A picture containing logo

Description automatically generated

Tesis para “Maestría en explotación de datos y descubrimiento del conocimiento”

Estudio de impactos de sequía en rendimientos de cultivos agrícolas mediante métodos de aprendizaje automático

Reporte de avance para Taller de Tesis I

Alumno: Ing. Santiago Luis Rovere

Director: Dr. Andrés Farall

Profesor de Taller de Tesis I: Dr. Ricardo Maronna

17 de julio de 2021

1. Sinopsis

El presente documento constituye el Reporte de avance N°1 correspondiente a la Tesis de Maestría a ser llevada a cabo por el autor. Este trabajo de tesis forma parte de un proyecto de investigación y desarrollo denominado “Diseño e implementación inicial de un Sistema de Información sobre Sequías para el Sur de América del Sur (SISSA)”, el cual es financiado por el programa de Bienes Públicos Regionales del Banco Interamericano de Desarrollo (Cooperación Técnica RG-T3308, <https://www.iadb.org/es/project/RG-T3308>).

El documento “Plan de Trabajo” presentado previamente incluyó una breve descripción del Proyecto SISSA, la descripción del tema de abordar en la tesis y los objetivos buscados. También se presentó un del plan de trabajo a llevar a cabo para concretar el trabajo de tesis en el término de un año, así como también un cronograma de actividades tentativo. Por tal motivo, esta información no será replicada en este documento, salvo aquello que sea necesario para el desarrollo del presente reporte.

Dado que la naturaleza de este documento es presentar el grado de avance del trabajo de investigación, se comenzará por realizar un resumen del flujo de tareas presentadas en el plan de trabajo. Luego, para cada una de las tareas se detallarán las actividades desarrolladas. Para aquellas tareas que aún no hayan sido abordadas, se presentarán las ideas tentativas que se tienen en mente para su ejecución en el futuro.

Por lo expuesto previamente, se proseguirá con un resumen de las tareas vinculadas al plan de trabajo (Sección 2). El resto de las secciones (Sección 3 a Sección 8) se corresponderán con las principales tareas del plan de trabajo. Finalmente, en la Sección 9 se indicarán las actividades más inmediatas a ser llevadas a cabo.

1. Plan de trabajo

Para llevar a cabo este trabajo de tesis han propuesto una serie de actividades basadas en al diagrama de la Fig. 1. Para poder abordar el estudio de impactos de sequías en rendimientos de cultivos, se deberán buscar regiones agrícolas importantes dentro del área abarcada por el CRC-SAS (<https://www.crc-sas.org>; Argentina, Bolivia, Brasil – debajo de 10°S, Chile, Paraguay y Uruguay). En estas regiones se determinarán ubicaciones puntuales para las cuales existan registros históricos largos (de al menos 30 años) de temperatura y precipitaciones.

Una vez definidas las ubicaciones puntuales que se utilizarán para el presente estudio, se deberá recabar información acerca de las actividades agrícolas más importantes, los manejos típicos y los tipos de suelos predominantes para cada zona. Esta información permitirá caracterizar y parametrizar las actividades agrícolas y los cultivos cuyos ciclos de vida serán simulados haciendo uso de DSSAT a partir de numerosas series temporales de precipitación y temperatura.

Como se mencionó en el párrafo previo, es necesario contar con numerosas series temporales de precipitación y temperatura para poder ejecutar las simulaciones de los ciclos de vida de los cultivos. Para ello se generarán series temporales estocásticas de precipitación y temperatura que tengan las mismas propiedades estadísticas que las series históricas originales. Esto se llevará a cabo haciendo uso del paquete de R *gamwgen* (<https://github.com/CRC-SAS/weather-generator>) que fue desarrollado, en parte, por integrantes del proyecto SISSA y de otros proyectos anteriores.

Haciendo uso de las series temporales estocásticas generadas y las parametrizaciones de actividades agrícolas (cultivos, manejos, tipos de suelo, etc.), se ejecutarán las simulaciones de los ciclos de vida de los cultivos correspondientes. Esto significa que para cada serie temporal de una campaña agrícola se asociará un rendimiento resultante, producto de la simulación.

Cada una de las series temporales generadas deberá transformarse en un conjunto de atributos que permitan definir condiciones de sequía para cada momento del ciclo de vida del cultivo. Este proceso se realizará mediante el cálculo de eventos basados en índices de sequía actualmente utilizados por el SISSA (<https://sissa.crc-sas.org/monitoreo/indices-de-sequia/>). A través de este proceso de *feature engineering* se podrá construir un conjunto de datos tabular con atributos y resultados.

Una vez que se haya logrado construir un conjunto de datos tabular con atributos y resultados, será posible aplicar diversos modelos estadísticos y de aprendizaje automático que permitan vincular los atributos (los cuales definen condiciones de sequía) con los rendimientos asociados. Como todo proceso de *data mining*, deberá ser llevado a cabo de manera iterativa e interactiva hasta lograr los objetivos propuestos.

Fig. 1. Pipeline conceptual del plan de trabajo propuesto

|  |
| --- |
| Diagram  Description automatically generated |
|  |

En las secciones subsiguientes se describirán cada uno de los componentes previamente mencionados. Cabe destacar que, dada la naturaleza colaborativa y multidisciplinaria del Proyecto SISSA, algunos de los componentes que se han listado requieren de la intervención de ciertos expertos de campo. Por ejemplo, para la parametrización de los tipos de suelos y de manejos de cultivos, será necesaria la intervención de ingenieros agrónomos que también participan en el proyecto.

También es importante destacar que las tareas no serán realizadas en un estricto orden secuencial, sino que, por cuestiones de interacción entre los integrantes del equipo, algunas de ellas se irán realizando de acuerdo con la dinámica de desarrollo del Proyecto SISSA. Para la ejecución de otras tareas también se hará uso de algunas herramientas y metodologías de cálculo desarrolladas en proyectos previos al SISSA (por ejemplo, para el cálculo de índices de sequía, como se describirá más adelante).

Finalmente, y antes de comenzar la exposición de cada uno de los componentes, debe mencionarse que el grado de avance alcanzado es muy heterogéneo. En algunos casos, aún no se han llevado a cabo actividades vinculadas a varios de los componentes. En otros casos, el grado de avance ha sido considerablemente mayor. Por este motivo, algunas de las secciones detalladas a continuación incluirán mucho menor contenido que otras.

1. Definición de zonas de estudio

Como se mencionó al inicio de este documento, el trabajo de tesis se encuentra enmarcado dentro del Proyecto SISSA. Este proyecto aborda, entre otras cosas, el estudio de la sequía dentro del área abarcada por el Centro Regional del Clima para el Sur de América del Sur (CRC-SAS). Por este motivo, es de interés para el proyecto SISSA, realizar un estudio de impactos de sequía para las zonas agrícolas más importantes de estos países.

De este modo, la primera etapa de esta tarea consta de una revisión de las actividades agrícolas para los 6 países definidos previamente con la finalidad de identificar las regiones y los cultivos agrícolas más importantes para cada país. Esta tarea está siendo liderada por un ingeniero agrónomo que forma parte del equipo del proyecto.

Como parte de la tarea de identificación de las zonas agrícolas importantes, también es necesario evaluar si existen datos suficientes para poder abordar el estudio de impactos de sequías en cada región. Los datos necesarios para llevar cabo el estudio son los siguientes: (a) series temporales diarias de precipitaciones, temperaturas máximas y mínimas y (b) información de actividades agrícolas, manejos típicos de cultivos y clasificación de suelos.

Los datos diarios de precipitaciones y temperaturas son necesarios para poder generar una cantidad importante de series temporales sintéticas de clima con las mismas propiedades estadísticas que las series históricas. Estas series sintéticas permitirán, a su vez, simular una cantidad considerable de rendimientos para estudiar el vínculo entre las condiciones de sequía y los impactos en términos de rendimientos resultantes.

Si bien se dispone de una gran cantidad de información climática para la región del CRC-SAS, la mayoría de ella proviene de sensores automáticos, los cuales cuentan con unos pocos años de datos. Debe tenerse en cuenta que, para caracterizar el clima de una región determinada, es necesario contar con series climáticas largas (de al menos 30 años, según la Organización Climática Mundial) para poder incluir la mayor variabilidad posible.

Esta restricción nos obliga a utilizar datos de estaciones climáticas convencionales, las cuales tienen registros que exceden la cantidad de años necesaria (en general, desde 1961 a la fecha). Sin embargo, las estaciones meteorológicas convencionales carecen de una buena cobertura espacial (son bastante escasas en relación con las estaciones automáticas) y además muchas veces tienen datos faltantes. La selección de las ubicaciones puntuales de estaciones meteorológicas convencionales constituye el primer filtro necesario dentro de la presente tarea.

Por otro lado, para poder realizar las simulaciones de rendimientos con el software DSSAT, también es necesario definir ciertos parámetros asociados al cultivo a simular, su manejo agronómico y el tipo de suelo donde transcurre su ciclo de vida. Esta búsqueda debe ser realizada para aquellas ubicaciones puntuales que hayan sido filtradas a partir de la selección previa (según los registros climáticos).

Hasta el momento, y dado que es el país con menor extensión territorial de los 6 enumerados, Uruguay es el único para el cual se han definido las ubicaciones puntuales en las que se llevará a cabo el estudio de impactos de sequía. Las ubicaciones seleccionadas corresponden a cinco estaciones meteorológicas convencionales y pueden visualizarse en el mapa de la Fig. 2.

Fig. 2. Mapa de ubicaciones puntuales de Uruguay seleccionadas para el estudio de impactos de sequía. Estas ubicaciones corresponden a estaciones meteorológicas convencionales.

|  |
| --- |
| Mapa  Descripción generada automáticamente |
|  |

1. Generación de series sintéticas de clima
2. Parametrización de tipos de suelo y manejos de cultivos
3. Simulación de rendimientos de cultivos

Como se observa en la Fig. 1, para poder simular los rendimientos de los cultivos, es necesario realizar una parametrización de los tipos de suelos y manejos utilizados en las ubicaciones definidas (y a definir) para este estudio. Sin embargo, esta tarea aún no se ha comenzado. Para ello es necesaria la coordinación con los ingenieros agrónomos que forman parte del Proyecto SISSA, dado que tal actividad requiere del conocimiento de expertos en esta área.

También es necesario contar con series temporales de precipitación, temperatura (mínima y máxima) y radiación solar diaria para cada una de dichas ubicaciones (por requerimiento del software que simula los rendimientos). Sin embargo, debe recordarse que el generador de clima sintético solamente es capaz de generar series temporales de temperatura y precipitación, pero no de radiación solar.

Sim embargo, debe destacarse que la radiación solar puede calcularse haciendo uso de varios modelos, de diversa complejidad y precisión. Pero debe tenerse en cuenta también, que los modelos más precisos requieren de información sobre nubosidad, heliofanía u otras variables meteorológicas cuya información no está disponible a partir de las series climáticas generadas. Por tal motivo, se ha seleccionado el modelo Bristow-Campbell (Bristow K. L., Campbell G. S., 1984), el único que permite estimar valores de radiación solar a partir de valores resultantes del proceso de generación estocástica (temperaturas máxima y mínima).

Este modelo además incluye ciertas constantes paramétricas que pueden ser modificadas mediante un proceso de calibración a partir de valores de radiación medidos *in situ*. La calibración de estas constantes deberá ser realizada, en general, para todas las zonas agrícolas donde se vayan a ejecutar los procesos de generación estocástica de series climáticas.

Esta calibración solamente ha sido llevada a cabo para las ubicaciones definidas para Uruguay, permitiendo de este modo el cálculo de las series de radiación diaria asociadas a las series temporales de precipitación y temperaturas generadas estocásticamente. Este mismo proceso deberá realizarse para las zonas agrícolas de Argentina, Brasil, Chile y Paraguay.

Una vez que se puedan calcular las series temporales de radiación diaria y se hayan parametrizado los tipos de suelos y manejos típicos de cada una de las zonas agrícolas, se podrá comenzar la ejecución masiva de procesos de simulación (uno por cada año calendario simulado para cada ubicación). Para ello, deberá considerarse la implementación de un proceso que permita paralelizar la ejecución haciendo uso de múltiples nodos, con el fin de disminuir considerablemente el tiempo de procesamiento.

1. Caracterización de series sintéticas de clima

Para llevar adelante un proceso de aprendizaje automático que permita vincular condiciones de sequía con rendimientos de cultivos, se deberán efectuar transformaciones sobre las series de clima simulado de modo que las mismas permitan caracterizar eventos o condiciones de sequía. Este proceso requiere, por lo tanto, que se puedan cuantificar las condiciones de sequía mediante ciertos indicadores numéricos. Sin embargo, en primera instancia debe indicarse qué se entiende por sequía.

No existe una única definición de sequía, debido a que este fenómeno se identifica por sus efectos o impactos sobre diferentes tipos de sistemas (agricultura, recursos hídricos, ecosistemas, economía, etc.). Los principales tipos de sequías son:

* Meteorológica: escasez de precipitación. Este tipo se sequías es el causante de otro tipo de sequías;
* Agrícola: escasez de agua para satisfacer las necesidades de un cultivo;
* Hidrológica: deficiencia de la disponibilidad de agua de superficie y/o subterránea. Se desarrolla más lentamente, debido a que hay un retraso entre la falta de lluvia y la reducción de agua en arroyos, ríos, lagos, embalses, etc.; y
* Socioeconómica: escasez hídrica con consecuencias sociales y económicas desfavorables. Es una consecuencia de los otros tipos de sequía y es claramente económica.

Los *indicadores de sequía* son variables o parámetros utilizados para describir las condiciones de las sequías. Cabe citar, por ejemplo, la precipitación, la temperatura, los caudales fluviales, los niveles de las aguas subterráneas y de los embalses, la humedad del suelo y el manto de nieve.

Los *índices de sequía* son medidas cuantitativas que se utilizan para caracterizar los niveles de sequía mediante la asimilación de uno o más indicadores de sequía. Suelen ser representaciones numéricas informatizadas de la gravedad de las sequías, determinadas mediante datos climáticos o hidrometeorológicos, entre los que se incluyen los indicadores enumerados. Tienen por objeto analizar el estado cualitativo de las sequías en el entorno en un período de tiempo determinado. Desde el punto de vista técnico, los índices también son indicadores.

Debido a la magnitud de sus impactos en la región, una de las principales líneas de trabajo del Centro Regional del Clima para el sur de América del Sur (CRC-SAS) ha sido el desarrollo de un sistema de vigilancia de sequías, para lo cual se ha implementado el cálculo de algunos índices de sequía, de los cuales pueden destacarse el SPI y SPEI por su uso e importancia.

## 7.1 Índice de Precipitación Estandarizado (SPI)

El Índice de Precipitación Estandarizado (IPE, o SPI por sus siglas en inglés) cuantifica las condiciones de déficit o exceso de precipitación en un lugar y para una escala determinada de tiempo. El SPI fue desarrollado por McKee et al. (1993) con la finalidad de mejorar la detección del inicio y el monitoreo de la evolución de las sequías meteorológicas (definidas únicamente en función de la precipitación). La principal ventaja de este índice es que su cálculo requiere una única variable climática para el cálculo: la precipitación. El SPI ha sido utilizado ampliamente a nivel mundial y es recomendado por la Organización Meteorológica Mundial.

El primer paso para el cálculo del SPI es el cálculo de los totales acumulados de precipitación correspondientes al mes/año y escala temporal deseada (ver discusión de escalas temporales más abajo). Luego, se ajusta una distribución teórica a los totales de cada mes en el período de referencia. En este proyecto se utiliza la distribución *Gamma*, debido a que ajusta adecuadamente las distribuciones empíricas de totales de precipitación para la mayoría de los meses y estaciones consideradas, y además porque requiere solamente dos parámetros para su especificación: *alfa* (α, parámetro de forma) y *beta* (β, parámetro de escala).

Los parámetros estimados se usan luego para calcular el percentil correspondiente a los distintos valores de precipitación acumulada cuyo índice se desea calcular. Finalmente, los valores de SPI resultantes son los cuantiles correspondientes a estos percentiles para una distribución normal estandarizada (con media = 0 y desvío estándar = 1).

## 7.2 Índice de Precipitación – Evapotranspiración Estandarizado (SPEI)

El Índice de Precipitación – Evapotranspiración Estandarizado (IPEE o SPEI por sus siglas en inglés), es un índice cuyo cálculo es similar al del SPI, pero incorpora el efecto de la evapotranspiración (i.e., la demanda atmosférica de agua) que influye en las condiciones de sequía. El SPEI fue desarrollado por Vicente-Serrano et al. (2010), pero ya ha sido utilizado para analizar distintas características de la sequía, como ser su variabilidad, impactos y mecanismos atmosféricos que la producen (Beguería et al., 2013; Hernandez and Uddameri, 2013; Vicente-Serrano et al., 2015; Xu et al. 2015).

El SPEI utiliza como valor de entrada al balance hídrico (o sea, la diferencia entre precipitación y evapotranspiración potencial o PET por sus siglas en inglés). El cálculo de la evapotranspiración potencial es complicado, debido a que involucra muchos parámetros (temperatura, humedad del aire, viento y radiación, entre otros). En este proyecto se ha utilizado la ecuación de Hargreaves-Samani (1985), que es eficiente en el cálculo de la evapotranspiración potencial utilizando sólo medias mensuales de temperatura máxima y mínima, y radiación solar.

Dado que el balance entre precipitación y evapotranspiración puede tomar valores negativos, el ajuste del mismo se realiza con la distribución teórica Log-Logística, que acepta valores nulos y negativos. El método para ajustar estos parámetros a la distribución es el de máxima verosimilitud (Beguería et al., 2013).

## 7.3 Implementación del cálculo de índices

Los efectos de las sequías se manifiestan en diferentes escalas temporales, ya que las respuestas de diferentes sistemas hidrológicos y biológicos a las anomalías de precipitación varían mucho (Ji and Peters, 2003). Es decir, puede haber grandes diferencias en la duración de los déficits hídricos necesarios para causar impactos negativos en diferentes sistemas.

Como la sequía es un fenómeno multiescalar, es necesario el uso de indicadores que puedan capturar adecuadamente las escalas temporales relevantes para detectar impactos negativos sobre los diferentes sistemas de interés. Por ejemplo, la sequía agrícola – definida como la escasez de agua para satisfacer las necesidades de un cultivo – puede ser bien representada por las escalas de 2 y 3 meses, mientras que déficits en los caudales de río o arroyos se reflejan mejor por medio de las escalas de 3 a 6 meses. Asimismo, se han encontrado asociaciones entre la variación del nivel de la napa freática y los valores de los índices con escalas de 6 a 24 meses.

Los índices de sequía que se han descripto en secciones anteriores están siendo calculados para múltiples escalas temporales: 1, 2, 3, 6, 9, 12, 18, 24, 36 y 48 meses. Es decir, los índices se basan en series de precipitaciones acumuladas para cada escala. Por ejemplo, para calcular un índice con una escala de 6 meses para junio de 2003, se considera la suma de los valores mensuales de precipitación desde enero hasta junio de 2003; este valor se pone en el contexto de los totales de lluvia para enero-junio en el registro histórico.

De la misma forma, se puede realizar el cálculo para otras escalas temporales. Para el cálculo del SPEI, que requiere temperaturas máximas y mínimas para estimar la evapotranspiración, se usa el promedio de estas temperaturas para cada escala temporal. Siguiendo con el ejemplo anterior, el cálculo de SPEI para una escala de 6 meses para junio de 2003 utilizará los promedios mensuales de temperaturas máximas y mínimas desde enero hasta junio de 2003.

Para el cálculo de los diferentes índices de sequía, generalmente es necesario estimar parámetros o cuantiles de la distribución empírica o teórica de precipitación. Esta estimación se puede repetir cada vez que se agregan datos nuevos a las series, es decir, al actualizar la base de datos del CRC-SAS. La desventaja de este enfoque es que los valores previos de un índice cambian a medida que las series se extienden, ya que cada vez el ajuste o estimación se hace con series diferentes (más largas). Una alternativa para evitar que los valores anteriores cambien constantemente es utilizar un período de referencia fijo para estimar parámetros o cuantiles necesarios. El uso de un período de referencia implica que, al agregar más registros a las series climáticas, los valores anteriores no cambiarán ya que los índices se calculan a partir de parámetros o cuantiles estimados para el período de referencia que no se modifica.

Es deseable que el período de referencia contenga la mayor parte de la variabilidad de las series climáticas. La ventaja de contar con registros largos es que permiten capturar oscilaciones de baja frecuencia (épocas secas y húmedas asociadas con la variabilidad climática multidecadal). Los registros más cortos, en cambio, podrían contener sólo parte de esa variabilidad, y los valores de un índice podrían estar sesgados por el uso de un período de referencia seco o húmedo. Para todos los índices de sequía calculados en base a datos in situ de estaciones meteorológicas convencionales, se utilizó el período 1971-2010 (40 años) como período de referencia.

1. Vinculación de condiciones de sequía con rendimientos

Una vez que se cuente con un conjunto de datos organizado de forma tabular (donde los atributos corresponden a condiciones de sequía en distintos momentos del cultivo y la variable dependiente a rendimientos resultantes) será posible llevar a cabo procesos de aprendizaje automático para encontrar patrones vinculantes entre condiciones de sequía y rendimientos de cultivos.

Si bien la ejecución de esta tarea aún está lejos en el tiempo, se ha comenzado a investigar acerca del estado del arte de esta área de estudio. En particular, se ha comenzado por (van Klompenburg, et al., 2020). Este trabajo es una revisión sistemática de 50 trabajos de investigación en esta temática. A partir de este trabajo (y de los otros trabajos a que hace referencia), se buscarán las alternativas óptimas para realizar esta tarea. Por el momento, solamente se puede adelantar que la gran mayoría de los trabajos de *machine learning* sobre rendimientos de cultivos, se han llevado a cabo mediante regresiones y clasificaciones basados en redes neuronales profundas y el ensamble *XGBoost*.

1. Próximas actividades

Referencias

Bristow K. L., Campbell G. S. (1984). *On the relationship between incoming solar radiation and daily maximum and minimum temperature*. Agricultural and Forest Meteorology, Volume 31, Issue 2, 159-166. <https://doi.org/10.1016/0168-1923(84)90017-0>.

van Klompenburg T., Kassahun A., Catal C. (2020). *Crop yield prediction using machine learning: A systematic literature review*. Computers and Electronics in Agriculture, Volume 177. <https://doi.org/10.1016/j.compag.2020.105709>.