
# 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*

Plan de trabajo

Alumno: Santiago Luis Rovere (Ingeniero en informática, UBA)

Director: Andrés Farall (Doctor en ciencias de la atmósfera, UBA)

8 de julio de 2021

# 1 Sinopsis

El presente documento constituye el Plan de Trabajo para 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 [1]).

Por lo expuesto previamente, y a fin de mejorar la motivación del tema de estudio (Sección 2), se proseguirá con una descripción del proyecto SISSA, su misión y principales objetivos (Sección 2.1). A continuación, se presentará formalmente el tema que se desarrollará en este trabajo de tesis y la importancia de su abordaje (Sección 2.2). Se concluirá la Sección 2 con la definición del objetivo del trabajo de tesis y una explicación sobre la transferencia de los resultados encontrados (Sección 2.3).

Una vez presentado el tema de estudio, en la Sección 3 se describirá el plan de trabajo propuesto para llevar a cabo la investigación. Finalmente, en la Sección 4 se definirá un cronograma de tareas con el fin estructurar el plan de trabajo propuesto en tareas que tengan estipuladas una duración aproximada.

# 2 Tema de estudio

## 2.1 ¿Qué es el SISSA?

El SISSA es el Sistema de Información sobre Sequías para el sur de Sudamérica. El SISSA provee herramientas e información (https://sissa.crc-sas.org) sobre las sequías y sus impactos a gobiernos, instituciones no gubernamentales y privadas, e individuos. Esta información permite:

- monitorear y predecir la ocurrencia de sequías;
- anticipar los impactos esperables en sectores económicos y comunidades; y
- fomentar la planificación y preparación anterior a la ocurrencia de sequías para mitigar sus daños, aumentar la resiliencia y reducir la vulnerabilidad.

El propósito último del SISSA es reemplazar las acciones *reactivas* (posteriores) a una sequía por un enfoque *proactivo* que permita gestionar los riesgos y reducir vulnerabilidades.

El SISSA es una institución virtual que funciona en el marco del Centro Regional del Clima para el sur de América del Sur (CRC-SAS, https://www.crc-sas.org). El CRC-SAS es una organización constituida en forma de red, según los principios definidos por la Organización Meteorológica Mundial (OMM, https://public.wmo.int/es). Se encuentra en fase operativa y ofrece servicios climáticos en apoyo a los Servicios Meteorológicos e Hidrológicos Nacionales (SMHN) y otros usuarios de los países situados en la región sur de América del Sur.

## 2.2 Estudio de impactos de la sequía en el rendimiento de los cultivos

Los eventos de sequía constituyen el factor más importante asociado a la disminución del rendimiento de los cultivos del sector agrícola. A su vez, esta disminución en el rendimiento de los cultivos se traduce en millonarias pérdidas para el sector y la economía en general. Es por ello por lo que resulta de interés el estudio de este fenómeno con el objetivo de definir políticas de mitigación y gestión de riesgos.

Para poder llevar a cabo acciones de gestión y mitigación, se requiere una comprensión de los vínculos entre las características de los eventos de sequía (por ejemplo, momento de ocurrencia, duración, intensidad, etc.) y sus impactos específicos sobre el rendimiento de los cultivos. El propósito de este trabajo será generar información cuantitativa que vincule condiciones de sequía con rendimientos obtenidos para cultivos comercialmente importantes.

Generar tal información será fundamental para diseñar un sistema de alerta temprana para el sector agrícola. Para emitir una alerta temprana es necesario, además, cuantificar o estimar los posibles impactos con cierto tiempo de antelación con el fin de gestionar los riesgos encontrados y llevar a cabo acciones de mitigación de dichos impactos. Para evaluar los impactos, se buscará vincular diferentes tipos de eventos secos (cortos pero intensos, largos y poco intensos, etc.) que ocurran en los distintos momentos del ciclo del cultivo con los rendimientos resultantes.

Un enfoque tradicional para abordar este estudio sería mediante la vinculación de variables meteorológicas históricas que definan condiciones de sequía y los resultados reales obtenidos. Sin embargo, abarcar todas las posibles combinaciones de tipos de eventos, momentos del ciclo de vida del cultivo y regiones geográficas (las cuales a su vez presentan distintos tipos de suelo y manejos de los cultivos) resulta una tarea imposible de ser realizada con series históricas, dado que éstas no incluyen la diversidad condiciones que se quieren estudiar.

La alternativa propuesta se basa en la utilización del modelo de simulación DSSAT (https://dssat.net). Este modelo representa el crecimiento y rendimiento del cultivo, y además provee información sobre variables intermedias (ej. porcentaje de agua útil, estrés del cultivo, etc.). El uso de DSSAT además permite ignorar efectos de confusión que pueden agregar variación al rendimiento (enfermedades, malezas, plagas, innovaciones tecnológicas, etc.) causadas por factores no relacionados directamente con la sequía.

DSSAT es una herramienta de software que está compuesta por un conjunto de módulos, cada uno de los cuales tiene el propósito de modelar un determinado fenómeno. A su vez, cada módulo está compuesto por varios submódulos que representan procesos biológicos relativamente simples. Estos procesos están descriptos por modelos basados en ecuaciones diferenciales en derivadas parciales.

Un primer enfoque para vincular condiciones de sequía (datos de entrada) con rendimientos resultantes (datos de salida), sería mediante el estudio de las ecuaciones diferenciales que gobiernan cada uno de los procesos que componen el DSSAT. Sin embargo, esta tarea sería operacionalmente imposible debido a la dimensionalidad de los datos de entrada y la complejidad del modelo integrado resultante (que, a su vez, contiene un considerable número de parámetros asociados a manejos del cultivo, tipo de suelos, etc.).

El problema de la dimensionalidad radica en el gran número de variables de entrada, dado que es necesario definir valores de temperatura (mínima y máxima) y precipitación para cada uno de los días del ciclo de vida del cultivo. Cada uno de estos valores representa un dato de entrada, por lo que el modelo resultante puede considerarse una composición de funciones dependientes de algunos cientos de variables. A su vez, esta composición de procesos (que son más de una decena) implica que el modelo, visto como un todo, resulta ser un sistema muy complejo como para ser estudiado por métodos analíticos tradicionales.

Por lo expuesto previamente, resulta más adecuado aplicar el *método de Montecarlo* para abordar el estudio de los efectos de la sequía en los rendimientos de los cultivos. Este método es utilizado para aproximar expresiones matemáticas complejas y costosas de evaluar con exactitud. Para ello se requiere la generación aleatoria de numerosos conjuntos de datos de entrada (variables climáticas

que permitan definir condiciones de sequía), los cuales puedan ser vinculados a datos de salida (rendimientos de cultivos) mediante procesos de simulación llevados a cabo mediante el uso de DSSAT.

Luego de obtener los resultados de las simulaciones, es necesario llevar a cabo tareas de análisis para vincular las condiciones de sequía derivadas de las series temporales de variables meteorológicas de entrada con los rendimientos obtenidos. Considerando la dimensionalidad de los datos de entrada y la cantidad de condiciones que se quieren simular, resulta adecuado aplicar técnicas de análisis basadas en procesos de aprendizaje automático.

## 2.3 Objetivo buscado y transferencia de resultados

El objetivo principal del trabajo propuesto es encontrar patrones que vinculen condiciones de sequía (definidos por su momento de ocurrencia, duración, intensidad, etc.) con rendimientos de cultivos. A través del conocimiento adquirido también se buscará, como objetivo secundario, cuantificar los impactos de las sequías en términos de resultados económicos vía los rendimientos de los cultivos.

Adquirir este tipo de conocimiento sobre los impactos de las condiciones de sequía y lograr una cuantificación de dichos impactos, sería un resultado de gran utilidad para el proyecto SISSA. Tales resultados podrán ser utilizados para la construcción de un sistema de alerta temprana de impactos de sequía. Este sistema podría implementarse, por ejemplo, mediante el desarrollo de una aplicación de visualización interactiva similar a las que actualmente existen en la sección *Monitoreo* del sitio web del SISSA (por ejemplo, https://sissa.crc-sas.org/monitoreo/estado-actual-de-la-sequia/).

Como se mencionó en la Sección 2.1, el propósito último del SISSA es reemplazar las acciones *reactivas* (posteriores) a una sequía por un enfoque *proactivo* que permita gestionar los riesgos y reducir vulnerabilidades. Por lo tanto, la implementación de un sistema de alerta temprana basado en los resultados derivados de este trabajo de tesis permitirá avanzar hacia la concreción de este objetivo.

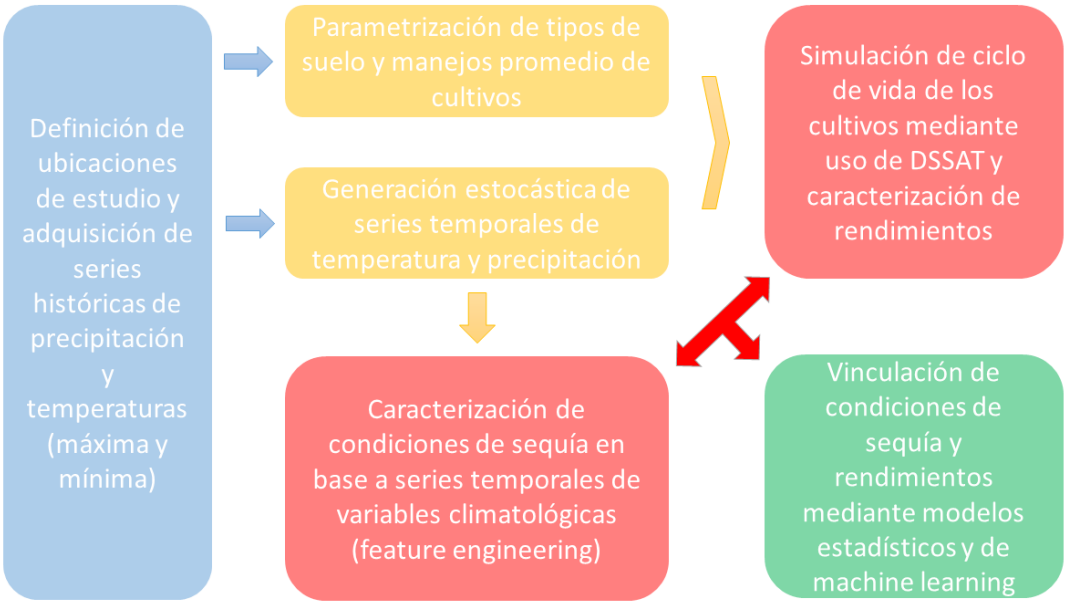
# 3 Plan de trabajo

Para llevar a cabo este trabajo de tesis se proponen una serie de actividades basadas en al diagrama de la Fig. 1. En primera instancia, se buscarán regiones agrícolas importantes dentro del área abarcada por el CRC-SAS (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* [2] que fue desarrollado, en parte, por integrantes del proyecto SISSA y de otros proyectos anteriores.

Fig. 1. Pipeline conceptual del plan de trabajo propuesto




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 [3]. 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. A través de todo el proceso de investigación, se irán documentando las tareas realizadas, que serán parte del trabajo de tesis a presentar.

# 4 Cronograma de tareas

Considerando el plan de trabajo descripto en la Sección 3, se propone el siguiente cronograma de actividades (Tabla 1) para completar el trabajo de tesis en el plazo de aproximadamente un año (con una fecha de finalización estimada para fines de julio de 2022). Debe tenerse en cuenta, sin embargo, la naturaleza interactiva y iterativa de los procesos de descubrimiento de conocimiento. Por este motivo, la mayoría de las actividades deberán ser revisitadas de acuerdo con los resultados que se vayan obteniendo.

Tabla 1. Cronograma de actividades propuestas como parte del plan de trabajo.

| # | Actividad | 2021 | | | | | | 2022 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *A* | *S* | *O* | *N* | *D* | *E* | *F* | *M* | *A* | *M* | *J* | *J* |
| *1* | Determinación de las ubicaciones puntuales (considerando zonas de importancia agrícola y disponibilidad de datos) para llevar a cabo los estudios de impactos de sequía. | | | | | | | | | | | | |
| *2* | Generación de series estocásticas de clima con propiedades estadísticas acordes a las series históricas. Elaboración de diagnósticos. | | | | | | | | | | | | |
| *3* | Caracterización y parametrización de actividades agrícolas, manejos de cultivos y tipos de suelo para las ubicaciones seleccionadas. | | | | | | | | | | | | |
| *4* | Ejecución de simulaciones de rendimientos para las ubicaciones seleccionadas haciendo uso de las series temporales generadas y los parámetros de actividades, manejos y tipos de suelos definidos. | | | | | | | | | | | | |
| *5* | Construcción de conjunto de datos tabular para caracterizar las condiciones de sequía y los rendimientos obtenidos (*feature engineering*). | | | | | | | | | | | | |
| *6* | Implementación de modelos de aprendizaje automático para vincular condiciones de sequía con rendimientos obtenidos. | | | | | | | | | | | | |
| *7* | Elaboración de diagnósticos en base a modelos de aprendizaje automático con el fin de encontrar patrones que permitan explicar los vínculos entre las condiciones de sequía y los impactos en el rendimiento. | | | | | | | | | | | | |
| *8* | Escritura del trabajo de tesis | | | | | | | | | | | | |

# Referencias


[1] RG-T3308: Diseño e Implementación Inicial de un Sistema de Información sobre Sequías para el sur de América del Sur, https://www.iadb.org/es/project/RG-T3308.

[2] Paquete *gamwgen*: https://github.com/CRC-SAS/weather-generator

[3] Índices de sequía: https://sissa.crc-sas.org/monitoreo/indices-de-sequia/