra dentro de un municipio. Al tener todas las veredas un municipio asociado, para datos geoespaciales se puede realizar la acción de 'disolver'. Disolver datos espaciales permite generar polígonos de mayor tamaño con respecto a una columna en común, por tal motivo, todas las verdeas que pertenezcan a un mismo municipio, se unirán y generarán un polígono más grande que corresponde a dicho municipio.

4.3. Extracción de Temperatura Superficial por Municipio

Teniendo el dataset georreferenciado de municipios, ahora es posibles extraer, por cada imagen georreferenciada, el promedio de temperatura superficial para cada uno de los municipios como el promedio de todos los pixeles de la grilla que se encuentren contenidos o que toquen el polígono georreferenciado de un municipio en concreto.

Esta extracción de temperatura superficial por municipio genera una columna adicional a la base de datos del dataset de municipios georreferenciados correspondiente al valor de temperatura superficial de ese municipio, y una columna adicional del semestre y el año correspondientes a dicha temperatura.

4.4. Unión con Datos de Producción

Los municipios se encuentran identificados por un código en la base de datos de producción, mientras que la columna de "periodo" nos indica el semestre y el año al que se encuentran asociados los datos de producción. En la base de datos de la temperatura superficial extraída, los municipios se encuentran identificados por el mismo código de municipio, mientras que existe una columna relativa a la columna "periodo" de la base de datos de producción.

Para unir ambas tablas, se creó un índice en cada una como la concatenación del código del municipio con el periodo al que corresponde los datos de producción o de temperatura superficial en el caso contrario. Dicho código fue utilizado como indicador del "Join" entre ambas tablas para obtener una tabla completa con los datos asociaodos de producción, municipios, departamentos y temperatura según el periodo de estudio y sobre la cual podrían empezarse todo tipo de análisis.

5. Plan de preservación

6. Análisis de los datos

7. Conclusiones

Referencias

[1] Cadena productiva del maíz: área de producción y rendimiento. Gobierno de Colombia. 2018. https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Cadena-Productiva-Ma-z-Area-Producci-n-Y-Rendimien/d968-yfb5

[2] Terra Land Surface Temperature and Emissivity Daily Global 1km. NASA LP DAAC at the USGS EROS Center. https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD11A1

Referencias

[FAO, 2006] (2006). Colombia. In *Calendario de cultivos: América Latina y el Caribe*, volume 186, pages 79–92. FAO, Roma.

[Commission European, 2019] Commission European (2019). Extension of the open research data pilot in Horizon 2020.

[Ospina Rojas and Duarte Pérez, 2011] Ospina Rojas, J. G. and Duarte Pérez, C. J. (2011). Fisiología de la planta del maíz. In *Aspectos técnicos de la producción de maíz en Colombia*, pages 33–59. Fenalce.