UNIVERSIDAD INTERNACIONAL DE LA RIOJA

Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos

Comparativa de Técnicas de Aprendizaje Automático para Análisis y Predicción de Sequía

Trabajo Fin de Máster

Tipo de trabajo: Comparativa de Soluciones

Presentado por: Andrade Zeballos, María Angélica

Director/a: Fernández Isabel, Alberto

Resumen

La sequía es un fenómeno natural complejo, que se constituye en un foco de interés a nivel mundial, sobre todo por las enormes pérdidas económicas y humanas que provoca en las regiones afectadas. Por ésta razón es importante comprender el comportamiento de la sequía y mucho más importante encontrar los mecanismos que permitan predecir su ocurrencia y así dotar a las autoridades de información para la elaboración de planes de prevención y mitigación. Para la medición de sequía generalmente se utilizan los índices, en el caso de éste proyecto se utilizan dos: el Índice Estandarizado de Precipitación y el Índice Estandarizado de Precipitación y Evapotranspiración, que requieren para su cálculo datos meteorológicos de precipitación y temperatura. Ambos índices se calculan a distintas escalas de tiempo: 3 y 12 meses. Los índices de sequía se constituyen en las variables a predecir. También se incluyen otras variables meteorológicas como ser temperatura, humedad relativa y el Índice Oceánico del Niño 3.4 y juntas se constituyen en variables exógenas que ayudan a predecir los índices mediante el uso de algoritmos predictivos de aprendizaje automático. Los datos provienen de 24 estaciones meteorológicas correspondientes al Altiplano Central de Bolivia, entre los años 1981 y 2020. Para la evaluación de la comparativa se utilizan las métricas: Error Medio Absoluto, Error Medio Absoluto Porcentual, Error Cuadrático Medio y Raíz Cuadrada del Error Cuadrático Medio.

Palabras Clave: Sequía, predicción, aprendizaje automático, machine learning, redes neuronales, SPI, SPEI

Abstract

Drought is a complex natural phenomenon and it is a focus of interest around the world, specially due to enormous economic and human losses it causes in affected regions. For this reason, it is important to understand its behavior, and more important to find mechanisms that allow predict its occurrence, in order to provide authorities with information to design prevention plans. Usually, indices are used to measure drought, two of them are used in this project: Standardized Precipitation Index and Standardized Precipitation Evapotranspiration Index, which require for their calculation meteorological data on precipitation and temperature. Both indexes are calculated at different time scales: 3 and 12 months. Drought indexes constitute the variables to be predicted. Also, others meteorological variables are included: temperature, relative humidity, and the Oceanic Niño Index 3.4, and together they constitute exogenous variables that contribute to predict the indices by using machine learning predictive algorithms. The data comes from 24 meteorological stations corresponding to the Bolivian Central Altiplano, between the years 1981 and 2020. The metrics used to evaluate the comparison are: Mean Square Error, Root Mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error

Keywords: Drought, predicting, forecasting, machine learning, neural networks, SPI, SPEI

Índice de Contenidos

Índice de tablas

Índice de figuras

1. Introducción

La sequía es un fenómeno natural bastante complejo y sus consecuencias negativas a nivel mundial no solo son de tipo económico, también genera pérdidas humanas. En la Figura 1 se aprecia el gran porcentaje de muertes que provocó la sequía en el mundo, considerando el periodo de 1970 a 2019. Las muertes por sequía fueron el 34% del total de muertes relacionadas con desastres, una cifra solo superada por el 38 % causada por ciclones tropicales (OMM, 2021).

1.1. Justificación

Al igual que en todo el mundo la problemática de la sequía afecta de gran manera a Sudamérica, causando grandes pérdidas económicas y humanas. En la región que comprende el altiplano de Perú y Bolivia en la década de los ochenta, específicamente entre los años 1982 y 1983 se registró una sequía severa que provocó un déficit de producción en cultivos de papa y quinua, también causó pérdidas en el sector ganadero. En la zona boliviana fue damnificada la cuarta parte de la población y los daños económicos fueron de aproximadamente 500 millones de dólares (Andrade et al., 2018, p. 74). Los efectos negativos provocados por sequías también alcanzan a zonas no altiplánicas, a saber, Maillard et al. (2020) en un estudio realizado en el oriente boliviano, específicamente en Santa Cruz, encontró la existencia de una relación entre eventos de extrema sequía y la propagación de los incendios forestales. Dicho estudio hace referencia a la sequía severa registrada el año 2010 en la Amazonía y los episodios de sequía del año 2019, en ambos casos las sequías están relacionados con una ocurrencia mayor de incendios forestales en la región (p. 11).

En consecuencia, es muy importante comprender el comportamiento de la sequía para así poder determinar de manera anticipada su impacto y predecir su ocurrencia. Es así que surgieron varios trabajos de investigación que se enfocaron primordialmente en el estudio del fenómeno.

Andrade et al. (2018) presenta un Atlas que estudia los eventos climáticos extremos, entre ellos la sequía, en el Altiplano de Perú y Bolivia. En dicho documento se presenta información climática que tiene el principal objetivo apoyar a las autoridades en la toma de decisiones, especialmente en los temas relacionados con el cambio climático (p.74).

En Bolivia destacan dos estudios relacionados con sequía. Satgé et al. (2019) utiliza datos de sensores remotos para determinar el efecto de las sequías severas y las actividades agrícolas en la disponibilidad de agua de la región del altiplano de Bolivia.

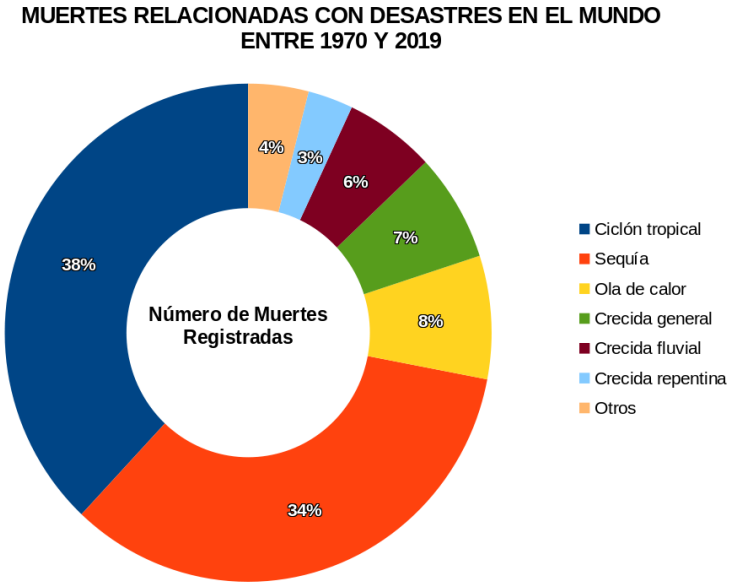


Figura 1. *Muertes provocadas con desastres entre los años 1970 y 2019*.
Elaboración propia a partir de OMM (2021)

Por otro lado Canedo-Rosso et al. (2021) realiza un análisis del impacto de la sequía en los cultivos , tomando como área de estudio el altiplano boliviano, para ello utiliza datos de derivados de imágenes satelitales. Los resultados del trabajo van dirigidos a la implementación de sistemas que proporcionen una alerta temprana ante desastres y gestión de riesgos relacionados con sequía.

Los trabajos señalados se enfocan principalmente en el análisis, seguimiento y monitoreo, y no abordan la temática de predicción de sequía.

1.2. Planteamiento del trabajo

Para poder comprender y caracterizar el fenómeno de sequía se requieren instrumentos que permitan su medición. Los más son los índices de sequía, ya que éstos tienen la capacidad de representar la complejidad del fenómeno de una manera cuantitativa y simplificada. Los índices que más destacan son el Índice Estandarizado de Precipitación y Índice Estandarizado de Precipitación y Evapotranspiración.

Tradicionalmente se han utilizado modelos estadísticos para la predicción de los índices de sequía; sin embargo, durante los últimos años se aumentó la utilización de algoritmos de aprendizaje automático para predicción climática y particularmente para la predicción de sequía, desde algoritmos de regresión como ser los Árboles de Decisión o Bosques Aleatorios, hasta algoritmos más avanzados como las Redes Neuronales Artificiales.

El presente trabajo tiene por objetivo realizar el análisis y predicción de sequía en el altiplano central boliviano a través de siete modelos de aprendizaje automático, para luego desarrollar una comparación del rendimiento de los mismos. Para ello en primera instancia se recolectan los datos de variables meteorológicas procedentes de 24 estaciones meteorológicas. Posteriormente se realiza la preparación y exploración de los datos, que luego deriva en el cálculo de los índices de sequía.

Los índices de sequía, además de otras variables a definir en el estudio se utilizan en el entrenamiento de los modelos de aprendizaje automático. Finalmente se lleva a cabo la evaluación y comparación de resultados utilizando las métricas: Error Medio Absoluto, Error Medio Absoluto Porcentual, Error Cuadrático Medio y Raíz Cuadrada del Error Cuadrático Medio.

1.3. Estructura de la memoria

Esta memoria está estructurada en 5 capítulos. El primero es la Introducción y los siguientes capítulos se describen en los siguientes párrafos:

En el Capítulo 2 se contextualiza la temática de sequía y los índices más utilizados para su caracterización. Luego se describen las series temporales y los modelos que se usan para tareas de predicción, a saber, modelos estadísticos y de aprendizaje automático. Por otra parte, se presentan trabajos relacionados a la temática planteada, analizando sus principales características.

En el Capítulo 3 se presentan los objetivos del trabajo y la metodología propuesta para el análisis y predicción de sequía, además de la comparativa entre los algoritmos utilizados. La metodología está basada en dos modelos, uno orientado a la minería de datos y el otro a la predicción de series temporales y se compone de 7 fases para alcanzar los objetivos planteados.

En el Capítulo 4 se desarrolla la contribución del trabajo, con la aplicación de la metodología propuesta en el capítulo 3, con la descripción de cada una de las fases definidas y los pasos detallados en cada fase.

En el Capítulo 5, para finalizar se presentan las conclusiones del trabajo , en relación a los objetivos planteados, también se realiza un análisis de futuros trabajos que se podrían originar a raíz del presente proyecto.

2. Contexto y estado del arte

En éste apartado se muestra una visión global del contexto del tema y estado del arte. Primero se contextualiza el tema sequía presentando algunas definiciones y los principales tipos de sequía. Luego se explican los índices más utilizados para su caracterización. Posteriormente se define una serie temporal explicando sus características principales y se describen los modelos utilizados para su predicción basados en: i) técnicas estadísticas y ii) técnicas de aprendizaje automático. En la última parte de la sección se hace un análisis de varios trabajos relativos a la predicción de índices de sequía, destacando su relación con este trabajo. Finalmente se resaltan las contribuciones del trabajo actual.

2.1. Caracterización de la Sequía

La sequía es un fenómeno natural que se produce por un largo período de tiempo sin precipitaciones y es una de las causas que provoca daños en la agricultura y la economía (Vicente-Serrano et al., 2010, p. 1696). Muchas veces se minimizan los efectos negativos de las sequías en relación a otros desastres naturales como los terremotos o inundaciones, porque aparentemente provoca una limitada mortalidad; pero las sequías afectan a regiones extensas y el tiempo de afectación es mucho mayor que en otros casos. (Eslamian et al., 2017, p. 50).

Hablar de un solo tipo de sequía no es posible ya que depende de muchos factores, para una mejor comprensión se clasifica en: meteorológica, hidrológica, agrícola y socio-económica. La *sequía meteorológica* es un periodo de sequedad que se produce cuando las precipitaciones están por debajo de lo normal (Eslamian et al., 2017, p. 50). Según Khan et al. (2018) la *sequía agrícola* se origina debido a la carencia de agua para los cultivos, lo que ocasiona la reducción del rendimiento de la producción agrícola , desembocando posteriormente en el problema social de desempleo en el sector (p. 108). Por otro lado, la *sequía hidrológica* es el déficit en la provisión de agua superficial y subterránea y se determina mediante la medición del nivel de agua en arroyos, lagos, embalses y aguas subterráneas (Khan et al., 2018, p. 108). Finalmente, la *sequía socio-económica* es ocasionada por una deficiencia de agua para el consumo doméstico, agrícola e industrial que provoca un impacto negativo en la sociedad ocasionando pobreza, enfermedad y desempleo. (Eslamian et al., 2017, p. 51). La figura 2 presenta los cuatro tipos de sequía y sus relaciones.

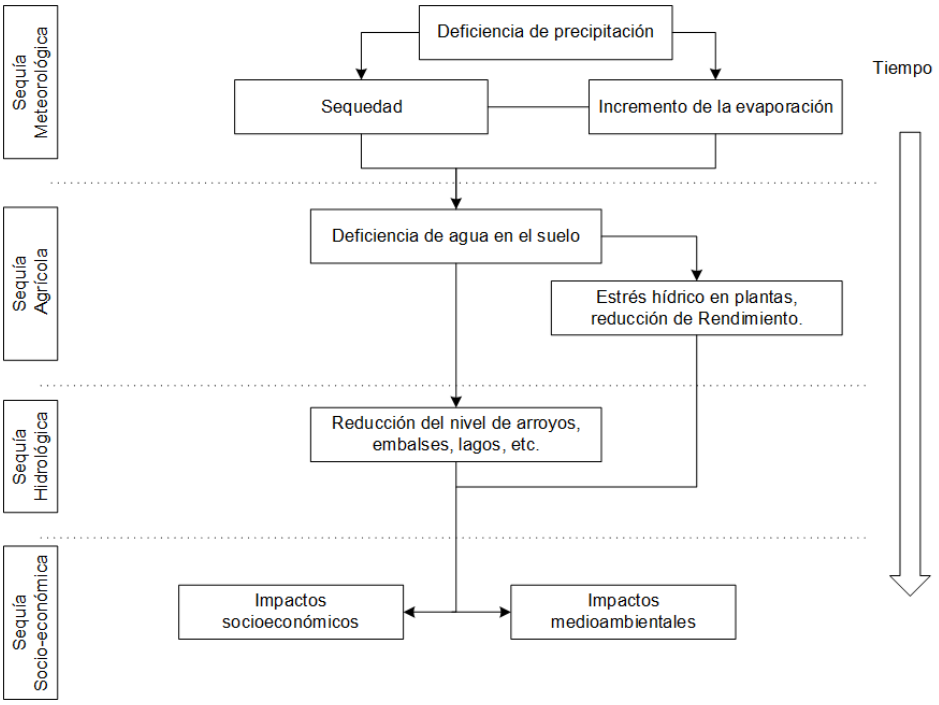


Figura 2. *Tipos de sequía*

Elaboración propia a partir de (Khan et al., 2018)

Es difícil determinar con claridad el momento en que se inicia, termina o la intensidad de un evento de sequía. Los índices de sequía se constituyen en un instrumento esencial para la medición de sequía, debido a que viabilizan la presentación de las diferentes variables involucradas en un único valor numérico (Eslamian et al., 2017). Según Maca & Pech (2016) los índices de sequía se expresan en forma de series temporales y permiten en primera instancia modelar y posteriormente coadyuvan a la tarea de predecir las sequías .

Se han utilizado muchos índices para modelar las sequías, entre los cuales destacan: el Índice Estandarizado de Precipitación y el Índice Estandarizado de Precipitación y Evapotranspiración. El primero solo requiere un parámetro: la precipitación y el segundo además de la precipitación incluye datos de temperatura. Ambos índices se pueden calcular para diferentes escalas de tiempo.

2.1.1. Índice Estandarizado de Precipitación

O Standardized Precipitation Index (SPI), es un índice de sequía que solo involucra a la variable precipitación y permite determinar las características de intensidad y duración de la sequía con una mayor claridad (McKee et al., 1993, p. 180).

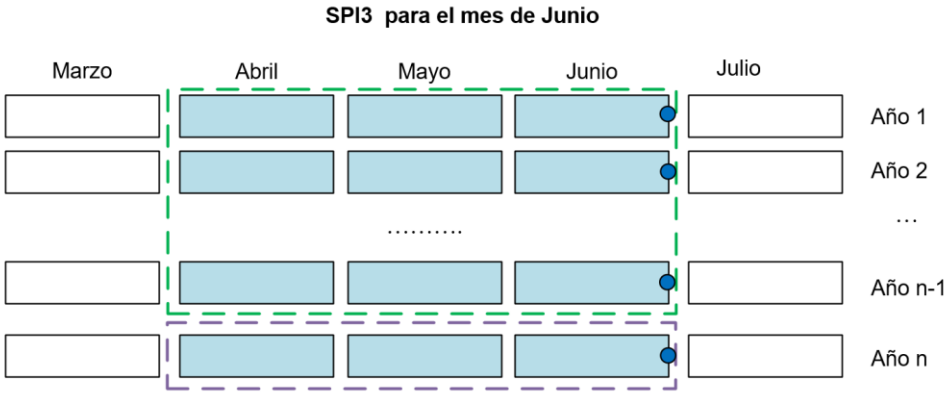


Figura 3. *Cálculo mensual de SPI3 para el mes de junio*

Para la obtención del SPI se utiliza la precipitación acumulada de una estación meteorológica en un determinado periodo de análisis (Por ejemplo 40 años). La medición se realiza durante una escala de tiempo definido que podría ser por ejemplo 3 meses. Por otra parte, se obtiene la media de la precipitación en el mismo periodo de tiempo (3 meses) pero de todo el periodo de análisis (40 años). La figura 3 muestra el cálculo del SPI de tres meses, representado como SPI3 para el mes de junio. Los datos de precipitación son ajustados a una función de distribución de probabilidad, generalmente Gamma; luego se realiza una estimación de la normal inversa para obtener la desviación de la precipitación para una densidad de probabilidad distribuida normalmente con media 0 y desviación estándar 1. (McKee et al., 1993; Eslamian et al., 2017)

Un evento de sequía es un intervalo de tiempo durante el cual el SPI toma valores continuamente negativos y menores o iguales a -1, es así que la sequía inicia cuando el SPI tiene el valor de -1 o inferior y finaliza con el primer valor positivo que toma el SPI después de un valor de -1 o menor (McKee et al., 1993, p. 180). Los distintos valores de SPI se asocian a diferentes categorías de sequía, dando lugar a la clasificación de la Tabla 1. Los valores negativos se asocian a eventos de sequía leve cuando el SPI es menor a -1, a partir de -1 la sequía es moderada, a partir de -1.5 se tiene una sequía severa y a partir de -2 se tiene una sequía extrema.

2.1.2. Índice Estandarizado de Precipitación y Evapotranspiración

O Standardized Precipitation Evapotranspiration Index (SPEI). Este índice se sustenta sobre los cálculos del SPI; pero en lugar de la precipitación toma como base el balance hídrico, que se calcula restando la Evapotranspiración Potencial o Potential Evapotranspiration (PET) de la precipitación, ese cálculo generalmente se efectúa a nivel mensual. (Vicente-Serrano et al., 2010).

Tabla 1. Categorías de Sequía en función del SPI

Valores de SPI	Categoría de Sequía
-0.99 ≤ SPI < 0	Sequía Leve
-1.49 ≤ SPI ≤ -1.0	Sequía Moderada
-1.99 ≤ SPI ≤ -1.5	Sequía Severa
SPI ≤ -2.0	Sequía Extrema

Fuente: (McKee et al., 1993, p. 180).

Se utilizan diferentes ecuaciones para la estimación del PET. La ecuación Thornthwaite solo requiere datos de precipitación y temperatura media, mientras que la ecuación de Hargreaves utiliza la temperatura máxima y mínima, en lugar de la temperatura media. Ambas ecuaciones además requieren el dato de la latitud de la estación analizada. Para estandarizar el balance hídrico se utiliza la distribución de probabilidad log-logística, obteniendo las series de SPEI como resultado.

Los valores del índice SPEI también se relacionan con intensidades de sequía, para lo cual se utiliza la misma clasificación que en el caso del SPI (Ali et al., 2017, p. 3) . Ver tabla 1.

2.1.3. Escalas de tiempo en índices de sequía

Las diferentes escalas de cálculo de los índices se asocian a distintos tipos de sequía. El estudio a escala de 1 mes se utiliza para analizar la sequía meteorológica, escalas de 3 meses y 6 meses para la sequía agrícola, se utiliza 12 meses como escala para analizar la sequía hidrológica (Li et al., 2015, p. 10927). Finalmente, la escala de 24 meses es útil en el análisis de sequías socio-económicas (Potop et al., 2014).

Cuando el periodo de tiempo es corto: 1,3 o 6 meses, tanto el SPI como el SPEI se mueven de valores positivos a negativos con mayor frecuencia. Por otro lado, cuando los periodos de tiempo son mayores: 12, 24 o más meses, la respuesta de ambos índices es más lenta ante cambios en la precipitación, los cambios de positivo a negativo son menos frecuentes y tienen una mayor duración (McKee et al., 1993, p. 181)

Para el presente trabajo se calculan los índices SPI y SPEI con 2 escalas: 3 y 12 meses, de esta manera se pueden analizar las sequías agrícolas e hidrológicas, además comparar el desempeño de ambos índices a corto y a largo plazo.

2.1.4. Caracterización de eventos secos

En la sección 2.1.1. se presenta la definición de un evento de sequía según McKee, utilizando los valores del índice SPI, la cual es aplicable también al índice SPEI. En la mencionada definición el inicio de un evento de sequía se define en -1; sin embargo, dependiendo la intensidad de sequía que se requiera analizar ese umbral se puede modificar. Por ejemplo, para detectar sequías de intensidad severa se podría utilizar como umbral -1.5.

Es importante identificar un evento de sequía; pero más importante aún es definir algunas métricas que permitan caracterizar dicho evento, las más importantes son: duración, severidad, intensidad y valor mínimo.

- **Duración (D).** Es el tiempo que pasa desde que comienza un evento de sequía hasta que éste termina. También se expresa "como el tiempo durante el cual el valor del índice está continuamente por debajo del umbral definido" (Mesbahzadeh et al., 2020, p. 4) .

- **Severidad (S)**. También denominado magnitud, es igual a la suma de los valores que alcanza el índice mientras está por debajo de un umbral definido (Mesbahzadeh et al., 2020, p. 4). Sea SPI el índice de sequía (también funciona para SPEI) una escala determinada, la severidad S se calcula con la fórmula (1) donde *D* representa el número de meses que dura la sequía.

$$S = -\left(\sum_{i=1}^{D} SPI_i\right) \quad (1)$$

- **Intensidad (I)**. Es la media de los valores del índice durante la duración del evento de sequía. La intensidad de sequía se obtiene con la fórmula (2) dividiendo la Severidad entre la Duración (Adhyani et al., 2017, p. 7).

$$I = \frac{S}{D} \quad (2)$$

- **Valor mínimo (P).** Es el valor mínimo (extremo) del índice que se alcanza durante el evento de sequía (Mokhtar et al., 2021, p. 65507).

En la Figura 4 se muestran las 4 métricas en una serie temporal SPI de escala n.

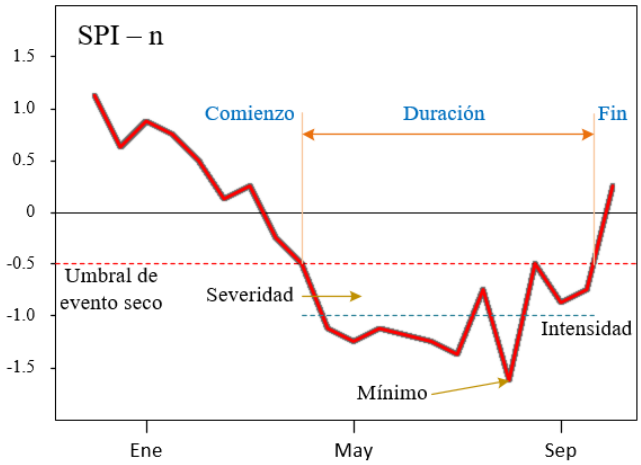


Figura 4. *Caracterización de un evento seco.*

2.2. Análisis y predicción de series temporales

Una serie temporal es una secuencia cronológica de observaciones respecto a una determinada variable de interés. Dichas observaciones se recopilan en periodos de tiempo igualmente espaciados, los periodos pueden ser diarios, mensuales , trimestrales o anuales (Montgomery et al., 2015, p. 2).

2.2.1. Patrones en series temporales

Las series de tiempo tienen algunos patrones que permiten analizarlas y comprender su comportamiento. Estos son: tendencia, estacionalidad y ciclo:

- **Tendencia.** Se produce cuando hay un aumento: tendencia creciente o disminución: tendencia decreciente, en los datos de la serie temporal (Hyndman & Athanasopoulos, 2021). Según Montgomery (2015) la tendencia puede tener larga duración o presentar un comportamiento más dinámico y una corta duración (p.14).

- **Estacionalidad.** Según Hyndman (2021) se produce cuando algunos factores estacionales (época del año, mes, día de la semana) afectan a una serie temporal. Este comportamiento se repite periódicamente y dicho periodo es fijo.

- **Ciclo:** Ocurre cuando hay fluctuaciones (subidas o bajadas) que, a diferencia de la estacionalidad, no se producen con una frecuencia fija. En general se relacionan con el ciclo económico (Hyndman & Athanasopoulos, 2021). Las series temporales cíclicas muestran un comportamiento recurrente, al igual que las estacionales, pero el periodo de tiempo es variable. (Nielsen, 2019, p. 78).

2.2.2. Métodos para analizar series temporales

Para entender el comportamiento y caracterizar una serie de tiempo se necesita realizar un análisis exploratorio, para lo cual se utilizan dos grupos de métodos: en el primero están los métodos tradicionales diseñados para datos no temporales, como ser: histogramas y gráficas de dispersión. En el segundo grupo están los métodos orientados series temporales como ser: funciones de autocorrelación y autocorrelación parcial. (Nielsen, 2019, p. 92).

Un concepto importante en el estudio de series temporales es la **estacionariedad** que según Nielsen (2019) es la característica de una serie temporal de mantener la media y la varianza estables en el tiempo.

- **Autocorrelación ACF.** "Mide la relación lineal entre los valores rezagados de una serie de tiempo" (Hyndman & Athanasopoulos, 2021). Es muy útil graficar la función ACF para visualizar cómo cambian las correlaciones respecto a un retardo o rezago (en inglés lag), a la gráfica generalmente se le denomina como correlograma. Según explica Hyndman (2021) cuando la serie de tiempo tiene tendencia, en la gráfica ACF los retardos tienden presentar valores grandes y positivos y estos disminuyen lentamente; de lo contrario (en una serie estacionaria) los descensos serán más bruscos. En el caso de la estacionalidad las autocorrelaciones son mayores para los retardos estacionales y se verá ese comportamiento también en los múltiplos del período estacional. La figura 5a permite visualizar una gráfica ACF de una serie temporal.

- **Autocorrelación Parcial PACF**

  La autocorrelación parcial mide la relación entre dos puntos en el tiempo y_t e y_{t-k} después de remover los efectos de los retardos, sean estos: 1,2, 3, ….., k-1. Por ésta razón el primer valor de PACF es similar al primer valor de ACF, porque no hay nada que remover. (Hyndman & Athanasopoulos, 2021). La figura 5b muestra la gráfica PACF de una serie de tiempo.

2.2.3. Predicción de series temporales

Cuando se realiza predicción de series temporales generalmente se utilizan dos tipos de técnicas: i) *técnicas de predicción cualitativas*, cuando no se tienen datos disponibles o los datos no son relevantes para realizar las predicciones y ii) *técnicas de predicción cuantitativas,* cuando se dispone de suficiente información numérica sobre el pasado, además se tienen argumentos para suponer que algunas características de los datos del pasado se repetirán en el futuro. (Hyndman & Athanasopoulos, 2021; Montgomery et al.,2015).

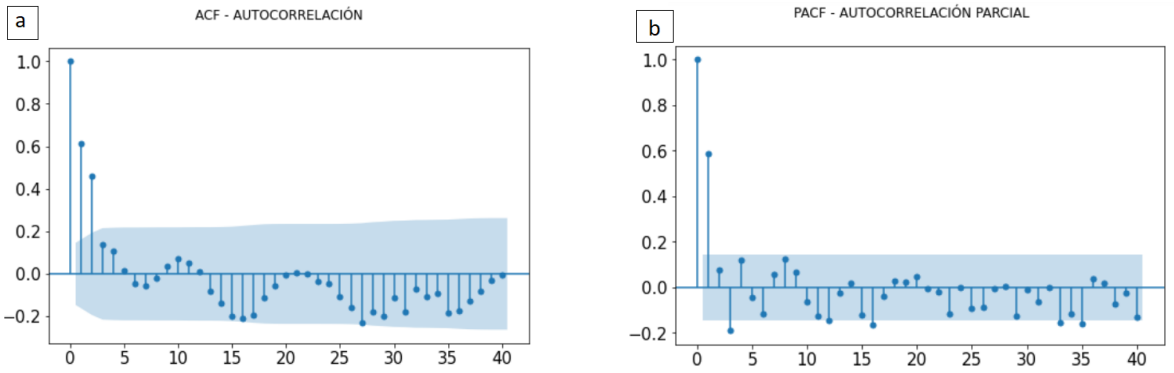


Figura 5. *Gráficas de ACF y PACF*

Autocorrelación (a) y Autocorrelación Parcial (b).

Dentro de la predicción cuantitativa se tienen varios tipos de modelos, que se diferencian por la forma como se presentan las *variables predictoras* en el modelo. Destacan tres tipos de modelos:

1) **Modelo Explicativo.** También denominado modelo de regresión o predicción causal (Montgomery et al., 2015, p. 5). En este tipo de modelo las variables predictoras, llamadas también externas o exógenas, ayudan a explicar las causas de la variación de la variable a predecir. Tomando como ejemplo la predicción del consumo de energía eléctrica ED, el modelo podrías representarse con la ecuación (3), teniendo como variables predictoras a: temperatura actual, población y día de la semana (Hyndman & Athanasopoulos, 2021).

$$ED = f(temperatura\ actual, población, día\ de\ la\ semana, error) \quad (3)$$

2) **Modelo de Serie Temporal.** En este tipo de modelo los valores futuros se basan en valores pasados de la variable a predecir y no se toman en cuenta variables externas (Hyndman & Athanasopoulos, 2021). Para el ejemplo de la predicción de demanda de energía, podría utilizarse la ecuación (4):

$$ED_{t+1} = f(ED_t, ED_{t-1}, ED_{t-2}, ED_{t-3}, error) \quad (4)$$

3) **Modelo Mixto.** También llamado modelo de Regresión Dinámica, modelo de panel o modelo longitudinal (Hyndman & Athanasopoulos, 2021). Este modelo combina las características de los dos modelos anteriores, la ecuación (5) muestra una posible representación para el ejemplo de demanda de energía:

$$ED_{t+1} = f(ED_t, temperatura\ actual, población, día\ de\ la\ semana, error) \quad (5)$$

Con relación a la cantidad de variables ,según Nielsen (2019) las series temporales objeto de análisis y predicción pueden ser:

1) **Univariantes,** se analiza una sola variable con relación al tiempo.

2) **Multivariantes,** son series con más de una variable que son medidas en cada instante de tiempo. Estas variables pueden estar interrelacionadas y mostrar dependencias temporales entre sí.

En este trabajo se utilizan las técnicas de predicción cuantitativas, y entre los tipos de modelo: el modelo de serie temporal y el modelo mixto. Respecto a las variables se realiza por un lado la predicción univariante y por otro lado la predicción univariante con la inclusión de variables exógenas.

2.3 Modelos estadísticos

Algunos modelos estadísticos fueron desarrollados específicamente para series temporales, los más sencillos para series de tiempo univariantes: como los modelos Autoregresivos o de Media Móvil, hasta modelos más complejos como los que incluyen además variables exógenas.

- **Modelo de Media Móvil Autoregresivo** o Autoregressive Moving Average (ARMA) es la integración de un modelo Autoregresivo (AR) y uno de Media Móvil (MA). Para AR el valor de la serie temporal en un punto en el tiempo *t* es una función de los valores de la serie en puntos anteriores en el tiempo. Para MA el valor de la serie temporal en cada punto en el tiempo es una función de los términos de “error” del valor pasado reciente, cada valor se considera independiente de los demás (Nielsen, 2019, p. 202).

- **Modelo de Media Móvil Integrado Autoregresivo o** Autoregressive Integrated Moving Average (ARIMA) además de los componentes AR y MA agrega el componente **I** que se refiere a la diferenciación, que elimina la tendencia en una serie y hace que ésta sea estacionaria (Nielsen, 2019, p. 207).

- **Modelo de Media Móvil Integrado Autoregresivo con Entrada Exógena** o Autoregressive Integrated Moving Average with Exogenous Input (ARIMAX). Este modelo es una extensión de ARIMA que incorpora en el análisis una o más variables independientes (exógenas) (Jalalkamali et al., 2015).

2.4. Modelos de Aprendizaje Automático

Se distinguen dos tipos principales de aprendizaje automático: El Aprendizaje Supervisado y el No Supervisado, aunque también existen el aprendizaje semisupervisado y reforzado.

- **Aprendizaje supervisado**, se caracteriza porque el conjunto de entrenamiento que sirve como entrada al algoritmo incluye las soluciones que se desean obtener, las cuales se conocen como etiquetas o labels. Los principales tipos de aprendizaje supervisado son clasificación y regresión (Géron, 2019, p. 8) .

- **Aprendizaje no supervisado,** se caracteriza porque los datos de entrenamiento del algoritmo no están etiquetados. Un modelo importante no supervisado es el de clusterización (Géron, 2019, p. 9).

En esta sección se describen algunos modelos de aprendizaje automático utilizados para la predicción series temporales.

- **Bosque Aleatorio** o Random Forest (RF). Un bosque aleatorio es un modelo basado en árboles de decisión; pero se calculan muchos en lugar de solo un árbol. La Regresión es el resultado de promediar las salidas de todos los árboles. Los bosques aleatorios buscan la "sabiduría de la multitud", que se compone de muchos modelos simples, tomando en cuenta que ninguno es bueno de manera independiente, pero todos juntos trabajan mejor (Nielsen, 2019). Muchos trabajos de predicción de sequía utilizaron como modelo principal los Bosques Aleatorios, obteniendo buenos resultados (Dikshit et al., 2020).

- **Redes Neuronales Artificiales** o Artificial Neural Network (ANN). Las redes neuronales son modelos de aprendizaje automático; pero por su importancia en el trabajo actual se crea una sección para su explicación.

2.5. Modelos de Redes Neuronales

Una red neuronal es un conjunto de entradas directamente asignadas a una salida, usando para ello una variación generalizada de una función lineal. La red neuronal, también conocida como Perceptrón, contiene una capa de entrada y un nodo de salida. La capa de entrada hace referencia a "valores observados", los cuales son parte de los datos recabados que sirven al entrenamiento; y el objetivo es predecir las variables que carecen de datos (Aggarwal, 2018). En la figura 6 se observa la arquitectura básica de un Perceptrón.

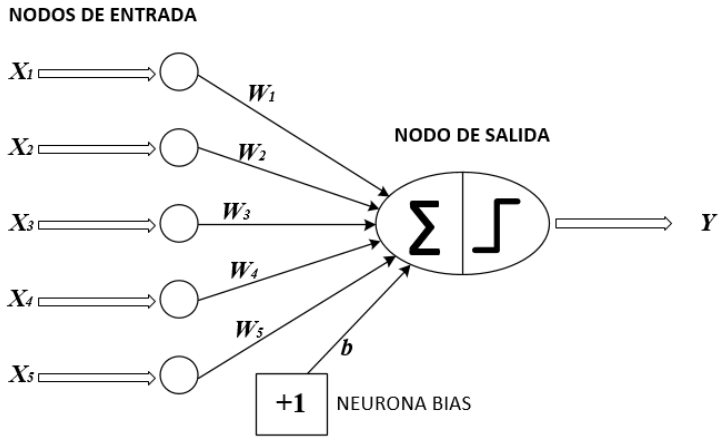


Figura 6. *Arquitectura básica del Perceptrón*
Basado en (Aggarwal, 2018)

- **Red Neuronal Perceptrón Multicapa** o MultiLayer Perceptron (MLP) Es la red neuronal más básica, que consiste en neuronas que se agrupan en capas. MLP es modelo de red neuronal más utilizado en el modelamiento de datos hidrológicos (Ali et al., 2017, p. 4). Este tipo de redes neuronales son realimentadas o Feed Forward, es decir que todos los nodos en una capa están conectados a otros en la siguiente capa. (Aggarwal, 2018).

- **Red Neuronal Recurrente de Memoria a Corto y Largo Plazo** o Long Short Term Memory (LSTM). Las primeras redes neuronales recurrentes fueron Elman. La característica principal de esta red es que intenta capturar el tiempo mediante una arquitectura diferente: cada una de las salidas de una capa se convierte en entrada a otra capa, con un patrón de entrada y salida, lo que ayuda a predecir los siguientes estados de la red, la salida y el tiempo final (Skansi, 2018). Las redes neuronales recurrentes (como la de Elman) tienen problemas asociados con los gradientes, es algo común que las sucesivas multiplicaciones las vuelvan inestables. Los resultados en la gradiente desaparecen durante la propagación hacia atrás (Back Propagation), o explotan hacia valores altos de manera descontrolada. Una red neuronal con esta característica de actualizaciones en base a la multiplicación es buena aprendiendo cuando aprende de periodos cortos, es decir que está dotada de buena memoria de corto plazo, pero su memoria a largo plazo es muy mala. (Aggarwal, 2018).

  Una solución para el problema descrito anteriormente es cambiar las ecuaciones de recurrencia por un vector oculto con el uso de las redes LSTM, en inglés Long Short Term Memory y el uso de memoria de largo plazo. Los cálculos y operaciones realizadas por la

red neuronal LSTM están diseñadas para tener un control granular muy fino sobre los datos que interactúan en la memoria de largo plazo (Aggarwal, 2018).

- **Red Neuronal Convolucional CNN**

  En inglés Convolutional Neural Network. Una red neuronal convolucional está diseñada para trabajar con entradas que tienen una estructura en cuadrícula. Además, tienen una fuerte dependencia espacial, lo cual hace que este tipo de red neuronal sea apropiada para trabajar con imágenes y cualquier tipo de estructura parecida, como texto, series temporales y secuencias, los cuales pueden ser considerados como casos especiales de este tipo de datos de estructura en cuadrícula por las varias relaciones entre valores adyacentes. Este tipo de redes neuronales también puede ser usado con todo tipo de datos temporales, espaciales y espaciotemporales (Aggarwal, 2018).

  En una red neuronal convolucional, los estados de cada capa son manejados acorde a la estructura de cuadricula espacial, las relaciones espaciales son recibidas desde una capa a la siguiente, esto debido a que la característica de cada valor está basada en una pequeña región espacial local de la capa anterior. Es importante mantener la relación espacial entre las celdas de la cuadricula debido a la operación de convolución y la transformación a la siguiente capa es excesivamente dependiente de esta relación. (Aggarwal, 2018)

2.6. División de los datos en Entrenamiento y Validación

El objetivo de entrenar un modelo de aprendizaje automático es que éste generalice bien los casos nuevos, para ello es necesario evaluar el modelo. Una opción consiste en separar los datos existentes en dos conjuntos: entrenamiento y validación. Se usará el conjunto de entrenamiento para entrenar al modelo, y el conjunto de validación para validarlo. La tasa de error para los casos nuevos se llama “error de generalización”. Ese error indica qué tan bien funcionará el modelo frente a casos que nunca ha visto antes y se puede estimar su valor evaluando el modelo con el conjunto de evaluación (Géron, 2019, p. 30).

Respecto al porcentaje adecuado para dividir los datos en entrenamiento y validación, es habitual utilizar el 80% de los datos para entrenamiento y 20% para las pruebas. No obstante, es posible que el tamaño del conjunto de datos sea grande y un porcentaje menor sea suficiente para obtener una buena estimación del error de generalización (Géron, 2019, p. 31). En el caso de las series temporales se recomienda realizar la división considerando periodos de tiempo completos como ser 24 horas, 7 días o 12 meses, dependiendo del problema.

2.7. Sobreajuste y subajuste de los modelos

Muchas veces los modelos, durante el entrenamiento, no adquieren la capacidad de generalización y podrían estar en uno de dos extremos:

Sobreajuste, se refiere al hecho que un modelo se desenvuelve muy bien con los datos de entrenamiento, pero no lo generaliza de la misma manera. Usualmente ocurre cuando el modelo es demasiado complejo en relación a la cantidad de ruido de los datos de entrenamiento. (Géron, 2019). Por mucho que el modelo se ajuste bien a los datos de entrenamiento, las predicciones para los nuevos valores tienen un error grande.

Subajuste, es lo opuesto de sobreajuste. El subajuste ocurre cuando el modelo es demasiado simple para aprender la estructura fundamental de los datos, y las predicciones son inexactas. (Géron, 2019). El modelo no tiene la capacidad suficiente para resolver el entrenamiento.

Es importante encontrar un modelo que esté en el punto intermedio, que sea capaz de aprender en el entrenamiento; pero también sea capaz de generalizar su conocimiento.

2.8. Métricas de Evaluación

Las métricas de evaluación miden la diferencia entre los valores resultado de la predicción del modelo y los valores reales.

Error Cuadrático Medio o Mean Square Error (MSE). Es la métrica más común para evaluar el rendimiento de la predicción de un modelo. Se calcula el cuadrado de los errores y luego se calcula su media. La razón por la que se eleva al cuadrado es para evitar que los valores positivos y negativos se cancelen unos a otros.

$$MSE(X,h) = \frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2 \qquad (6)$$

Raíz Cuadrada del Error Cuadrático Medio o Root Mean Square Error (RMSE). Esta métrica muestra cuán grande es el error del sistema debido a las predicciones. La medida de cálculo de los errores es de la misma escala que los errores originales, para ellos se obtiene la raíz cuadrada del MSE (Géron, 2019). Su fórmula es:

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2} \qquad (7)$$

Error Medio Absoluto o Mean Absolute Error (MAE), Por lo general, es usado en casos donde existe varios grupos de valores atípicos. Mide la distancia ente dos vectores: el vector de predicción (Géron, 2019) y el de valores reales. Para evitar la cancelación de valores

positivos y negativos usa valor absoluto en lugar del cuadrado, es decir que no penaliza los errores grandes tanto como lo hace el MSE. Su ecuación es:

$$MAE(X,h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(x^{(i)}\right) - y^{(i)}\right| \qquad (8)$$

Error Porcentual Medio Absoluto o Mean Absolute Percentage Error (MAPE). Esta medida es muy parecida al MAE, pero el cálculo del error es a nivel porcentual, éste error permite medir el tamaño de los errores en comparación con los valores.

$$MAPE(X,h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(x^{(i)}\right) - y^{(i)}\right| \qquad (9)$$

En éste trabajo se calculan las cuatro métricas en la etapa de evaluación del modelo.

2.9. Herramientas computacionales

Se utilizaron dos principales herramientas para el desarrollo del trabajo: Jupyter Notebooks y Google Colab Pro. Para las primeras fases del trabajo se codificó en lenguaje Python en cuadernos de Jupyter y en las últimas etapas, para las pruebas a los modelos predictivos se utilizó Google Colab la versión Pro, sobre todo por sus características técnicas superiores a la versión gratuita en RAM y GPU.

- **Jupyter Notebook**, se caracteriza por ser un software transparente y con alta adaptabilidad, una de sus principales fortalezas es que los modelos desarrollados con la herramienta son fáciles de implementar y reproducir.(Peñuela et al., 2021).

  Para el análisis y exploración de datos se utilizan las bibliotecas: pandas y numpy para manejo de los dataframes y cálculo, y matplotlib y seaborn para la visualización. Para el modelado con Redes Neuronales se utiliza la Biblioteca Scikit-learn y la Biblioteca Sktime (*Sktime*, 2021), ésta última enfocada específicamente a series temporales.

- **Bibliotecas en lenguaje R**, se utilizan dos bibliotecas escritas en Lenguaje R : “Climatol” (Guijarro, 2019b) que permite el relleno de datos faltantes y homogenización de series temporales y el paquete SPEI (Beguería & Vicente-Serrano, 2017) utilizado para el cálculo de los dos índices de sequía: SPI y SPEI.

  Para incluir código en R dentro de los cuadernos de Jupyter Notebook utilizó el Kernel de R denominado IRKernel (IRkernel, 2021).

2.10. Trabajos relacionados

Muchos investigadores han abordado temas de predicción climática utilizando técnicas de aprendizaje automático, algunos de ellos estuvieron particularmente enfocados a la predicción de sequía y específicamente a la estimación de índices, en su mayoría del SPEI. En la tabla 2 se describen los principales trabajos que se estudiaron por su relación con el trabajo actual, considerando aspectos básicos como el área y el periodo de estudio, las variables involucradas, el modelo de aprendizaje automático utilizado, las métricas de evaluación y los principales resultados alcanzados.

Tabla 2. Resumen comparativo de trabajos relacionados

Título del trabajo de Investigación	**Área y Periodo de Estudio**	**Índice de Sequía a predecir**	**Variables Predictoras**	**Modelo de Aprendizaje Automático**	**Métricas de Evaluación**	**Resultados Alcanzados**
Predicción de índices de sequía SPEI y SPI utilizando las redes neuronales artificiales integradas (Maca & Pech, 2016)	Estados Unidos 2 estaciones meteorológicas (1948-2002)	SPI SPEI	3 combinaciones: 1) Retardos de 2,3 y 4 meses de SPI/SPEI 2) Retardos de 2,3,4 meses de SPI/SPEI y Temperatura Media 3) Retardos de 2,..10 meses de SPI/SPEI	2 modelos de redes neuronales: - 1 MLP Simple - 5 MLP integrados	MAE MSE Otras	Ambos modelos se desempeñaron bien pero el modelo de red neuronal integrado fue superior al Perceptrón Multicapa Simple. Respecto a las variables predictoras el conjunto 3) de 9 retardos es el que dio mejores resultados
Predicción de sequía con el modelo de red neuronal artificial de Perceptrón Multicapa (Ali et al., 2017)	Pakistán 17 estaciones climatológicas (1975 – 2012)	SPEI 1,3,12 meses	Utiliza 3 retardos de la serie SPEI	Red Neuronal Perceptrón Multicapa	MAE, RMSE, R	El modelo Perceptrón Multicapa tiene una buena capacidad para la predicción del SPEI, obteniendo coeficientes de correlación entre 0.780 y 0.985 en las estaciones.
Predicción de sequía basada en SPI y SPEI con escalas de tiempo variables utilizando la red neuronal recurrente LSTM. (Poornima & Pushpalatha, 2019).	Región de Hyderabad 1 estación meteorológica (1980 – 2014)	SPI SPEI 1,6,12 meses	Precipitación, temperatura y humedad relativa	Red Neuronal Recurrente LSTM	RMSE	El trabajo compara ARIMA con una red neuronal LSTM, utilizando entrada multivariable. El modelo ARIMA da resultados buenos con predicciones a corto plazo (1 mes). LSTM proporciona mejores resultados para 6 y 12 meses

Predicción de sequía espacio-temporal a corto plazo utilizando el modelo de bosques aleatorios en Nueva Gales del Sur, Australia (Dikshit et al., 2020)	Nueva Gales del Sur, Australia Toda la región (1901-2018)	SPEI 1,3 meses	Precipitación, temperatura media, evapotranspiración, presión de vapor, cobertura nubosa. Datos satelitales CRU(Climate Research Unit)	Bosques Aleatorios (RF)	RMSE, R^2	El modelo produjo buenos resultados de predicción, con un R^2 de 0.73 para SPI-1 y 0.76 para SPI-3. Los autores recomiendan el uso del modelo de Bosques Aleatorios para predicciones a corto plazo.
Estimación de sequía meteorológica del SPEI mediante algoritmos de aprendizaje automático. Mokhtar et al. (2021)	Tibetan Plateau 30 estaciones meteorológicas (1980 – 2019)	SPEI 3,6 meses	Precipitación, temperatura,velocidad del viento , humedad relativa y radiación solar.	Bosques Aleatorios Extreme Gradient Boost Redes Convolucionales Redes recurrentes	MSE, MAE	Los autores concluyen que el mejor modelo encontrado fue XGB para SPEI3. con todas las variables menos radiación solar y RF con todas las variables también en la estimación de SPEI3. LSTM dio buenos resultados con precipitación, temperatura y velocidad del viento para SPI6

Fuente: Elaboración Propia

2.11. Conclusiones del Estado del Arte

Después de haber descrito los principales conceptos asociados al tema y revisado los trabajos relacionados se concluye que gran parte de los autores que abordan la temática eligieron modelos de aprendizaje automático de Redes Neuronales , siendo lo más utilizados los modelos de Redes Recurrentes LSTM y Red Neuronal Perceptrón Multicapa , también se concluye que los principales índices elegidos para la predicción de sequía son el SPI y SPEI que requieren como dato de entrada solo la precipitación , en el caso de SPI y precipitación y temperatura en el caso de SPEI. Asimismo, las métricas para evaluación más utilizadas son: MAE, MSE y RMSE.

Es importante destacar que ninguno de los trabajos descritos tiene como área de estudio Sudamérica, tampoco Bolivia, lo que representa un aspecto relevante para este trabajo por el dominio de aplicación. También se plantea la comparativa de varios modelos, muchos de los cuales fueron estudiados en los distintos trabajos, pero de forma separada y en distintos ámbitos de aplicación. Finalmente se plantea la estimación de dos índices: SPI y SPEI lo que permite tener parámetros de comparación entre ellos y además con relación a dos escalas de tiempo: 3 meses y 6 meses.

3. Objetivos concretos y metodología de trabajo

En éste capítulo se enuncian: el objetivo general, los objetivos específicos del presente trabajo y la metodología que permite alcanzar dichos objetivos.

3.1. Objetivo general

El objetivo general del trabajo es realizar un análisis y predicción de sequía con diferentes algoritmos de aprendizaje automático, para determinar, mediante una comparativa, las ventajas y desventajas de los algoritmos con relación a la temática y además evaluar la eficiencia de los mismos.
Además de realizar la comparativa entre los modelos se plantea una metodología que permita analizar y predecir sequías en una determinada región, partiendo de la recolección de los datos, su preparación y procesamiento, elección del algoritmo para el modelado hasta el entrenamiento y evaluación de los modelos.

3.2. Objetivos específicos

- Determinar las fuentes principales de información meteorológica para la recolección de los datos de precipitación, temperatura y humedad relativa correspondientes al área de estudio.
- Realizar la recolección, análisis descriptivo y exploratorio de los datos para tener una comprensión mayor de los mismos.
- Calcular dos índices de sequía: SPI y SPEI que servirán como entrada para los modelos de aprendizaje automático.
- Analizar los índices SPI y SPEI obtenidos a fin de determinar la severidad, intensidad y duración de sequías pasadas.
- Seleccionar los modelos de aprendizaje automático orientados a series temporales, específicamente para la predicción de sequía.
- Entrenar los modelos de aprendizaje automático seleccionados con los datos de los índices de sequía: SPI y SPEI
- Evaluar los modelos utilizando las métricas: MAE, MAPE, RMSE y MSE, para que mediante una comparativa entre ellos se pueda determinar las ventajas y desventajas de los mismos y su eficiencia respecto a la predicción de sequía.

3.3. Metodología del trabajo

Existen diferentes formas de abordar el problema de predicción de sequía, que varían respecto al alcance y a las metas definidas. En ésta sección plantea un modelo específico para la predicción de sequía, que se fundamentó en dos modelos base para su construcción. A continuación, se describen los dos modelos:

- **Descubrimiento de conocimiento en bases de datos**

  O Knowledge Discovery in Databases (KDD). Según Maimon (2010) este modelo permite la comprensión de un fenómeno mediante el análisis y predicción de los datos. Se focaliza en la minería de datos ya que permite identificar patrones relevantes partiendo de los datos. Se compone de 9 pasos detallados en la Figura 7a

- **Pasos básicos para la predicción de series temporales**

  Según Hydman (2021) una tarea de predicción se compone de 5 pasos básicos, detallados en la Figura 7b. Se comienza con la definición del problema de predicción, pasando por la recopilación y análisis exploratorio de los datos. Posteriormente la selección y ajuste de los modelos para finalizar con la evaluación y el uso de los modelos.

3.3.1 Metodología propuesta

Para el trabajo actual se proponen 7 fases, que permiten una retroalimentación a la fase anterior en caso de ser necesario.

1) **Definición del problema y objetivos**.

   Se define con claridad el problema de predicción relacionado con el tema de sequía, entendiendo la manera en la que se utilizarán las predicciones y los usuarios que requieren dichas predicciones. Entre los usuarios más importantes están los funcionarios del Servicio Meteorológico Nacional, Ministerios de Medio Ambiente y Agua, Gobiernos Departamentales o Municipales o también entidades privadas por ejemplo del ámbito agrícola. Se requiere determinar cómo ellos utilizarán las predicciones, por ejemplo, para elaborar reportes de monitoreo y prevención de sequía o para la gestión de riesgos a nivel local y nacional. En base a toda la información encontrada se definen los Objetivos de la Predicción.

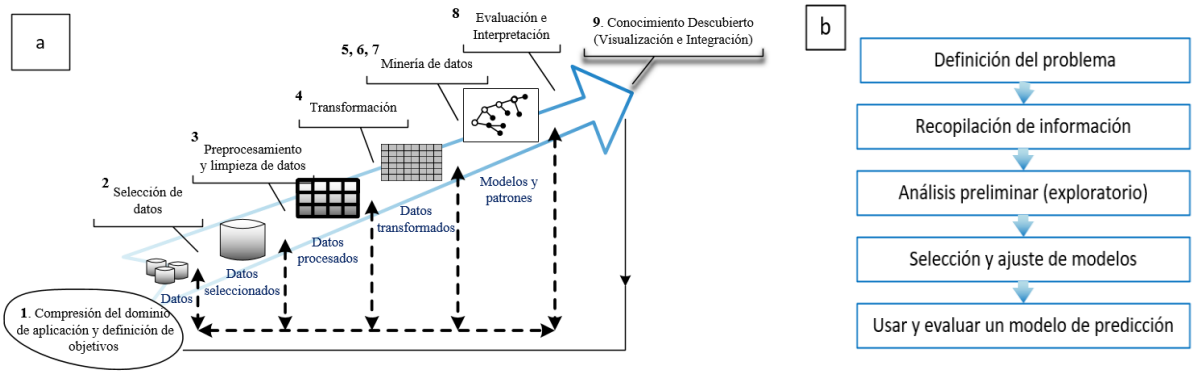


Figura 7. *Modelos de Referencia para el trabajo*

a.KDD (Maimon & Rokach, 2010) b. Predicción de Series Temporales (Hyndman et al., 2021)

2) Recolección y Selección de Datos

Se realiza la recopilación de información relevante para cumplir con el objetivo definido en la fase inicial. Para el tema de sequía generalmente se trabaja con datos meteorológicos. La fuente más importante de información meteorológica son los Servicios Nacionales de Meteorología; pero también se tiene información meteorológica y climática en múltiples repositorios en línea por ejemplo de National Aeronautics and Space Administration (NASA).

3) Preprocesamiento y Limpieza de datos

Se efectúa un análisis de tipo descriptivo de los datos recolectados, se verifica la cantidad de instancias y atributos. En ésta fase se determinan los datos atípicos y los datos faltantes, también es importante definir un periodo de análisis y determinar la cantidad de datos disponibles en cada serie respecto a ese periodo.

Luego se realiza un análisis exploratorio, generalmente mediante herramientas de visualización. Se utilizan las gráficas tradicionales como histogramas o gráficos de caja y también las técnicas específicas de series temporales para determinar patrones, tendencia, estacionalidad. Si en ésta fase se llega a determinar que los datos con los que se cuenta no cumplen los criterios de calidad definidos se puede volver a la fase 2 y recurrir a otras fuentes de datos.

4) Transformación de Datos

En ésta fase se selecciona y se prepara los datos más importantes según el objetivo definido. Se seleccionan tanto las instancias, registros u observaciones y atributos o características. También se crean nuevos datos, si es necesario, por ejemplo, mediante la transformación de atributos o se pueden generar nuevas instancias. Se pueden utilizar algunos métodos de trasformación como reducción de dimensionalidad. La etapa de

transformación tiene por propósito adecuar y formatear de datos de acuerdo al modelo que se utilizará en las siguientes fases.

5) Selección e Implementación del modelo

Se selecciona el modelo específico que permita alcanzar los objetivos planteados. La predicción va relacionada con el aprendizaje automático supervisado de tipo Regresión. Luego se define el modelo específico, en el caso de predicción se podría elegir un modelo de regresión como Regresión Lineal o Bosques Aleatorios para regresión. Se determinan los conjuntos de entrenamiento, validación y pruebas y se entrena el modelo. En base al modelo elegido se va configurando un modelo de predicción, para lo cual se deben ajustar los hiperparámetros del mismo. Es esta fase es común seleccionar y entrenar más de un modelo. Para que la implementación del modelo sea efectiva se debe cuidar que las relaciones entre las variables explicativas y la variable a predecir sean sólidas.

6) Evaluación e Interpretación

Se evalúan los resultados del modelo en cuanto a su exactitud y confianza, para ello se utilizan métricas predefinidas que dependen del tipo específico de modelo elegido en la fase 5. En el caso de regresiones las métricas más utilizadas son MSE y MAE. En éste trabajo se plantean además el RMSE y MAPE.

7) Visualización e Integración

Consiste en hacer una transición de las “condiciones de laboratorio” con las que se trabajó en el proceso a un entorno real. El conocimiento descubierto debe ser utilizado según el objetivo definido en la primera fase. Se requiere visualizar los resultados obtenidos de una manera entendible e interpretable, que permita integrar el conocimiento adquirido a otro sistema para la toma de decisiones futuras.

3.3.2. Protección de datos

Para el desarrollo del trabajo se utilizan varias fuentes de datos, las cuales se pueden agrupar en dos tipos:

- Datos meteorológicos descargados de plataformas de acceso abierto por ejemplo NASA o NOAA que son datos públicos y pueden ser utilizados sin necesidad de solicitar un permiso explícito.

- Datos meteorológicos de estaciones de Oruro y La Paz proporcionados por el Servicio Nacional de Meteorología e Hidrología de Bolivia (SENAMHI) mediante solicitud formal. Los datos también están disponibles en la página de SENAMHI mediante el Sistema

SISMET que proporciona un acceso público a los mismos (SENAMHI, 2021) . Una de las líneas estratégicas del Gobierno Boliviano es la Transparencia y los Datos Abiertos cuyo objetivo es: “Promover la publicación, uso y reutilización de datos abiertos de las entidades públicas, para la generación de información con valor agregado para la población” (AGETIC, 2017).

En ambos casos los datos son meteorológicos y no corresponden a datos de carácter personal.

4. Desarrollo específico de la contribución

En éste apartado se muestran las contribuciones del proyecto, tomando como base la metodología para predicción de sequía planteada en la sección anterior. Se detallan las tareas realizadas en cada fase y los resultados alcanzados.

4.1. Definición del problema y objetivos.

Las sequías causan impactos negativos en el sector del altiplano boliviano, provocando grandes pérdidas económicas a los pobladores de la región afectada. Muchas instituciones gubernamentales como el Ministerio de Medio Ambiente, las Gobernaciones departamentales, Gobiernos Municipales o SENAMHI demandan información acerca de las sequías que les permita conocer su comportamiento actual mediante el seguimiento y monitoreo; pero también contar con herramientas que permitan tener un conocimiento de su comportamiento futuro. Dicho conocimiento podría ayudarles a generar planes de alerta temprana y prevención de desastres.

4.1.1. Definición del Objetivo.

En función del problema se define el siguiente objetivo: Realizar un análisis y predicción de sequía con la estimación de los índices SPEI y SPI mediante modelos de aprendizaje automático. Las escalas de tiempo para ambos índices son 3 y 12 meses y el área geográfica es el altiplano central de Bolivia, con datos de 24 estaciones meteorológicas en periodo de 1981 a 2020. Para mejorar las predicciones primero se realiza una clusterización de las estaciones para que los modelos posteriormente puedan obtener un resultado para cada clúster en lugar de crear un modelo para cada estación. Para evaluar y comparar los modelos se utilizan las métricas: MAE, MAPE, y RMSE y MSE
Una vez obtenidas las predicciones de sequía se realiza un análisis de los eventos secos en el periodo pronosticado, para lo cual se determina la intensidad, severidad y duración de las sequías identificadas.

4.1.2. Definición del área de estudio.

El área de estudio es el altiplano central de Bolivia, la zona está comprendida entre 17°5’ y 19° 20’ de latitud sud y entre 66°40’ y 69°29’ de longitud oeste. Comprende el departamento de Oruro y la parte sur del Departamento de La Paz.

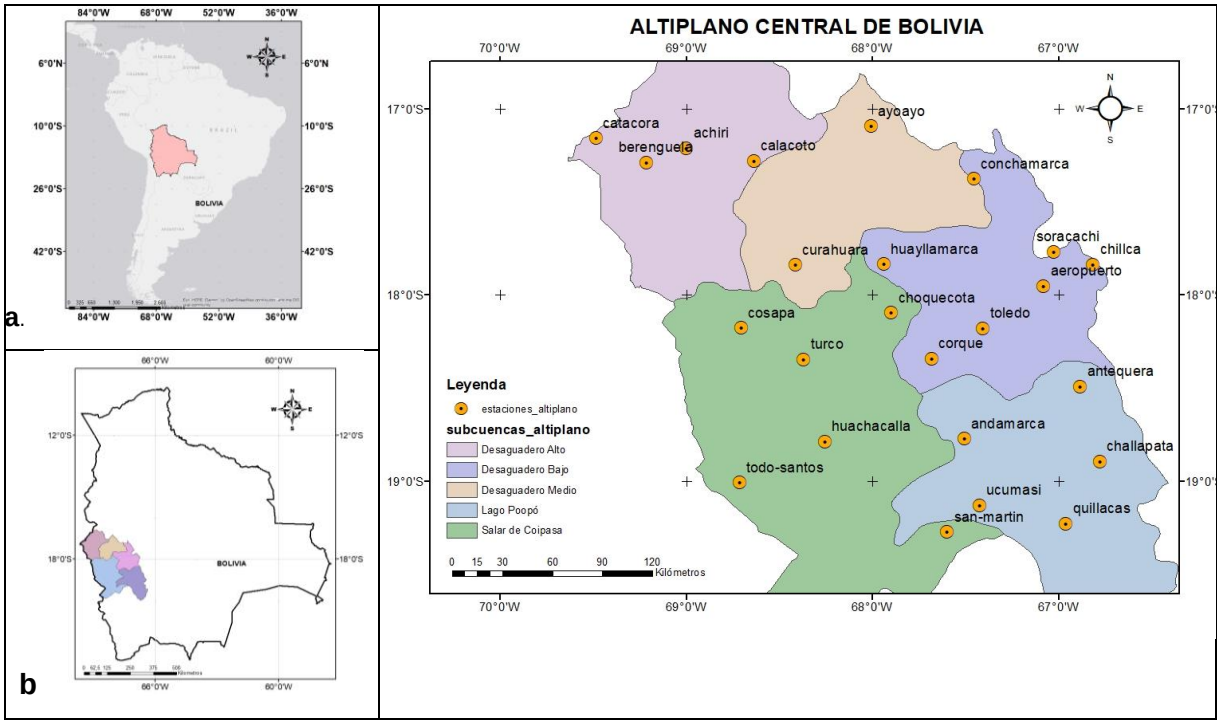


Figura 8. *Área de Estudio: Altiplano Central de Bolivia*

a. Mapa de Bolivia en Sudamérica, b. Mapa de Área de estudio en Bolivia c. Mapa de estaciones del altiplano central de Bolivia

El área de estudio comprende 5 subcuencas: Desaguadero Alto, Desaguadero Medio, Desaguadero Bajo, Cuenca Lago Poopó y Cuenca Salar de Coipasa. Además, se identificaron un total de 24 estaciones meteorológicas, 6 en el departamento de La Paz y 18 estaciones en el departamento de Oruro (Figura 8).

Según el Atlas Climatológico de Bolivia presentado por Campero (2013) la precipitación media anual en el altiplano de Bolivia está en el intervalo de 0 a 415 mm (p.11). Respecto a la temperatura máxima media anual está en el intervalo de 12°C a 20°C (p.63), la temperatura mínima media anual está entre los -9°C y 4°C (p.115) y la humedad relativa media anual está en el intervalo de 20 a 70 % (p.167). Respecto a la altura, la zona de estudio está en el intervalo 3850 y 4300 metros sobre el nivel del mar.

4.2. Recolección y Selección de datos

En primer lugar, se hace la recolección de datos necesarios para alcanzar el objetivo planteado en el apartado anterior, luego se realiza una selección de los mismos de acuerdo a criterios establecidos.

Según el objetivo planteado para el análisis de sequía se requiere el cálculo de dos índices: SPI y SPEI. Para el SPI solo se necesita la variable precipitación y para el SPEI, las variables

precipitación y evapotranspiración, que a su vez se calcula con datos de temperatura media (Thornthwaite) o temperatura máxima y temperatura mínima (Hargreaves). En el trabajo actual se utiliza la ecuación de Hargreaves; pero en el código se tiene implementada también la ecuación de Thornthwaite. Además de las variables mencionadas se considerarán otras variables meteorológicas, que podrían estar relacionadas con las sequías como ser: humedad relativa media y adicionalmente la variable Índice Oceánico del Niño 3.4.

- **Datos de Estaciones Meteorológicas**

  De acuerdo a la zona de estudio definida, se identificaron un total de 24 estaciones meteorológicas e hidrológicas administradas por el SENAMHI de Bolivia (Ver Tabla 3).

  De todas las principales variables meteorológicas que registran diariamente las estaciones, muchas de ellas solo se registran en algunas estaciones y periodos de tiempo. Las cinco variables que están presentes en la mayoría de las estaciones son: precipitación, temperatura máxima, temperatura mínima, temperatura media y humedad relativa media a escala diaria.

  Los datos proporcionados por SENAMHI inicialmente estaban en tres formatos distintos (Anexo I), los cuales fueron uniformados mediante código en Python (Anexo II). Finalmente se generaron archivos .csv para cada estación con la estructura uniforme de la Tabla 4.

- **Datos del Índice Oceánico del Niño 3.4**

  Oceanic Niño Index (ONI). Es utilizado para identificar sucesos relacionados con El Niño y La Niña en la zona tropical del Pacífico. El niño está relacionado con eventos cálidos y la Niña con eventos fríos. Es la anomalía de la Temperatura superficial del mar o Sea Surface Temperature (SST) en la zona 3.4. Tanto el evento del Niño como de la Niña se identifican a partir de 5 periodos consecutivos superpuestos de tres meses de +0.5 °C para eventos cálidos y por debajo de -0.5°C de anomalía en el caso de eventos fríos (NOAA, 2021). Algunos estudios resaltaron la relación entre el índice ONI y los episodios de sequía (Canedo-Rosso et al., 2021, p. 999). En este trabajo se plantea utilizar el índice como una variable exógena para la estimación del SPI y el SPEI. Los datos fueron uniformados al formato de la Tabla 4 (Ver Código en Anexo III).

Tabla 3. Estaciones meteorológicas del área de estudio

ID	ESTACIÓN	DEPARTAMENTO	ALTITUD	LONGITUD	LATITUD
100	Aeropuerto	Oruro	3702	-67.080	-17.953
101	Andamarca	Oruro	3762	-67.506	-18.772
102	Antequera	Oruro	4057	-66.883	-18.493
105	Chillca	Oruro	4025	-66.814	-17.837
106	Choquecota	Oruro	3867	-67.899	-18.097
107	Corque	Oruro	3758	-67.679	-18.344
108	Cosapa	Oruro	3906	-68.706	-18.178
111	Huachacalla	Oruro	3746	-68.257	-18.787
112	Huayllamarca	Oruro	3873	-67.940	-17.836
113	Quillacas	Oruro	3724	-66.960	-19.230
114	San-Martin	Oruro	3712	-67.599	-19.275
115	Soracachi	Oruro	3802	-67.025	-17.768
116	Todo-Santos	Oruro	3805	-68.715	-19.008
117	Toledo	Oruro	4038	-67.406	-18.179
118	Turco	Oruro	3842	-68.369	-18.346
119	Ucumasi	Oruro	3764	-67.425	-19.130
120	Challapata	Oruro	3733	-66.778	-18.896
121	Curahuara	Oruro	3917	-68.414	-17.839
200	Achiri	La Paz	3880	-68.999	-17.212
201	Ayo Ayo	La Paz	3888	-68.008	-17.094
202	Berenguela	La Paz	4120	-69.214	-17.289
203	Calacoto	La Paz	3826	-68.636	-17.281
204	Catacora	La Paz	4253	-69.486	-17.159
205	Conchamarca	La Paz	3965	-67.455	-17.377

Elaboración propia en base a (SENAMHI, 2021).

Tabla 4. Datos meteorológicos en formato uniforme

fecha	pp	tmax	tmin	tmed	hmed
1/7/2010	0	17.6	-8.2	9.2	49
2/7/2010	0	17.8	-7.8	9.3	43
3/7/2010	0	20.1	-7	10.8	40
4/7/2010	0	17.6	-6.4	9.6	41
5/7/2010	0	20.4	-6.6	10.2	42
6/7/2010	0	17	-6.8	8.6	42
7/7/2010	0	15.4	-8.6	7.4	38
8/7/2010	0	17.4	-8.4	8.9	47
9/7/2010	0	16.8	-9.2	7.6	39
10/7/2010	0	16.5	-6.3	8.8	44
11/7/2010	0	14.1	-14	5	39
12/7/2010	0	18.2	-13.6	6.6	36
13/7/2010	0	17.6	-13	13.3	22
14/7/2010	0	17.2	-13.2	7.2	32
15/7/2010	0	15.8	-10	7.8	32
16/7/2010	0.8	16.2	-7.3	8	41
17/7/2010	2.3	11.4	2.6	6.7	70

4.3. Preprocesamiento y limpieza

4.3.1. Descripción de los datos

Se verifica la cantidad de instancias y atributos con los que se cuenta. Los datos registrados en las estaciones meteorológicas no son completos, algunas estaciones están activas solo por un periodo de tiempo. En consecuencia, se hizo un análisis descriptivo de los datos faltantes respecto a las variables que se utilizan para el cálculo de los índices SPI y SPEI.

- **Verificación de Datos Faltantes de Precipitación**

La variable precipitación es la principal en el análisis de sequía. La figura 9 (posición superior izquierda) muestra los datos faltantes por estación, el color claro indica un dato nulo o faltante y el color negro los datos presentes. Las estaciones: Aeropuerto de Oruro (id 100), Achiri (id 200), Ayo Ayo (id 201), Berenguela (id 202), Calacoto (id 203) y Conchamarca (id 205) son las estaciones con menor cantidad de faltantes y las estaciones Corque (id 107) y Soracachi (id 115) son aquellas con más valores faltantes.

Tomando en cuenta el periodo de estudio: 1981 a 2020, se determinó que solo 6 estaciones cumplen con el porcentaje de datos faltantes aceptable menor o igual a 10%, utilizado en trabajos relacionados (Canedo-Rosso et al.,2019; Cerano P. et al., 2020). La Tabla 5 muestra el conteo de datos faltantes en la variable precipitación en las 6 estaciones meteorológicas con menos de 10% de datos faltantes, asimismo se pueden ver la cantidad de datos faltantes respecto a las otras variables (Ver el código en Anexo IV).

- **Verificación de Datos Faltantes de Temperatura Mínima, Máxima y Media**

La figura 9 (posición superior derecha y central) muestra los datos faltantes por estación de las variables temperatura mínima, máxima y media. Las estaciones: Aeropuerto de Oruro (id 100), Ayo Ayo (id 201) y Calacoto (id 203) son las estaciones con menor cantidad de faltantes y las estaciones Ucumasi (id 119), Berenguela (id 202) y Conchamarca (id 205) son aquellas con más valores faltantes, ya que son estaciones hidrológicas, que solo registran la precipitación.

Tabla 5. *Estaciones con datos faltantes de precipitación menores a 10%*

ID	precipitacion	temp_max	temp_min	temp_med	hum_med	estacion
201	0.99	1.55	1.49	100.00	10.81	ayoayo
203	1.10	1.62	1.51	1.66	13.05	calacoto
100	1.33	2.95	2.07	76.83	37.42	aeropuerto
202	3.19	100.00	100.00	100.00	100.00	berenguela
205	5.50	100.00	100.00	100.00	100.00	conchamarca
200	6.19	35.54	35.48	35.56	37.05	achiri

- **Verificación de Datos Faltantes de Humedad Relativa Media**

La figura 9 (posición inferior) muestra los datos faltantes por estación de la variable humedad relativa media. Las estaciones: Aeropuerto de Oruro (id 100), Ayo Ayo (id 201) y Calacoto (id 203) son las estaciones con menor cantidad de faltantes y las demás estaciones tienen una gran cantidad de datos faltantes. Tomando en cuenta la gran cantidad de datos faltantes en las estaciones meteorológicas respecto a las variables más importantes en el estudio, especialmente precipitación, se busca otras fuentes alternativas de datos. Para ello se retorna a la Fase 1 Recolección de datos y se encuentra dos fuentes de datos que proveen la información requerida.

- **Datos Satelitales de Precipitación de CHIRPS v.2**

Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) almacena datos de precipitación, su cobertura espacial es de 0.05° (aproximadamente 5 kilómetros). Su cobertura es amplia: desde 50° de latitud Sud a 50°de latitud Norte, respecto a la longitud tiene una cobertura total. (Funk et al., 2014). La precipitación disponible para descarga es diaria, a nivel de péntada o mensual, desde el año 1981 hasta la actualidad están disponibles para descarga (CHC, 2021). La base de Datos CHIRPS es ampliamente utilizada para trabajos de investigación porque los datos satelitales son corregidos con datos observados de estaciones hidrológicas o meteorológicas.

Para este trabajo se realizó la descarga de los archivos diarios de precipitación agrupados por gestión en formato NETCDF[1], se realizó una extracción del área de estudio para reducir su tamaño y luego se los procesó mediante Python (Anexo V) para obtener los datos respecto a las 24 estaciones meteorológicas definidas en el anterior punto, dando como resultado archivos en formato csv con las mismas características que los datos procesados de las estaciones.

[1] NETCDF es un contenedor que incluye tanto datos científicos orientados a matrices como sus metadatos, de tal manera que se consideran auto descriptivos(Car et al., 2017).

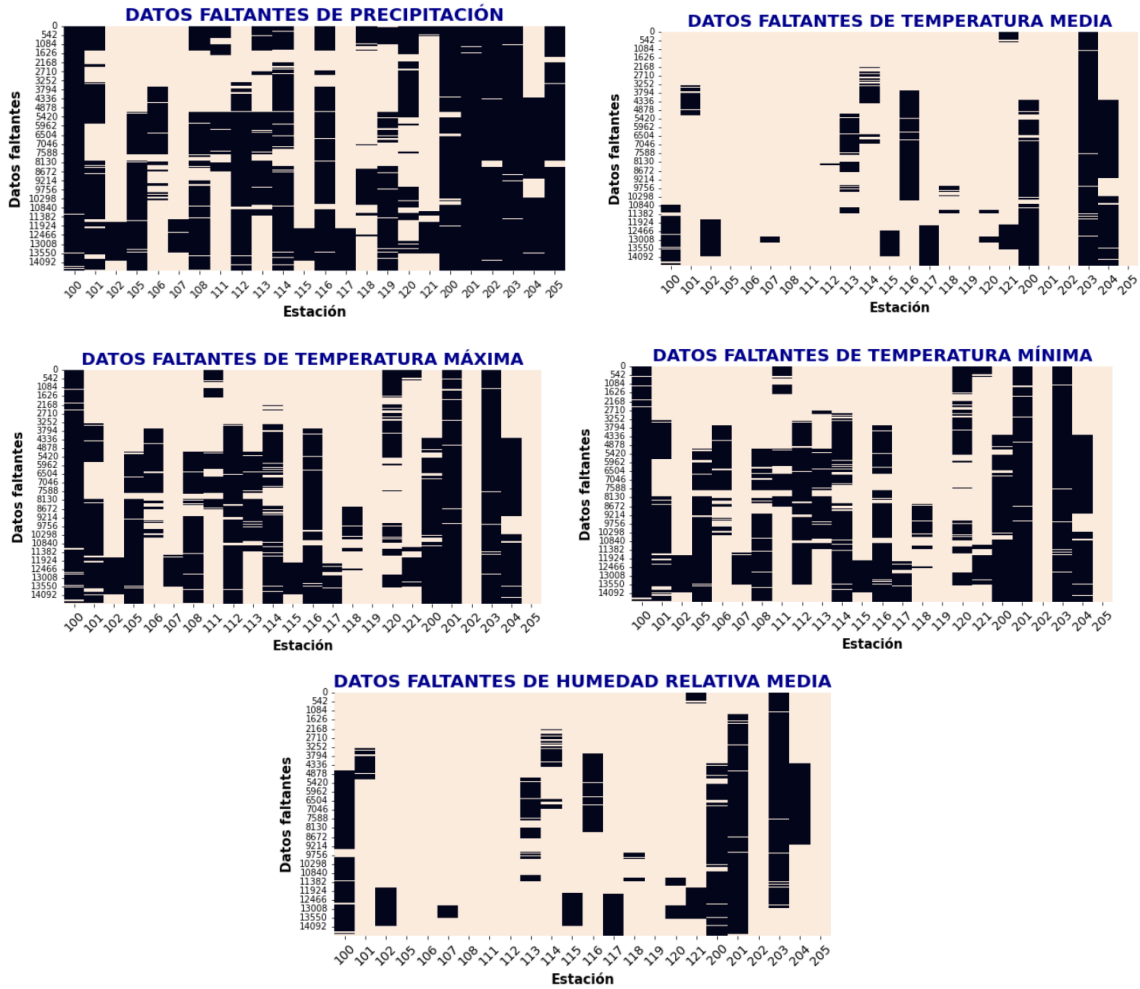


Figura 9. Datos faltantes de precipitación, temperatura y humedad relativa

- **Datos Satelitales de Humedad y Temperatura de NASA**

  Para los datos de temperatura y humedad se utilizó la base de datos POWER (NASA, 2021) administrada por la National Aeronautics and Space Administration (NASA). Se descargaron datos de precipitación, temperatura mínima, temperatura máxima, temperatura media y humedad relativa media, todos a escala diaria. De la misma manera que en el caso de los archivos CHRIPS, éstos fueron procesados respecto a las 24 estaciones definidas para el estudio y se obtuvo como resultado archivos .csv con el mismo formato que los otros datos. El proceso se llevó a cabo mediante Python (Ver Anexo VI)

4.3.2. Correlación entre datos observados y satelitales

Para utilizar los datos satelitales en lugar de los datos observados se debe verificar que existe una correlación entre los mismos y para dicha verificación se siguen los siguientes pasos:

- **Verificación de datos atípicos y rellenado de datos faltantes en las series observadas**

  Antes de verificar la correlación de las series temporales observadas con las derivadas de satélite, se rellenan los datos faltantes y se homogenizan valores atípicos, para ello se utiliza el paquete "Climatol" (Guijarro, 2019b) creado específicamente para el tratamiento de datos climáticos. Según su documentación para la homogenización de los datos se basa en un modelo de regresión ortogonal (Guijarro, 2019a).

  Climatol está desarrollado en lenguaje R y para utilizarlo en el presente proyecto se instaló el paquete IRKernel disponible en Github (IRkernel, 2021), y a continuación se instaló el Kernel R en Jupyter Notebook. Con ese procedimiento se pudo ejecutar comandos R dentro del entorno de Jupyter Notebook (Anexo VII).

  Para la homogenización de los datos se realizaron dos tareas principales:

  - **Rellenado de datos faltantes.** Para el llenado de datos faltantes, el paquete Climatol permite verificar la similitud entre los datos de las diferentes estaciones meteorológicas, tomando como principal factor la cercanía entre las mismas (latitud, longitud) para el cálculo de series completas que luego sirven de referencia para el llenado de los datos.

  - **Detección y eliminación de datos atípicos.** Cuando se analiza la precipitación es importante que no se tengan valores negativos, tampoco valores muy altos, ya que el área de estudio corresponde al altiplano, lo mismo pasa con la temperatura que está dentro de un rango y un valor muy alejado podría significar un error de registro. En la Tabla 6 se muestra un reporte de anomalías encontradas por Climatol en las series de precipitación diaria y en la Tabla 7 las anomalías encontradas en las series de precipitación mensual.

    Finalmente, después de corregir los datos atípicos y rellenar las series diarias y mensuales se exportaron los archivos csv con el mismo formato de todos los datos generados en el trabajo (Anexo VIII).

    Para verificar la calidad de los datos homogenizados por Climatol se realizaron las pruebas de correlación, para lo cual se utilizaron las pruebas de Pearson y de Spearman, entre ambas series temporales originales y las corregidas (Anexo IX).

Tabla 6. Datos atípicos de precipitación y temperatura diaria

Estación	Fecha	Variable	Valor Observado	Valor Sugerido	Anomalía en Desviaciones Estándar
108	19/02/2009	Precipitación	131.6	3.3	32.76
116	21/12/2008	Precipitación	0	59.0	-19.86
118	21/12/2008	Precipitación	161.1	34.3	26
107	15/12/2012	Temperatura mínima	-20	5.8	-13.45

Fuente: Reporte de Climatol

Tabla 7. Datos atípicos de precipitación y temperatura mensual

Estación	Fecha	Variable	Valor Observado	Valor Sugerido	Anomalía en Desviaciones Estándar
202	1/01/1984	Precipitación	617	145.3	11.03
202	1/02/1984	Precipitación	580.4	148.0	11.78
116	1/11/2013	Temperatura máxima	31.3	19.9	6.56
106	1/08/2003	Temperatura mínima	9	-9.3	8.66

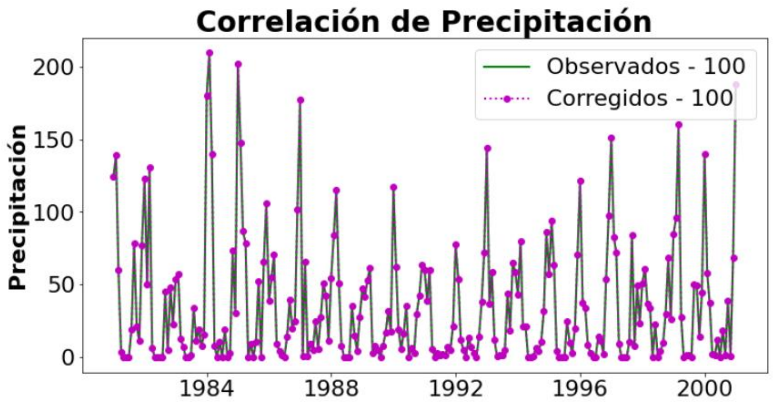
Fuente: Reporte de Climatol

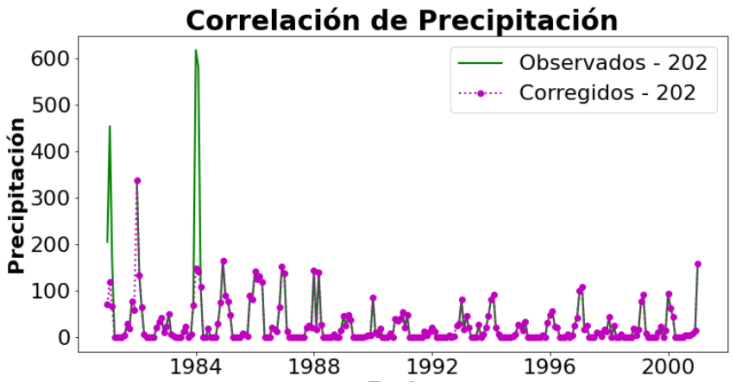
La Figura 10 muestra las gráficas de las series temporales mensuales de precipitación, temperatura máxima y temperatura mínima observadas, es decir originales, en comparación con las homogenizadas por Climatol para algunas estaciones representativas de la zona de estudio, se puede apreciar que solo se corrigieron los datos atípicos identificados en las Tablas 6 y 7 y se rellenaron los datos faltantes. También se observa a la derecha de las gráficas los coeficientes de correlación obtenidos para cada variable que algunos casos muestran un valor de 1 con una correlación perfecta y en la mayoría superan el valor de 0.97.

El siguiente paso consiste en verificar la correlación entre los datos observados corregidos con Climatol y los datos derivados de satélite CHIRPS y NASA, para lo cual se utilizó una vez más las correlaciones de Pearson y Spearman, además de las gráficas de series temporales (Anexo X).

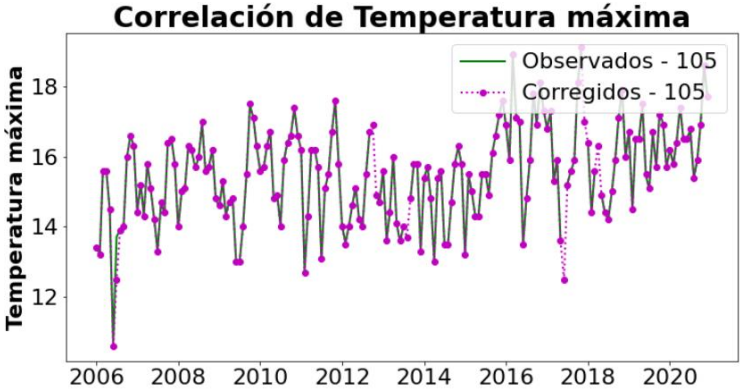
- **Verificación de correlación entre precipitación observada corregida y CHIRPS**

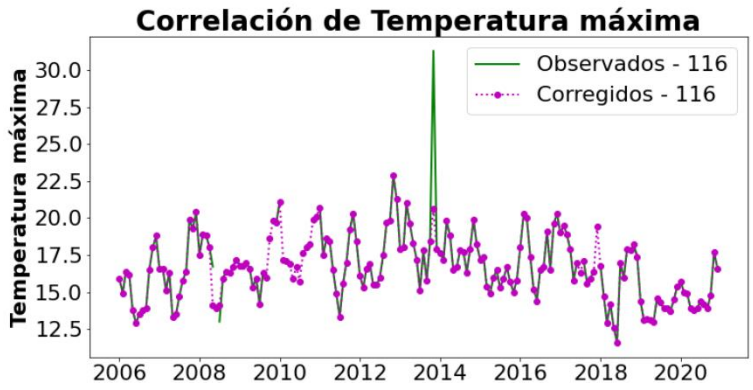
La Tabla 8 muestra los coeficientes de correlación Pearson y Spearman obtenidos en la comparación de las series temporales mensuales de precipitación observada corregida de algunas estaciones meteorológicas con los datos mensuales de precipitación CHIRPS. Se observa claramente que la correlación es alta con valores por encima de 0.71.



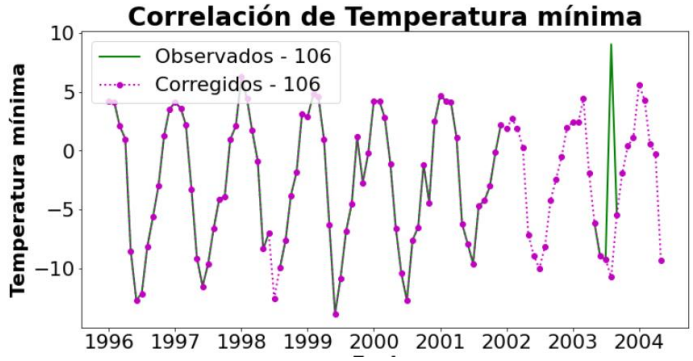


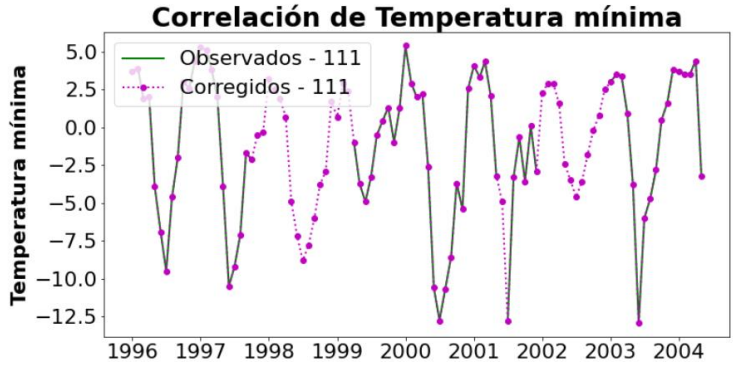
	spearman	pearson
100	0.99974	0.999984
101	1.0	1.0
102	0.999264	0.999957
105	1.0	1.0
111	1.0	1.0
118	0.994543	0.998207
121	0.680845	0.955173
202	0.99965	0.816457
204	1.0	1.0





	spearman	pearson
100	0.997491	0.996912
101	0.998776	0.997641
102	0.999507	0.999845
105	0.99935	0.998756
111	0.999685	0.99991
116	0.996697	0.96119
120	0.987544	0.989624
200	0.999543	0.999765
204	0.719786	0.733181





	spearman	pearson
100	0.99856	0.998204
101	0.998654	0.999271
102	0.996561	0.992704
106	0.97156	0.963617
111	0.999926	0.999982
116	0.999883	0.999948
120	0.995701	0.997045
200	0.999948	0.999978
204	0.999939	0.999969

Figura 10. Series Observadas vs Series Observadas corregidas

Tabla 8. Correlación entre precipitación observada corregida y CHIRPS

	100	101	102	108	112	116	118	120	121	200
spearman	0.911596	0.772802	0.911686	0.777544	0.82552	0.719997	0.826588	0.84815	0.871865	0.850473
pearson	0.909629	0.80923	0.925648	0.766194	0.802389	0.861891	0.851161	0.857728	0.772747	0.884211

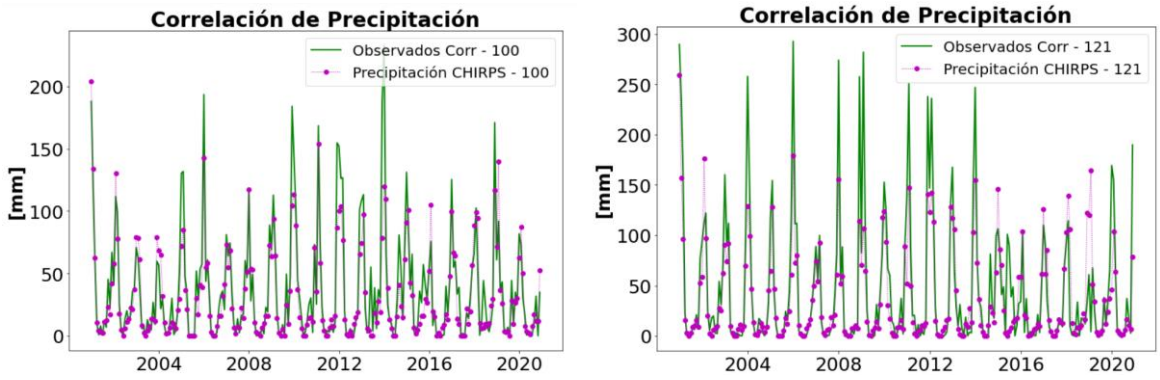


Figura 11. Precipitación mensual observada vs CHIRPS

La Figura 11 muestra las gráficas de ambas series donde se aprecia la buena correlación mencionada; sin embargo, para valores altos de precipitación observada, CHIRPS muestra valores menores, es decir que subestima los valores extremos de precipitación.

- **Verificación de correlación entre temperatura observada corregida y NASA**

La correlación entre los datos observados corregidos y los datos obtenidos de temperatura máxima y temperatura mínima de la NASA se presenta en la Tabla 9. Algunas estaciones muestran una muy buena correlación; pero otras estaciones la correlación es menor, llegando a un valor aproximado de 0.70 lo que se considera aceptable.

La figura 12 muestra algunas gráficas de las series temporales de temperaturas comparadas. Los resultados para la temperatura mínima son mejores que para la temperatura máxima.

- **Verificación de correlación entre humedad relativa observada y NASA**

La correlación entre los datos observados homogenizados y los datos obtenidos de humedad relativa media de la NASA se presentan en la Tabla 9. Respecto a la variable humedad, muy pocas estaciones cuentan con datos a largo plazo, la mayoría cuenta con solo algunos años de registro. Los valores obtenidos de correlación no son tan buenos como en el caso de la precipitación y temperatura.

La figura 12 muestra algunas gráficas de las series temporales de humedad relativa. En la derecha se aprecia una serie con una gran cantidad de datos disponibles y el caso de la derecha se tienen muchos datos faltantes.

Tabla 9. Correlación entre variables observadas con NASA

	100	101	102	106	107	113	114	115	117
spearman	0.880642	0.845153	0.86563	0.810065	0.832906	0.872432	0.815254	0.821056	0.845328
pearson	0.883745	0.832686	0.85338	0.821876	0.820135	0.868829	0.797875	0.803288	0.840419

Temperatura máxima

	100	101	102	106	107	113	114	115	117
spearman	0.950162	0.944795	0.938601	0.931005	0.95037	0.88279	0.901091	0.93883	0.943865
pearson	0.969579	0.951117	0.95494	0.949113	0.966129	0.897961	0.88362	0.956549	0.959564

Temperatura mínima

	100	102	107	115	117	118	120	121	201
spearman	0.808655	0.930596	0.923077	0.930813	0.734373	0.576623	0.895632	0.553162	0.795645
pearson	0.791637	0.931289	0.923638	0.936275	0.712276	0.614846	0.903723	0.501146	0.787499

Humedad relativa

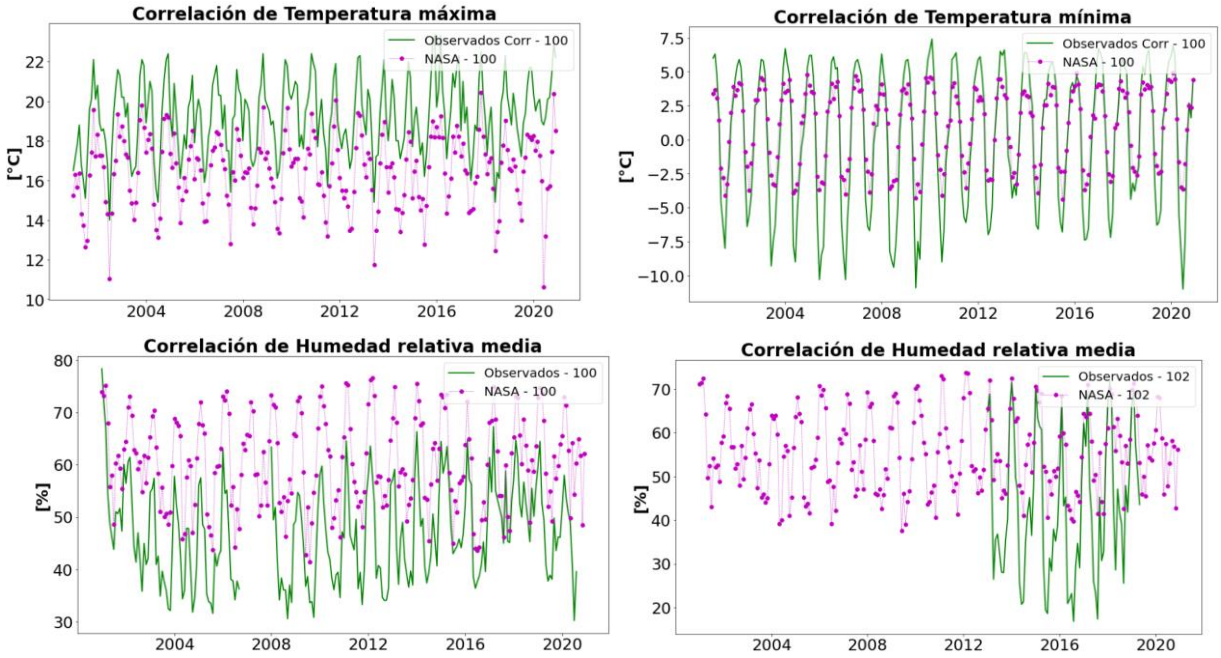


Figura 12. Correlación entre Temperatura - Humedad observada y NASA

4.3.3. Corrección BIAS de los datos satelitales

Si bien es cierto que la correlación entre los datos observados de precipitación, temperatura máxima, temperatura mínima con relación a los datos satelitales es buena, también se evidenció en las gráficas que en muchos casos los valores difieren, sobre todo cuando se trata de extremos mínimos o máximos, es decir los datos satelitales subestiman los valores extremos. En el tema de sequía los valores extremos son importantes, en consecuencia, se realizó una corrección Bias para las variables precipitación y temperatura. La corrección Bias

para la precipitación se basa en la ecuación (10), adaptada de (Teutschbein & Seibert, 2012, p. 16), que utiliza el método de escalado lineal. Donde $P_{sat}(d)$ es la precipitación satelital, $\mu_m(P_{obs}(d))$ es la media mensual de precipitación observada y de manera similar $\mu_m(P_{sat}(d))$ es la media mensual de precipitación satelital.

$$P^*_{sat}(d) = P_{sat}(d) \cdot \left[\frac{\mu_m(P_{obs}(d))}{\mu_m(P_{sat}(d))}\right] \quad (10)$$

La corrección Bias para la temperatura se basa en la ecuación (11). Donde $T_{sat}(d)$ es la temperatura satelital, $\mu_m(T_{obs}(d))$ es la media mensual de temperatura observada y de manera similar $\mu_m(T_{sat}(d))$ es la media mensual de temperatura satelital.

$$T^*_{sat}(d) = T_{sat}(d) + \mu_m(T_{obs}(d)) - \mu_m(T_{sat}(d)) \quad (11)$$

La corrección Bias se implementó mediante código en Python (Anexo XI) y posteriormente se realizaron las pruebas de correlación entre los datos observados y los datos satelitales corregidos (Anexo XII) con resultados muy buenos. Los resultados de correlación que se muestran en la tabla 10 son mejores que los obtenidos en la tabla 8 respecto a precipitación y que la tabla 9, respecto a temperatura, antes de la corrección Bias. En la figura 13 se puede visualizar la mejoría de los datos de CHIRPS de precipitación con la corrección Bias y de los datos de NASA de temperatura con la corrección Bias.

Después de realizadas las pruebas de correlación correspondientes se concluye que los datos descargados de CHIRPS y NASA y corregidos mediante la corrección Bias, guardan una buena correlación con los datos observados por lo que se utilizarán a partir de la siguiente sección en lugar de los datos observados.

4.3.4. Cálculo de SPI y SPEI

Se integra los datos de corregidos de precipitación de CHIRPS y temperatura máxima, mínima, media y humedad relativa media de la NASA en una sola base de datos y luego se realiza el cálculo de SPI y SPEI.

Para el cálculo de ambos índices se utilizó el paquete SPEI que está implementado en R (Beguería & Vicente-Serrano, 2017). Para tener todos los procedimientos del trabajo integrados en Jupyter se utilizó el Kernel de R para utilizar el paquete SPEI en un cuaderno de Jupyter Notebook (Código en Anexo XIII). Las escalas temporales elegidas para el cálculo de los índices fueron: 3 y 6 meses para un análisis a corto y mediano plazo y la escala de 12 meses para análisis a largo plazo.

Tabla 10. Correlación entre datos observados y corregidos BIAS

	100	101	102	105	106	107	108	111	112	113
spearman	0.921657	0.780728	0.927977	0.899558	0.866346	0.882808	0.804799	0.833357	0.825944	0.855226
pearson	0.920984	0.815312	0.931058	0.894135	0.863008	0.875573	0.788866	0.811254	0.805112	0.885146
MAE	11.387575	13.398337	10.317383	11.493381	22.598494	8.186473	14.79905	15.914481	18.101065	9.866554

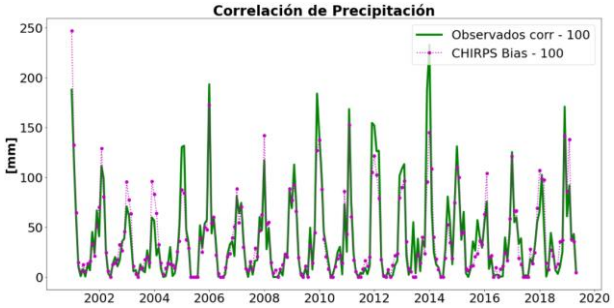
Precipitación

	100	101	102	105	106	107	108	111	112	113
spearman	0.890727	0.858777	0.852183	0.506378	0.813431	0.853394	0.501006	0.669577	0.669473	0.886049
pearson	0.893466	0.839603	0.837409	0.506828	0.825596	0.840176	0.509937	0.646741	0.633571	0.881322
MAE	0.724302	0.875471	0.584056	1.00896	0.726785	0.674592	1.608152	1.020421	1.083838	0.907896

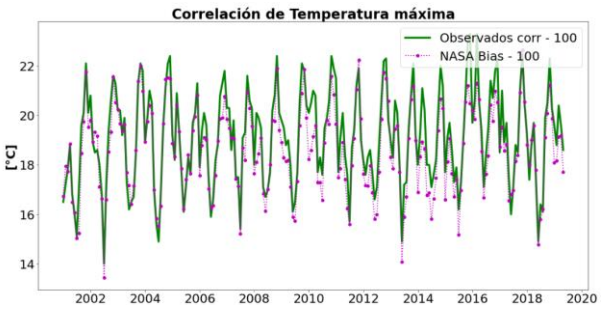
Temperatura máxima

	100	101	102	105	106	107	108	111	112	113
spearman	0.966064	0.950327	0.927659	0.838765	0.94927	0.95776	0.81724	0.920249	0.852498	0.895344
pearson	0.976521	0.956674	0.935987	0.870402	0.960914	0.97077	0.802008	0.919425	0.863511	0.908227
MAE	0.859362	0.90164	0.526042	1.729104	1.282056	0.732306	2.153231	1.359138	1.07135	1.631225

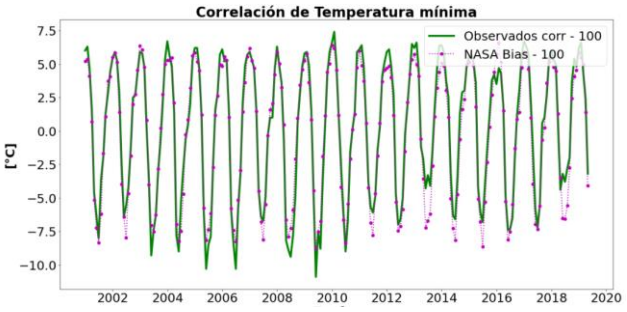
Temperatura mínima



Precipitación observada y CHIRPS Bias



Temperatura Máxima observada y NASA Bias



Temperatura Mínima y NASA Bias

Figura 13. Correlación de precipitación -temperatura observadas y corregidos BIAS

La figura 14 permite comparar los índices SPI y SPEI en escalas de 3 y 12 meses para cuatro estaciones meteorológicas representativas. (Anexo XIV). Se puede apreciar la diferencia entre las escalas de tiempo, claramente en el caso de SPI3 muestra datos con una mayor variabilidad. También se puede visualizar las diferencias entre SPI y SPEI, uno de los aspectos importantes a destacar es que el índice SPEI captura de mejor manera los valores extremos negativos que van relacionados con sequía.

4.3.3. Análisis Exploratorio

Una vez calculados los índices ya se cuenta con todas variables meteorológicas planificadas para la predicción. En consecuencia, se realiza la descripción de todos los datos de la base de datos, donde se aplica en primer lugar estadística descriptiva para conocer las características de cada variable. (Anexo XV).

La figura 15 muestra los diagramas de caja (boxplot) para las variables precipitación, Temperatura Máxima, Temperatura Mínima, Humedad Relativa Media, comparando su comportamiento en los 12 meses del año. Se realiza éste análisis debido a que los comportamientos de las variables meteorológicas varían respecto al mes. Por ejemplo, en los meses de enero y diciembre la precipitación es mucho mayor que en los meses de junio y julio.

También se realizó un análisis de la distribución de todas las variables. En el caso de la precipitación claramente no se sigue una distribución normal, en la temperatura máxima, mínima y humedad relativa la distribución se asemeja más a la normal.

Respecto a las series temporales también se tienen métodos específicos diseñados para realizar un análisis exploratorio. Los más importantes son los diagramas de línea, la figura 16 visualiza las gráficas de la variable precipitación y variable temperatura máxima.

- **Índice Oceánico del Niño ONI.** La Tabla 11 muestra las características de la variable global ONI. La media es 0.275, la desviación estándar es 0.65. El valor mínimo es -.13 y el máximo 1.6

  La figura 17 visualiza los diagramas de caja y la distribución del índice ONI.

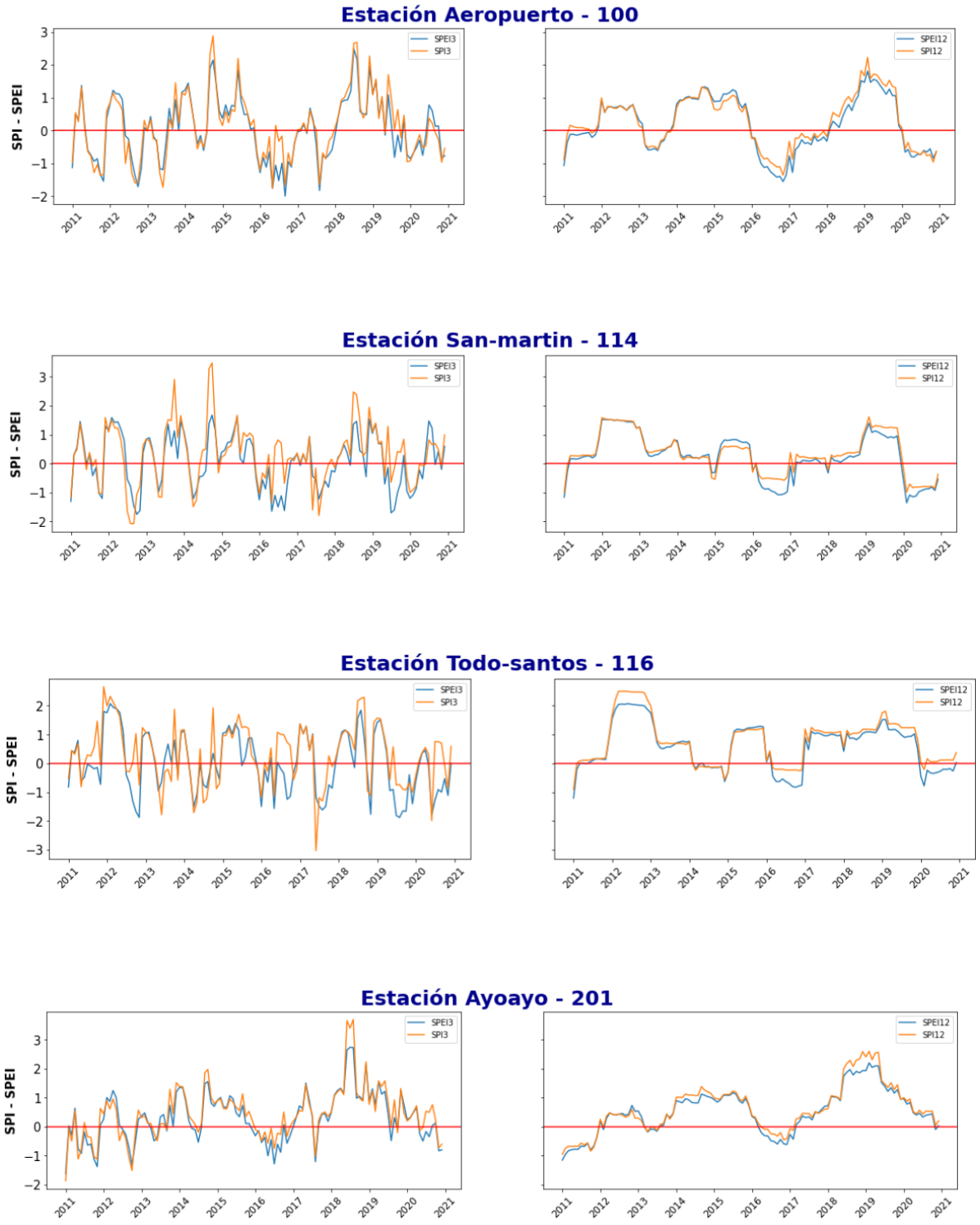


Figura 14. Comparación entre SPI y SPEI para escalas de 3 y 12 meses

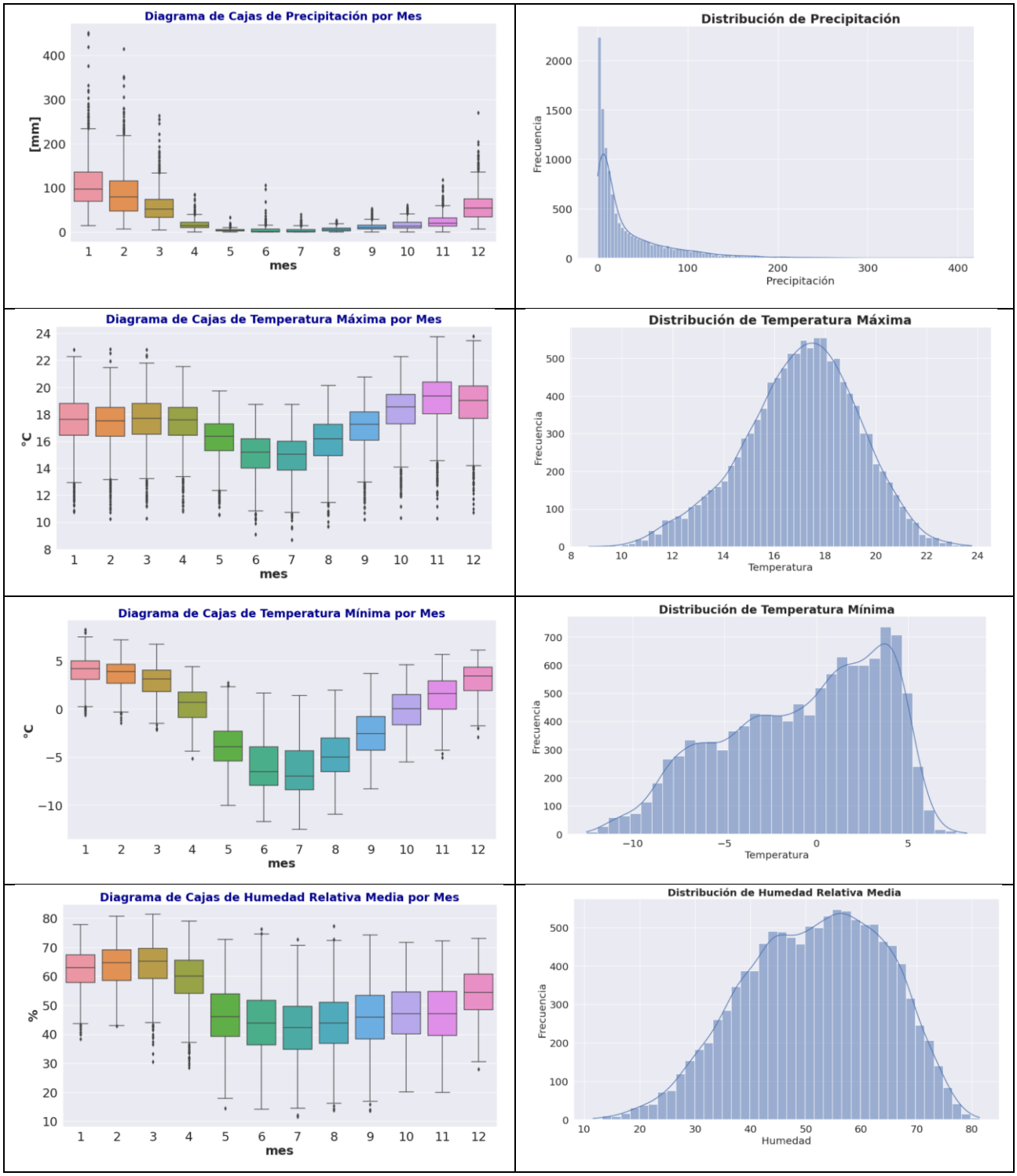


Figura 15. *Diagramas de caja y Distribución mensual de las variables*

Tabla 11. Análisis descriptivo de la variable ONI.

	count	mean	std	min	25%	50%	75%	max
oni	40.0	0.0275	0.658665	-1.3	-0.4	0.05	0.325	1.6

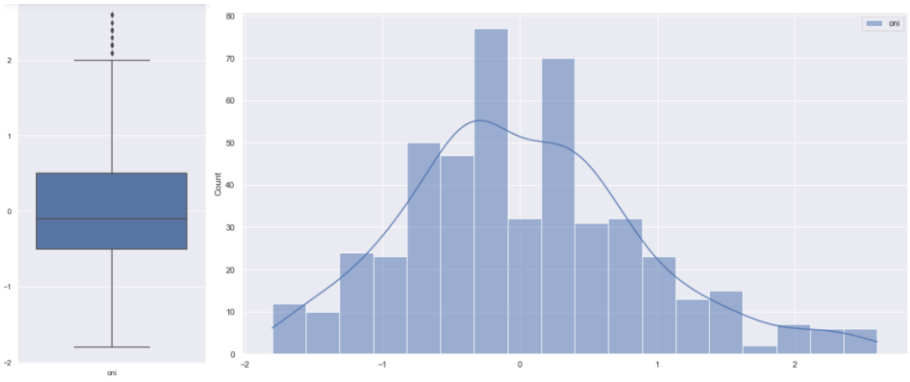


Figura 16. Diagrama de caja y Distribución de ONI

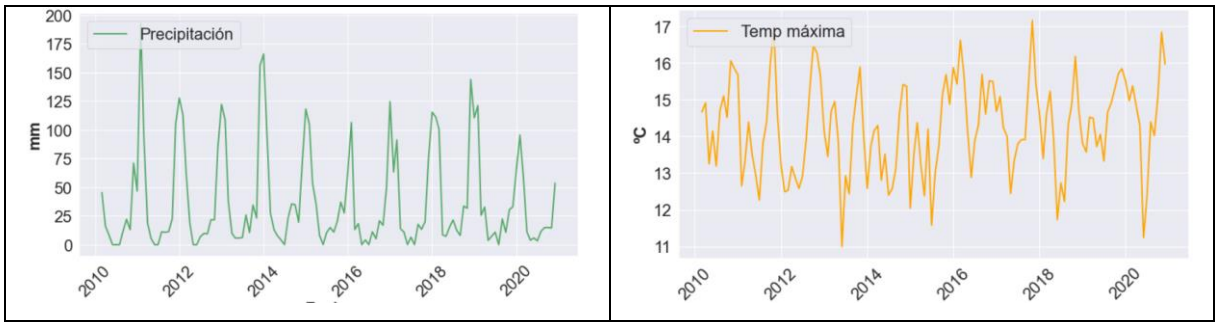


Figura 17. Serie de tiempo Precipitación y temperatura máxima

También las gráficas de Autocorrelación ACF y Autocorrelación Parcial PACF son muy importantes para visualizar características intrínsecas a las series temporales. Este análisis se realizó en otro cuaderno de Jupyter Notebook (Anexo XVI). Antes de realizar las gráficas de ACF y PACF se realiza una prueba de Estacionariedad denominada ADF o Augmented Dickey–Fuller. La prueba ADF aplicada a todas las series de SPI y SPEI dio como resultado que las series de SPI Y SPEI en la escala 3 son estacionarias. Pero en el caso de los índices de 12 meses a pesar de que el resultado de la prueba ADF que las series eran estacionarias, en las gráficas de ACF y PACF se pudo verificar que no era así.

Respecto a las gráficas de ACF y PACF, éstas son muy útiles para determinar algunas características de las series, por ejemplo, la estacionariedad o estacionalidad. La gráfica de PACF ayuda a determinar el parámetro “p” para los modelos ARIMA y la gráfica ACF para el parámetro “q”. En primer lugar, se grafican el ACF y PACF para el SPI tomando la escala de 3 meses. La figura 18 muestra el resultado. Según los resultados en ninguno de los casos el comportamiento de las auto correlaciones sugiere un modelo AR o MA sencillos, el valor de “p” podría ser 2, 3 o 4 y el valor de q es 2 en todas las estaciones. No se observa estacionalidad.

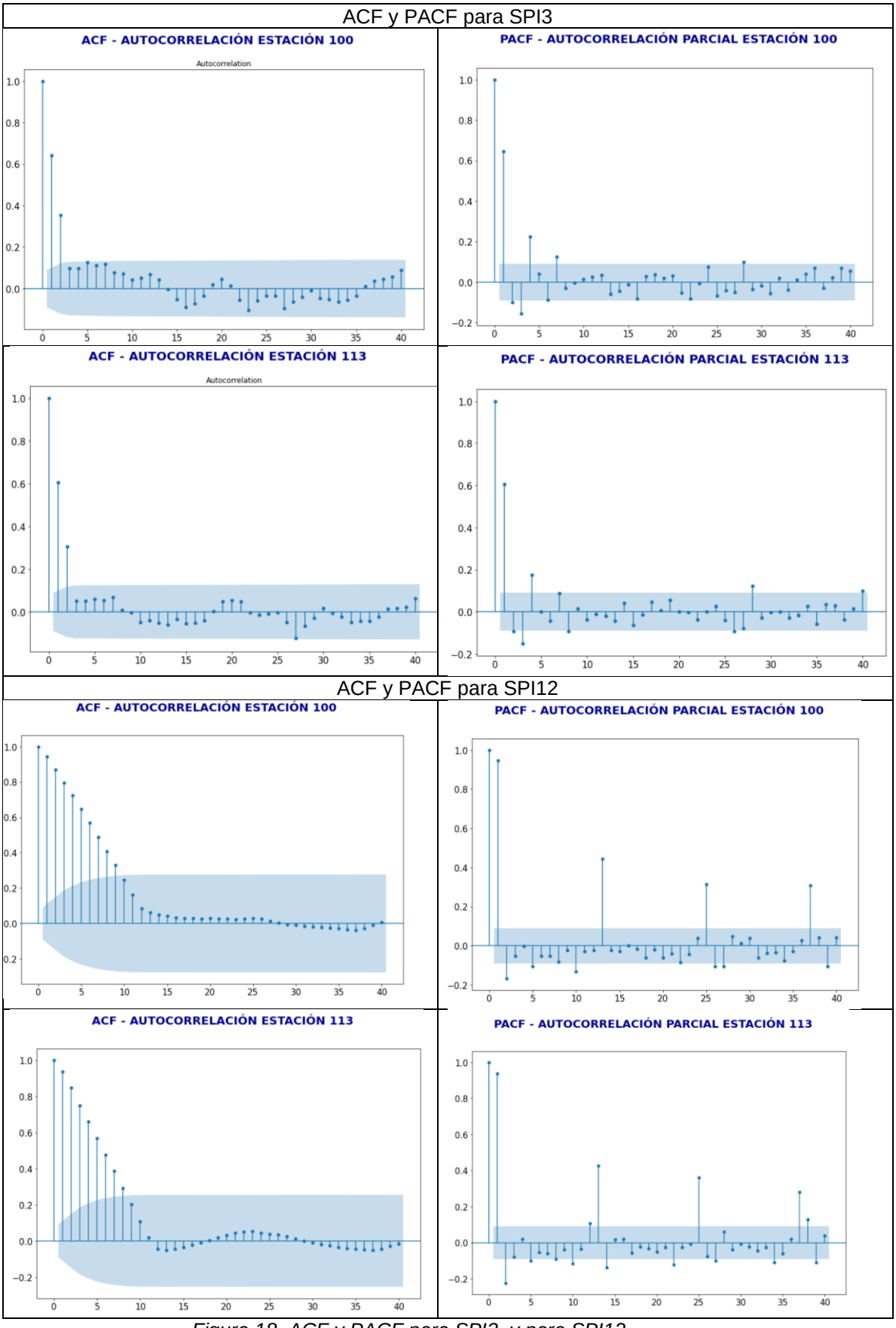


Figura 18. ACF y PACF para SPI3 y para SPI12

En la misma Figura 18 se aprecia las gráficas ACF y PACF para el caso de SPI12 y SPEI12, según los resultados se determina que las series presentan estacionalidad y claramente tiene un valor de 12 meses. Es decir que las series son no estacionarias. En éste caso para la etapa de modelado se tomará en cuenta la estacionalidad en el caso de éstas series.

- **Verificación de datos faltantes, atípicos y no válidos**

Se realiza una exploración de todas las variables, con la función describe de pandas. La Tabla 12 muestra las características descriptivas de las 9 variables para las 24 estaciones y para todo el periodo de análisis 1981 a 2020.

Se tiene un total de 11520 observaciones para pp, tmax, tmin, tmed y hmed que representan la precipitación, temperatura máxima, temperatura mínima, temperatura media y humedad relativa media respectivamente.

Sin embargo, se evidencian NaN en todos los índices de sequía, la tabla 13 muestra el detalle: Respecto a SPI3 y SPEI3 se tienen 48 faltantes que corresponden a los dos primeros meses de cada estación. En el caso de SPI12 y SPEI12 son 264 meses faltantes que conciernen a 11 meses de cada estación.

Todos los datos faltantes se generaron en el cálculo de los índices y no es correcto realizar la remoción de los mismos afectando a todo el Dataframe ya que la predicción se realizará respecto a cada índice y a cada escala de tiempo. La remoción de los datos faltantes se realizará antes de entrenar los modelos en las siguientes fases.

Además de los datos faltantes, en la tabla 12 se identifican datos no válidos en la variable SPEI3, específicamente es un valor –inf. Es importante realizar un tratamiento de ese tipo de datos y no removerlos directamente. Se realiza una gráfica de líneas para visualizar de mejor manera el dato erróneo y se verifica que en la estación 200 el dato corresponde a Julio de 2005

Para reemplazar el dato erróneo se aplica una función que busca en los valores históricos del mismo índice y de la misma estación y por la tendencia que se visualiza en la figura 19 no se calcula la media, en su lugar se calcula el valor mínimo (Anexo XV).

Tabla 12. Verificación de datos faltantes, atípicos y no válidos

	pp	tmax	tmin	tmed	hmed	spi3	spi12	spei3	spei12
count	11520.000000	11520.000000	11520.000000	11520.000000	11520.000000	11472.000000	11256.000000	11472.000	11256.000000
mean	33.598820	17.037100	-0.729409	7.806132	51.602942	0.008031	0.002627	-inf	0.005164
std	45.131442	2.265388	4.281310	2.793143	12.601594	0.975905	0.984134	NaN	0.980756
min	0.000000	8.696000	-12.520000	0.604000	11.573000	-3.568000	-2.980000	-inf	-2.461000
25%	5.101250	15.624750	-3.952000	5.488500	42.473500	-0.688000	-0.637000	-0.759	-0.733250
50%	14.466500	17.184500	0.023500	8.275000	52.351000	0.007500	0.017000	-0.004	-0.010000
75%	45.846750	18.607500	2.924000	9.939000	61.476500	0.661000	0.666250	0.754	0.732000
max	450.919000	23.780000	8.213000	15.486000	81.417000	3.993000	3.032000	2.741	2.473000

Tabla 13. Datos Faltantes en las variables índices de sequía

Variable		Total NaN	Meses/estación
SPI3	SPEI3	48	2
SPI12	SPEI12	264	11

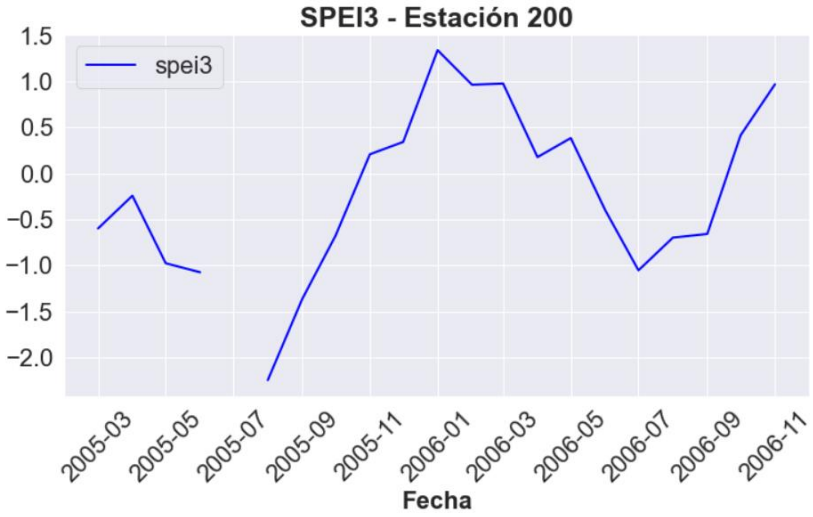


Figura 19. *Tratamiento de Datos no válidos (–inf) en estación 200*

4.4. Transformación

En ésta fase se preparan los datos que alimentan los modelos de Aprendizaje Automático. Se realiza tareas de extracción de características, reducción de dimensionalidad, preprocesamiento de variables, si se tiene variables categóricas estas se pueden llevar a dummies, etc.

Las 24 estaciones meteorológicas distribuidas en el altiplano central de Bolivia analizadas por separado darían lugar a 24 modelos diferentes, entonces para reducir la dimensión del problema se aplica una técnica de aprendizaje no supervisado denominada clusterización. El algoritmo elegido para realizar la clusterización es KMeans; pero antes se realiza una prueba

de determinación de cantidad óptima de clúster, con el método Elbow y el método Silhoute (Ver código en Anexo XVII). El número óptimo de clústeres obtenido por ambos métodos es 4, en la Figura 20 se observa ambos resultados. En el primer método (Elbow) el análisis es más de tipo visual y no está tan claro; pero el número óptimo está entre 4 y 5. Utilizando el método Silhoute es más sencillo visualizar el número óptimo de clústeres ya que es igual al mayor puntaje obtenido, que en el caso actual es 0.2547 para 4 clústeres, confirmando el resultado.

Luego de obtener el número de clústeres se utiliza el algoritmo KMeans para la clusterización dando como resultado los grupos que se visualizan en la Figura 21.

Los modelos que se definen y entrenan en las siguientes fases toman en cuenta las agrupaciones definidas, es decir no se realiza la predicción para cada estación, dando lugar a 24 predicciones para cada modelo, en cambio se toman como base los clústeres obtenidos y las predicciones se realiza para cada clúster.

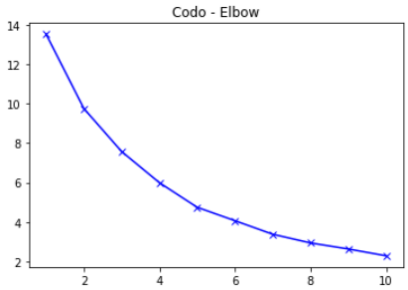
4.5. Selección e implementación del algoritmo

4.5.1. Selección del algoritmo

De acuerdo al problema planteado en la fase 1 la tarea a resolver es la predicción y dentro de la predicción se toma como “base de comparación” el modelo estadístico denominado ARIMA. Al ser la base de comparación el modelo ARIMA estará presente en las tablas comparativas al inicio de la lista de modelos.

Respecto a los algoritmos de Aprendizaje Automático se utilizan dos grupos principales:

1) Algoritmos basados en Árboles: Bosques Aleatorios
2) Algoritmos de Redes Neuronales:
    - Perceptrón Multicapa MLP
    - Red Recurrente LSTM
    - Red Convolucional TCN




```
Puntaje Silhouette para 2 clusters es  0.20816921004046277
Puntaje Silhouette para 3 clusters es  0.20852907939956203
Puntaje Silhouette para 4 clusters es  0.2547738908961459
Puntaje Silhouette para 5 clusters es  0.24641887975927745
Puntaje Silhouette para 6 clusters es  0.2410148594908131
Puntaje Silhouette para 7 clusters es  0.2390911505792083
```

Figura 20. Resultado de la clusterización

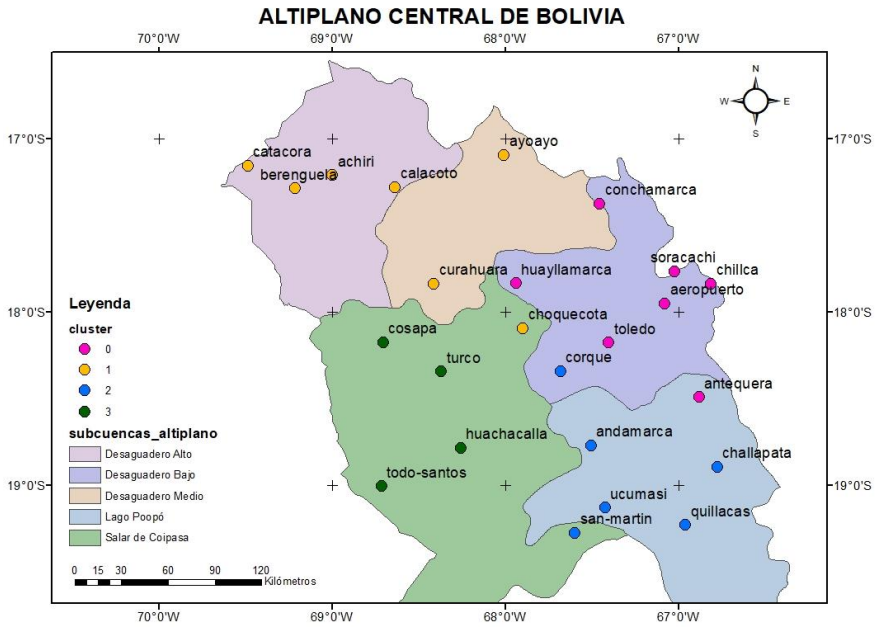


Figura 21. Resultado de la clusterización

Además, es importante establecer que se definen dos tipos de modelos predictivos, tomando como referencia la clasificación presentada en el apartado 2.2.3 Predicción de series temporales.

1) **Modelo de Serie Temporal Univariante**, ya que para la predicción solo se toman en cuenta los retardos o rezagos de la propia serie y no así otras variables externas. En el presente trabajo se llamará a éste modelo como **Univariante.**

2) **Modelo Mixto** o también llamado de regresión dinámica, ya que combina las características de los modelos de regresión o causales en los que participan variables exógenas con los modelos de Series temporales. En éste trabajo se llamará a este modelo **Univariante con variables exógenas.**

4.5.2. Determinación de Datos de entrenamiento y evaluación

Antes de aplicar los modelos de Aprendizaje automático se determina dos conjuntos de datos: entrenamiento y evaluación, para ello se seleccionó la siguiente relación: 12 meses para evaluación (testing) y 468 meses para entrenamiento(training).

Para el enfoque de series temporales los “features” y “target” no son directamente columnas seleccionadas del dataset. Para determinar ambos elementos se utiliza la estrategia conocida como “Ventana móvil” que consiste en seleccionar un grupo de N valores de la variable a predecir y se utiliza como X (features) y los siguientes M valores se constituyen en el Y Target a predecir, realizando un recorrido por todo el dataset de entrenamiento. Ver figura 22.

4.5.3. Definición de Hiperparámetros para cada modelo

Habiendo definido los parámetros para en entrenamiento y testeo de los modelos se implementaron los modelos de aprendizaje automático en notebooks de Jupyter. En ambos casos se utilizó el Lenguaje Python, en el entorno Colab, se importaron las librerías pandas, numpy , sklearn, tensorflow y keras para el segundo modelo de redes neuronales.
Se aplicaron 7 modelos de aprendizaje automático para predecir el SPI (3 y 12 meses) y el SPEI con las mismas escalas.

Los hiperparámetros que proporcionaron el modelo base ARIMA y para Bosques Aleatorios se pueden visualizar en la tabla 14

Los hiperparámetros que proporcionaron el mejor resultado para las redes neuronales Perceptrón Multicapa se visualizan en la Tabla 15 para SPI y la tabla 16 para SPEI

En el caso del Modelo LSTM Multicapa los hiperparámetros que proporcionaron el mejor resultado para se visualizan en la Tabla 17 para SPI y la tabla 18 para SPEI.

En el caso del Modelo LSTM Multicapa con una Convolución los hiperparámetros que proporcionaron el mejor resultado se visualizan en la Tabla 19 para SPI y la tabla 20 para SPEI

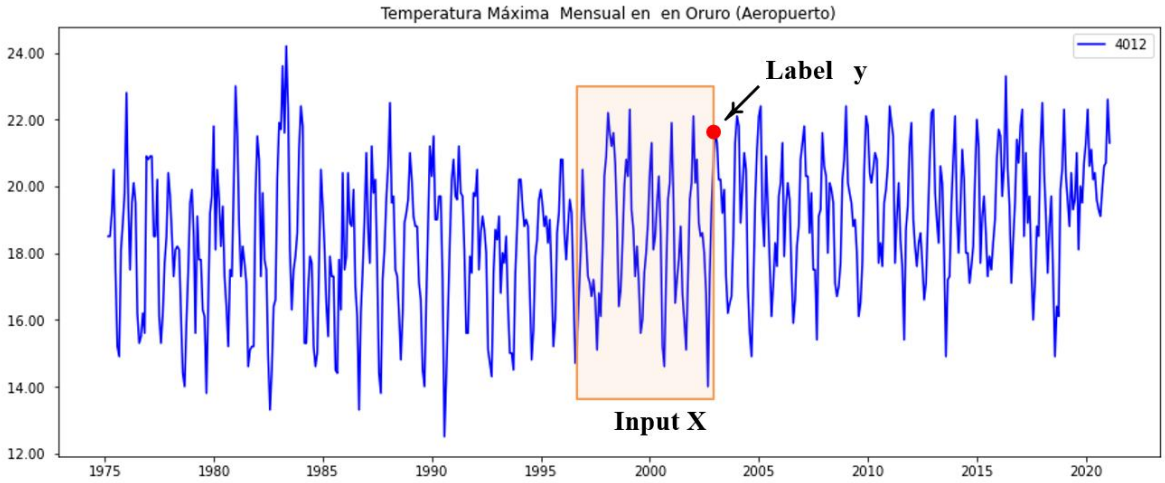


Figura 22. Estrategia conocida como “Ventana móvil

Tabla 14. Hiperparámetros del modelo ARIMA

MODELO	HIPERPARÁMETROS
ARIMA	estacionalidad = 12, ventana = 48 , estrategia = 'mean'
Bosques Aleatorios	ventana = 48, estrategia = 'recursive' estimadores=30, vecinos=11

Tabla 15. Hiperparámetros de Perceptrón multicapa para SPI

		Ventana	Epoca	Batch	Nodos Capa 1	Nodos Capa 2
Cluster 0	**spi3**	36	60	64	30	20
	spi12	72	100	32	30	24
Cluster 1	**spi3**	36	100	128	30	30
	spi12	72	100	64	40	40
Cluster 2	**spi3**	72	60	64	50	40
	spi12	72	100	64	40	40
Cluster 3	**spi3**	72	100	64	50	30
	spi12	48	60	64	40	30

Tabla 16. Hiperparámetros de Perceptrón multicapa para SPEI

		Ventana	Epoca	Batch	Nodos Capa 1	Nodos Capa 2
Cluster 0	**spei3**	36	80	128	30	20
	spei12	72	100	32	40	40
Cluster 1	**spei3**	36	100	128	30	20
	spei12	48	100	128	24	24
Cluster 2	**spei3**	36	80	128	30	20
	spei12	72	100	64	40	40
Cluster 3	**spei3**	36	100	128	30	30
	spei12	24	60	32	24	24

Tabla 17. Hiperparámetros de LSTM multicapa para SPI

		Ventana	Epoca	Batch	Capa 1 LSTM	Capa 2 LSTM	Capa 3 Densa
Cluster 0	**spi3**	48	140	128	20	12	12
	spi12	72	100	32	48	48	48
Cluster 1	**spi3**	48	100	128	30	30	30
	spi12	72	100	64	48	48	12
Cluster 2	**spi3**	72	60	64	30	24	12
	spi12	72	150	32	48	48	48
Cluster 3	**spi3**	48	60	128	20	12	12
	spi12	48	100	32	24	24	12

Tabla 18. Hiperparámetros de LSTM multicapa para SPEI

		Ventana	Epoca	Batch	Capa 1 LSTM	Capa 2 LSTM	Capa 3 Densa
Cluster 0	**spei3**	36	100	128	30	20	20
	spei12	72	100	32	48	32	32
Cluster 1	**spei3**	36	64	128	30	30	30
	spei12	72	100	32	32	12	12
Cluster 2	**spei3**	36	60	128	30	30	30
	spei12	72	100	32	48	48	32
Cluster 3	**spei3**	72	60	128	50	40	40
	spei12	72	100	32	32	32	12

Tabla 19. Hiperparámetros de LSTM multicapa con una capa convolucional para SPI

		Ventana	Epoca	Batch	Capa 1 LSTM-CNN	Capa 2 LSTM-CNN	Capa 3 Densa
Cluster 0	**spi3**	72	100	32	24	24	24
	spi12	72	100	32	48	48	32
Cluster 1	**spi3**	72	100	32	12	12	12
	spi12	72	100	64	48	48	24
Cluster 2	**spi3**	70	100	32	32	12	12
	spi12	72	150	32	48	48	12
Cluster 3	**spi3**	72	100	32	32	32	32
	spi12	48	100	32	24	24	24

Tabla 20. Hiperparámetros de LSTM multicapa con una capa convolucional para SPEI

		Ventana	Epoca	Batch	Capa 1 LSTM-CNN	Capa 2 LSTM-CNN	Capa 3 Densa
Cluster 0	**spei3**	36	100	128	30	30	20
	spei12	72	100	32	48	48	32
Cluster 1	**spei3**	36	64	128	50	40	30
	spei12	72	100	32	32	24	12
Cluster 2	**spei3**	36	60	128	30	10	10
	spei12	72	100	32	48	32	32
Cluster 3	**spei3**	72	60	128	40	40	40
	spei12	72	100	32	32	12	12

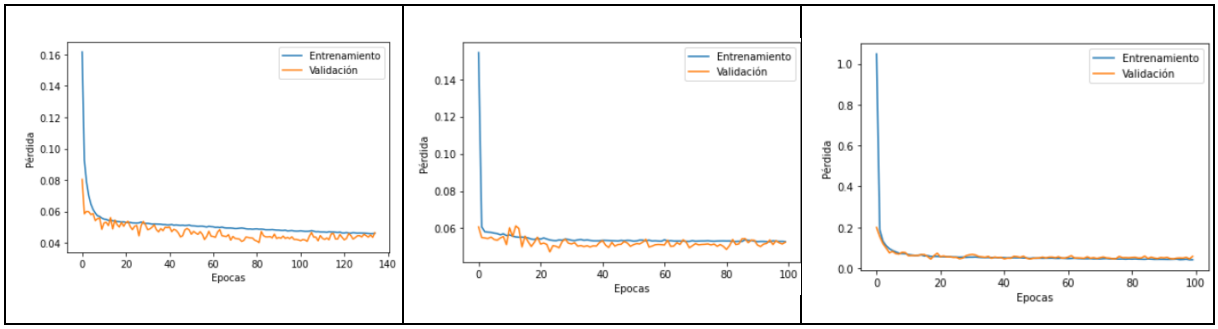


Figura 23. Rendimiento de algunos los modelos de Redes Neuronales

Las gráficas de rendimiento de los modelos se muestran en la figura 23, para los 3 modelos de redes neuronales.

Los modelos de Redes Neuronales no se ejecutaron como los anteriores pasos del trabajo en Jupyter Notebook. Se utilizó la herramienta Colab pro por sus mayores capacidades de procesamiento de los modelos por sus recursos GPU.

El código para ejecutar los modelos se encuentra la carpeta Github del proyecto:

https://github.com/MandbeZ/TFM_sequia).

Los notebooks corresponden a los diferentes modelos de series temporales de dos tipos: univariantes, tomando como única variable el SPI o el SPEI, en el cual se entrenan y evalúan los modelos ARIMA y Bosques Aleatorios. El segundo corresponde a los Modelos Univariantes con variables exógenas), en el cual se entrenan y evalúan los modelos ARIMA y Bosques Aleatorios, pero incluyendo en los análisis variables exógenas como la humedad y el índice ONI. El tercer tipo corresponde a los tres modelos de Redes Neuronales: Perceptrón Multicapa, Redes recurrentes LSTM y finalmente LSTM con Redes Convolucionales CCN

4.6. Evaluación e Interpretación

En total se entrenaron y validaron 112 modelos: 14 para SPI modelos aplicados a 4 clústeres y otros 14 modelos para SPEI también aplicados a 4 clústeres. Los resultados obtenidos por los modelos se resumen en las tablas 21 para SPI y Tabla 22 para SPEI. Es importante destacar que se entrenaron dos tipos de modelos: i) Los modelos enfocados a series temporales, es decir la entrada univariante: solo el índice SPI o SPEI y ii) Los modelos Mixtos, que tomaron en cuenta el tiempo, pero también variables exógenas.

Tabla 21. Comparativa de resultados de los modelos para el índice SPI

		SPI3				SPI12			
		MAE	***MAPE***	***RMSE***	***MSE***	***MAE***	***MAPE***	***RMSE***	***MSE***
CLUSTER 0	***ARIMA ST***	0.049674	0.131113	0.072302	0.005227	0.158549	0.335941	0.159311	0.025380
	ARIMA EX	0.150928	0.057130	0.085863	0.007372	0.546046	0.256813	0.259372	0.067274
	RF ST	0.059156	0.157283	0.083714	0.007008	0.178584	0.379760	0.180911	0.032728
	RF EX	0.143074	0.059449	0.068992	0.004760	0.148364	0.069639	0.071117	0.005058
	MLP	0.072834	0.200677	0.095919	0.009200	0.028766	0.134901	0.048962	0.002397
	LSTM	0.066884	0.199191	0.100986	0.010198	0.043316	0.203278	0.077166	0.005955
	LSTM-CNN	0.078298	0.228624	0.109708	0.012036	0.052941	0.221811	0.055057	0.005811
CLUSTER 1	***ARIMA ST***	0.073090	0.190572	0.090058	0.008110	0.039595	0.075164	0.045052	0.002030
	ARIMA EX	0.155129	0.060245	0.074369	0.005531	0.069017	0.035291	0.041440	0.001717
	RF ST	0.102867	0.233376	0.118625	0.014072	0.044433	0.089582	0.049045	0.002405
	RF EX	0.115508	0.053681	0.064762	0.004194	0.081268	0.043202	0.052372	0.002743
	MLP	0.070250	0.219831	0.087313	0.007624	0.032756	0.082634	0.043901	0.001927
	LSTM	0.074754	0.231924	0.093271	0.008699	0.045211	0.120362	0.065077	0.004235
	LSTM-CNN	0.087412	0.299214	0.112731	0.013901	0.048293	0.142142	0.080071	0.006010
CLUSTER 2	***ARIMA ST***	0.128498	0.443699	0.151341	0.022904	0.137050	0.608050	0.144178	0.020787
	ARIMA EX	0.483752	0.142215	0.164838	0.027172	0.838579	0.186655	0.197499	0.039006
	RF ST	0.118826	0.401608	0.144613	0.020912	0.157920	0.717925	0.170196	0.028966
	RF EX	0.440286	0.126926	0.158214	0.025032	0.176140	0.039582	0.042621	0.001817
	MLP	0.091488	0.268055	0.116606	0.013597	0.026276	0.083974	0.033806	0.001143
	LSTM	0.072760	0.212776	0.095706	0.009160	0.037370	0.125211	0.055669	0.003099
	LSTM-CNN	0.082623	0.356345	0.099989	0.010061	0.049041	0.232112	0.066751	0.005002
CLUSTER 3	***ARIMA ST***	0.061217	0.132420	0.075390	0.005684	0.085527	0.157060	0.086203	0.007431
	ARIMA EX	0.147478	0.068834	0.089195	0.007956	0.247207	0.134473	0.135420	0.018338
	RF ST	0.059321	0.123163	0.072591	0.005269	0.086607	0.159727	0.088680	0.007864
	RF EX	0.107199	0.056184	0.071695	0.005140	0.065912	0.035676	0.040797	0.001664
	MLP	0.067739	0.182523	0.083992	0.007055	0.031950	0.109587	0.049499	0.002450
	LSTM	0.093007	0.262877	0.115021	0.013230	0.043978	0.146451	0.061395	0.003769
	LSTM-CNN	0.103711	0.371134	0.120223	0.015445	0.050654	0.163284	0.055282	0.004618

Tabla 22. Comparativa de resultados de los modelos para el índice SPEI

		SPEI3				SPEI12			
		MAE	***MAPE***	***RMSE***	***MSE***	***MAE***	***MAPE***	***RMSE***	***MSE***
CLUSTER 0	***ARIMA ST***	0.07765	0.19664	0.09733	0.00947	0.17241	0.51428	0.17287	0.02989
	ARIMA EX	0.10920	0.27343	0.13586	0.01846	0.22410	0.67205	0.22740	0.05171
	RF ST	0.16832	0.06385	0.08972	0.00805	0.45612	0.21420	0.21698	0.04708
	RF EX	0.03484	0.08010	0.03923	0.00154	0.11511	0.34714	0.12059	0.01454
	MLP	0.09705	0.28960	0.12172	0.01482	0.03385	0.17798	0.06311	0.00398
	LSTM	0.10102	0.30718	0.12696	0.01612	0.05535	0.26638	0.09232	0.00852
	LSTM-CNN	0.11403	0.32973	0.12328	0.02173	0.07446	0.30879	0.07699	0.01027
CLUSTER 1	***ARIMA ST***	0.07537	0.19361	0.09369	0.00879	0.10600	0.29472	0.10896	0.01596
	ARIMA EX	0.05621	0.14121	0.07039	0.00496	0.03208	0.05701	0.03974	0.00158
	RF ST	0.21883	0.09678	0.11163	0.01246	0.06111	0.03010	0.03922	0.00154
	RF EX	0.06867	0.14579	0.07398	0.00547	0.14127	0.24615	0.14496	0.02101
	MLP	0.07355	0.19925	0.08900	0.00792	0.04051	0.13117	0.05577	0.00311
	LSTM	0.10096	0.30052	0.13191	0.01740	0.05750	0.18689	0.07857	0.00617
	LSTM-CNN	0.14149	0.36372	0.15318	0.02836	0.06234	0.19951	0.06999	0.00661
CLUSTER 2	***ARIMA ST***	0.12323	0.45369	0.15604	0.02435	0.16950	0.94802	0.17559	0.03083
	ARIMA EX	0.11533	0.41089	0.14154	0.02003	0.19726	1.11528	0.20373	0.04151
	RF ST	0.42378	0.12438	0.14225	0.02023	0.68697	0.15153	0.16222	0.02632
	RF EX	0.13719	0.50822	0.17764	0.03156	0.11025	0.63078	0.11956	0.01429
	MLP	0.09214	0.21383	0.12756	0.01627	0.03287	0.10660	0.04295	0.00184
	LSTM	0.08349	0.18841	0.12257	0.01502	0.04800	0.15452	0.06531	0.00426
	LSTM-CNN	0.19119	0.44025	0.20358	0.05229	0.05770	0.18424	0.06192	0.00571
CLUSTER 3	***ARIMA ST***	0.09222	0.29306	0.11572	0.01502	0.12751	0.55254	0.13090	0.01913
	ARIMA EX	0.13532	0.46321	0.16020	0.02566	0.12077	0.29390	0.12316	0.01517
	RF ST	0.12180	0.05885	0.07596	0.00577	0.17400	0.09442	0.09606	0.00923
	RF EX	0.08071	0.27580	0.09957	0.00991	0.05130	0.12528	0.05551	0.00308
	MLP	0.11975	0.37439	0.13890	0.01929	0.04329	0.17617	0.06325	0.00400
	LSTM	0.16589	0.50183	0.18696	0.03495	0.04764	0.18231	0.06445	0.00415
	LSTM-CNN	0.16007	0.45741	0.17399	0.03884	0.05745	0.19976	0.06165	0.00534

Para el primer tipo se desarrollaron 5 modelos: ARIMA ST (ARIMA Series Temporales), RF ST (Bosques Aleatorios Series Temporales), MLP (Perceptrón Multicapa), LSTM (Red Recurrente con tres capas ocultas), LSTM-CNN (Red Recurrente con tres capas ocultas y una capa Convolucional). Dos de los modelos fueron entrenados utilizando variables exógenas como la humedad relativa media, temperatura media y el índice oceánico del Niño, ARIMA EX (ARIMA con exógenas) y RF EX (Bosques Aleatorios con exógenas). Es importante destacar que el modelo ARIMA se toma como una base de comparación de los modelos estadísticos tradicionales con los modelos de Machine Learning.

Discusión y Análisis de Resultados

Resultados Generales

Tomando como base las dos tablas comparativas (Tabla 22 y Tabla 23) se puede comenzar el análisis comparando en rendimiento de los modelos con ambos índices SPI y SPEI, en términos generales comparando los mínimos errores generados por los modelos, las estimaciones de los índices SPI son mejores que las estimaciones del SPEI; pero la diferencia es muy pequeña. Tomando la media de los mínimos errores globales de todos los modelos y de todas las escalas, para SPI se tiene un MAE de 0.02627 y un y MSE de 0.0011 que corresponde a SPI12 para el modelo Perceptrón Multicapa, en segundo lugar, está el modelo recurrente LSTM con un MAE de 0.0373 y un MSE de 0.0031 que corresponde a SPI12.

Por otro lado, el SPEI con el menor error registra un MAE de 0.03208 y MSE de 0.0016 para un SPI12 para el modelo Bosques aleatorios con variables exógenas y en segundo lugar está ARIMA con variables exógenas con un MAE de 0.0328 y MSE de 0.0018, también para el índice SPEI12.

Resultados Para SPI

El mejor de los modelos para SPI3 es ARIMA Series Temporales con un MAE de 0.4967 y un MSE de 0.005, el segundo mejor es Bosques Aleatorios ST con un MAE de 0.5915 y MSE de 0.007 ambos para el clúster 0. En el caso de SPI12 el mejor de los modelos es Perceptrón Multicapa con un MAE de 0.02627 y MSE de 0.0011 para el clúster 2 y el segundo LSTM con un MAE de 0.0373 y un MSE de 0.0031 también para el clúster 2

Resultados Para SPEI

El mejor de los modelos para SPEI3 es Bosques Aleatorios con un MAE de 0.03484 y MSE de 0.0015 para el clúster 0 y el segundo mejor es ARIMA con variables exógenas con un MAE de 0.05621 y MSE de 0.0049 para el clúster 1.

El mejor de los modelos para SPEI12 es ARIMA EX con un MAE de 0.03208 y el segundo mejor es Perceptrón Multicapa con un MAE de 0.03287.

El modelo ARIMA y Bosques Aleatorios obtiene mejores resultados en escalas a corto plazo; pero a largo plazo los modelos de Redes Neuronales se desempeñan mucho mejor.

Realizando una comparativa respecto a los clústeres los modelos en términos generales pronosticaron mejor el clúster 0 en escala de 12 meses.

El mejor de los modelos para el Clúster 0 tomando en cuenta el SPI es Perceptrón Multicapa con un MAE de 0.02876 y el segundo mejor es LSTM con un MAE de 0.0433

El mejor de los modelos para el Clúster 0 tomando en cuenta el SPEI es Random Forest con un MAE de 0.03484 y el segundo mejor es LSTM con un MAE de 0.0553

Respecto al desempeño de los modelos de Aprendizaje Automático el modelo de Bosques aleatorios con variables exógenas obtuvo bueno resultados en comparación con las redes neuronales, confirmando los estudios mencionados en el estado del Arte, además es importante resaltar el papel que jugaron las variables exógenas en los modelos, algunas de tipo local como ser temperatura y humedad relativa y otras de tipo Global como el índice Oceánico del Niño.

En lo que se refiere a los Modelos de Redes Neuronales en el estado del arte se destacó dos principales: Perceptrón Multicapa y Redes recurrentes LSTM; sin embargo, respecto al resultado que se obtuvo del entrenamiento y evaluación de ambos modelos, el modelo de Perceptrón Multicapa tuvo un desempeño muy bueno, llegando a obtener el mejor resultado en algunas escalas y clústeres; sin embargo, los resultados de error de LSTM tampoco estuvieron muy distanciados de los modelos con mejor desempeño.

Es importante destacar que a pesar de que los modelos de Redes Neuronales, especialmente LSTM tuvieron importantes avances en los últimos años, para la obtención de mejores resultados se requieren una mayor cantidad de datos, en el caso del estudio de la sequía se podría tener un periodo mayor de análisis, que en el presente proyecto fue de 40 años o también se podría utilizar escalas temporales menores a un mes, como ser el caso de péntadas que se refiere a 5 días.

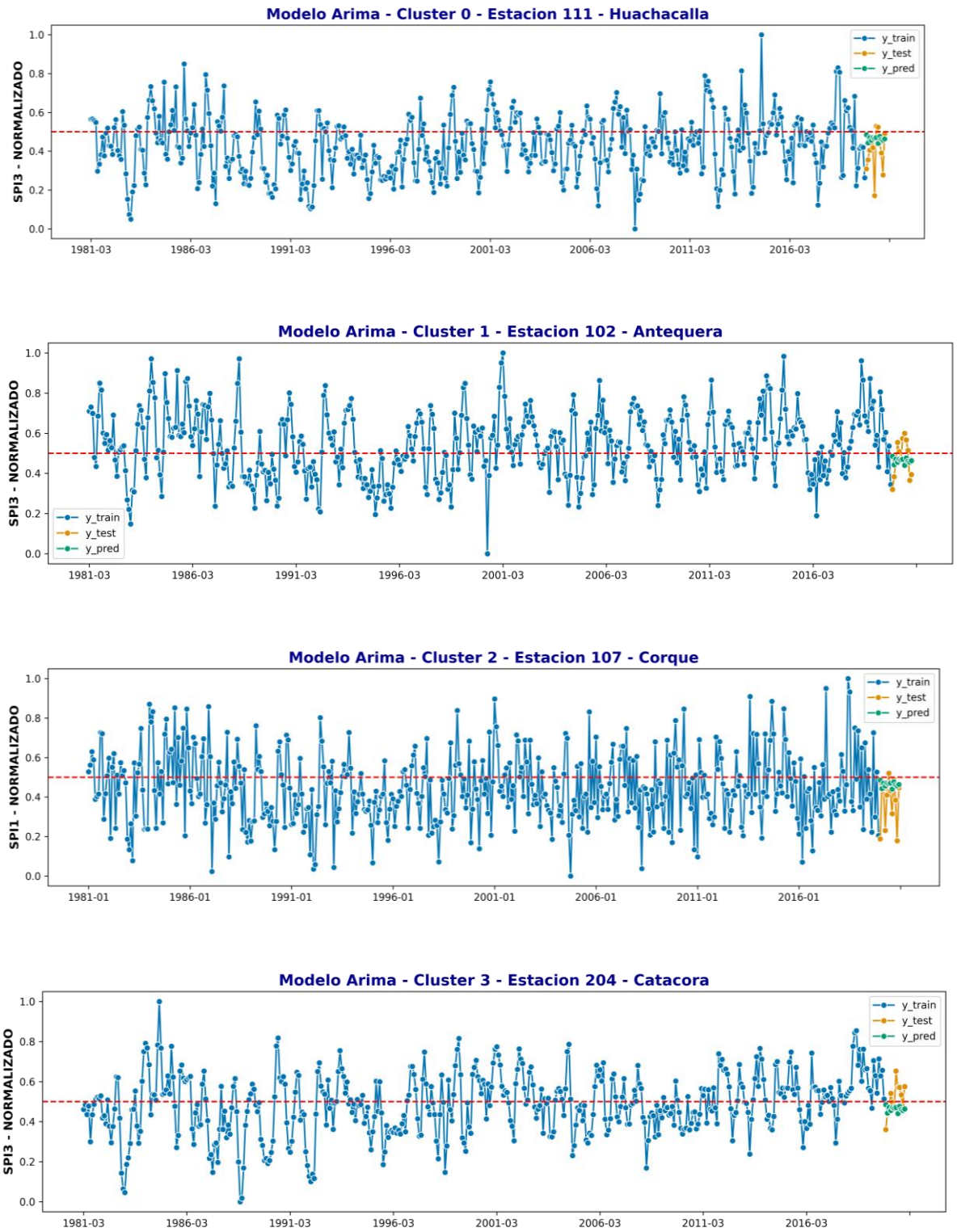


Figura 24. Predicciones ARIMA para los 4 clústeres

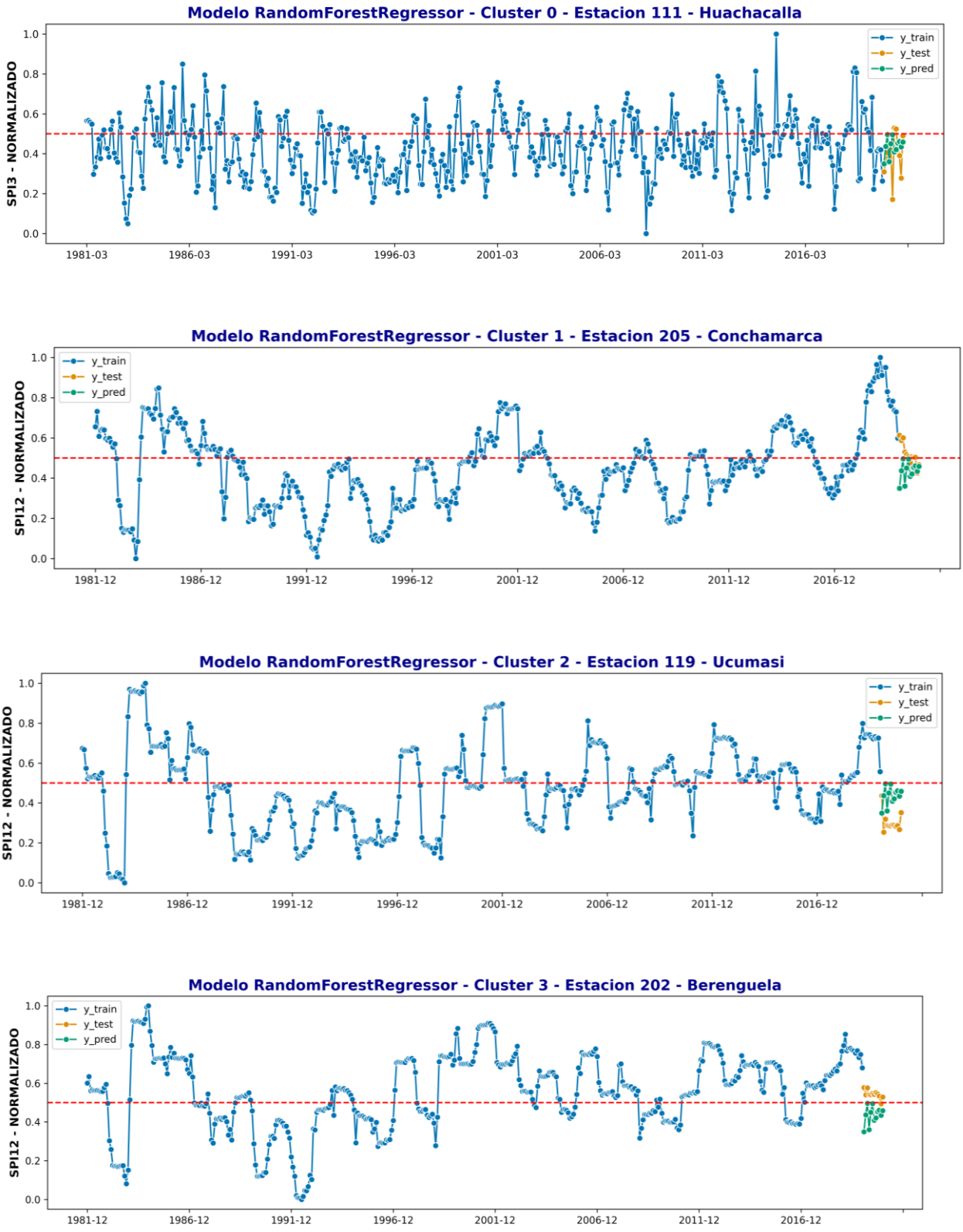


Figura 25. Predicciones Bosques Aleatorios para los 4 clústeres

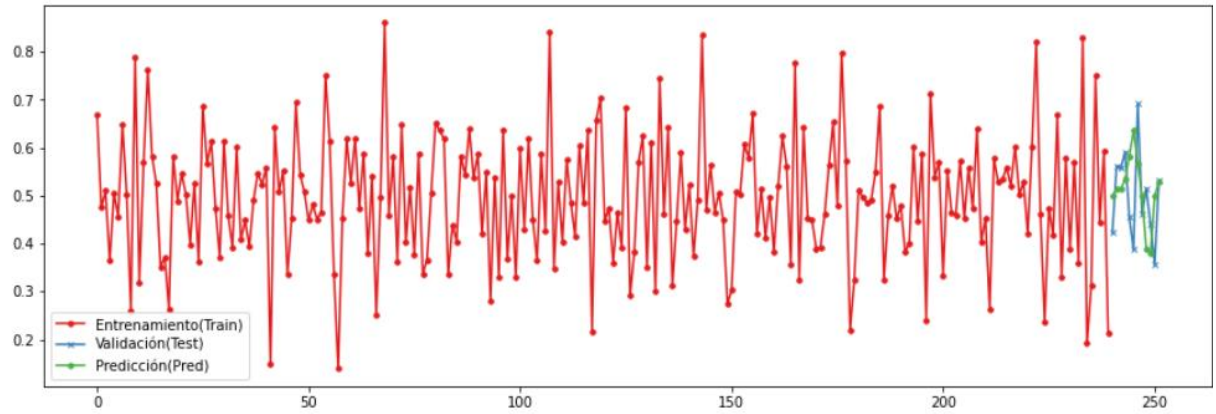


Figura 26. Predicciones con MLP

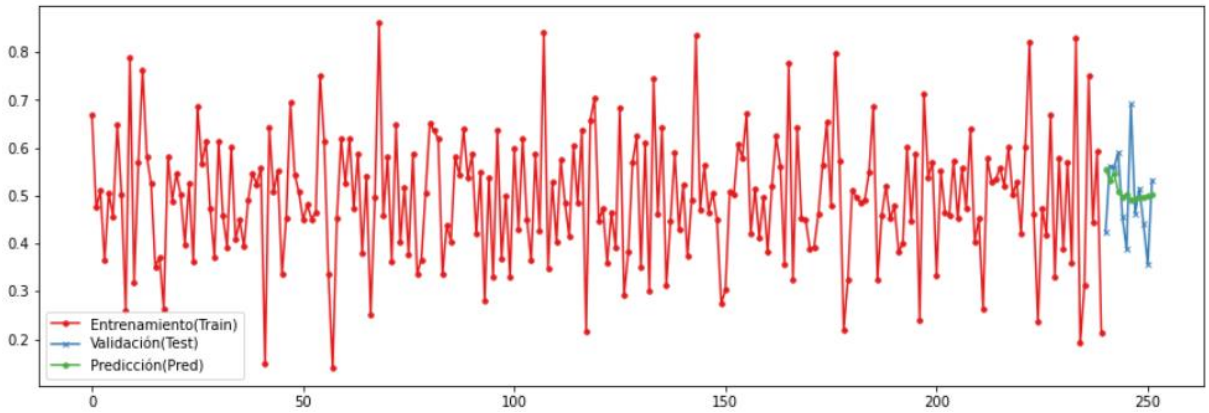


Figura 27. Predicción con LSTM de varias capas ocultas

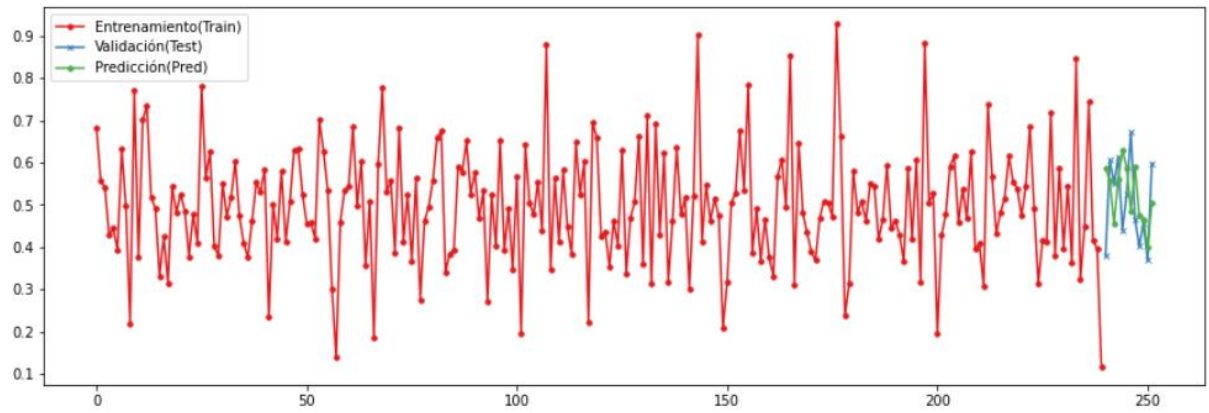


Figura 28. Predicción con LSTM-CCN

4.7. Visualización e Integración

Se presentan las visualizaciones de las predicciones realizadas por los modelos en escalas de 3 y 12 meses y para los índices SPI y SPEI. Resulta un número muy grande de gráficas generadas por los modelos; sin embargo, se presentarán algunas a modo de ejemplo en la Figuras 25,26,27 y 28, cada una relativa a uno de los Modelos utilizados, orientados a series de tiempo.

Una vez que ya se tienen los resultados de las series es necesario realizar un análisis de los eventos de la sequía, para lo cual se utilizan las métricas descritas en el apartado 2.1.4 Caracterización de eventos secos, sobre todo para determinar los tipos de sequía y también la duración, para ello también se plantea una herramienta desarrollada en Python, incluida en los cuadernos del proyecto, que permite visualizar las características de un evento seco. En la Figura 29 se pueden apreciar las sequías más importantes registradas a nivel mundial, relacionadas con el índice ONI (Canedo-Rosso et al., 2021, p. 999) entre los años 1980 y 2020.

- Sequía Extrema, en la que destacan las registradas entre los años1982 y1983, la segunda 1997 y 1998 y la tercera entre 2015 y 2016. Además, se pueden visualizar dos eventos de sequía extrema a nivel local 1992 y 2004
- Sequía Fuerte, en la que destacan las registradas entre los años1987 y1988, la segunda entre 1991 y 1992
- Sequía Moderada: La primera entre 1986 y 1987, la segunda entre 1994 y 1995, la tercera 2002–2003 y la cuarta entre 2009 y 2010

Todas las gráficas resultantes se encuentran en el enlace Github del proyecto:

https://github.com/MandbeZ/TFM_sequia

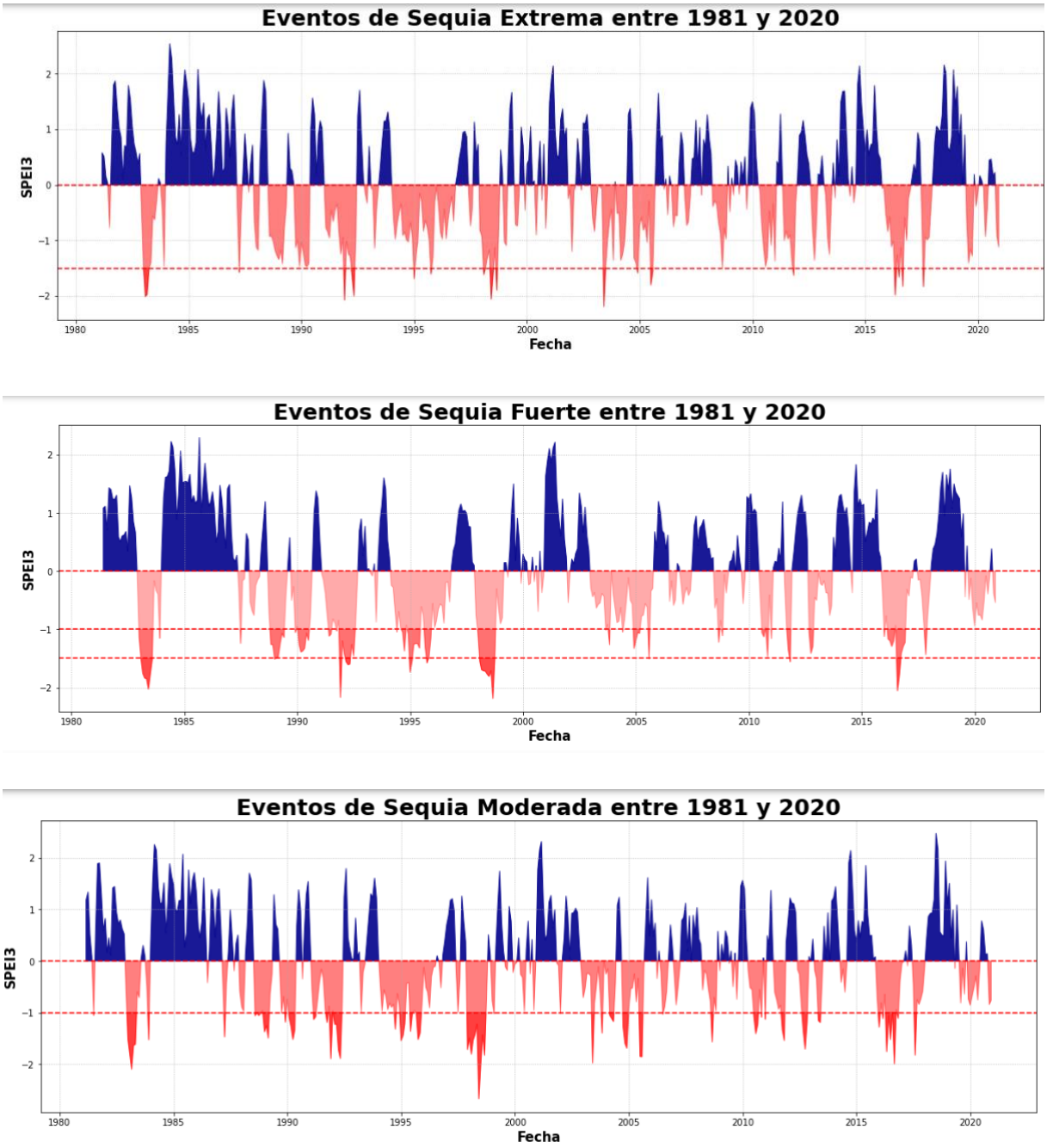


Figura 29. Identificación de Eventos secos

5. Conclusiones y trabajo futuro

En éste apartado se presentan las conclusiones del trabajo realizado en relación con los objetivos planteados. A continuación, se realiza un análisis de futuros trabajos que se podrían originar a raíz del presente proyecto.

5.1. Conclusiones

La sequía, por ser un fenómeno muy complejo generan muchas pérdidas en sistemas agrícolas, económicos y sociales. Los daños causados por éste fenómeno natural son muy grandes, ya que su ocurrencia no depende de la zona geográfica ni la época del año. En consecuencia, es importante tener mecanismos que permitan identificar las sequías y los más utilizados en el mundo para dicha tarea son los índices de sequía.

En virtud de lo mencionado el trabajo se planteó un objetivo principal y varios objetivos específicos, los cuales se alcanzaron satisfactoriamente:

Respecto al objetivo general se realizó un análisis del fenómeno de sequía utilizando los índices estandarizados SPI y SPEI. La predicción de sequía se pudo obtener mediante el entrenamiento de 6 diferentes algoritmos de aprendizaje automático, y el Modelo ARIMA que pertenece a los Modelos estadísticos se utilizó como base de comparación. Además, mediante una comparativa se determinaron algunas ventajas y desventajas de los algoritmos con relación a la temática de sequía y los datos que se utilizaron para el entrenamiento y además se evaluó su eficiencia con 4 métricas: MAE, MAPE, MSE y RMSE.

Además de realizar la comparativa entre los modelos se planteó una metodología a manera de contribución que permitió analizar y predecir sequías, ya que sus fases estaban directamente enfocadas a la temática de sequía.

Se definieron varias fuentes de información meteorológica para recolectar los datos de precipitación, temperatura y humedad relativa para la región del altiplano central de Bolivia. En primera instancia se recurrió a la fuente primaria: SENAMHI que brindó datos de 24 estaciones meteorológicas las cuales se tomaron como base para el desarrollo del trabajo; sin embargo, se determinó una gran cantidad de datos faltantes por lo que se recurrió a dos fuentes adicionales de datos satelitales como ser la NASA con datos de temperatura y humedad y CHIRPS con datos de precipitación.

Se realizó un análisis descriptivo y exploratorio de los datos en dos oportunidades: la primera con los datos observados provenientes de SENAMHI y posteriormente con los datos

satelitales. Es importante resaltar que se realizaron varias pruebas de correlación, para justificar el uso de los datos satelitales en lugar de los datos observados. Se utilizaron las pruebas de correlación de Pearson y Spearman, además de la corrección Bias que produjo muy buenos resultados en la corrección de los datos satelitales.

Se realizó el cálculo de dos índices de sequía: SPI y SPEI mediante el paquete en R SPEI. Se resalta el uso del Kernel de R en Jupyter Notebook que permite una flexibilidad para aquellos usuarios de R que están migrando a Python, ya que evita que se tenga que reescribir todo código en el otro lenguaje. Las escalas elegidas: 3 y 12 para el cálculo del SPI y SPEI resultaron ser las adecuadas para mostrar una mayor amplitud en el análisis del comportamiento de la sequía, tanto para corto como para largo plazo. Ambos dos índices permiten realizar mediciones y comparaciones efectivas entre las ocurrencias de sequía en una zona y otra porque ambos son estandarizados.

Se implementó un modelo de clusterización debido a la gran cantidad de las estaciones meteorológicas estudiadas. Los grupos resultantes de dicha clusterización resultaron ser muy próximos a la segmentación de tipo cuenca que generalmente se utiliza en hidrología.

Se seleccionaron los modelos de aprendizaje automático más adecuados a series temporales con las consideraciones necesarias para la obtención de mejores resultados. Los modelos fueron entrenados posteriormente para la predicción de sequía. El resultado de esta fase fue una gran cantidad de modelos generados, debido a que se tenían varias características comparables, en primer lugar, dos índices de sequía: el SPI basado solo en precipitación y el SPEI que además utiliza temperatura. Ambos índices se calcularon en 2 escalas temporales: 3 meses para un corto plazo y 12 meses para un análisis de largo plazo. Luego realizó una comparativa de los modelos utilizando las métricas: MAE, MAPE, RMSE y MSE, dando como resultado a los modelos más destacados: Bosques aleatorios con variables exógenas, Perceptrón Multicapa., Red Neuronal Recurrente LSTM. Los Modelos ARIMA, a pesar del avance de aprendizaje automático todavía se desempeñan muy bien en comparación a los demás modelos cuando se sigue el enfoque de series de tiempo, sobre todo en las escalas temporales de 3 meses. Cuando se trabaja con variables exógenas claramente el modelo de Bosques Aleatorios obtuvo mejores resultados.

Finalmente, una vez realizada una caracterización de un evento seco mediante sus principales métricas como ser su duración o su severidad para dotar a los tomadores de decisiones de información útil en la planificación de acciones futuras.

5.2. Líneas de trabajo futuro

Los resultados obtenidos en el presente trabajo serán de mucha utilidad a los organismos encargado de tomar decisiones relacionadas con prevención y gestión de desastres naturales, sobre todo en la zona estudiada: el altiplano central de Bolivia; pero también en otras zonas del país y del mundo, ya que solo se requieren datos de precipitación y temperatura básicamente y si se desea otras variables exógenas para realizar la estimación de los índices y el posterior modelado.

Respecto a las escalas de tiempo de las variables meteorológicas analizadas Para futuros proyectos se podría utilizar una escala menor al mes para el estudio de las variables meteorológicas, por ejemplo, péntadas que son periodos de 5 días, eso permitiría que cada año en lugar de tener 12 meses de análisis se tuvieran 72 péntadas.

Con la comparativa se tiene una mejor noción sobre la eficiencia de los modelos en diferentes escenarios como ser la escala temporal, según los resultados expuestos se tiene una idea de qué modelos se desempeñan mejor a corto, mediano o largo plazo.

Con la utilización de los cuadernos de Jupyter, que se encuentran en un repositorio abierto de Github, cualquier investigador que esté interesado en la temática de sequía podrá explorar las herramientas generadas el para análisis y predicción de sequía, además podrá aportar con ideas para su mejoramiento.

.

6. Bibliografía

Adhyani, N. L., June, T., & Sopaheluwakan, A. (2017). Exposure to Drought: Duration, Severity and Intensity (Java, Bali and Nusa Tenggara). *IOP Conference Series: Earth and Environmental Science*, *58*, 012040. https://doi.org/10.1088/1755-1315/58/1/012040

AGETIC. (2017). *PLAN DE IMPLEMENTACIÓN DE GOBIERNO ELECTRÓNICO 2017 – 2025*. COPLUTIC. https://coplutic.gob.bo/IMG/pdf/plan_gobierno_electronico_.pdf

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning. A Textbook*. Springer.

Ali, Z., Hussain, I., Faisal, M., Nazir, H. M., Hussain, T., Shad, M. Y., Mohamd Shoukry, A., & Hussain Gani, S. (2017). Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model. *Advances in Meteorology*, *2017*, e5681308. https://doi.org/10.1155/2017/5681308

Andrade, M. F., Moreno, I., Calle, J. M., Ticona, L., Blacutt, L., Lavado-Casimiro, W., Sabino, E., Huerta, A., Aybar, C., Hunziker, S., & Brönnimann, S. (2018). *Climate and extreme events from the Central Altiplano of Peru and Bolivia 1981-2010Atlas—Clima y eventos extremos del Altiplano Central perú-boliviano* [Application/pdf]. https://doi.org/10.7892/BORIS.108978

Beguería, S., & Vicente-Serrano, S. M. (2017). *SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index* (1.7) [Computer software]. https://CRAN.R-project.org/package=SPEI

Campero Marin, S. A. (2013). Atlas Climatológico De Bolivia. En *Programa Piloto de Resiliencia Climática—PPCR Bolivia Fase I* (p. 243). SENAMHI. http://senamhi.gob.bo/agromet/investigaciones//AtlasClimatologicosBolivia_final.pdf

Canedo-Rosso, C., Hochrainer-Stigler, S., Pflug, G., Condori, B., & Berndtsson, R. (2021). Drought impact in the Bolivian Altiplano agriculture associated with the El Niño–Southern Oscillation using satellite imagery data. *Natural Hazards and Earth System Sciences*, *21*(3), 995-1010. https://doi.org/10.5194/nhess-21-995-2021

Car, N. J., Ip, A., & Druken, K. (2017). netCDF-LD SKOS: Demonstrating Linked Data Vocabulary Use Within netCDF-Compliant Files. *Environmental Software Systems. Computer Science for Environmental Protection*, 329-337. https://doi.org/10.1007/978-3-319-89935-0_27

CHC. (2021). *Base de Datos CHIRPS 2.0—The Climate Hazards Center*. https://data.chc.ucsb.edu/products/CHIRPS-2.0/global_daily/netcdf/p05/

Dikshit, A., Pradhan, B., & Alamri, A. M. (2020). Short-Term Spatio-Temporal Drought Forecasting Using Random Forests Model at New South Wales, Australia. *Applied Sciences*, *10*(12). http://dx.doi.org/10.3390/app10124254

Eslamian, S., Ostad-Ali-Askari, K., Singh, V., Dalezios, N., Woldeyohannes, D.-Y., & Matouq, M. (2017). A Review of Drought Indices. *International Journal of Constructive Research in Civil Engineering*, *3*, 48-66. https://doi.org/10.20431/2454-8693.0304005

Funk, C. C., Peterson, P. J., Landsfeld, M. F., Pedreros, D. H., & Verdin, J. P. (2014). *A quasi-global precipitation time series for drought monitoring* (USGS Numbered Series N.º 832; Data Series, p. 12). U.S. Geological Survey. http://pubs.er.usgs.gov/publication/ds832

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

Guijarro, J. A. (2019a). *Homogeneización de series climáticas con Climatol*. https://www.climatol.eu/homog_climatol-es.pdf

Guijarro, J. A. (2019b). *climatol: Climate Tools (Series Homogenization and Derived Products)* (3.1.2) [Computer software]. https://CRAN.R-project.org/package=climatol

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd ed)* (3rd edition). OTexts. https://Otexts.com/fpp3/

IRkernel. (2021). *IRkernel*. GitHub. https://github.com/IRkernel

Jalalkamali, A., Moradi, M., & Moradi, N. (2015). Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized

Precipitation Index. *International Journal of Environmental Science and Technology*, *12*(4), 1201-1210. https://doi.org/10.1007/s13762-014-0717-6

Khan, M. M. H., Muhammad, N. S., & El-Shafie, A. (2018). A Review of Fundamental Drought Concepts, Impacts and Analyses of Indices in Asian Continent. *Journal of Urban and Environmental Engineering*, *12*(1), 106-119. http://dx.doi.org/10.4090/juee.2017.v12n1.106-119

Lara-Benítez, P., Carranza-García, M., Luna-Romera, J. M., & Riquelme, J. C. (2020). Temporal Convolutional Networks Applied to Energy-Related Time Series Forecasting. *Applied Sciences*, *10*(7), 2322. https://doi.org/10.3390/app10072322

Li, X., He, B., Quan, X., Liao, Z., & Bai, X. (2015). Use of the Standardized Precipitation Evapotranspiration Index (SPEI) to Characterize the Drying Trend in Southwest China from 1982–2012. *Remote Sensing*, *7*(8), 10917-10937. https://doi.org/10.3390/rs70810917

Maca, P., & Pech, P. (2016). Forecasting SPEI and SPI Drought Indices Using the Integrated Artificial Neural Networks. *Computational Intelligence and Neuroscience*, *2016*, 17. https://doi.org/10.1155/2016/3868519

Maillard, O., Vides-Almonacid, R., Flores-Valencia, M., Coronado, R., Vogt, P., Vicente-Serrano, S. M., Azurduy, H., Anívarro, R., & Cuellar, R. L. (2020). Relationship of Forest Cover Fragmentation and Drought with the Occurrence of Forest Fires in the Department of Santa Cruz, Bolivia. *Forests*, *11*(9), 910. https://doi.org/10.3390/f11090910

Maimon, O., & Rokach, L. (2010). Introduction to Knowledge Discovery in Databases. *Data Mining and Knowledge Discovery Handbook*, 1-17. https://doi.org/10.1007/0-387-25465-X_1

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). THE RELATIONSHIP OF DROUGHT FREQUENCY AND DURATION TO TIME SCALES. En *Proceedings of the 8th Conference on Applied Climatology* (Vol. 17, pp. 179-183).

Mesbahzadeh, T., Mirakbari, M., Mohseni Saravi, M., Soleimani Sardoo, F., & Miglietta, M. M. (2020). Meteorological drought analysis using copula theory and drought indicators under climate change scenarios (RCP). *Meteorological Applications*, *27*(1), e1856. https://doi.org/10.1002/met.1856

Mokhtar, A., Jalali, M., He, H., Al-Ansari, N., Elbeltagi, A., Alsafadi, K., Abdo, H. G., Sammen, S. Sh., Gyasi-Agyei, Y., & Rodrigo-Comino, J. (2021). Estimation of SPEI Meteorological Drought Using Machine Learning Algorithms. *IEEE Access*, *9*, 65503-65523. https://doi.org/10.1109/ACCESS.2021.3074305

Montgomery, D. C., Jennning, C. L., & Kulahci, M. (2015). Introduction to Forecasting. En *Introduction to Time Series Analysis and Forecasting* (2nd edition, pp. 1-24). Wiley.

NASA. (2021). *POWER | Data Access Viewer*. https://power.larc.nasa.gov/data-access-viewer/

Nielsen, A. (2019). *Practical Time Series Analysis: Prediction with Statistics and Machine Learning* (First Edition). O'Reilly Media, Inc.

NOAA. (2021). *Historical El Nino / La Nina episodes (1950-present)*. https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php

OMM. (2021). *Atlas de la OMM(Organización Meteorológica Mundial) sobre mortalidad y pérdidas económicas debidas a fenómenos meteorológicos, climáticos e hidrológicos extremos (1970–2019)* (Vol. 1267). OMM. https://library.wmo.int/doc_num.php?explnum_id=10927

Peñuela, A., Hutton, C., & Pianosi, F. (2021). An open-source package with interactive Jupyter Notebooks to enhance the accessibility of reservoir operations simulation and optimisation. *Environmental Modelling & Software*, *145*, 105188. https://doi.org/10.1016/j.envsoft.2021.105188

Poornima, S., & Pushpalatha, M. (2019). Drought prediction based on SPI and SPEI with varying timescales using LSTM recurrent neural network. *Soft Computing*, *23*(18), 8399-8412. https://doi.org/10.1007/s00500-019-04120-1

Potop, V., Boroneanţ, C., Možný, M., Štěpánek, P., & Skalák, P. (2014). Observed spatiotemporal characteristics of drought on various time scales over the Czech Republic. *Theoretical and Applied Climatology*, *115*(3-4), 563-581. https://doi.org/10.1007/s00704-013-0908-y

Satgé, F., Hussain, Y., Xavier, A., Zolá, R. P., Salles, L., Timouk, F., Seyler, F., Garnier, J., Frappart, F., & Bonnet, M.-P. (2019). Unraveling the impacts of droughts and agricultural intensification on the Altiplano water resources. *Agricultural and Forest Meteorology*, *279*, 107710. https://doi.org/10.1016/j.agrformet.2019.107710

SENAMHI. (2021). *SENAMHI - SISMET*. http://senamhi.gob.bo/index.php/sismet

Sktime. (2021). [Python]. The Alan Turing Institute. https://github.com/alan-turing-institute/sktime

Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456-457*, 12-29. https://doi.org/10.1016/j.jhydrol.2012.05.052

Vicente-Serrano, S. M., Beguería, S., & López-Moreno, J. I. (2010). A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. *Journal of Climate*, *23*(7), 1696-1718. https://doi.org/10.1175/2009JCLI2909.1

Anexos

Anexo I. Datos iniciales proporcionados por SENAMHI

Los datos fueron proporcionados en 3 formatos distintos. En el formato 1 (Figura A1) se tienen los años y meses en filas y los días en columnas. Cada variable en una hoja del libro Excel.

Estación:		San Martin							Latitud Sud:		19º 16' 30"	
Departamento:		Oruro							Longitud Oeste:		67º 35' 57"	
Provincia:		Ladislao Cabrera							Altitud m/s/n/m:		3712	
DATOS DE : PRECIPITACIÓN DIARIA (mm)												
AÑO	**MES / DIA**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1975	11											
1975	12	3,5	7,4	7,0	6,2	11,5	21,7	12,5	0,0	0,0	0,0	9,2
1976	1	0,0	0,2	0,8	0,0	0,3	0,0	0,0	0,0	0,0	0,0	0,0
1976	2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	18,0	2,5	0,0	0,0
1976	3	0,0	0,0	0,0	4,0	0,0	15,0	9,0	0,0	0,0	0,0	0,0
1976	4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
1976	5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Figura A1: Estación San Martín en Formato 1

En el formato 2 (Figura A2) se tienen los años en el encabezado, meses en columnas y días en filas. Cada variable en una hoja Excel.

Estación:	Achiri								Latitud Sud:	17º 12' 42"			
Departamento:	La Paz								Longitud Oeste:	68º 59' 58"			
Provincia:	Pacajes								Altitud m/s/n/m:	3880			
DATOS DE : PRECIPITACIÓN DIARIA (mm)													
AÑO: 1975													
DIA	**ENE**	**FEB**	**MAR**	**ABR**	**MAY**	**JUN**	**JUL**	**AGO**	**SEP**	**OCT**	**NOV**	**DIC**	**TOTAL**
1	****	****	****	****	0,0	2,0	0,0	0,0	0,0	9,8	0,0	0,0	11,8
2	****	****	****	****	0,0	7,1	0,0	0,0	0,0	2,0	0,0	0,0	9,1
3	****	****	****	****	0,0	3,1	0,0	0,0	0,0	0,0	0,0	0,0	3,1
4	****	****	****	****	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5	****	****	****	****	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
6	****	****	****	****	0,0	0,0	0,0	0,0	1,1	0,0	0,0	2,9	4,0
7	****	****	****	****	0,0	0,0	0,0	0,0	1,8	0,3	0,0	1,7	3,8
8	****	****	****	****	0,0	0,0	0,0	0,0	0,0	0,5	0,0	0,0	0,5

Figura A2: Estación Achiri en Formato 2

En el formato 3 (Figura A3) se tiene la fecha completa en filas y las variables en columnas.

DATOS METEOROLOGICOS

Estación:	*SORACACHI*			Departamento:	ORURO								
Departamento:	*Oruro*												
	PRECIP. (mm)	TEMP. (º C)			HUM. REL. (%)			PRES. BAR. (mbar)			RAD. SOLAR (W/m2)		
Fecha		MAX. MEDIA	MIN. MEDIA	MED	MED	MAX	MIN	MED	MAX	MIN	MED	MAX	MIN
1/1/2015	0,0	17,4	2,8	10,8	47,4	75,4	21,3	646,0	647,6	644,2	293,0	1243,0	0,0
2/1/2015	4,6	16,5	6,0	9,8	69,7	91,6	43,3	646,5	648,4	644,2	211,5	1238,0	0,0
3/1/2015	0,0	16,8	6,8	10,5	66,3	90,7	35,2	647,0	648,5	645,4	229,6	1075,0	0,0
4/1/2015	9,0	16,8	5,0	10,1	67,8	85,5	34,4	647,9	649,2	646,1	225,9	1314,0	0,0
5/1/2015	0,2	14,2	5,5	8,7	79,1	95,2	49,6	648,1	649,5	646,4	248,3	1365,0	0,0
6/1/2015	13,8	14,6	4,1	8,0	79,5	93,7	50,3	647,7	649,5	645,6	150,2	1205,0	0,0
7/1/2015	13,4	16,4	4,3	9,7	72,2	93,5	44,2	646,7	648,4	644,6	276,0	1231,0	0,0

Figura A3: Estación Soracachi en Formato 3

Anexo II. Código Python para uniformar datos

Los modelos en Jupyter Notebook y todos los datos utilizados en el proyecto están en el siguiente enlace: https://github.com/MandbeZ/TFM_sequia

Convertir archivos excel de SENAMHI a archivos CSV

```
'''instalar openpyxl para abrir varias hojas de un libro de excel'''
%pip install openpyxl
%pip install xlrd
```

```
import pandas as pd
import numpy as np
import datetime
from calendar import monthrange
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
```

```
t0 = datetime.datetime.now()
```

Convertir de formato obtenido a fomato CSV

Leer los datos del archivo CSV 'estaciones_senamhi'. Las columnas que nos interesan:

- path_format contiene las rutas relativas de los archivos,
- path_csv contiene las rutas relativas para guardar los arvhivos resultantes
- f contiene el tipo de formato que tiene cada archivo <1 , 2 , 3>

Formato 1, 2, 3 a formato CSV

Definimos los metodos para las diferentes operaciones y cálculos Para un dataframe de la forma

anio = 2003

	DIA	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	TOTAL
	1	1	2	3	4		2	3	3	5	1			x
	2	0	0	0	0	0	9	8	5	3				y

.....

```
'''mostrar dataframe de precipitacion'''
print(dfpp[['100','101','109','111','117','120','124','125']].head(5))
```

```
    100  101  109  111   117  120   124   125
0   0.0  0.0  0.0  1.8   7.5  0.0   0.0   0.0
1   0.0  1.5  2.0  0.0   0.0  0.0   5.8  19.0
2   0.0  0.0  5.0  1.3   0.0  2.3   1.4   0.0
3   2.8  0.0  2.0  0.0  14.5  1.2   4.0  18.0
4  23.9  0.8  2.0  1.2   6.5  0.0  10.5   0.0
```

```
'''mostrar dataframe de temperatura máxima'''
print(dftmx[['100','116','117','120','201','203']].head(5))
```

```
    100   116   117   120   201   203
0  19.8  19.5  22.8  23.0  17.2  20.2
1  20.0  21.0  29.4  20.0  17.9  21.8
2  19.8  23.0  23.9  19.5  17.8  16.4
3  18.6  16.4  22.8  18.5  14.8  18.0
4  16.8  17.0  26.1  20.0  12.5  17.0
```

```
'''mostrar dataframe de temperatura mínima'''
print(dftmn[['100','116','117','120','201','203']].head(5))
```

```
   100  116  117  120  201  203
0  6.8  5.6 -6.1 -2.0  3.0  3.6
1  7.9  7.5 -5.6  1.0  2.9  3.0
2  5.1  6.5 -8.3  2.0  5.3  4.9
3  5.3  8.1 -2.8  3.0  4.1  6.2
4  7.1  5.2 -1.7  3.5  4.5  5.3
```

```
'''datos de las estaciones'''
print(dflista[['id','estacion','depto','altitud','lon','lat']].head(5))
```

```
    id    estacion  depto  altitud        lon        lat
0  100  aeropuerto  oruro     3702 -67.079722 -17.952778
1  101   andamarca  oruro     3762 -67.506389 -18.771944
2  102   antequera  oruro     4057 -66.882900 -18.492900
3  104     chillca  oruro     4025 -66.813889 -17.836944
4  105      cosapa  oruro     3906 -68.706389 -18.177778
```

Anexo III. Código Python para procesar NOAA

Uniformar_satelitalesNOA

```
import pandas as pd
import numpy as np
```

```
'''Cargar datos de NOA '''
oni=pd.read_csv('../datos/1.4.formato6_noaa/noaa_oni.csv', sep = ',', parse_dates=True)
oni.head()
```

	Year	12	1	2	3	4	5	6	7	8	9	10	11
0	1981	-0.3	-0.5	-0.5	-0.4	-0.3	-0.3	-0.3	-0.2	-0.2	-0.1	-0.2	-0.1
1	1982	0.0	0.1	0.2	0.5	0.7	0.7	0.8	1.1	1.6	2.0	2.2	2.2
2	1983	2.2	1.9	1.5	1.3	1.1	0.7	0.3	-0.1	-0.5	-0.8	-1.0	-0.9
3	1984	-0.6	-0.4	-0.3	-0.4	-0.5	-0.4	-0.3	-0.2	-0.2	-0.6	-0.9	-1.1
4	1985	-1.0	-0.8	-0.8	-0.8	-0.8	-0.6	-0.5	-0.5	-0.4	-0.3	-0.3	-0.4

```
'''Aplicar Melt '''
oni1=pd.melt(oni,id_vars=["Year"],value_vars=["12","1","2","3","4","5","6","7","8","9","10","11"],value_name="oni").sort_valu
oni1['fecha']=pd.to_datetime(oni1['Year'].astype(str)+'-'+oni1['variable'].astype(str))
oni1.head()
```

	Year	variable	oni	fecha
0	1981	12	-0.3	1981-12-01
1	1981	1	-0.5	1981-01-01
2	1981	11	-0.1	1981-11-01
3	1981	2	-0.5	1981-02-01
4	1981	3	-0.4	1981-03-01

```
oni1.drop(['Year','variable'],axis=1,inplace=True)
oni1=oni1.sort_values(by="fecha").reset_index(drop=True)
oni1= oni1[['fecha','oni']]
oni1.to_csv('../datos/1.0.variables/noaa_mensual_oni.csv', index = False)
```


```
oni1.drop(['Year','variable'],axis=1,inplace=True)
oni1=oni1.sort_values(by="fecha").reset_index(drop=True)
oni1= oni1[['fecha','oni']]
oni1.to_csv('../datos/1.0.variables/noaa_mensual_oni.csv', index = False)
```


```
print(oni1.head(10))
```

```
       fecha  oni
0 1981-01-01 -0.5
1 1981-02-01 -0.5
2 1981-03-01 -0.4
3 1981-04-01 -0.3
4 1981-05-01 -0.3
5 1981-06-01 -0.3
6 1981-07-01 -0.2
7 1981-08-01 -0.2
8 1981-09-01 -0.1
9 1981-10-01 -0.2
```

Anexo IV. Código Python para determinar datos faltantes

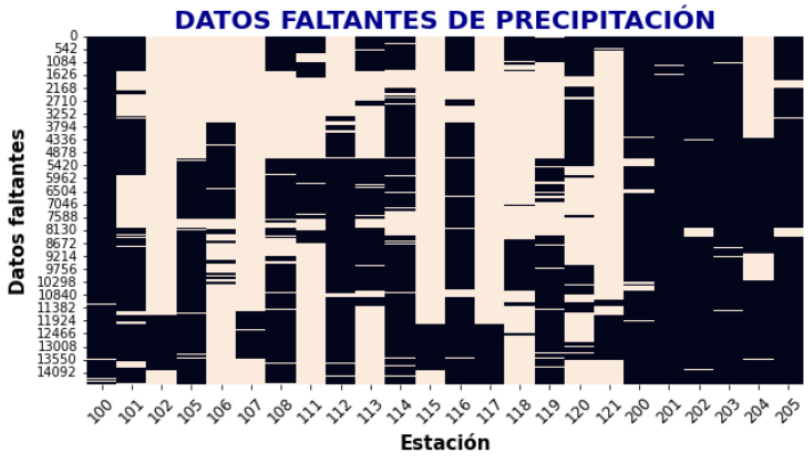
DETERMINAR CANTIDAD DE DATOS FALTANTES ¶

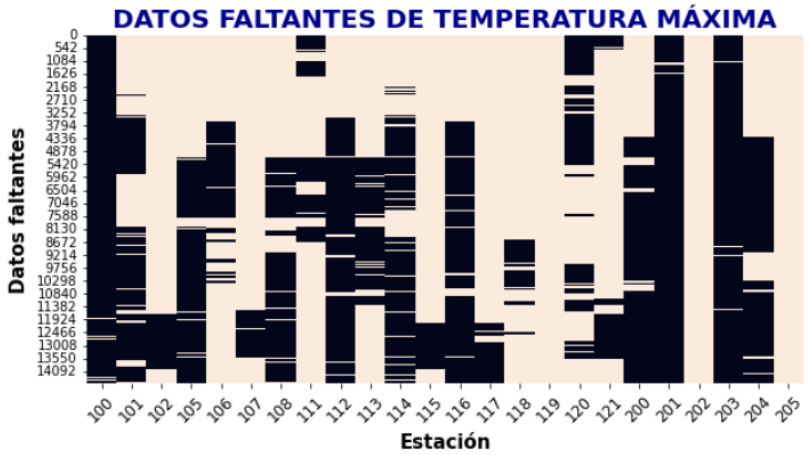
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

```
'''lista de archivos a leer'''
lis_arch = ['obs_diario_pp.csv','obs_diario_tmax.csv','obs_diario_tmin.csv','obs_diario_tmed.csv',
            'obs_diario_hmed.csv']
nom = ['precipitacion','temp_max','temp_min','temp_med','hum_med']
'''Se crea un data frame vacío para almacenar los resúmenes de todas las variables '''
temp = pd.read_csv('../datos/1.0.variables/obs_diario_pp.csv', sep=',')
```

```
'''Datos faltantes respecto a una variable'''
def revisar_null(data):
    data = data.drop(columns='fecha', axis=0)
```





Anexo V. Código Python para procesar CHIRPS

Uniformar los datos satelitales de CHIRPS - diarios por estación

```
from netCDF4 import Dataset
import numpy as np
import pandas as pd
from datetime import datetime, timedelta
import time
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

```
'''Definir periodo de análisis: fecha de inicio y fin.
   Comprendido entre [1981-01-01, 2021-11-30]'''
fecha_inicio = '1981-01-01'
fecha_fin = '2020-12-31'

'''devuelve los datos de una variable'''
def variable_dato(data, var):
    v = data.variables[var][:]
    return v

def creardiario(cad):
    anio = cad[2].split('.')
    md_i = '-01-01'
    if anio[0] == '2021':
        md_f = '-11-30'
    else:
        md_f = '-12-31'
    d1 = anio[0]+md_i
    d2 = anio[0]+md_f
```

.....

```
pp_diario.columns = nom_col
pp_diario = pp_diario.reset_index()
pp_diario.rename(columns={'index':'fecha'},inplace=True)
pp_diario['fecha'] = pd.to_datetime(pp_diario['fecha'])
pp_diario = pp_diario[(pp_diario['fecha'] >= fecha_inicio) & (pp_diario['fecha'] <= fecha_fin)]
guardar(pp_diario, 'chirps_diario_pp.csv')

for i in range(dflista.shape[0]):
    nombre = '../datos/1.0.formato_csv_corregido/chirps_'+dflista.iloc[i,0].astype(str)+'_'+dflista.iloc[i,4]+'.csv'
    estacion = pd.read_csv(nombre, header=None, usecols = [i for i in range(2)])
    estacion.columns = ['fecha', 'pp']
    estacion['fecha'] = pd.to_datetime(estacion['fecha'])
    estacion = estacion[(estacion['fecha'] >= fecha_inicio) & (estacion['fecha'] <= fecha_fin)]
    estacion.to_csv(nombre, index=False)
```

```
'''Datos del dataframe generado'''
print(estacion.head(5))
```

```
       fecha     pp
0 1981-01-01  0.000
1 1981-01-02  3.296
2 1981-01-03  5.145
3 1981-01-04  0.000
4 1981-01-05  9.486
```

Anexo VI. Código Python para procesar NASA

Uniformar los datos satelitales de NASA - diarios por estación

```
import pandas as pd
import numpy as np
import os
```

Funciones

```
'''Definir perido de los datos'''
fecha_ini = '1981-01-01'
fecha_fin = '2020-12-31'

'''devolver la fecha en base al año y el numero de dia'''
def aniodia(anio, dia=1):
    fecha = pd.to_datetime(dia-1, unit='D', origin=str(anio))
    return fecha.date()

'''devolver el rango de fecha en base al año, dia inicial y periodo'''
def rango_fecha1(anio, dia, periodo):
    rango = pd.date_range(start = aniodia(anio, dia), freq = 'D', periods = periodo)
    return rango

'''leer los archivos en un carpeta dada SI son archivos CSV y si su nombre comienza con NASA'''
def leer_dir(path):
    with os.scandir(path) as ficheros:
        ficheros = [fichero.name for fichero in ficheros if fichero.is_file() \
                    and fichero.name.endswith('.csv') and fichero.name.startswith('nasa')]
```

Generar archivos con datos uniformados

```
dir_f4 = '../datos/1.3.formato4_nasa/'
archivos = leer_dir(dir_f4)
```

```
for p in archivos:
    ruta = dir_f4 + p
    data = pd.read_csv(ruta, skiprows=13)

    '''leer anio y dia de inicio'''
    anio = data.iloc[0,0]
    dia = data.iloc[0,1]
    dias = data.shape[0]

    '''adicionar la columna fecha'''
    data['fecha'] = rango_fecha1(anio, dia, dias)

    '''cambiar de nombre las columnas'''
    data = data.rename(columns={"fecha": "fecha", "PRECTOTCORR": "pp", "T2M_MAX":\
                                "tmax", "T2M_MIN": "tmin", "T2M": "tmed", "RH2M": "hmed"})

    '''reordenar las columnas y eliminar columnas YEAR y DOY'''
    data = data.reindex(['fecha','pp','tmax','tmin','tmed','hmed'], axis=1)

    '''periodo de datos'''
    data = data[(data['fecha'] >= fecha_ini) & (data['fecha'] <= fecha_fin)]

    '''guardar archivo'''
    dir_guardar = '../datos/formato_csv_corregido/' + p
    data.to_csv(dir_guardar, index = False)
```

Anexo VII. Código en R para homogenización

HOMOGENIZAR DATOS CON CLIMATOL - R

```
#Instalar Climatol -  solo instalar una vez
#install.packages("climatol")
```

```
#Cargar climatol
library(climatol)
```

```
#Configurar ruta del archivo
setwd("../datos/2.1.estaciones_climatol")
getwd()
```

'C:/Users/Angélica Andrade/Dropbox/documentosA/TFM_UNIR_2022/datos/2.1.estaciones_climatol'

```
anio_min<-1981 #año mínimo para el análisis
anio_max<-2020 #año máximo para el análisis
```

PARA PRECIPITACION

```
#1.Convertir los datos a formato climatol
daily2climatol(stfile = "stations.txt", stcol = 1:6, datcol = c(1:3,4), varcli="pre",anyi = anio_min,anyf = anio_max, mindat
```

```
14610 days between 1981-01-01 and 2020-12-31

Generating pre input files for Climatol from daily files...:
```

.....

```
#5.Ajuste mensual  std = 2 (El valor 2 se utiliza para variables como precipitaci?n)
#dz.min y dz.max Histogram of normalized anomalies
#Histogram of maximum windowed SNHT para snht1  ,  Histogram of maximum global SNHT  snht2
homogen("pre-m",anio_min, anio_max, dz.min = -9, dz.max = 10, snht1 = 30, snht2 = 44, std =2, cutlev = 2.3, vmin = 0 )
```

```
#6.Ajuste diario (metad = TRUE toma los break mensuales)
homogen("pre",anio_min, anio_max,  dz.min = -16, dz.max = 26, snht1 = 220, snht2 = 250, std =2, vmin = 0, metad = TRUE)
```

```
#7.Generar las series homogenizadas diarias
dahstat('pre',anio_min,anio_max, stat = 'series')
```

```
Homogenized values written to pre_1981-2020_series.csv,
with flags in pre_1981-2020_flags.csv:
  0: Observed data
  1: Missing data (filled in)
  2: Corrected data
```

```
#8.Generar las series homogenizadas mensuales
dahstat('pre-m',anio_min,anio_max, stat = 'series')
```

```
Homogenized values written to pre-m_1981-2020_series.csv,
with flags in pre-m_1981-2020_flags.csv:
  0: Observed data
  1: Missing data (filled in)
  2: Corrected data
```

PARA TEMPERATURA MAXIMA

```
#1.Convertir los datos a formato climatol
daily2climatol(stfile = "stations.txt", stcol = 1:6, datcol = c(1:3,5), varcli="tmax",anyi = anio_min ,anyf = anio_max, minda
```

```
#2.Aplicar homogen en modo exploratorio diario
homogen("tmax",anio_min, anio_max, expl = TRUE)
```

Anexo VIII. Resultados de Precipitación homogenizada

Precipitación Diaria:

Date	100	101	102	105	106	107	108	111	112	113	114	115	116	117	118	119	120	121	200	201	202	203	204	205
1/1/1981	0	0	0	0.9	11.3	4.3	7.1	0	8.7	0	0	0.9	1.8	4	19	0	0	7.5	0	0.8	3.9	5	5.8	4.1
2/1/1981	0	1.5	4.4	0.3	0.6	0.4	0	0	0.3	19	2	0.2	0	0.4	0	0	0	0	26.6	1.1	9.5	8.5	15.5	0
3/1/1981	0	0	3.5	5.1	0.9	0.5	0.7	2.3	7.7	0	5	4.5	1.3	1.9	0	0	12	0	0.1	17.4	3	5.2	4.6	11.6
4/1/1981	2.8	0	8.4	3.8	6.9	1.7	4.1	1.2	9.3	18	2	3	0	1.5	4	15	13.3	14.5	2	16.2	6.1	9.1	9.3	0
5/1/1981	23.9	0.8	5.4	10.7	3.8	5.6	2.3	0	4.3	0	2	9.4	1.2	8.4	3	0	0	6.5	0	0.9	1.5	2.1	2.2	6.1
6/1/1981	11.4	0	3.2	4.9	4.1	3.4	2.4	0	2.9	0	15	4.1	0	4	5	20	2.5	5.5	2.5	0	2.4	2.7	3.6	0
7/1/1981	0	0.4	6.8	4	7	2.6	4.4	10.8	3.9	16	9	3.4	2.8	1.3	0	28	11	8.5	10.5	0	4.9	5	7.8	7.2
8/1/1981	3.5	0.4	5.1	4.4	4.2	3.2	2.3	10.4	0.8	0	9	3.5	0.5	1.1	0	11	14.5	0	14.5	3	5.6	5.4	9.1	0
9/1/1981	5.1	0	2.7	3.1	10.2	6.4	6.2	12.8	4.8	0	1	2.5	1.8	3.3	11	13	5.5	0	0	5.5	2.5	3.5	3.7	0
10/1/1981	4.6	0	1	1.8	0.5	0.9	0.3	0	1.7	0	0	1.5	0	1.3	0	10	0	1.5	0	4.8	1	1.7	1.6	0
11/1/1981	15.6	0	3.4	7.4	0.8	3.1	0.4	0	2.2	0	0	6.5	0	5.4	0	0	0	2.4	3.1	0.4	1.5	1.6	2.4	6.2
12/1/1981	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	10.9	1.9	3.2	2.8	0
13/1/1981	0	0	0	1.2	0	0	0	0	1.9	0	0	1.2	0	0.9	0	0	0	0	0	2.1	0.4	0.6	0.5	5.4
14/1/1981	10.5	0	2.3	4	1.4	2.7	0.9	0	2.7	0	0	3.5	0	3.4	3	0	0	0	1.2	6.4	1.9	2.8	3	0
15/1/1981	7	0	1.5	5	8.9	2.6	5.1	5.4	7.9	0	0	4.7	0	3.7	0	0	0	21.5	1	0	3.5	4.7	5.1	10.4
16/1/1981	9.1	0	2	3.5	0	1.8	0	0	4.5	0	0	3	0	2.5	0	0	0	0	0.8	16	3	5	4.6	0
17/1/1981	3.7	3	1.5	6.4	1.2	1.6	0	0	5	0	0	6.2	0	5.2	0	0	0	0	0	0.3	0.1	0.1	0.1	20
18/1/1981	2.3	3	1.2	1.5	1.2	1.3	0	0	0	0	0	1.2	0	1.4	0	0	0	0	0	0	0	0	0	0
19/1/1981	0	0	1.9	4.5	4.7	2.3	2.8	2.3	6	0	4	4.2	0	3.8	8	0	6.5	0	0	0.6	1.2	1.5	1.8	14.2
20/1/1981	2	2	1.5	1.5	3.8	1	2.4	0	5	0	11	1.3	3.4	1.1	0	2	1.9	9.5	3	9.4	4	5.7	6.2	0
21/1/1981	1.6	1.4	0.7	2.5	10.4	6.3	5.5	20.5	5.6	0	0	2.4	0	2.8	4	0	0	0	11.8	9.2	6.2	7.1	10	7.4
22/1/1981	0	1.2	0.3	0.2	1.4	0.8	3.9	0	0.6	0	0	0.2	17.3	0.7	2	0	0	0	1.8	0	0.9	0.9	1.5	0
23/1/1981	0	1.1	0.3	0.2	3.9	2.3	4.1	8.8	0	0	0	0.2	11.4	0.3	0	0	0	0	3.1	0	1.1	1	1.8	0
24/1/1981	0	1	1	0.7	10.3	5.9	9.4	20.5	1.5	0	0	0.6	20.5	1	4	0	2.6	0	2.3	1.2	1.6	1.7	2.5	0
25/1/1981	2.2	2.2	1	1.3	18.4	5.4	10.5	18.8	8.1	0	10	1.1	2.6	1.2	0	0	0	32.5	1.5	0	5.3	7.1	7.7	0
26/1/1981	0.8	0	0.6	2.4	12.2	5.4	10.7	15.8	9.8	0	8	2.3	19.5	2.8	7	12	1.5	8.8	8.3	12.5	7.3	9.3	11.4	8.1
27/1/1981	5.3	0	4.9	2.4	10.8	5.8	9	20.9	3.5	15	0	2.1	15.5	1.5	0	9	2	8.5	5.8	5.1	4.1	5	6.4	0
28/1/1981	5.6	2.1	1.7	4.1	8.9	6.5	5.1	20.6	1.8	0	0	3.8	3.8	3.3	0.1	10	0	0	1.4	0.1	0.5	0.5	0.9	7.2

Precipitación Mensual:

Date	100	101	102	105	106	107	108	111	112	113	114	115	116	117	118	119	120	121	200	201	202	203	204	205
1/1/1981	124.3	26.2	88.2	94.4	164.3	67.5	103.8	184.7	117.9	68	78	79.9	105.9	82.6	101.1	130	73.3	129.6	101.3	123.9	69.7	106.7	135.8	113.2
1/2/1981	139	68.3	119	125.3	395.5	122.5	136.9	315	236.5	74	93	105.9	180.2	133.7	257	87.2	105.1	441.2	132.6	112.8	119.3	183.6	228.3	153.9
1/3/1981	59.7	1.4	63.1	52.4	123.8	19.3	44.4	66.9	108.3	85	27	44.4	7.7	32	0	60	57.8	318.2	69.9	55.9	66.6	104.9	126.6	76.2
1/4/1981	3.6	0	7.7	19.1	23.7	9.4	0	58	21.2	0	14	17.4	28.8	12.1	0	0	19.1	7	25.7	16.1	0	40.5	21.8	61.7
1/5/1981	0	0	0	4.5	0	0	0	0	7.1	0	0	4.4	0	3.6	0	6	0	0	0	7.6	0	0	0	20.2
1/6/1981	0	0	0	0.9	0	0	0	0	1	0	0	0.8	0	0.7	0	0	0	0	0	0	0	0	0	3.9
1/7/1981	0	0	0	0.5	0	0	0	0	0.5	0	0	0.5	0	0.4	0	0	0	0	0	0	0	0	0	2.1
1/8/1981	19	1.2	22.5	27.2	7.8	5.6	18.7	2.3	25.2	26	9	24	1.5	20	15	22	25.4	0	25.9	20.4	4.1	18.5	17.6	64.2
1/9/1981	78.5	3	39.5	50.4	48.5	16.7	3.2	0	46.8	0	9	41.7	10.2	36.7	32	24	51.3	108.4	14.9	14.1	28.9	29.5	47.4	39.1
1/10/1981	21.3	0	6.6	21.3	17.1	7	0.1	21	35.7	0	0	19.4	0	17.2	5	2	2.9	23.2	12.5	54.8	19.7	4	22.2	56
1/11/1981	11	9	15.8	25.2	43	13.5	31.5	21	41.4	0	4.6	22	4.5	24	45	0	28.4	37.9	12.2	24.9	76.4	27	69.1	58.2
1/12/1981	77	125.5	88.2	90.1	166.3	78.7	91.7	118.2	78.7	14	59	75.5	72.5	100.5	122	56	63.3	32.9	119.1	69.7	58.9	79.5	116.7	85
1/1/1982	122.7	146.9	103.2	130.2	213.2	97.1	100.3	140.5	143.9	7.6	112	111.5	86.9	139.8	142	61	57.9	97	195.9	157.8	337	132.4	352.6	169.2
1/2/1982	50	7	31.5	42.2	285.3	18.3	27.7	15	226.3	7.9	41	35.8	28.4	35.8	49	33.2	38.9	827.6	16.7	25.1	133.9	69.4	223.1	59.2
1/3/1982	130.9	0	61.8	98.6	158	42.5	67	96	140.9	0	38	84.1	42.3	74.4	61	76	78.9	312.2	80.6	65.2	64.9	86.2	138.5	141.7
1/4/1982	6	0	2.3	9.7	20.8	7.9	8.3	3.4	28.6	0	17	9	3.5	14.8	46	19	2.2	0	40.7	33.4	6.4	6.2	21.3	31.2
1/5/1982	0	0	0.5	0.8	0	0	0	0	0.6	0	0	0.7	0	0.4	0	0	1.3	0.3	0	0	0.4	1	0.6	2.3
1/6/1982	0	0	0	1.6	0	0	0	0	2.6	0	0	1.6	0	1.3	0	0	0	0.7	0	2.1	0	2	0.6	7.2
1/7/1982	0	0	0	1.8	0	0	0	0	2.4	0	0	1.7	0	1.4	0	0	0	0	0	0.9	0	0	0	7.9
1/8/1982	0	0	0.4	1.3	0	0	0	0	2.7	0.2	0	1.3	0	0.9	0	0	1	0	0	5.3	0	0	0	5.1
1/9/1982	45.4	13.3	36.5	32.7	20.7	13.6	5.6	1.4	26.5	32	0	27.2	8.5	26.9	24	19	29	27.1	33.3	35	21.6	21.2	34.5	27.4
1/10/1982	5	17.2	16.7	15.7	42.8	13.2	23	22.8	37.7	25	25	14.1	23.6	19.6	32	91	5.3	42.7	57.7	35.5	34.1	14.7	54.3	39.5
1/11/1982	48.1	1.3	35.6	38.7	32.4	17.7	3	27.9	38.3	18	8	32.2	11.5	29.6	43	0	48.2	38.5	19	44.7	41.3	48.6	49.7	43.7
1/12/1982	22.4	13	36	27	16.7	6.7	20.9	0.9	20.9	37	9	22.3	8.5	16.7	0	30.2	40.5	31	10.6	19	10.3	50.7	29.9	31.3
1/1/1983	53.7	4.9	33.9	34.3	31.4	13.2	2.9	6.6	24.6	24	14	28.5	0	25.6	18	7	30.5	23.6	12	36.7	23.3	36.6	31.4	27.5
1/2/1983	57.2	8	40.3	48.9	58.5	28.2	4.4	27.7	51.4	24	9	41.6	16	43.9	70	41	43.3	61.1	44.9	37.1	50.6	65.5	70.9	71.6
1/3/1983	12.7	14	19.8	17.2	19.4	11.5	4.5	12.2	11.9	13	0	14.4	9.4	14.2	12.3	23	22.2	19.7	45	5.3	7	5.2	24.2	20.4

Anexo IX. Correlación Observados y Homogenizados

Correlación entre Datos observados y datos corregidos por climatol

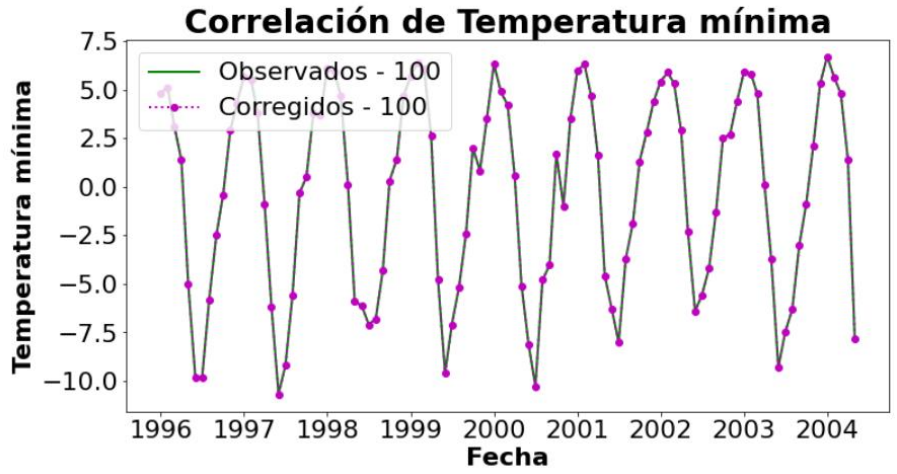
```
import pandas as pd
import numpy as np
from sklearn.metrics import mean_absolute_error
import matplotlib.pyplot as plt
```

```
'''definir lista de estaciones'''
lista_estaciones = ['100','101','102','105','106','107','108','111','112','113','114','115',
                    '116','117','118','119','120','121','200','201','202','203','204','205'] #Completa
#lista_estaciones = ['100','101','102','106','111','116','120','200','204']

lista_var = ['pp','tmax','tmin']

'''definir cantidad de datos a mostrar d1=Fecha Inicio , d2=Fecha Fin'''
d1=180
d2=280

'''funcion para leer datos y anexar informacion en la salida'''
def leer_datos(direccion):
    data = pd.read_csv(direccion)

    if (direccion.count('series')): data=data.rename(columns={'Date':'fecha'})
    data['fecha']=pd.to_datetime(data['fecha'])

    if (direccion.count('climatol')): origen='Corregidos'
    elif (direccion.count('obs')): origen='Observados'
    elif (direccion.count('nasa')): origen='NASA'
    elif (direccion.count('bc')): origen='chirps bc'
    else: origen='Chirps'
```

...

```
'''Carga de Datos Mensuales  de todas las variables'''
obs = leer_datos('../datos/1.0.variables/obs_mensual_tmin.csv')
corr = leer_datos('../datos/2.1.estaciones_climatol/tmin-m_1981-2020_series.csv')
```

```
graficar_datos(obs, corr, lista_estaciones)
```




```
corr_columnas(obs, corr, lista_estaciones).T
```

	spearman	pearson
100	0.99856	0.998204
101	0.998654	0.999271
102	0.996561	0.992704
106	0.97156	0.963617
111	0.999926	0.999982
116	0.999883	0.999948
120	0.995701	0.997045
200	0.999948	0.999978
204	0.999939	0.999969

Anexo X. Correlación entre Observados y CHIRPS - NASA

Correlación entre Datos observados corregidos y datos de CHIRPS y NASA

```
import pandas as pd
import numpy as np
from sklearn.metrics import mean_absolute_error
import matplotlib.pyplot as plt
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

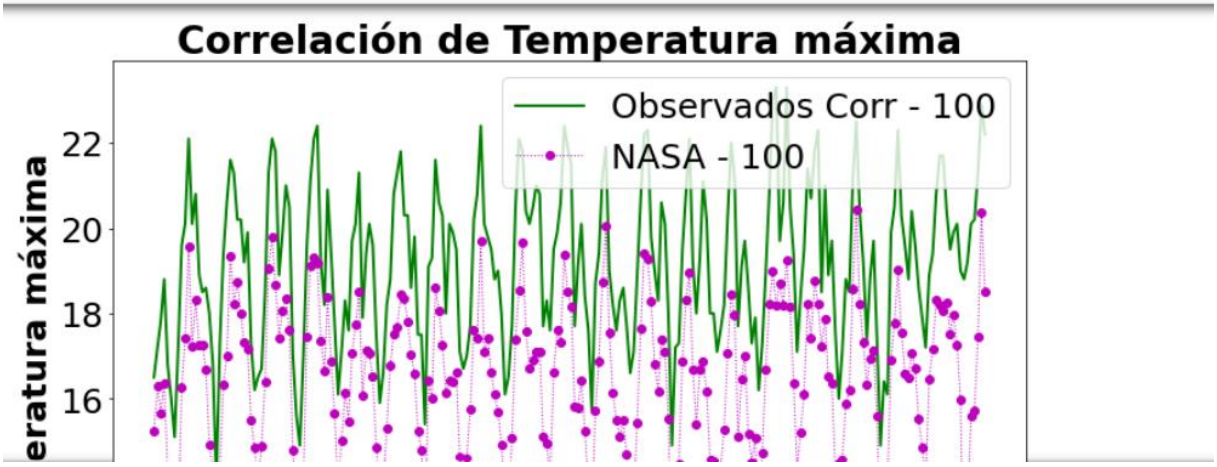
```
'''definir lista de estaciones'''
# lista_estaciones = ['100','101','102','105','106','107','108','111','112','113','114','115',/
#                     '116','117','118','119','120','121','200','201','202','203','204','205'] #Completa
lista_estaciones = ['100','101','102','108','112','116','118','120','121','200']

'''definir cantidad de datos a mostrar'''
# MENOR O IGUAL a 480
d1=240
d2=480

'''funcion para leer datos y anexar informacion en la salida'''
def leer_datos(direccion):
    data = pd.read_csv(direccion)

    if (direccion.count('series')): data=data.rename(columns={'Date':'fecha'})
    data['fecha']=pd.to_datetime(data['fecha'])
```

....

```
corr_columnas(obsc_tmax, tmax, lista_estaciones)
```

	100	101	102	108	112	116	118	120	121	200
spearman	0.880642	0.845153	0.86563	0.501619	0.667161	0.693651	0.756583	0.873625	0.757634	0.729931
pearson	0.883745	0.832686	0.85338	0.507623	0.626786	0.686772	0.746452	0.869011	0.723555	0.74173
MAE	2.172335	1.09464	2.302561	2.088998	1.262064	1.661005	1.383828	1.187318	1.231933	1.375527

```
graficar_datos(obsc_tmax, tmax, lista_estaciones)
```



Anexo XI. Código Python Corrección BIAS

Correccion BIAS

```
'''Instalar Sklearn para usar mean_squared_error '''
# %pip install sklearn
```

```
'Instalar Sklearn para usar mean_squared_error '
```

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error,mean_absolute_percentage_error, r2_score,mean_absolute_error
from sklearn.metrics import mean_absolute_error
from math import sqrt
from datetime import datetime
import matplotlib.pyplot as plt
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

```
def trans_fecha(data):
    if 'Date' in data.columns:
        data.rename(columns={'Date':'fecha'}, inplace=True)
    data['fecha'] = pd.to_datetime(data['fecha'])
    return data

def mes(data, mes):
    data = data[data['fecha'].dt.month == mes]
    data.reset_index(drop=True, inplace=True)
    return data
```

PARA PRECIPITACION

```
'''Datos Mensuales Sin Correcion'''
ppobs = pd.read_csv('../datos/2.1.estaciones_climatol/pre-m_1981-2020_series.csv')   #Corregidos climatol
ppchirp = pd.read_csv('../datos/1.0.variables/chirps_mensual_pp.csv')
```

```
'''Cálculo de correciones BIAS'''
ppobs = trans_fecha(ppobs)
ppchirp = trans_fecha(ppchirp)

ppchirp_correg = pd.DataFrame()

for m in range(1,13):
    obs_m = mes(ppobs, m)
    media_mes_obs = media_mensual(obs_m)

    chirp_m = mes(ppchirp, m)
    media_mes_chirp = media_mensual(chirp_m)

    mensual_correg = pd.DataFrame(chirp_m['fecha'], columns=chirp_m.columns)

    res = pd.Series(index = media_mes_chirp.index, dtype='float64')
    for indice in media_mes_chirp.index:
        if indice in media_mes_obs:
            res[indice] = media_mes_obs[indice] / media_mes_chirp[indice]
        else:
            res[indice] = 1.0

    for estacion in media_mes_chirp.index:
        mensual_correg.loc[:,estacion] = chirp_m.loc[:,estacion]*res[estacion]
    ppchirp_correg = pd.concat([ppchirp_correg, mensual_correg], axis = 0)
ppchirp_correg.reset_index(drop=True, inplace=True)
ppchirp_correg = ppchirp_correg.sort_values('fecha', ascending = True)
ppchirp_correg.reset_index(drop=True, inplace=True)
```

Anexo XII. Correlación datos observados corregidos-Bias

Correlación entre Datos Observados corregidos - CHIRPS y NASA-BIAS

```
import pandas as pd
import numpy as np
from sklearn.metrics import mean_absolute_error
import matplotlib.pyplot as plt
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

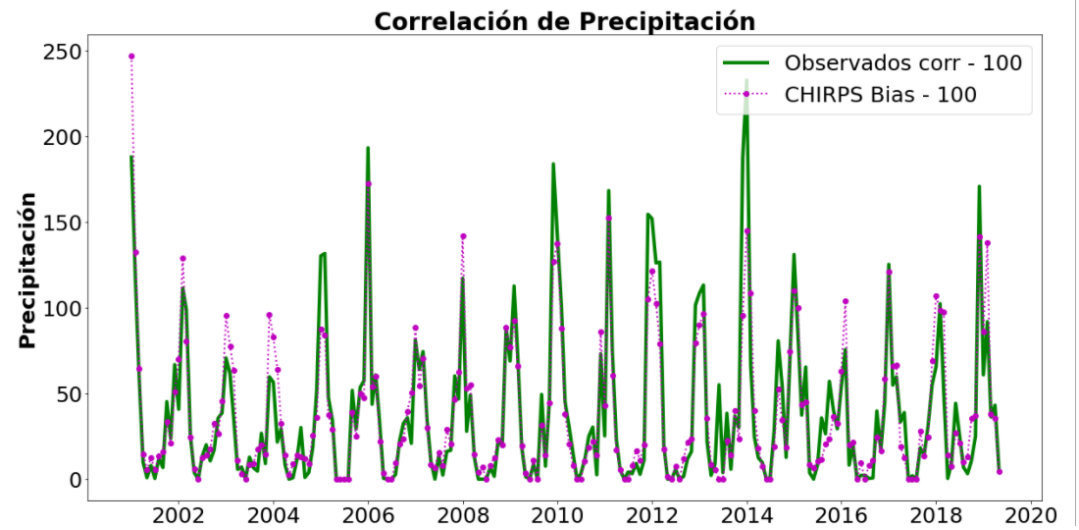
```
'''definir lista de estaciones'''
lista_estaciones = ['100','101','102','105','106','107','108','111','112','113','114','115',
                    '116','117','118','119','120','121','200','201','202','203','204','205'] #Completa
# lista_estaciones = ['100','101','102','108','111','112','116','118','120','121','200']

'''definir cantidad de datos a mostrar'''
# MENOR a 480
d1=240
d2=460

'''funcion para leer datos y anexar informacion en la salida'''
def leer_datos(direccion):
    data = pd.read_csv(direccion)

    if (direccion.count('series')): data=data.rename(columns={'Date':'fecha'})
    data['fecha']=pd.to_datetime(data['fecha'])

    if (direccion.count('climatol')): origen='Observados corr'
    elif (direccion.count('nasa')): origen='NASA Bias'
    else : origen='CHIRPS Bias'
```

.....

	100	101	102	105	106	107	108	111	112	113	...	118	119	120
spearman	0.921657	0.780728	0.927977	0.899558	0.866346	0.882808	0.804799	0.833357	0.825944	0.855226	...	0.85533	0.803402	0.884192
pearson	0.920984	0.815312	0.931058	0.894135	0.863008	0.875573	0.788866	0.811254	0.805112	0.885146	...	0.861817	0.859453	0.887714
MAE	11.387575	13.398337	10.317383	11.493381	22.598494	8.186473	14.79905	15.914481	18.101065	9.866554	...	15.509421	11.689996	11.2407

3 rows × 24 columns

```
graficar_datos(obsc_pp, pp, lista_estaciones)
```



Anexo XIII. Código en R para cálculo de SPI y SPEI

```r
#Instalar paquete SPI - SPEI solo instalar una vez
# install.packages("SPEI")
# install.packages("readr")
```

```r
#Cargar librería SPEI y SPEI
library(SPEI)
library(readr)
```

```r
setwd("../datos/por_variables")
getwd()
```

'C:/Users/Angélica Andrade/Dropbox/documentosA/TFM_UNIR_2022/datos/por_variables'

Funciones para calcular SPI y SPEI en todas las escalas

```r
spi_escalas <- function(df,escala){
  df1 = data.frame()
  df1 = cbind(df['fecha'])
  spi_es<-spi(df[2:length(df)], scale = escala , distribution = 'Gamma')
  vals=spi_es$fitted
  df1$spi= vals
  return(df1)
}


guardar_csv <- function(data, directorio){
  write.csv(data, file = directorio, row.names = FALSE) # guarda un archivo csv

  df <- read.csv(directorio, header = TRUE, sep = ",", dec = ".")
  colnames(df) <- c("fecha","spi100","spi101","spi102","spi104","spi105","spi107","spi108","spi109","spi110","spi111","spi112
  write.csv(df, file = directorio, row.names = FALSE) # guarda un archivo csv
}
```

```r
spei_calc <- function(dfstac, escala, dfprec, dftm, dftmin = data.frame(), metodo = 2){

    dfspei <- data.frame(cbind(dfprec['fecha']))
    columnas = colnames(dfprec[,2:ncol(dfprec)])

    for (est in 1:30){
        # Asignar los datos de la estacion "est" a una Serie de Tiempo
        tspp <- ts(dfprec[est+1], end=c(2020,12), frequency=12)
        tstm = ts(dftm[est+1], end=c(2020,12), frequency=12)

        if (metodo == 1){

            tsmn = ts(dftmin[est+1], end=c(2020,12), frequency=12)

            # Calcular PET, por el metodo 1:hargreaves
            tspet <- hargreaves(Tmin = tsmn ,Tmax = tstm, lat = dfstac[est,3])

        } else {

            # Calcular PET, por el metodo 2:thornthwaite
            tspet <- thornthwaite(Tave = tstm, lat = dfstac[est,3])
        }

        # Calcular el balance hídrico: pp - PET
        tsbh = tspp - tspet

        # Calcular el SPEI para una escala "escala"
        tspei = spei(tsbh, scale = escala , distribution = 'log-Logistic')

        # Generar el nombre de la columna
        columna <- paste('spei', substr(columnas[est], 2, 4), sep = '')

        # Guardar en los valores obtenidos en dfspei
        fitted_tspei <- tspei$fitted
        dfspei[columna] <- fitted_tspei[,1]

    }

    return (dfspei)
}
```

Anexo XIV. Comparación entre SPI y SPEI

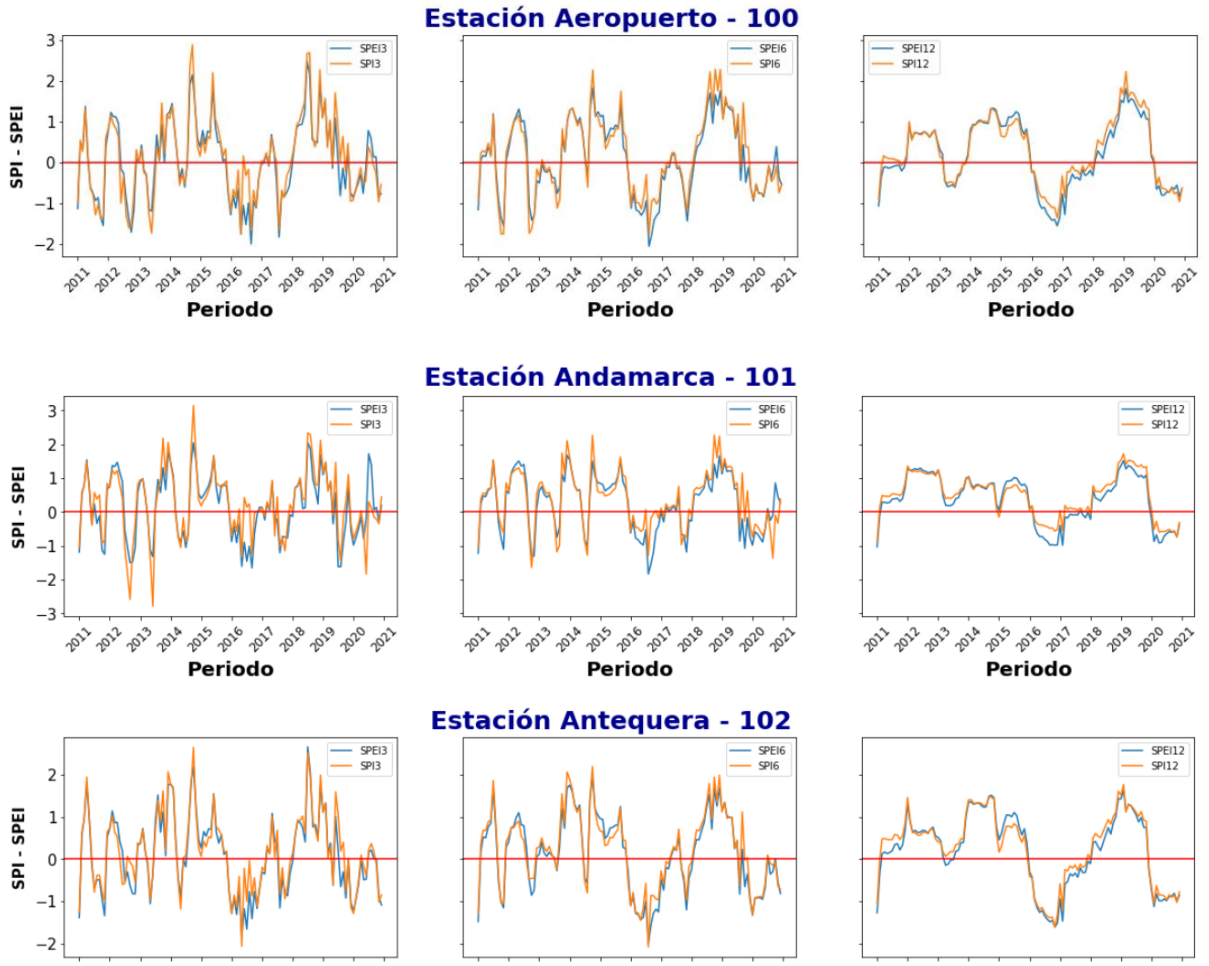
Comparación SPI - SPEI

```
import os
import pandas as pd
from matplotlib import pyplot as plt
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

```
'''Leer el contenido de un directorio en base a una escala de SPI/SPEI'''
def leer_dir(escala, path='../datos/3.0.spi_spei/'):
    with os.scandir(path) as ficheros:
        ficheros = [fichero.name for fichero in ficheros if fichero.is_file()
                    and fichero.name.endswith(str(escala)+'.csv')]
    ficheros.sort()
    return ficheros

'''Leer los archivos CSV de los indices'''
def leer_csv (archivo, ruta = '../datos/3.0.spi_spei/'):
    data = pd.read_csv(ruta + archivo,  sep = ',', parse_dates=True)
    return data
```

```
'''graficar SPI-SPEI por estacion'''
lista_arch = lista_archivo_esc([3,6,12])
for e in range(estaciones.shape[0]):
    plot_grafica_esta(lista_arch, estaciones.iloc[e,0],estaciones.iloc[e,1], 120)
```



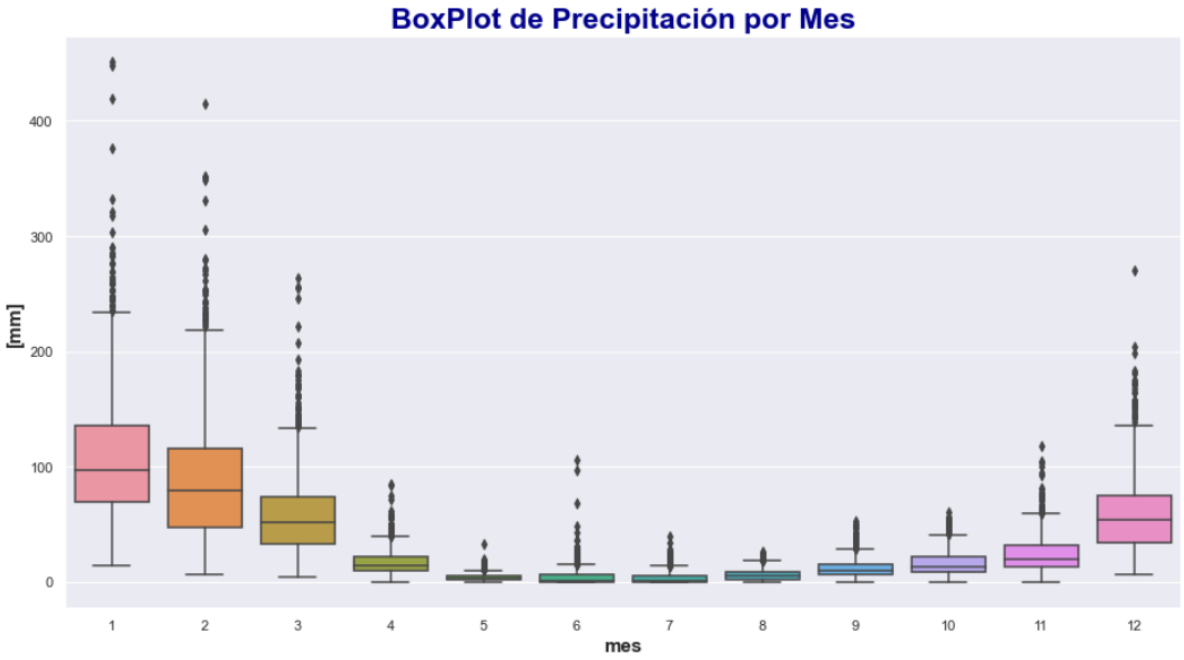
Anexo XV. Descripción y exploración de Datos(Parte1)

Descripción y Exploración de datos

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
def por_mes(data, mes):
    data['fecha']=pd.to_datetime(data['fecha'])
    return data[data['fecha'].dt.month == mes]

''' Leer archivos'''
def leer_archivo(archivo):
    if (archivo.count('indices') > 0):
        ruta = '../datos/3.0.spi_spei/'
    else:
        ruta = '../datos/1.0.variables/'
    data = pd.read_csv(ruta + archivo)
    data['fecha'] = pd.to_datetime(data['fecha'])
    return data

''' Melt por variables'''
def melt_var(data):
    valores_melt = data.columns.difference(pd.Index(['fecha']))
    data = pd.melt(data,id_vars=['fecha'], value_vars=valores_melt, var_name='estacion', value_name='valor')
    data = data.sort_values(by=['fecha','estacion']).reset_index(drop=True)
    return data
```

```
'''Diagrama de cajas por Variable mensual de todas las estaciones'''

graficar_caja_total('mes', datos=datos)
```



Anexo XVI. Descripción y exploración de Datos(Parte2)

Descripción y Exploración de datos

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_pacf, plot_acf
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
```

```
df = pd.read_csv('https://raw.githubusercontent.com/MandbeZ/TFM_sequia/main/datos/spi_spei/indices_spi3.csv', sep = ',', pars
# df = df.loc[:,["fecha","101"]]
df.dropna(inplace = True)
df.set_index('fecha',inplace = True)
```

Prueba de estacionariedad de Dickey Fuller

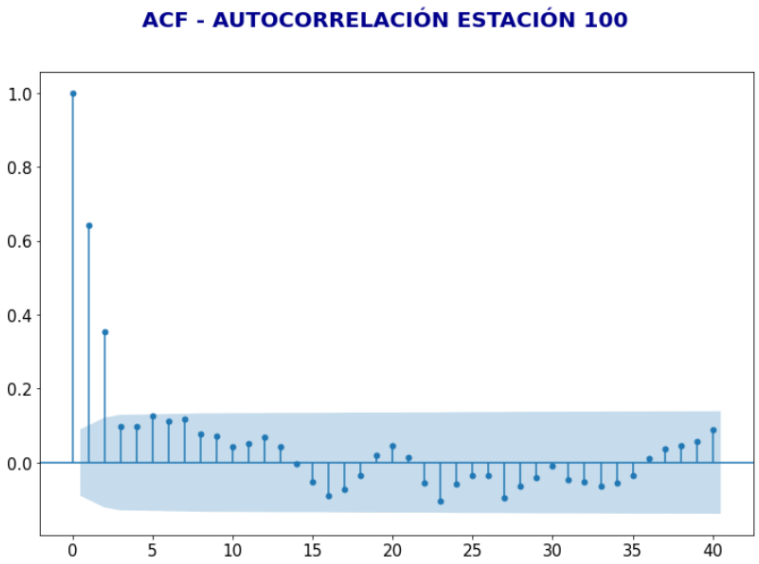
```
#Función para realizar la Prueba de Dickey-Fuller aumentada (  ADF augmented Dickey-Fuller)
def adf(x):
  res = adfuller(x)
  print("Test-Statistic:", res[0])
  print("P-Value:", res[1])
  if res[1] < 0.05:
    print("La serie es estacionaria")
  else:
    print("La serie es NO estacionaria")
```

```
for i in df.columns.values:
    print(f'Estación:{i} ')
    adf(df[i])
```

```
Estación:100
Test-Statistic: -6.023725279500043
P-Value: 1.472821729720382e-07
La serie es estacionaria
Estación:101
Test-Statistic: -8.837515886398338
P-Value: 1.703251797336152e-14
La serie es estacionaria
```

Prueba de Autocorrelación ACF y Autocorrelación Parcial PACF

```
'''Autocorrelación ACF'''
for i in df.columns.values:
    fig, ax = plt.subplots(figsize=(12, 8))
    plt.suptitle('ACF - AUTOCORRELACIÓN ESTACIÓN '+str(i), fontsize=20, color = 'darkblue', weight='bold')
    plt.xticks(fontsize=15)
    plt.yticks(fontsize=15)
    plot_acf(df[i].dropna(), ax=ax,lags=40.,title='');
```



Anexo XVII. Clusterización

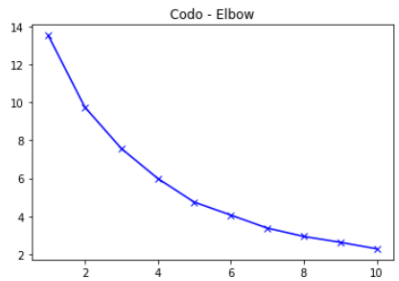
Generar clusters

```
import os
import pandas as pd
from sklearn.cluster import KMeans
import sklearn.metrics as metrics
import matplotlib.pyplot as plt
```

```
from warnings import simplefilter
simplefilter(action="ignore", category=RuntimeWarning)
simplefilter(action="ignore", category=FutureWarning)
```

```
'''leer los archivos en un carpeta dada SI son archivos CSV'''
def leer_dir(path, cond1 = 'resumen_anual'):
    with os.scandir(path) as ficheros:
        ficheros = [fichero.name for fichero in ficheros if fichero.is_file()
                                                            and fichero.name.endswith('.csv')
                                                            and fichero.name.count(cond1)>0]
    ficheros.sort()
    return ficheros

def resumen(df,col):
    res=pd.DataFrame()
    aux=df[df['anio']>=2011].reset_index(drop=True).mean()
    res[col]=aux.drop('anio')
    return res.reset_index(drop=True)
```

```
plt.plot(range(1,11),wss,'bx-')
plt.title("Codo - Elbow")
plt.show()
```




```
'''Método de Silhoute'''
for i in range(2,8):
    labels=KMeans(n_clusters=i,init="k-means++",random_state=200).fit(df_norm).labels_
    print ("Puntaje Silhouette para "+str(i)+" clusters es  "
           +str(metrics.silhouette_score(df_norm,labels,metric="euclidean",sample_size=1000,random_state=200)))
```

```
Puntaje Silhouette para 2 clusters es  0.20816921004046277
Puntaje Silhouette para 3 clusters es  0.20852907939956203
Puntaje Silhouette para 4 clusters es  0.2547738908961459
Puntaje Silhouette para 5 clusters es  0.24641887975927745
Puntaje Silhouette para 6 clusters es  0.2410148594908131
Puntaje Silhouette para 7 clusters es  0.2390911505792083
```

```
'''Para ambos métodos el número de clusters óptimo es 7'''
clustering = KMeans(n_clusters=4,max_iter=300)
clustering.fit(df_norm)
clustering.labels_
```