UNIVERSIDAD DE CANTABRIA

Proyecto de Ciclo de Vida de los Datos

Análisis de la relación entre la temperatura 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 (Gobierno de Colombia (2018)).
- Imágenes de temperatura a escala global (NASA LP DAAC at the USGS EROS Center (2019))
- Polígonos de veredas de Colombia, donde un conjunto de veredas define un municipio (Departamento Aministrativo Nacional de Estadística de Colombia (2017))

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.

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 se espera un rendimiento menor llegando a una producción nula para temperaturas muy bajas.
Además, esperamos una dependencia lineal creciente de la producción en toneladas como función del área cosechada.

2.3. Requisitos

A continuación, desarrollaremos las funcionalidades 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 diaria 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:

Región	Período de siembra	Referencia
ZAE1	Abril; septiembre	
ZAE2	Marzo	
ZAE3	Marzo; septiembre	(modificado de FAO (2006))
ZAE4; ZAE5; ZAE6	Febrero; septiembre	

Tabla 1: Calendario de siembra de *Zea mays* en Colombia

  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 y Duarte Pérez (2011)).

- 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

El maíz constituye uno de los pilares más importantes en la producción agrícola de Colombia, encontrándose en prácticamente todo el territorio (Ospina Rojas y Duarte Pérez (2011)). Es por esto que consideramos de interés la optimización de su producción, concretamente queremos centrarnos en los pequeños agricultores, pretendemos proporcionarles un método para asegurar condiciones óptimas desde que empieza la fase del cultivo. Así con una producción inicial estable sería posible ampliar la producción, y podrían hacerlo aplicando de nuevo el método usado.

Por tanto, el problema a enfocar en este proyecto es encontrar la temperatura óptima para la producción de *Zea Mays*, además de la temperatura asociada para dicha producción óptima.

La razón de la elección de la temperatura como variable principal en el análisis se basa en que es esencial para el desarrollo del maíz (Lafitte (2001)).

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.

2.5.3. 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 (Gobierno de Colombia (2018)). Concretamente los datos de producción han sido obtenidos 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) (NASA LP DAAC at the USGS EROS Center (2019)).

- Geoportal del Departamento Administrativo Nacional de Estadística de Colombia (DANE), en particular tomamos los datos de polígonos de veredas que dividen por regiones a Colombia (Departamento Aministrativo Nacional de Estadística de Colombia (2017)).

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. Y finalmente, se especificará cómo serán almacenados y si serán comparados.

Para ello, seguiremos el plan desarrollado en la convocatoria del H2020 (Commission European (2019)). Siguiendo el template desarrollado en la convocatoria del documento del H2020 crearemos unos datos FAIR : Findable, Accessible, Interoperable, Re-usable (GO FARE (2018)). Por tanto, los datos creados podrán ser identificados con metadatos accesibles y legibles tanto por máquinas como por personas, gracias al uso de protocolos estándarizados. Y finalmente, estarán bien descritos mediante unos metadatos bien explicados para su posible reúso.

3.1. Data summary

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

El propósito de la recogida y generación de datos es encontrar una relación entre temperatura y rendimiento de la producción de maíz que permita decidir cuál es la temperatura óptima de producción.

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

Los tipos de datos recogidos y generados para el proyecto encajan dentro del ámbito de la ciencia de datos, concretamente en las secciones de curación y análisis.

En cuanto a los formatos en que se han recogido los datos:

- Producción de maíz en Colombia: formato csv.
- Imágenes de temperatura georreferenciadas para Colombia: formato tiff.
- Polígonos de veredas en Colombia: formato shp.

Por otra parte el formato de los datos generados es de csv.

¿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 y cosechado del maíz, pasando por la contratación de los servicios de las estaciones meteorológicas que nos permitirán la toma de datos de temperatura y de la obtención de imágenes vía satélite de los municipios de Colombia con el gobierno colombiano.

Los paquetes de trabajo y el personal implicados en la obtención y generación de los datos del proyecto aparecen detallados en el diagrama de Gantt.

¿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.

¿Cuál es el tamaño esperado de los datos?

Para calcular el tamaño esperado de los datos hemos partido de la base de que si cada imagen pesa 2.6 Mb, obteniendo una imágen diaria a lo largo de 3 años y medio, solamente con los datos asociados a las imágenes se generán 3.32 Gb. Por otra parte los datos de producción iniciales ocuparon 1Mb y los datos de veredas ocuparon 336 Mb. Si sumamos el tamaño de todos los datos comentados hasta ahora hace un total de 3.65Gb.

Teniendo en cuenta que el análisis de los datos se basa mayormente en su gráficación, el tamaño de los datos derivados no debería ser superior a 10 Mb, el tamaño esperado de los datos es de 3.75Gb.

¿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.

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 siguiente repositorio de github , junto con este informe en el que se describe su obtención, uso y análisis. También es posible encontrar los datos publicados en Zenodo en el repositorio: Datos curados de variables de producción por departamento colombiano (2015-2018).

¿Qué convenciones de nombres se han seguido?

La convención adoptada para el nombre de los metadatos se basa en el formato de metadatos Dublin Core. Dado que es un formato estandarizado, bien estructurado y sencillo de manejar.

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

Las palabras 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.

¿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.

Los datasets usados, i.e. imágenes georreferenciadas de temperatura, polígonos de veredas en Colombia y la producción y rendimiento de maíz en el país son públicos y de libre acceso. En el repositorio se publican el dataset original de la producción y rendimiento de maíz, además de los datos transformados de la temperatura en los distintos municipios del país y dataset final que agrupa a todos los datos (todos los data sets usados son referenciados a sus repositorios).

¿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 (Creative Commons Corporation (2019)).

¿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.

Los datos están listos de forma libre e inmediata para su reutilización gracias a la plataforma web de GitHub. Los procedimientos realizados para obtener los resultados no están patentados (son scripts de python usando librerías de uso libre). En el caso de que fuera necesario un embargo, sería para búsquedas de patentes de nuestros procedimientos.

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

Los datos estarán el mayor tiempo posible disponibles, bien en GitHub o Zenodo o almacenados en cintas magnéticas como posteriormente se detalla en este informe (sección 5).

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

La calidad de los datos está fuertemente arraigada a la calidad de las fuentes originales. Como tratamos con datos gubernamentales por un lado y datos de satélite de la compañía Google por otro, podemos asegurarnos que los datos son de muy buena calidad y fiables.

Por otra parte, los datos se encuentran disponibles y de libre acceso. La comunidad interersada puede participar y ayudar para la curación y calidad de los datos a largo periodo.

3.3. Localización de los recursos

¿Cuál es el coste de hacer los datos FAIR?

Como los datos se encuentran en un repositorio abierto. El acceso es fácil y gratuito. Además, se encuentran en formato *csv*, muchos lenguajes y entornos de programación están preparados para tratar con este formato. Además, los datos estarán disponibles en la base de Zenodo que lleva preparado APIs para conseguir el recurso.

3.4. Seguridad de los datos

¿Qué medidas preventivas se encuentran impuestas a los datos?

Los datos en la nube sólo podrán ser sobreescritos por los contribuidores de este proyecto. De todas formas, se levanta una arquitectura de hardware con discos duros que realizan back ups de forma periódica sobre el repositorio además del uso de cintas magnéticas para el repositorio completo en caso de que los recursos se pierdan en la nube.

Tanto los discos como las cintas se encuentran a disposición de los contribuidores del proyecto exclusivamente, para evitar ataques externos.

¿Los datos se encuentran almacendos un repositorio certificado para la preservación y curación a largo plazo?

Los datos estarán disponibles en la plataforma de Zenodo. Esta será la localización recomendada para obtenerlos. Sin embargo, como también publicamos los scipts que permiten llegar a crear el recurso, una versión más volátil se puede encontrar en el repositorio de Gitub.

3.5. Aspectos éticos

No existen conflictos de interés en este proyecto.

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 veredas 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 asociados 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

Los archivos de datos serán almacenados en 3 dispositivos de memoria distintos. Cada dispositivo representará un nivel distinto de accesibilidad y rapidez de lectura y sobrescritura. Se creará un plan para almacenar los datos durante los próximos 20 años. En este periodo de tiempo los datos están sujetos a cambios según surjan mejoras en su calidad.

El primer repositorio será de acceso público y disponible en la nube para que cualquier usuario de la comunidad pueda acceder a los datos *on the go*. Usaremos la plataforma de GitHub para almacenar el repositorio. De esta forma, la comunidad puede participar libremente en la conservación, calidad y contraste de los datos con otras fuentes. Los datos se actualizarán frecuentemente (cada 6 meses) según se actualizen las bases de datos fuente y contribuciones por parte de la comunidad.

En el segundo nivel utilizaremos discos duros para almacenar los datos y los resultados. Utilizaremos la interfaz SATA puesto que es el más rentable en cuanto a espacio. Se montarán sobre una arquitectura RAID 6. De esta forma podemos garantizar la protección de los datos por imperfecciones en los discos utilizando paridad por bloques de memoria en el disco.

Este dispositivo será de acceso exclusivo para los colaboradores del proyecto y recogerá los cambios realizados al repositorio en la nube. También sirve como un backup a los datos recogidos en la nube si estos se pierden.

El último nivel serán cintas magnéticas para el almacenaje. Es el método de más rentabilidad para largo periodo. Como no es un repositorio muy extenso, la prioridad es encontrar el tipo de cinta más rentable (una cinta LTO-1 son $\approx$ 20 euros (Backupworks.com (2020))). En esta cinta se almacenarán todos los datos de tal forma que si los datos de la nube se pierden, los backups realizados en el disco duro junto con la información en las cintas pueden recuperar el repositorio actualizado.

Por otra parte, los datos obtenidos en este informe serán subidos a la plataforma web de Zenodo. Gracias a ella, se podrán acceder a sus metadatos de forma limpia y además también con el uso de scripts y la API de Zenodo. Estos son los datos recomendados a usar ya que los encontrados en GitHub son producidos por los scripts del mismo repositorio y sujetos a cambios.

6. Análisis de los datos

6.1. Relación entre la producción y el área cosechada

Comenzamos por comparar la producción que se obtiene de una cosecha según el área cosechada. Esto sirve para observar si existe una regla de proporcionalidad entre las dos variables o existe otra tendencia producida por otros factores (i.e. la calidad del suelo implica mayor eficiencia en función de la zona y no se cumple una proporción tal que a mayor área mayor producción).

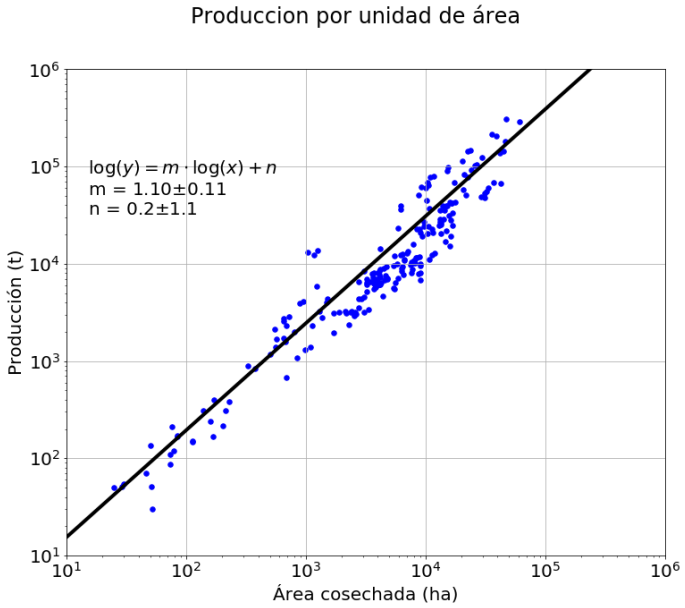

Figura 1: Distribución de los cultivos según el área cosechada en hectáreas y la producción toneladas. La figura se encuentra en escala logarítmica.

En la figura (1) observamos un comportamiento lineal con una pendiente $1{,}10 \pm 0{,}11$ por lo que nos encontramos en el régimen lineal. Se da una relación de proporcionalidad directa entre la producción en toneladas y el área cosechada. No observamos ninguna tendencia que se desvíe de la proporcionalidad.

6.2. Estudio del rendimiento de las cosechas en función de la temperatura

Por otro lado, comparamos el rendimiento, esto es, la masa de maíz generado por unidad de área de maíz cosechado. Esperamos que dependiendo de la temperatura, el rendimiento sea variado y que exista una temperatura que produzca la mayor cantidad de rendimiento de la cosecha.

Un vistazo previo a los datos nos revela que la cantidad de ruido es muy elevado y debemos simplificar enormemente el problema. Decidimos coger intervalos de 2,5ºC y encontramos la media de temperatura y de rendimiento de ese grupo. Luego los representamos en una gráfica (figura (2)).

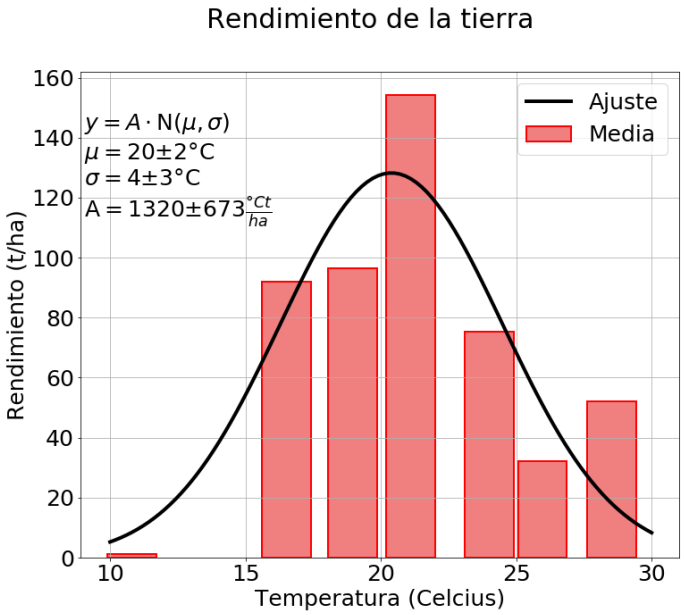

Figura 2: Rendimiento de los cultivos en función de la temperatura en esa época. Los datos han sido agrupados para observar mejor la tendencia y encontrar un pico de rendimiento alrededor de los 20ºC. Las barras agrupan cosechas en intervalos de 2,5ºC y están centradas según sus temperaturas, su altura es el rendimiento medio del grupo.

Observamos que se produce un pico de rendimiento y ajustamos a una forma gaussiana. Aunque no es un ajuste muy adecuado (no aparecen datos alrededor de los 12ºC) encontramos que la temperatura óptima es de 20ºC. El rendimiento potencial de nuestro modelo hallado es de 131 toneladas por hectárea.

7. Conclusiones

Los datos sugieren una temperatura óptima inferior a la sugerida en la literatura (Ospina Rojas y Duarte Pérez (2011)). Esto puede deberse a otros factores que no se han tenido en cuenta, como la precipitación, la presión atmosférica, la temperatura de la tierra (nuestra variable independiente es la temperatura del aire en la superficie) entre otras.

Por otro lado, se ha comprobado la linealidad de la relación entre la producción y el área de cosecha. Con lo que podemos excluir como factor determinante la calidad del suelo en las diferentes regiones colombianas, no habiendo diferencias significativas en el rendimiento de la cosecha.

La arquitectura levantada para el mantenimiento y disposición de los datos juega un papel fundamental en el ámbito de los datos en abierto en la comunidad científica. Permite ayudar a futuros proyectos y colaboraciones para conseguir resultados de interés común.

Referencias

Backupworks.com. (2020). *LTO-1 Ultrium Data Cartridges - 100/200GB LTO1 Tape Media.* Descargado 2020-01-18, de https://www.backupworks.com/LTO1-Ultrium-Tape-Media.aspx

Commission European. (2019). *Extension of the open research data pilot in Horizon 2020.* Descargado 2020-01-12, de https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Creative Commons Corporation. (2019). *Creative Commons — Reconocimiento 4.0 Internacional — CC BY 4.0.* Descargado 2020-01-16, de https://creativecommons.org/licenses/by/4.0/deed.es_ES

Departamento Aministrativo Nacional de Estadística de Colombia, D. (2017). *Nivel de referencia de veredas.* Descargado 2019-12-04, de https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-nivel-de-referencia-de-veredas/

FAO. (2006). Colombia. En *Calendario de cultivos: América latina y el caribe* (Vol. 186, pp. 79–92). Roma: FAO.

GO FARE. (2018). *FAIR Principles - GO FAIR.* Descargado 2020-01-18, de https://www.go

-fair.org/fair-principles/

Gobierno de Colombia. (2018). *Cadena productiva del maíz: área de producción y rendimiento.* Descargado 2019-12-04, de https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Cadena-Productiva-Ma-z-Area-Producci-n-Y-Rendimien/d968-yfb5

Lafitte, H. R. (2001). Fisiología del maíz tropical. En R. L. Paliwa, G. Granado, H. R. Lafitte, y A. D. Violic (Eds.), *El maíz en los trópicos: Mejoramiento y producción.* Roma: FAO. Descargado de http://www.fao.org/docrep/003/x7650s/x7650s00.htm

NASA LP DAAC at the USGS EROS Center. (2019). *Imágenes de temperatura.* Descargado 2019-12-04, de https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD11A1 doi: https://doi.org/10.5067/MODIS/MOD11A1.006

Ospina Rojas, J. G., y Duarte Pérez, C. J. (2011). Fisiología de la planta del maíz. En *Aspectos técnicos de la producción de maíz en colombia* (pp. 33–59). Fenalce.

Diagrama de Gantt del proyecto

Paquete de trabajo	Responsable(s)	2015												2016												2017												2018												2019												2020											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Cultivo de *Zea mays*																																																																									
Contratación de personal	Junta directiva del proyecto	■												■												■												■																																			
Sembrado ZAE1	Agricultores de la región geográfica correspondiente				■					■							■					■							■					■							■					■																											
Sembrado ZAE2	Agricultores de la región geográfica correspondiente			■												■												■												■																																	
Sembrado ZAE3	Agricultores de la región geográfica correspondiente			■						■						■						■						■						■						■						■																											
Sembrado ZAE4	Agricultores de la región geográfica correspondiente		■							■					■							■					■							■					■							■																											
Sembrado ZAE5	Agricultores de la región geográfica correspondiente		■							■					■							■					■							■					■							■																											
Sembrado ZAE6	Agricultores de la región geográfica correspondiente		■							■					■							■					■							■					■							■																											
Cosechado ZAE1	Agricultores de la región geográfica correspondiente									■					■							■					■							■					■							■																											
Cosechado ZAE2	Agricultores de la región geográfica correspondiente								■												■												■												■																												
Cosechado ZAE3	Agricultores de la región geográfica correspondiente								■						■						■						■						■						■						■																												
Cosechado ZAE4	Agricultores de la región geográfica correspondiente							■							■					■							■					■							■					■																													
Cosechado ZAE5	Agricultores de la región geográfica correspondiente							■							■					■							■					■							■					■																													
Cosechado ZAE6	Agricultores de la región geográfica correspondiente							■							■					■							■					■							■					■																													
Recogida de datos de producción	Agricultores y división de Data Science del proyecto																																																																								
Estaciones metereológicas																																																																									
Contratación del servicio con el gobierno de Colombia	Junta directiva del proyecto	■												■												■												■																																			
Recogida de datos de temperatura	División de Data Science del proyecto	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																								
Obtención de imágenes de satélite de Colombia																																																																									
Contratación del servicio con el gobierno de Colombia	Junta directiva del proyecto	■												■												■												■																																			
Procesado de las imágenes por municipios	División de Data Science del proyecto																																																												■												
Integración de los datos de producción, localización geográfica y temperatura																																																																									
Procesado de los datos en R	División de Data Science del proyecto																																																												■												
Procesado de los datos en Python	División de Data Science del proyecto																																																												■												
Análisis de los resultados obtenidos																																																																									
Elección de la mejor región para la producción	División de Data Science del proyecto																																																													■											
Publicación del informe desarrollado																																																																									
Redacción del informe	División de Data Science del proyecto																																																													■	■										
Data Management Plan	División de Data Science del proyecto																																																													■	■										
Creación de los metadatos correspondientes	División de Data Science del proyecto																																																													■	■										
Presentación de la convocatoria a la comisión del Horizonte 2020	Junta directiva del proyecto																																																													■											
Desarrollo de licencia y derechos de uso sobre los datos creados																																																																									
Definición de los términos de uso	Junta directiva del proyecto																																																													■											
Adquisición de una licencia que se adecuade a los términos	Junta directiva del proyecto																																																													■											
Deposito de los datos generados																																																																									
Depósito de los datos e informe en github	División de Data Science del proyecto																																																													■	■										
Depósito de los datos e informe en el repositorio del gobierno de Colombia	División de Data Science del proyecto																																																														■										