



Modelamiento de Sequía en el departamento de La Guajira, Colombia

Juan Sebastián Vinasco Salinas

Universidad del Valle
Facultad de Ingeniería, Escuela Ingeniería Civil y Geomática
Santiago de Cali, Colombia
2022

Modelamiento de Sequía en el departamento de La Guajira, Colombia

Juan Sebastián Vinasco Salinas

Trabajo de grado presentado como requisito para optar al título de:
Ingeniero Topográfico

Director:
MSc. Francisco Luis Hernandez Torres

Línea de Investigación:
Modelamiento y monitoreo de fenómenos biofísicos
Grupo de Investigación en Percepción Remota

Universidad del Valle
Facultad de Ingeniería, Escuela Ingeniería Civil y Geomática
Santiago de Cali, Colombia
2022

Alea iacta est

If you did not understand the nature of the beasts,

it would be of little use to know the mechanics of
their anatomy.

Agradecimientos

Para poder llegar a los resultados aquí descritos, he seguido un proceso arduo y complejo y en cada pequeño paso hubo personas que me extendieron su mano para ayudarme, así que a continuación enumeraré algunas de esas personas.

Mi familia, sin cuyo apoyo a mi poco común y quizás algo irreflexiva empresa de lanzarme a hacer un trabajo de grado como el presente.

A mis compañeros algunos que ejercieron como mentores Frank, Rafael, Mauricio y Wilmer. Otros que me inspiraron a continuar a través del ejemplo de sus proyectos, July, Sergio, Camilo, Daniel, Evelin, Laura, Stephany, Jorge, Diego, Valentina, Angie y otros que se me escapan.

A mis profesores, que me enseñaron no solo conocimiento práctico sino también humano.

Al profesor Francisco que me enseñó el valor de soñar y perseguir los sueños.

A la profesora Jackeline que me empujó para que pudiese llegar al presente resultado.

Al profesor Mauricio con quien empecé esta aventura llamada investigación.

A otros como Eduardo Caicedo, Wilfredo Morales que compartieron su conocimiento en modelos impulsados por datos, al profesor John Sanabria que me abrió las puertas para a través de la tecnología extender los límites de los análisis que hacemos en la profesión, al menos en Univalle.

A todos los anteriores, gracias totales !!!

Abreviaturas

NDVI : Índice Normalizado de Vegetación.

NDWI : Índice de Agua Normalizado.

EVI : Índice de Vegetación Mejorado.

LST : Temperatura Superficial.

SPI: Índice de Sequía Estandarizado.

SPI No Paramétrico: Índice de Sequía Estandarizado No Paramétrico.

Random Forest : *Random Forest* o Bosque Aleatorio.

Extrem Gradient Boosting : *Extrem Gradient Boosting* o *Xtrem Gradient Boosting* o Aumento Extremo del Gradiente.

NN : *Neural Networks* o Redes Neuronales Artificiales\CART : *Classification And Regression Trees* o Árboles de Clasificación y Regresión.

MOD11A2 : Producto MODIS Temperatura Superficial

MOD16A2 : Producto MODIS Evapotranspiración

MCD43B3 : Producto MODIS índices de albedo

MYD13A2 : Producto MODIS índices de vegetación

MOD16A2 : Producto MODIS índices de evapotranspiración

MCD15A2 : Producto MODIS índices de area foliar y fracción de la radiación fotosintéticamente activa.

Resumen

Las sequías son procesos hidrometeoro-lógicos que eliminan o disminuyen la cantidad y distribución espacial de la precipitación limitando la disponibilidad de agua superficial y/o subterránea, de modo que se presentan condiciones mucho más secas de lo usual, que se convierten en potencialmente perjudiciales. En palabras de la Organización Meteorológica Mundial en su Manual de Indicadores de Sequía este fenómeno natural es uno de los que más daños económicos causan año tras año, afectando de manera significativa a los bienes y la calidad de vida de las personas. En particular en Colombia una de las zonas más vulnerables a este fenómeno es el departamento de La Guajira, que por sus condiciones geográficas, y sociales ha sido históricamente impactado por este fenómeno.

Ahora bien, con respecto a lo anterior, es preciso aclarar que las sequías a diferencia de otros fenómenos naturales se caracterizan por usualmente cubrir áreas más extensas que sus contrapartes, además que su ocurrencia gradual y paulatina hace difícil determinar su duración absoluta, haciendo de la sequía un fenómeno especialmente susceptible al monitoreo constante. Igualmente la estrecha relación existente entre los fenómenos de sequía y su impacto en las condiciones usuales de las diferentes variables del ciclo hidrológico, se hace necesario evaluar la ocurrencia de sequías con el comportamiento de variables como la precipitación, la temperatura superficial, el comportamiento de la vegetación, y la humedad presente en el suelo.

Sin embargo y a pesar de la necesidad de monitoreo constate, la evaluación de los fenómenos de sequía se puede verse limitado por la cantidad y calidad de los registros necesarios para evaluarla, además de las limitaciones propias de las aproximaciones utilizadas, donde los índices basados en rachas de datos hidrometeorológicos de estaciones en terreno son más precisos pero tienen una cobertura espacial pobre, por su parte las metodologías basadas en sensores remotos cubren grandes áreas utilizando un muestreo espacial mucho más detallado pero adolecen de imprecisión.

En este orden de ideas en el trabajo aquí descrito adapta una metodología existente en la que a partir del cálculo del índice de precipitación estandarizada no paramétrico en un conjunto de estaciones meteorológicas, se implementan una serie de algoritmos impulsados por datos para la integración e interpolación no lineal de este índice con productos satelitales referentes a diferentes variables del ciclo hidrológico para la derivación de productos del índice de precipitación estandarizado no paramétrico para el departamento de La Guajira consistentes espacial y temporalmente, en las escalas de 3, 6 y 9 meses.

Se obtuvo una serie temporal de mapas de SPI para la duración total del estudio, el modelo que mejor se aproximó fue el de Extrem Gradient Boosting con un error cuadrático medio estimado de 0.7 y un error absoluto medio de 0.8; frente a los 0. y 0.9 de Random Forest, mientras que las redes neuronales no consiguieron captar el comportamiento del fenómeno. Estos errores se concentran en la estimación de los valores máximos y mínimos locales del SPI. Además de un efecto pixelado en la predicción espacial. Finalmente, se midió la importancia de las variables utilizadas en la predicción espacial, obteniendo como variable más importante la precipitación estimada por satélite (producto CHIRPS), seguida de la temperatura superficial y el índice de sequía normalizado (NDDI).

Palabras clave: (Sequía, Aprendizaje Automático, Modelos Espacio Temporales).

Abstract

Droughts are hydrometeorological processes that eliminate or diminish the amount and spatial distribution of precipitation limiting the availability of surface water and/or groundwater, so that much drier conditions than usual occur and are potentially harmful. In the words of the World Meteorological Organization in its Drought Indicators Manual, this natural phenomenon is one of the ones that causes the most economic damage year after year, significantly affecting people's assets and quality of life. Particularly in Colombia, one of the most vulnerable areas to this phenomenon is the department of Guajira, a department that due to its geographical and social conditions has been historically impacted by this phenomenon.

Now, with respect to the above, it is necessary to clarify that droughts, unlike other natural phenomena, are usually characterized by covering more extensive areas than their counterparts, and that their gradual and gradual occurrence makes it difficult to determine their absolute duration, making drought a phenomenon that is especially susceptible to constant monitoring. Likewise, the close relationship between drought phenomena and disturbances in the usual conditions of the different variables of the hydrological cycle, it is necessary to evaluate the occurrence of droughts with the behavior of variables such as precipitation, surface temperature, vegetation behavior, and the humidity present in the soil.

However, and in spite of the need of constant monitoring, the evaluation of drought phenomena can be limited by the quantity and quality of the necessary registries to evaluate it, in addition to the limitations of the used approximations, where the indexes based on historical registries of stations in the field are more precise but have a poor spatial cover, on the other hand, the methodologies based on remote sensors cover great areas using a much more detailed spatial sampling but they suffer from imprecision.

In this order of ideas in the work described here, a methodology is proposed in which from the calculation of the nonparametric standardized precipitation index in a set of meteorological stations, a series of algorithms driven by data are implemented for the integration and nonlinear interpolation of this index with satellite products referring to different variables of the hydrological cycle for the derivation of products of the nonparametric standardized precipitation index for the department of Guajira consistent spatially and temporally, in the scales of 3, 6 and 9 months.

Obtaining a time series of SPI maps for the total length of the study, the model that best approximated was the Extrem Gradient Boosting model with an estimated mean squa-

re error of 0.7 and a mean absolute error of 0.8; compared to 0.8 and 0.9 for the random forest, while the neural networks failed to capture the behavior of the phenomenon. These errors are concentrated in the estimation of the highest and lowest SPI values. In addition to a pixelated effect in the spatial prediction. Finally, the importance of the variables used in the prediction was measured, obtaining as the most important variable the precipitation estimated by satellite (CHIRPS product), followed by the surface temperature and the NDDI indicator.

Key words (Drought, Machine Learning, Spatio-Temporal Models).

Índice general

I	Introducción	18
II	Planteamiento del Problema	21
III	Objetivos	24
1	Objetivo General:	25
2	Objetivos Específicos	26
IV	Justificación	27
V	Marco Conceptual	30
3	Sequía	31
3.1	Sequía Meteorológica	31
3.2	Sequía Agrícola	32
3.3	Sequía Socio-económica	32
4	Rachas de datos hidrometeorológicos	33
5	Sensores Remotos	34
6	Modelamiento de Sequías	37
7	Técnicas basadas en modelos impulsados por datos para el modelamiento de Sequías	40
VI	Marco Teórico	41
8	Sequía	43

9 Modelos	48
9.1 Datos <i>in situ</i>	48
9.1.1 Series de Tiempo	48
9.1.2 Geoestadística	50
9.2 Sensoramiento Remoto	51
9.2.1 Índices de basados en teledetección	51
9.2.2 Productos derivados de datos satelitales	53
9.3 Modelos Impulsados por datos	53
9.3.1 Árboles de Decisión	55
9.3.2 Bosque Aleatorio	56
9.3.3 Aumento Extremo del Gradiente	57
9.3.4 Redes Neuronales Artificiales	58
9.3.5 Validación Cruzada	59
 VII Marco de Referencia	 61
 VIII Metodología	 65
10 Zona de Estudio	66
11 Conjunto de Datos	67
12 Métodos	69
12.1 Pre-procesamiento de datos <i>in situ</i>	71
12.2 Pre-procesamiento de datos raster	72
12.3 Procesamiento	76
12.4 Entrenamiento y ajuste de hiperparámetros	76
12.4.1 Random Forest	76
12.4.2 Extrem Gradient Boostings	76
12.4.3 MLP	77
12.5 Validación Cruzada	77
12.6 Importancia de las Variables	78
 IX Resultados	 79
13 Pre-procesamiento de datos <i>in situ</i>	81
14 Pre-procesamiento de datos raster	83
14.1 Precipitación	83

14.2 Evapotranspiración	85
14.3 Temperatura Superficial	87
14.4 NDWI	89
14.5 NDVI	91
14.6 EVI	93
14.7 LAI	95
14.8 NDDI	97
15 SPI no paramétrico estimado	100
15.1 Ajuste de los Valores	100
15.1.1 Redes Neuronales	100
15.1.2 Bosque Aleatorio (Random Forest)	101
15.1.3 Aumento Extremo del Gradiente (Extrem Gradient Boosting)	103
15.2 Métricas de Error	105
15.3 Comportamiento Temporal	107
15.3.1 Random Forest	107
15.3.2 Extrem Gradient Boostings	109
15.4 Comportamiento Espacial	111
15.4.1 Random Forests	111
15.4.2 Extrem Gradient Boostings	114
15.5 Importancia de los predictores	116
X Discusión	118
XI Conclusiones	123
XII Observaciones y Recomendaciones	126
XIII Anexos	128
16 AED Estaciones Metereológicas	129
17 Imputación de datos	134
18 Ciencia Reproducible	135
XIV Bibliografía	136

Índice de figuras

1	Ejemplo del muestreo de información realizado por un sensor remoto.	35
2	Resolución espacial de un sensor remoto	35
3	Espectro electromagnético	36
4	Representación del algoritmo Random Forest	56
5	Representación gráfica Validación cruzada	59
6	Validación cruzada en estaciones meteorológicas	59
7	Estrategia de validación cruzada espacio-temporal	60
8	Zona de estudio del proyecto	66
9	Flujograma General del Proyecto	70
10	Flujograma de preprocesamiento raster	72
11	Recorte a la zona de estudio	73
12	Cálculo NDDI	75
13	Imputación de datos faltantes en una estación de muestra	81
14	Comportamiento del SPI en una estación de muestra entre el año 2000 y el año 2014.	82
15	Comportamiento del SPI en una estación de muestra entre el año 2000 y el año 2014.	82
16	Comportamiento promedio mensual de la precipitación	84
17	Gráfico de densidad promedio mensual de la precipitación	84
18	Gráfico de caja-violín promedio mensual de la precipitación	85
19	Comportamiento promedio mensual de la evapotranspiración	86
20	Gráfico de densidad promedio mensual de la evapotranspiración	86
21	Gráfico de caja-violín promedio mensual de la evapotranspiración	87
22	Comportamiento promedio mensual de la temperatura Superficial	88
23	Gráfico de densidad promedio mensual de la temperatura Superficial	88
24	Gráfico de caja-violín promedio mensual de la temperatura Superficial	89
25	Comportamiento promedio mensual del NDWI	90

26	Gráfico de densidad promedio mensual del NDWI	90
27	Gráfico de caja-violín promedio mensual del NDWI	91
28	Comportamiento promedio mensual del NDVI	92
29	Gráfico de densidad promedio mensual del NDVI	92
30	Gráfico de caja-violín promedio mensual del NDVI	93
31	Comportamiento promedio mensual del EVI	94
32	Gráfico de densidad promedio mensual del EVI	94
33	Gráfico de caja-violín promedio mensual del EVI	95
34	Comportamiento promedio mensual del LAI	96
35	Gráfico de densidad promedio mensual del LAI	96
36	Gráfico de caja-violín promedio mensual del LAI	97
37	Comportamiento promedio mensual del NDDI	98
38	Gráfico de densidad promedio mensual del NDDI	98
39	Gráfico de caja-violín promedio mensual del NDDI	99
40	Correlacion entre los valores reales y predichos Red Neuronal 3 Meses	100
41	Correlacion entre los valores reales y predichos Random Forest 3 meses	101
42	Correlacion entre los valores reales y predichos Random Forest 6 meses	102
43	Correlacion entre los valores reales y predichos Random Forest 9 meses	102
44	Correlacion entre los valores reales y predichos Extrem Gradient Boosting	103
45	Correlacion entre los valores reales y predichos Extrem Gradient Boosting	104
46	Correlacion entre los valores reales y predichos Extrem Gradient Boosting	104
47	Métricas de Error Validación Cruzada el eje x representa 20 subconjuntos de datos y el eje y representa los errores en cada ronda de entrenamiento	106
48	Perfil temporal entre los valores reales y predichos Random Forest 3 Meses (En negro el SPI en verde la estimación del modelo)	108
49	Perfil temporal entre los valores reales y predichos Random Forest 6 Meses (En negro el SPI en verde la estimación del modelo)	108
50	Perfil temporal entre los valores reales y predichos Random Forest 9 Meses (En negro el SPI en verde la estimación del modelo)	109
51	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses (En negro el SPI en verde la estimación del modelo)	110
52	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 6 Meses (En negro el SPI en verde la estimación del modelo)	110
53	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 9 Meses (En negro el SPI en verde la estimación del modelo)	111
54	Comportamiento promedio mensual de la predicción SPI por Random Forest	112
55	Gráfico de densidad promedio mensual de la predicción SPI por Random Forest	112

56	Gráfico de caja-violín promedio mensual de la predicción SPI por Random Forest	113
57	Comportamiento promedio mensual de la predicción SPI por Extrem Gradient Boosting	114
58	Gráfico de densidad promedio mensual de la predicción SPI por Extrem Gradient Boosting	115
59	Gráfico de caja-violín promedio mensual de la predicción SPI por Extrem Gradient Boosting	115
60	Importancia de las variables predictoras Random Forest 3 Meses	117
61	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	129
62	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	130
63	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	130
64	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	131
65	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	131
66	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	132
67	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	132
68	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	133
69	Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses	133

Índice de tablas

1	Valores de Referencia del SPI	43
2	Variables derivadas del sensor MODIS	67

3	Tabla de errores MAE promedio.	107
4	Tabla de errores MSE promedio	107

Parte I

Introducción

Las sequías como fenómenos hidrometeoro-lógicos extremos dependiendo de las condiciones de la zona afectada pueden llegar a tener una incidencia mayor sobre las comunidades que la padecen, sin embargo la caracterización de esta es difícil debido a su comportamiento espacio-temporal sobre amplias zonas y la ausencia de una estrategia de muestro de información que capture la integridad del fenómeno.

En el presente trabajo se propone la estimación y monitoreo de parámetros biofísicos relacionados con la sequía utilizando modelos de inteligencia artificial. Se utilizaron datos mensuales desde 1982 hasta 2014, de precipitación provenientes de estaciones meteorológicas y datos entre el año 2000 y el 2014 de diferentes plataformas espaciales.

Provenientes del sensor Espectrorradiómetro de Imagen de Resolución Moderada (*MODIS* por sus siglas en ingles) se seleccionaron los productos satelitales de temperatura superficial, índice de vegetación de diferencia normalizada, índice de vegetación mejorando, índice de área foliar y evapotranspiración.

Del producto CHIRPS se obtuvo valores de precipitación en (mm) derivados de satélites con sensores infrarrojos y un conjunto de datos en tierra.

El proceso metodológico comprendió la identificación, evaluación y selección de imputación de datos, recorte y homogenización de las series de tiempo de todos los datos utilizados y extracción de datos de píxeles en los que se encontraban ubicadas las estaciones meteorológicas. Con los datos de las estaciones meteorológicas se calculó el índice de precipitación estandarizado (SPI) no paramétrico. Además con los datos derivados de los productos satelitales se buscó espacializar el indicador SPI mediante experimentos con tres modelos de inteligencia artificial diferentes, red neuronal perceptron multicapa, bosque aleatorio, Extrem Gradient Boostings.

Para tener en cuenta las particularidades de los datos espaciales y temporales, como los son sus propiedades de autocorrelación espacial y temporal se implementó una metodología llamada validación cruzada espacio-temporal, de modo que los resultados obtenidos fueran realistas y no sobre optimistas.

De modo que el presente documento, adicional a este capítulo se encuentra compuesto por los siguientes capítulos : Introducción I, Planteamiento del Problema II, Objetivos III, Justificación IV, Marco Conceptual V, Marco Teórico VI, Marco Referencial VII, Marco Conceptual

VIII, Resultados IX, Discusión X, Conclusiones XI, Observaciones y Recomendaciones XII, Anexos XIII, Bibliografía XIV.

Parte II

Planteamiento del Problema

Uno de los retos colectivos más desafiantes a futuro, que la humanidad deberá enfrentar es el cambio climático y las consecuencias que este ha traído y traerá; uno de los más preocupantes es que a raíz de este cambio global se están alterando los regímenes de precipitación en todas sus escalas, que conforme a los pronósticos y la opinión general de los expertos es la ocurrencia cada vez más frecuente y más extrema de las sequías [1] .

Como lo menciona la Convención de las Naciones Unidas para la lucha contra la Desertificación y la Sequía (UNCCD) las temperaturas más altas, los cambios en los patrones de precipitaciones y la mayor escasez de agua debido al cambio climático alterarán la idoneidad de vastas regiones para la producción de alimentos y la vivienda humana [2]. Esto se evidencia en la constante batida de records de temperatura global, año a año.

A nivel mundial la situación es preocupante, la ocurrencia de sequías en todo el globo, ha impactado los recursos en países africanos como Somalia, Sudan del Sur, Nigeria y Yemen, afectando un estimado de 20 Millones de personas [3]. Pero la sequía también afecta a países altamente industrializados, en el estado Norteamericano de California recientemente finalizó una sequía de seis años (2018), la pero desde que se tienen registros, impactando el estilo de vida y la producción pecuaria de este estado Norte Americano[3].

En Colombia, una de las zonas que ha sido gravemente afectada por este fenómeno es el Departamento de La Guajira, zona del país en la que la prolongada sequía de varios años aumento el desabastecimiento de agua y disminuyó los niveles de seguridad alimentaria, afectando unas 63.000 personas [4], causando además la muerte de unos 4.770 niños entre 2007 y 2015 [5] .

Este territorio tiene dos particularidades que lo hacen altamente vulnerable a las sequías, una es la alta tasa de pobreza que llega al 53.3 %, lo que lo convierte en el segundo departamento del país en este sentido [6], y sus condiciones geográficas de ser en buena medida una zona semi-desértica, su clima es árido seco [7], asimismo su hidrografía presenta una muy baja regulación hídrica (es decir muchos causes son temporales) [8], limitando la recuperación que puedan tener tanto los ecosistemas como las personas a los fenómenos de sequía.

Ahora bien el estudio de la sequía en La Guajira se basa en redes meteorológicas e hidrológicas, en palabras de la Organización Meteorológica Mundial en países en desarrollo suelen tener una densidad de estaciones inadecuada (insuficiente representatividad espacial) para medir los principales parámetros climáticos y de abastecimiento de agua, además la calidad

de los datos también es un problema, debido a las lagunas temporales de que adolencen o a la inadecuada longitud de los registros [9]. Esto se evidencia en la presencia de solo dos estaciones meteorológicas automáticas en el departamento, que se suman a unas 100 estaciones no automáticas, para cubrir un área aproximada de unos 20.848 km².

En este orden de ideas la problemática ambiental que genera la ocurrencia de sequías y otras problemáticas ambientales en general, requieren de un intenso trabajo de levantamiento de datos y generación de información que permitan la construcción adecuada y oportuna de medidas de adaptación, y manejo tendientes a tratar el problema.

Parte III

Objetivos

1 Objetivo General:

Representar el comportamiento espacio temporal de la sequía en el departamento de La Guajira, por medio de técnicas de aprendizaje automático, y haciendo uso de variables bio-físicas.

2 Objetivos Específicos

- Caracterizar las variables que permitan evaluar la sequía en las condiciones semi-desérticas de La Guajira.
- Modelar las condiciones de Sequía de La Guajira, por medio de información espacio temporal, entre los años 2000 y 2012.
- Validar los resultados obtenidos con información de estaciones meteorológicas.

Parte IV

Justificación

El secretario general de la Organización Meteorológica Mundial, M. Jarraud menciona «A lo largo de la historia de la humanidad, la sequía ha sido uno de los problemas que han afectado a nuestro bienestar y a la seguridad alimentaria» [9]. Sin embargo es necesario precisar que la sequía no es en si misma un desastre, sino que puede llegar a serlo en función de sus efectos sobre la población local, sobre la economía y sobre el ambiente y en función de la capacidad de estos últimos para hacer frente al fenómeno y recuperarse de tales efectos [9].

Para el caso del departamento de La Guajira, este ultimo aspecto, toma preponderancia, pues la población de este departamento tiene unos altos índices de pobreza que llegan al 53 % [6]. Además la habitan alrededor de 267.000 indígenas wayuu [6] que se concentran en las zonas más áridas y secas del departamento por estas mismás condiciones geográficas, este pueblo tiene condiciones de vida nómadas y semi-nómadas, lo anterior es preocupante pues habla de una población con problemás y vulnerabilidades grandes para hacerle frente a los fenómenos de sequía.

En este punto cobra importancia la gestión de los riesgos de sequía que tiene por objeto mejorar la capacidad de la sociedad para hacer frente a ese fenómeno, donde la vigilancia y alerta temprana de la sequía son dos componentes importantes en la gestión del riesgos de sequía [9] .

En este orden de ideas, los beneficios que trae la implementación de este proyecto son diversos, pues la información generada y comunicada tendrá incidencia en la toma de decisiones.

Decisiones sobre la gestión del recurso hídrico, van a permitir la adopción oportuna de medidas para mitigar la desertificación y la sequía impactando a sectores como la agricultura y la ganadería tanto a gran escala como la de subsistencia, además permitir el adecuado manejo de las concesiones de agua de la industria minera del departamento.

Permitirá además una mejoría significativa en la planeación y ejecución de obras para la captación y potabilización de agua [10], pues sectorizar las zonas más afectadas por la sequía, en conjunto con otra información como la hidrografía superficial y subterránea (acuíferos), posibilitará que estas obras se ubiquen en los lugares con las mejores condiciones, impactando a su vez la calidad y oferta continua del recurso hídrico.

Por otro lado decisiones en cuanto a política pública, guiados por este estudio de la mano

de la previsión y las alertas sobre las condiciones de sequía facultaría al Estado para mitigar los efectos de las sequías en la salud de la población por medio de campañas que minimicen cifras como la de menores muertos, igualmente políticas públicas tienen el potencial de impactar positivamente sobre los medios de subsistencia de la población para que se adapten mejor a las condiciones secas.

Parte V

Marco Conceptual

3 Sequía

La sequía es un fenómeno hidrometeorológico, en el que a raíz de un déficit en la disponibilidad del recurso hídrico desencadenado por una baja relativa respecto a los niveles promedio de la precipitación, se genera una perturbación generalizada en todas las partes constituyentes del ciclo hidrológico, afectando una región geográfica particular, durante un intervalo de tiempo acotado.

Un aspecto importante a tener en cuenta es que a diferencia de otros fenómenos naturales causantes de desastres, las sequías son acontecimientos que se desarrollan de manera lenta en el tiempo y extendida en el espacio, es decir su formación se ve en términos de semanas y meses, en casos extremos en años (no confundir con desertificación).

En consecuencia, la sequía es un fenómeno hidrológico incluido en el contexto del ciclo hidrológico, y dada su ocurrencia, sus efectos se van propagando en cada parte del sistema, tomando diferentes denominaciones según la clase de recurso hídrico impactado, a continuación se describen los tipos de sequía más aceptados en la literatura y en particular se señala la variable del ciclo hidrológico afectada [9].

3.1. Sequía Meteorológica

Este tipo de sequía es el más común y se caracteriza por ser la primera en manifestarse, podría definirse como un déficit de precipitación prolongado por cierto periodo de tiempo respecto al comportamiento normal o promedio de la zona geográfica de estudio [11].

3.2. Sequía Agrícola

Esta se presenta como consecuencia de la anterior, y en esta se evidencia como la falta de agua precipitada disminuye la cantidad de agua almacenada en el suelo (humedad del suelo), afectando así la disponibilidad del recurso hídrico en la zona de raíces para las plantas y cultivos; lo clave en este tipo de sequías es la modelación de la humedad del suelo, y el estrés hídrico de las plantas ¹ [11].

3.3. Sequía Socio-económica

Finalmente, la afectación sufrida por la sociedades humanas generada por un fenómeno de sequía se denomina sequía socio-económica, y su impacto se puede vislumbrar en como la falta del recurso hídrico altera el desempeño normal de las personas, es la más difícil de modelar por que afecta temás como la provisión de servicios públicos, limitaciones a las industrias en la producción por falta de agua, entre muchos otros. Y no solo depende de las condiciones naturales sino que también inciden variables como la resiliencia de las comunidades o el sector económico estudiado.¹

Los sensores remotos y las rachas de datos hidrometeorológicos son la forma es que se recolecta la información necesaria para modelar la sequía, en este contexto estas se definirán a continuación.

¹Las sequías agrícolas y socio-económicas queda por fuera del alcance de este trabajo, aunque algunas de las variables que describen el estrés hídrico son usadas como predictores.

4 Rachas de datos hidrometeorológicos

Las rachas de datos hidrometeorológicos se refieren a datos climáticos u ambientales tomados en campo, con el objetivo de estudiar o monitorear un fenómeno bio-físicos que pueden que representar distintas mediciones con diferentes estrategias de medición. Eso sí, dadas las características de la sequía como se menciona en la sección 6, es preciso usar variables ambientales relacionadas con el ciclo hidrológico, más aún dados los largos periodos de retorno de la sequía, es preponderante que los registros sean de la más larga duración posible registros de variables como la precipitación, la temperatura y la evapotranspiración (por poner algunos ejemplos) [9] de las estaciones climatológicas e hidrometeorológicas de una zona geográfica en particular.

5 Sensores Remotos

Los sensores remotos son en su definición más general todos aquellos instrumentos capaces de obtener información sobre un objetivo de manera indirecta o a distancia; no obstante la clase específica de sensores remotos útiles para el estudio de la sequía se refiere usualmente a sensores transportados en vehículos satelitales, el estudio de las diferentes variables bio-físicas a través de dichos sensores se conoce por diferentes nombres, sensores remotos o teledetección en la literatura norteamericana u observación de la tierra en la literatura europea¹.

Ahora bien, formalmente la **Teledetección** es el proceso de detección y monitoreo de características físicas de un cuerpo mediante la medición de la radiación reflejada, emitida o dispersada por este^{2, 3, 4}.

Una característica importante del tipo de información básica derivada de sensores remotos es que por sus características intrínsecas como la distancia entre el sensor y el objetivo, el intervalo de tiempo entre observaciones o las longitudes de onda utilizadas para caracterizar la cubierta observada se suele clasificar los datos de acuerdo a las siguientes categorías.

- **Resolución Espacial** se refiere a el tamaño mínimo del objetivo que un sensor puede caracterizar individualmente, es expresada usualmente como el tamaño del pixel sobre el terreno, cuyas unidades son metros.

¹Se utilizará los términos **Sensores Remotos**, **Teledetección** u **Observación de la Tierra** de manera indistinta y siempre entendiéndose la teledetección como el estudio de variables bio-físicas por medio de sensores sobre plataformas satelitales.

²<https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/remote-sensing>

³<https://oceanservice.noaa.gov/facts/remotesensing.html>

⁴https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used?qt-news_science_products=0#qt-news_science_products

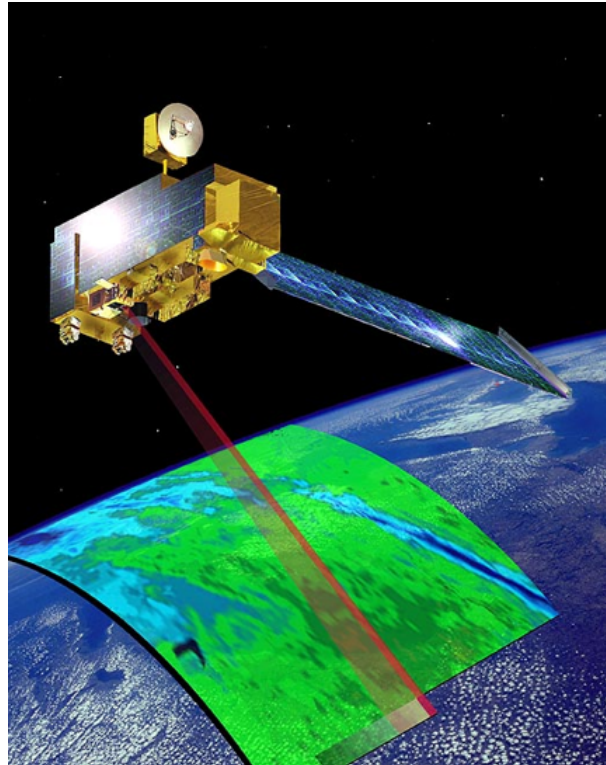


Figura 1: Ejemplo del muestreo de información realizado por un sensor remoto.

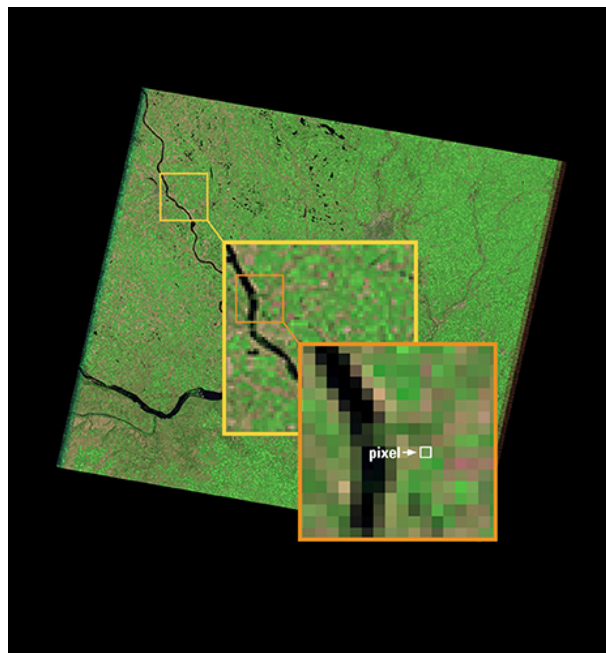


Figura 2: Resolución espacial de un sensor remoto

- **Resolución Temporal** se refiere, al intervalo de tiempo entre tomas por parte del sensor, dada la ubicación de estos sensores en órbita terrestre solo es posible sensar un objetivo con las mismas condiciones cada cierto intervalo de tiempo, cuyas unidades son días o minutos.

En particular la resolución espacial y temporal tienen una relación inversamente proporcional, entre más alta la resolución espacial, más baja es la temporal; sin embargo esto evoluciona con los nuevos desarrollos en instrumentación.

- **Resolución Espectral** se refiere a la parte del espectro electromagnético que es observada por el sensor, usualmente son unas cuantas franjas estrechas, cuyas unidades son micrómetros.

En particular los sensores usados en el presente trabajo toman información en longitudes de onda de la zona visible, infrarrojo cercano e infrarrojo térmico.

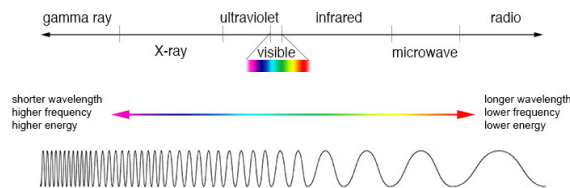


Figura 3: Espectro electromagnético

- **Resolución Radiométrica** se refiere al nivel de detalle con el que el sensor es capaz de discernir entre dos diferentes respuestas, es decir que tan sensible es a los cambios en la señal, se suele medir en el número de bits en los que el sensor graba la información captada.

Esta resolución también va evolucionando en el tiempo y usualmente los sensores más modernos tienen mayores capacidades.

6 Modelamiento de Sequías

Es preciso aclarar que las sequías se constituyen en uno de los peligros naturales más costosos económicamente hablando, además de ser uno de los técnicamente más desafiantes [12]; puesto que las zonas afectadas usualmente cubren áreas mucho más extensas que la de otros fenómenos naturales considerados desastres (por ejemplo un inundación), sumado a el hecho de que su ocurrencia es gradual o paulatina, esto la hace especialmente susceptible al monitoreo constante [12].

Ahora bien, desde el punto de vista del modelamiento del fenómeno, y haciendo énfasis en sus características como un comportamiento altamente variable tanto en el espacio, como en el tiempo; Dada la complejidad señalada, y con el objetivo de usar una métrica resumen del comportamiento del fenómeno se han desarrollado y adoptado según el caso diferentes índices para medir el comportamiento de la sequía.

Es posible, Dividir estos índices en dos grandes familias, los índices relacionados a las rachas de datos hidrometeorológicos y aquellos que tienen su origen en datos derivados de teledetección espacial.

Para empezar los índices basados en rachas de datos hidrometeorológicos, usualmente desarrollan diferentes estrategias de transformación, ponderación, y estandarización de variables relacionadas al ciclo hidrológico [11]. Entre los indicadores más utilizados se encuentra el Índice de Precipitación Estandarizado o **SPI**, y el *Índice de Sequía de Palmer*. El primero es un tipo de estandarización de la precipitación, y es usado para el modelamiento de sequías meteorológicas, el segundo más complejo que el anterior usa una aproximación a un balance hídrico sobre la zona de estudio.

Ahora bien, este tipo de aproximaciones tienen varias limitaciones, la principal recae sobre la disponibilidad y calidad de los datos necesarios para calcular los índices, y se refiere a los periodos de retorno propios del ciclo hidrológico, por lo que es necesario que las rachas de

datos tenga un mínimo de 30 años continuos [12] .

Por otro lado, dado que estos índices de sequía son calculados, con base a datos de estaciones meteorológicas o hidrológicas, también es cierto que las estimaciones son confiables en el punto exacto en el que se muestreo la información. Por lo que la información de la sequía carece de una dimensión espacial.

Para esto último, se han aplicado estimaciones basadas en una rama de la estadística llamada geo-estadística, que tiene entre otros como objetivo estimar los valores de una variables en el espacio dado un muestro espacial. Sin embargo la precisión de estas aproximaciones no es la deseable [11].

Por otro lado, y desde los años ochentas, se han desarrollado, pero sobretodo adaptado índices basados en la teledetección como aproximación a la sequía, dada la capacidad de la teledetección para capturar información relacionada con el contenido de clorofila, humedad de la hoja o temperatura de la superficie [11].

Uno de los indicadores clásicos para seguir el comportamiento del estado vegetativo de las plantas en el *Índice de Vegetación Normalizado* o NDVI, variaciones de este como el *Índice de Vegetación Mejorado* también son ampliamente usados. Otros como el *Índice de Área Foliar* o el *Índice de Agua Normalizado* o el *Índice de Sequía Normalizado* pueden ser buenos descriptores del comportamiento de la sequía.

Sin embargo , los datos remotamente sensados, pueden ser usados como entrada para complejos modelos de inversión de la transferencia radiativa y calcular otro tipo de variables de interés como lo son la Temperatura Superficial o LST por sus siglas en ingles (*Land Surface Temperature*) o estimaciones de la *Evapotranspiración* (ET), además de estimaciones de la precipitación.

Sin embargo estos métodos usualmente adolecen precisiones altas para caracterizar el comportamiento de la sequía.

Ahora bien, el paso lógico sería mezclar información proveniente de las rachas hidrológicas y de los datos remotamente sensados para obtener productos de sequía operacionales con una alta precisión y una distribución homogénea en el espacio. Sabido que no existe una

manera obvia de combinar estas fuentes de datos y de su comportamiento no lineal, el uso de métodos impulsados por datos puede ser una manera factible de combinar esta información [13].

7 Técnicas basadas en modelos impulsados por datos para el modelamiento de Sequías

Los modelos impulsados por datos son una familia de modelos y/o algoritmos capaces de establecer relaciones entre distintas variables y adaptarse de acuerdo al fenómeno estudiado, esta flexibilidad los hace atractivos para entre otras cosas servir de puente entre los diferentes tipos de datos recolectados para modelar la sequía y fusionarlos de manera que se explote los puntos fuertes de cada estrategia y se cubran los puntos débiles.

Genéricamente los modelos impulsados por datos son conocidos como «*Inteligencia Artificial*», y se corresponde con el estado actual del intento de la humanidad por sintetizar su propia inteligencia, ahora bien, los modelos impulsados por datos son un conjunto enorme de algoritmos y modelos matemáticos, alimentado por otras ramás del saber, entre sus grandes a portantes tenemos a el *Aprendizaje Estadístico* y el *Aprendizaje de Automático*.

Actualmente varios de los algoritmos desarrollados en esta rama del saber se encuentran en desuso, y la mayoría de los esfuerzos se centran en dos grandes familias, los algoritmos basados en árboles de decisión, y los algoritmos basados en redes neuronales artificiales, en particular un sub-grupo de estos modelos conocido como *Aprendizaje Profundo*, ha recibido mucha atención desde el 2014 año en el que se empezó a popularizar en el campo de la visión por computador; sin embargo, este se encuentra por fuera del alcance de este trabajo.

Parte VI

Marco Teórico

Continuando con el capítulo anterior la Parte V donde se presentan algunas definiciones sobre los temas que se tocan en este documento, se procede a formalizar teóricamente dichas definiciones relevantes.

8 Sequía

Para empezar y recapitulando, en la sección 3 la sequía es básicamente una perturbación sobre el ciclo hidrológico, la aproximación más obvia a la sequía meteorológica es medir la perturbación del agua precipitada sobre la región de estudio.

Existen distintas maneras de modelar la perturbación o el déficit de la precipitación, sin embargo, la Organización Meteorológica Mundial recomienda el uso de entre otros el Índice de Precipitación Estandarizado o SPI, este fue introducido por investigadores de la Universidad Estatal de Colorado [14]. y fue diseñado para cuantificar los déficits de precipitación a partir de una serie o racha de datos de precipitación medido continuamente por al menos 30 años.

Los posibles valores del SPI y su interpretación se listan en la Tabla 1.

Tabla 1: Valores de Referencia del SPI

Valor SPI	Intensidad de la Sequía
$SPI > 2.0$	Severamente Húmedo
$1.5 < SPI \leq 2.0$	Moderadamente Húmedo
$1.0 < SPI \leq 1.5$	Anormalmente Húmedo
$1.0 \leq SPI \leq -1.0$	Normal
$-1.5 \leq SPI < -1.0$	Anormalmente Seco
$SPI < -2.0$	Severamente Seco

Para obtener el SPI original, se aplica el siguiente algoritmo a partir de los registros de precipitación, agregados en ventanas de 3, 6 o 9 meses comúnmente, luego se realiza la visualización en frecuencia de la precipitación acumulada.

Luego, se verifica empíricamente que la precipitación no sigue una distribución normal o gaussiana, por lo que se aplica la ecuación 1 que corresponde a la función de densidad gamma [15].

$$g(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, (x > 0) \quad (1)$$

donde α es un parámetro de ajuste ($\alpha > 0$), β es un parámetro de escala ($\beta > 0$) y x la cantidad de precipitación acumulada ($x > 0$). De manera que la probabilidad acumulada de precipitación para una escala de tiempo dada es de (2):

$$G(x); \int_0^x g(x) dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} = \int_0^x x^{\alpha-1} e^{-\frac{x}{\beta}} dx \quad (2)$$

Los parámetros de forma α y escala β (3, 4) se calculan usando las siguientes ecuaciones.

$$\alpha = \frac{1 + \sqrt{1 + \frac{4 * A}{3}}}{4 * A} \quad (3)$$

$$\beta = \frac{\bar{x}}{\alpha} \quad (4)$$

Que a su vez dependen de una variable auxiliar definida por (5):

$$A = \ln(\bar{x}) - \frac{\sum \ln(x)}{n} \quad (5)$$

Donde n es el número de precipitaciones observadas y $\bar{x}$ es el promedio de la precipitación bajo la escala de interés [15].

Dado que es posible que exista una precipitación de cero y la función gamma es indefinida para este valor, se aplica un factor de corrección al ajuste que depende de la precipitación nula. Donde la probabilidad acumulada total es de: (6)

$$H(x) = q + (1 - q)G(x) \quad (6)$$

Donde q es la probabilidad de que ocurra un cero y este se calcula como $q = m/n$, siendo m el número de ceros en la serie de tiempo n . Por otra parte $(1 - q)$ es la probabilidad de que no ocurra un cero [15].

Ahora como para transformar esta función de densidad de probabilidad acumulada a la distribución normal, se aplica la siguiente ecuación (7):

$$H(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \quad (7)$$

Dándose dos posibles casos de solución:

- Caso 1: $0 < H(x) \leq 0,5$.

$$SPI = -\left(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 d_3^t}\right) \quad (8)$$

$$t = \sqrt{-2 * \ln(H(x))} \quad (9)$$

- Caso 2: $0,5 < H(x) \leq 1$.

$$SPI = t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 d_3^t} \quad (10)$$

$$t = \sqrt{-2 * \ln(1 - H(x))} \quad (11)$$

Con valores de constantes de: $c_0 = 2,515517$, $c_1 = 0,802853$, $c_2 = 0,010328$, $d_1 = 1,432788$, $d_2 = 0,189269$, $d_3 = 0,001308$ [15].

Sin embargo, este indicador adolece de algunas propiedades necesarias para su aplicación sobre extensas áreas geográficas. Entre estas la precipitación se constituye en el único dato de entrada del indicador, dejando de lado variables relevantes como lo son la temperatura, o la humedad del suelo [12].

Otra variable importante a tener en cuenta, es la longitud mínima de registro necesaria para la aplicación de este índice que viene a ser de no menos de 20 años, y se recomiendan al menos 30 sin embargo en países en desarrollo como el que nos atañe estos registros son más bien escasos y además de eso la consistencia de los datos usualmente no es la idónea y se presentan períodos de tiempo sin registros, o registros erróneos debido a daños en los sensores.

Finalmente, la debilidad más importante del SPI en su cálculo ordinario presupone que el registro en frecuencia de la precipitación se ajusta a una distribución de probabilidad usualmente la distribución gamma, sin embargo, y especialmente cuando existen áreas extensas y comportamiento climático variado se puede dar el caso que diferentes distribuciones de probabilidad se ajusten a diferentes condiciones climato-geográficas o sencillamente que el supuesto no se cumpla.

Esta última dificultad fue abordada en el año 2014 por dos investigadores de la universidad del California Irvine [16], que propusieron un marco general para la derivación de indicadores de sequía no paramétrica estandarizada.

Para esto los autores proponen sustituir el ajuste de una distribución de probabilidad (gamma en el ejemplo anterior), por una función de probabilidad empírica, por medio de un método llamada empírica gringorten plotting position. Lo anterior expresado de una forma más formal, se describe en la siguiente sección:

Partiendo de la ecuación (5-8), el método propone reemplazar la función gamma por la posición de graficación de gringorten denotada por (12):

$$p(X_i) = \frac{i - 0,44}{n + 0,12} \quad (12)$$

donde n denota el tamaño de la muestra, i denota el rango de la precipitación no cero, y $p(x_i)$ corresponde a la probabilidad empírica. Usando esta aproximación empírica no necesita de las ecuaciones (5-14, 5-15, 5-16, 5-17, 5-18) (7,8,9,11) para derivar probabilidades empíricas la salida de la ecuación $p(X)$ puede ser transformada en un índice estandarizado mediante la siguiente ecuación (13):

$$SI = \phi^{-1}(p) \quad (13)$$

donde ϕ corresponde a la función de distribución normal y p es la probabilidad derivada de (5-19) (12) También se pueden estandarizar los percentiles utilizando la siguiente aproximación comúnmente utilizada (14) [15].

$$SI = \begin{cases} -(t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 d_3^t}) si 0 < p \leq 0,5 \\ t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 d_3^t} si 0,5 < p \leq 1 \end{cases} \quad (14)$$

donde $c_0 = 2,515517$; $c_1 0,802583$; $c_2 = 0 : 010328$; $d_1 1,432788$; $d_2 = 0,189269$; $d_3 = 0,001308$

9 Modelos

Ahora bien, las aproximaciones matemáticas a la hora de modelar el comportamiento de la sequía dependen en buena medida del tipo de mediciones que se estén utilizando para su caracterización, se distinguen dos tipos de datos, los datos de rachas hidrometeorológicas en este caso de precipitación, con dos sub casos el análisis temporal y el análisis espacial de estos; por otro lado se distinguen los datos provenientes de sensores remotos.

9.1. Datos *in situ*

Las rachas de datos hidrometeoro lógicos *in situ* por sus características intrínsecas es posible realizar ciertas modelaciones

9.1.1. Series de Tiempo

Primeramente una **Serie de Tiempo** es posible definirla como una serie de datos indexados en el tiempo, que cumplen la propiedad de que los intervalos temporales en los que se toman las medidas son constantes y dotados de una dependencia estadística entre unos y otros.

La dependencia puede ser definida como la relación lineal entre una serie y una versión retrasada de sí misma en el tiempo, también conocida como auto-correlación temporal. Esta ultima puede ser expresada matemáticamente por medio de la Función de Auto-correlación (*ACF* 15 por sus siglas en ingles), que es descrita matemáticamente mediante la ecuación [17][18]:

$$p(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, t) \cdot \gamma(t, t)}} \quad (15)$$

En la ecuación anterior (15) $p(s, t)$ representa la predictibilidad lineal de una serie de x_t tiempo al instante t usando solo el valor de otra serie de tiempo x_s [17].

Dada la auto-correlación temporal de las series de tiempo, estas tienden a tener comportamientos más o menos cíclicos, dichos comportamientos pueden dividirse en tres elementos constituyentes para facilitar su modelamientos, y pueden ser:

$$Y(t) = T(t) + S(t) + e(t) \quad (16)$$

Donde $T(t)$ se refiere a la tendencia, $S(t)$ se refiere a la estacionalidad y $e(t)$ se refiere a un componente aleatorio residual.

La tendencia se refiere a el crecimiento, decrecimiento o estancamiento de la serie a lo largo de toda la ventana temporal de estudio y se puede definir matemáticamente como :

$$\hat{T}_{tendencia} = \frac{1}{m} \cdot \sum_{k=-k}^k y_t + j, \quad (17)$$

Por su parte la estacionalidad, se refiere a los ciclos cuasi-repetitivos de una serie temporal dada, definiéndose matemáticamente como:

$$\hat{T}_{estacional} = \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1} + \frac{1}{4}y_{t+2} \quad (18)$$

El último componente es la aleatoriedad de la serie temporal, definido como $e(t)$ y como su nombre lo dice tiene un comportamiento aleatorio.

El planteamiento anterior de las series temporales, puede ser usado para la imputación de datos faltantes, en particular para el método *Seasonally Decomposed Missing Value Imputation*.

Este método funciona bajo el supuesto de que la serie puede ser caracterizada por los tres anteriores elementos constituyentes, y estima un valor faltante de una serie temporal dado en un tiempo dado, según el valor de la tendencia y la estacionalidad, y calcula el error de esta estimación como:

$$\hat{T}_{error} = \hat{T}_t - (\hat{T}_{tendencial} + \hat{T}_{testacional}) \quad (19)$$

Finalmente asigna o imputa el valor faltante según la siguiente ecuación:

$$X(t) = T(t) + S(t) + e'(t) \quad (20)$$

Donde T corresponde a la tendencia de la serie, S corresponde a la estacionalidad y e' es un valor aleatorio para un tiempo (t) .

9.1.2. Geoestadística

Similar a auto-correlación temporal, las mediciones tomadas en el espacio también tienden a correlacionarse entre ellas, esto se conoce como auto-correlación espacial y es una propiedad de la que se infiere que los valores medidos geográficamente más próximos tienden a ser más similares entre sí que aquellos tomados a distancias muy largas.

Aunque existen distintas maneras de modelar este comportamiento en este documento se referirá únicamente a el **Índice de Moran**. Formalmente el coeficiente de Auto-correlación espacial Global de Moran o **I** es un índice que evalúa la extensión de la autocorrelación espacial entre un conjunto de celdas o píxeles $n = x_i$ localizado en áreas contiguas, donde i es el rango i -ésimo o el valor de X .

$$I = \frac{\sum_i \sum_j W_{ij} C_{ij}}{s^2 \sum_i \sum_j W_{ij}} \quad (21)$$

Donde $w_{ij} = 1$ si la celda i y j son vecinas, de otro modo $w_{ij} = 0$; y $c_{ij} = (X_i - \bar{X})(X_j - \bar{X})$ son variables, en particular y otra locación respectivamente y s^2 corresponde a la desviación estándar.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (22)$$

Los valores positivos del Índice de Moran indican similitud entre los vecinos, los valores negativos indican que los valores de puntos cercanos son disímiles y finalmente los valores cercanos a cero indican que se trata de un proceso aleatorio no modelable por medio de técnicas de interpolación.

Existen distintas metodologías que explotan la propiedad de la auto-correlación espacial para la predicción de locaciones desconocidas en el espacio, sin embargo, esto no se abordará en el presente documento.

9.2. Sensoramiento Remoto

Como se mencionó anteriormente, (en la sección 5) los sensores remotos son útiles para derivar indicadores relacionados con la sequía, en particular hay dos tipos de variables derivadas de la teledetección que son útiles, los índices de vegetación relacionados con el estado vegetativo de las plantas y otras variables relacionadas con el ciclo hidrológico.

9.2.1. Índices de basados en teledetección

La primera familia de variables descriptoras de la sequía una de las ramas de la teledetección más clásicas y con más historia los índices de vegetación.

Estos son combinaciones inteligentes usualmente normalizadas de diferentes longitudes de onda usadas para caracterizar el estado vegetativo de las plantas.

Uno de los más antiguos y ampliamente utilizado es el índice normalizado de vegetación [19], este índice explota el hecho de que el estado vegetativo de la planta está íntimamente ligado a la respuesta de la reflectancia en los rangos del rojo e infrarrojo cercano. Definiéndose matemáticamente como:

$$NDVI = \left(\frac{\rho_{nir} - \rho_{rojo}}{\rho_{nir} + \rho_{rojo}} \right) \quad (23)$$

Donde ρ_{nir} corresponde a una reflectancia dada en el infrarrojo cercano y ρ_{rojo} a una reflectancia en el rojo.

Ampliamente utilizado dada su sencillez, y su sensibilidad a los cambios en el contenido de clorofila de las plantas [15].

En el NDVI en particular se han detectado algunos problemás de saturación con la vegetación densa, es decir el indicador pierde la capacidad de discriminar cuando las concentraciones de material vegetal son altas, existen distintas propuestas para reemplazar este indicador y evitar estos problemás de saturación, el más importante es quizás el índice de vegetación mejorado, descrito con la ecuación 24.

$$EVI = \left(\frac{\rho_{nir} - \rho_{rojo}}{L + \rho_{nir} + c1 * \rho_{rojo} + c2 * \rho_{azul}} \right) \quad (24)$$

Donde ρ_{nir} , ρ_{rojo} , ρ_{azul} se refieren a las reflectancias en las longitudes de onda roja, infrarroja y azul respectivamente y L corresponden a unas constantes, L es un factor de corrección asociado a la contaminación de la señal por la reflectividad del suelo, y los parámetros $c1$ y $c2$ son parámetros de corrección de los efectos atmosféricos sobre la longitud de onda roja y azul.

Por otro lado, algunos índices relacionados la humedad de las cubiertas observadas, por ejemplo el índice de agua normalizado, definido por la siguiente ecuación.

$$NDWI = \left(\frac{\rho_{nir} - \rho_{swir}}{\rho_{nir} + \rho_{swir}} \right) \quad (25)$$

Donde ρ_{nir} y ρ_{swir} corresponden a la reflectancia en la longitud de onda del infrarrojo cercano y el infrarrojo de onda corta.

Finalmente, realizando una combinación normalizada del comportamiento de la vegetación derivada del NDWI y de la humedad de la cubierta caracterizada por NDWI, se ha propuesto el índice de sequía normalizado, definido por la ecuación.

$$NDDI = \left(\frac{NDVI - NDWI}{NDVI + NDWI} \right) \quad (26)$$

9.2.2. Productos derivados de datos satelitales

Por otro lado es también posible derivar información de sensores remotos utilizando relaciones más complejas que las estandarizaciones entre longitudes de onda vistas anteriormente. Estos métodos usualmente se derivan de modelos de transferencia radiativa y la inversión de los mismos, en algunos casos usan estos modelos como entrada a modelos de balance hídrico o de flujos de carbono, dependiendo de la aplicación.

En este trabajo se utiliza la temperatura superficial (LST), Evapotranspiración (ET), Índice de Área Foliar (LAI) y el producto de precipitación CHRIPS como datos derivados de productos satelitales.

Son métodos bastante complejos, por lo que estos solo se mencionan y se recomienda la lectura del respectivo Documento de Bases Teóricas del Algoritmo (ATBD por sus siglas en inglés *Algorithm Theoretical Basis Document*) de cada producto si se desea profundizar.

9.3. Modelos Impulsados por datos

Los algoritmos utilizados en el presente trabajo, estos algoritmos hacen parte de una rama del conocimiento llamada *Machine Learning* o Aprendizaje automático, una sub-rama de la inteligencia artificial.

Para empezar el *Machine Learning* es un conjunto de estrategias, técnicas y métodos, tanto teóricos como prácticos que tratan de abordar problemas complejos que *a priori* no son fácilmente programables o pueden ser llevados a cabo por un sistema experto. El ejemplo clásico es el reconocimiento de rostros en una imagen, es sumamente complicado imaginar una serie de reglas absolutas para generar un algoritmo capaz de reconocer caras en una fotografía.

La forma de abordar los problemas por parte del *machine learning* también es diferente, usualmente en las metodologías clásicas se optaba por construir el modelo más simple posible que pudiese resolver un problema dado; dando como resultado la preponderancia de modelos lineales por su simplicidad y por sobre todo la capacidad explicativa de esta familia de modelos.

Regla de Predicción

Esta idea es, sin embargo, descartada por completo por la aproximación basada en *machine learning* puesto que la idea aquí es encontrar una regla de predicción entre un conjunto de variables predictoras y una o varias variables a predecir. Según la ecuación

$$f : \mathbb{R}^d \rightarrow \mathbb{R} \quad (27)$$

$$*x* \mapsto y \quad (28)$$

Aprendizaje Supervizado

La estrategia para encontrar esta regla de predicción se basa en minimizar una función de pérdida entre el resultado obtenido y el resultado deseado, ajustando un conjunto de hiperparametros θ que controlan el comportamiento de la función de predicción hallada.

La estrategia anteriormente descrita se conoce como aprendizaje supervisado, y es la estrategia de *machine learning* más relevante hoy en día, más no la única¹.

Existen una enorme variedad de algoritmos desarrollados en el marco anteriormente mencionado, sin embargo existen dos grupos o familias de modelos que resaltan y que se describirá a continuación.

¹Otras estrategias de *machine learning* son aprendizaje no supervisado, aprendizaje semi-supervisado y aprendizaje de refuerzo.

9.3.1. Árboles de Decisión

Los árboles de decisión, o CART es un algoritmo diseñado para la generación automática de un grafo, capaz de representar la complejidad de un conjunto de datos mediante su división a diferentes niveles, en dos nodos hijos, a partir de escisiones numéricas repetidamente. Ahora bien, la idea del crecimiento de un árbol, es escoger la división ente todas las posibles divisiones en la que los datos del nodo hijo sean los más puros posibles [20] .

La formulación de este algoritmo también denominado el proceso de crecimiento de un árbol puede dividirse en cuatro pasos y estos son:

1. Selección de características:

En este paso se selecciona una característica o variable descriptor que será usada en este nivel para dividir el conjunto de datos, de acuerdo a la varianza de cada variable independiente.

2. Condiciones de división:

En el siguiente paso se una el error medio cuadrático de todas las muestras que caen en dicho nodo para determinar qué tan válido ha sido la división escogida.

3. Condiciones de Parada:

- Sí un nodo se convierte en puro, es en todos los casos en el nodo tiene un valor idéntico a la variable dependiente, así que el nodo no será dividido.
- Sí en todos los casos tienen valores idénticos para cada predictor, el nodo no será dividido.
- Sí el tamaño del nodo es menor que el especificado por el usuario, el nodo no será dividido.
- Sí el nodo resultado es hijo de un nodo cuya talla es menor que la especificada por el usuario, el nodo no será dividido.

4. Poda

Un árbol mal configurado o sin criterios de parada adicionales, puede llegar a sobre ajustarse a los datos de entrada, por lo que algunos autores usan técnicas de podado de árboles, sin embargo estas no son consideradas en el presente trabajo.

9.3.2. Bosque Aleatorio

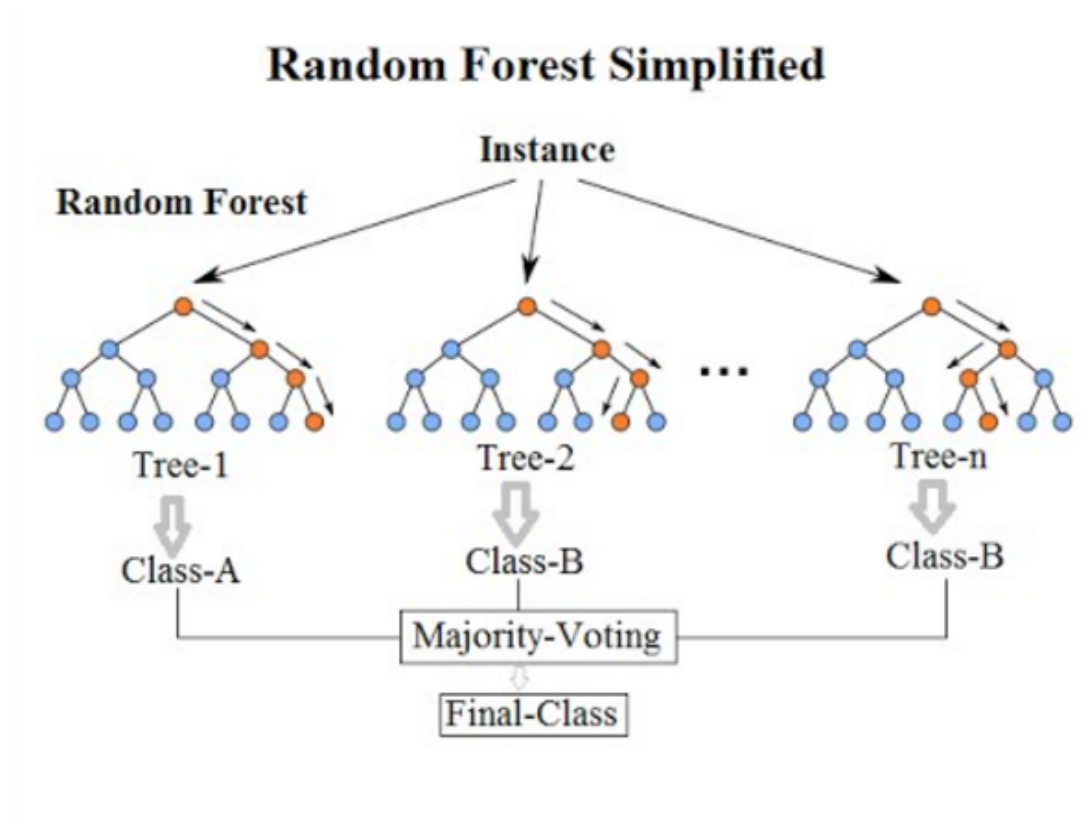


Figura 4: Representación del algoritmo Random Forest

Los árboles de decisión por su diseño mismo tienden a sobre entrenarse sobre los datos de entrenamiento, este comportamiento también conocido como «memorización» de los modelos.

Para evitar esto, una propuesta interesante que mejora a los CART, es el bosque aleatorio o *Random Forest* este algoritmo introduce dos cambios importantes para evitar el sobre entrenamiento.

El primero y más importante es el Impulso o *Boosting* se basa en la idea de que la combinación de muchos modelos débiles puede dar como resultado un modelo fuerte. Estos modelos débiles son de hecho CART individuales.

Ahora bien, para evitar que estos modelos débiles sean muy similares entre sí, se introduce el otro gran cambio y este es el Embolsado o *Bagging* consiste en entrenar cada uno de los modelos débiles sobre un sub-conjunto único de los datos de entrenamiento.

La generación de un conjunto de árboles entrenados en sub-conjuntos de datos diferentes y que las decisiones se realizan en base a el voto de la mayoría, es lo que consiste el algoritmo de bosque aleatorio.

9.3.3. Aumento Extremo del Gradiente

El algoritmo de Bosque Aleatorio es un algoritmo famoso y ampliamente utilizado en distintos contextos, no solo a nivel experimental, sino que también en producción, ejemplos de esto son la cadena de clasificación de imágenes de satélite Iota-2 [21], sin embargo otro algoritmo basado en CART ha obtenido resultados interesantes.

A diferencia del modelo de Bosque Aleatorio, el algoritmo Aumento Extremo del Gradiente, variación de las *Gradient Boosting Machine* en este algoritmo el impulso o *Boosting* se usa de una manera diferente, se genera un conjunto usualmente pequeño de árboles de decisión y estos son mejorados iterativa mente usando una función de pérdida con regularización descrito matemáticamente así: [22]

Se un conjunto de datos de n ejemplos y m numero de predictores, un conjunto de CART se define como los k -ésimos

$$\hat{y} = \phi(x_i) = \sum_{k=1}^k f_k(x_i), f_k \in \mathcal{F} \quad (29)$$

donde $\mathcal{F}$ es el espacio de árboles de regresión conocido también como CART. Aquí cada f_k representa un CART distinto e independiente [22].

Estos árboles son optimizados minimizando la siguiente función objetivo:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (30)$$

donde :

$$\Omega(f) = +\frac{1}{2}\lambda||w||^2 \quad (31)$$

Donde l es la función de perdida convexa diferenciable entre la predicción $\hat{y}_i$ y el valor real y_i . El segundo termino Ω penaliza la complejidad del modelo² [22].

9.3.4. Redes Neuronales Artificiales

Son una familia modelos de inteligencia artificial que intenta reproducir el comportamiento del cerebro, principalmente la relación entre la neuronas y sus conexiones. Modelos se constituyen por un grupo de elementos conocidos como neuronas que trabajan conjuntamente. Cada una de estas partes de la red recibe información y esta es enviada por medio de interconexiones hacia otras neuronas. Esta concepción es posible utilizar distintas configuraciones para realizar clasificaciones, regresiones, agrupaciones, entre otros.

Perceptrón

La forma más simple y la base de las redes neuronales es el perceptrón, un modelo matemático que trata de reproducir el comportamiento individual de una neurona. Se define matemáticamente según la ecuación ??

$$f(x) = \text{sign} \left[\sum_{i=1}^n w_j \cdot x_{i \cdot j} + B \right] \quad (32)$$

Esta ecuación esta constituida por una composición de funciones, la primera es una función lineal constituida por un los pesos w_j y el sesgo B y por una función de activación sign que aplica una transformación afín a la función anterior.

Red Neuronal

$$f(x) = W_0 + k \cdot \sum_{i=1}^n w_i \cdot x_i \quad (33)$$

Donde w_0 es el vector de B , w_i es el vector de pesos, x_i la matriz de variables independientes y k la transformación no lineal.

²Para más información sobre Extrem Gradient Boostings consulte : [22]

Propagación hacia atrás

$$\nabla w_{ij}(n) = \eta \cdot \delta \cdot J \cdot x_{ij} \cdot + \alpha \cdot \nabla w_{ij}(n-1) \quad (34)$$

9.3.5. Validación Cruzada

La validación cruzada también conocida como *K-Fold Cross Validation*³, es una técnica desarrollada para probar o testear la estabilidad y capacidad de los modelos entrenados sobre conjuntos con pocos datos, la idea general es dividir el conjunto completo de datos en paquetes más pequeños conocidos como *folds*, que a su vez se subdividen, en dos conjuntos, uno para entrenamiento y otro para prueba, de manera que los algoritmos a evaluar serán entrenados de manera independiente en cada uno de los subconjuntos de entrenamiento y se testearán sobre los conjuntos de prueba.

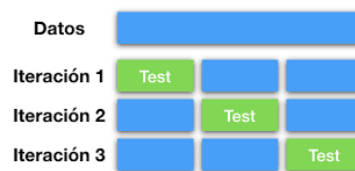


Figura 5: Representación gráfica Validación cruzada

LLTO CV

El **LLTO CV** o *Leave Location and Time Out - CV* es un caso especial de la validación cruzada, propuesto inicialmente por [MEYER20181] para enfrentar una apropiada validación en modelos espacio-temporales, teniendo en cuenta las propiedades de auto-correlación espacial y auto-correlación temporal, donde su combinación es conocida como

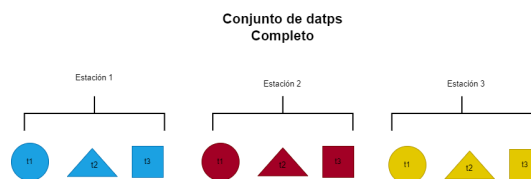


Figura 6: Validación cruzada en estaciones meteorológicas

La idea detrás del uso de la estrategia de validación cruzada **LLTO** es dada la limitada cobertura espacial de las estaciones y la imposibilidad de tener datos independientes para la

³ CV: Cross Validation

evaluación, es subdividir los datos de manera, que se incite a los modelos entrenados a generalizar y capturar la auto-correlación tanto espacial como temporal[23].

Como ejemplo dadas las estaciones azul, roja y amarilla, para las fechas circulo, cuadrado y triangulo, definidas en la figura 6 ⁴ la estrategia **LLTO CV** divide los datos con los criterios espaciales y temporales mencionados, obteniendo el resultado visualizado en la Figura 7 ⁴.

LLTO-CV	subconjunto 1	 			 	
	subconjunto 2	 				  
	subconjunto 3	 	 			

Figura 7: Estrategia de validación cruzada espacio-temporal

⁴Figuras tomadas de: [MEYER20181]

Parte VII

Marco de Referencia

La gestión de los riesgos de sequía tiene por objetivo mejorar la capacidad de la sociedad para hacer frente a eses fenómeno, donde la vigilancia y alerta temprana de la sequía son dos componentes importantes [9].

El monitoreo de la sequía utilizando índices espectrales puede contribuir a diversos procesos de toma de decisiones y al desarrollo de sistemás de alerta temprana de sequía [24], impactando positivamente las zonas afectadas por este fenómeno. Por este motivo, diferentes autores han abordado esta temática desde perspectivas diferentes.

Como por definición, la sequía es una disminución temporal y espacial de la precipitación pues lógicamente los estudios de sequía más clásicos se centran en este parámetro, valiéndose inicialmente de estaciones meteorológicas, pasado un tiempo se han integrado otro tipo de variables para el estudio de este fenómeno, que esta estrechamente relacionado con un cambio en la concepción de la sequía.

Un caso de este tipo de aproximación, que se puede denominar clásica es la descrita por [25], donde se evalúa la adaptabilidad que puede tener los agricultores de una cuenca en la región de Dagua, Colombia, por medio del estudio de las estaciones meteorológicas de la zona, evaluando la posición y continuidad temporal de los datos que proveen estas estaciones para realizar su análisis [25].

Por su parte en [11] se hace un recuento de las diferentes técnicas tanto clásicas como contemporáneas para la evaluación de sequías, resaltando a su vez las técnicas basadas en el análisis estadístico mediante índices de series históricas de estaciones de campo, técnicas e índices sustentados en la teledetección, y la poderosa combinación que puede resultar de ambas [11].

Sin embargo la complejidad de este fenómeno ha requerido la combinación de diferentes variables para su caracterización, es el caso de [26] se evaluó la cantidad y la distribución espacial de la precipitación, y a partir de esto se determinó los factores que el fenómeno de la sequía [26]. Entre los factores evaluados estan:

La precipitación, la temperatura, la altitud, la vegetación y la humedad del suelo. Aplicando por su parte, *Precipitation Anomaly Percentage (Pa)*, y análisis de correlación [26].

Ahora bien, los autores que a continuación se describen, utilizaron unas aproximaciones más contemporáneas, entre las que resalta la implementación de técnicas de inteligencia artificial (I.A.) aplicadas a datos obtenidos con sensoramiento remoto de alta frecuencia temporal, con el propósito de para mejorar la resolución espacial de conjuntos de datos que son muy precisos en cuanto a su resolución temporal, y de este modo hacer un control y monitoreó más efectivo.

Es el caso de [13], centraron sus esfuerzos en la determinación de la humedad del suelo como indicador de la presencia o ausencia de la sequía, bajo el supuesto de que la humedad del suelo es un factor clave en el monitoreo de la sequía pues se relaciona con la precipitación, evapotranspiración y rendimiento de cultivos [13]

Para esto se valieron de datos del sensor AMSR-E a bordo del satélite AQUA, además de productos de MODIS, como el MOD11A2, MOD16A2, MCD43B3, MCD15A2 MYD13A2, que constituyen temperatura superficial, evapotranspiración, albedo, índice de área foliar. NDVI, y EVI. Por otro lado, para la precipitación usaron datos de la misión TRMM, y datos *in situ* para la validación del modelo de *downscaling* o de desagregación utilizado. La zona de estudio fue la península de Corea. En particular en su flujo de trabajo resalta la implementación del modelo de *Machine Learning* Bosque Aleatorio con el fin de realizar la reducción de la escala de las variables utilizadas y mejorar la resolución espacial de los datos sin perder la resolución temporal, el coeficiente de correlación R^2 ascendió a entre 0.93 y 0.95 con un RMSE asociado de entre 0.32 y 0.035.

Por otro lado [27], los autores pretende utilizar datos de teledetección y predicción de un largo registro temporal para implementar un modelo de aprendizaje de maquina para las áreas no evaluadas anteriormente en su zona de estudio. Afirmando que un monitoreo de la sequía en tiempo real o casi real se constituye en una herramienta de gran valor para los sistemas de alerta temprana de sequía, y a partir de estos potenciar la toma de decisiones apropiadas que en ultima instancia disminuya los daños que ocasionan las sequías[27].

Por lo que es importante evaluar las metodologías para realizar los pronósticos de sequía que incluyen modelos de regresión, modelos de series de tiempo, redes neuronales artificiales, y modelos híbridos; estos métodos se aplican con el fin de determinar aspectos de los pronósticos como la estimación del inicio y fin del fenómeno, la severidad de este, su probabilidad de ocurrencia, entre otros.

Para esto [27], centra su estudio en el uso del SPI y el SPEI (índice estandarizado de precipitación y evapotranspiración), apuntando sus esfuerzos en determinar sequía meteorológica. Evaluando cuatro modelos de aprendizaje de máquina para identificar cuál obtiene los mejores resultados, estos tres son árboles de decisión (CART), Bosque Aleatorio (Random Forest) y árboles extremadamente aleatorios, comparando los anteriores con una interpolación espacial tipo Kriging.

Entre los resultados destaca un mejor comportamiento de los modelos de aprendizaje de máquina tipo regresión sobre los modelos de clasificación, centrando sus resultados en analizar los primeros. Donde los árboles extremadamente aleatorios obtuvieron los resultados más relevantes.

En un sentido similar se dirigió [13], productos de precipitación TRMM aplicaron tres enfoques de aprendizaje automático basados en árboles; Bosque Aleatorio (Random Forest), Árboles de Regresión Potenciados, y Cubista; modelos que han demostrado ser robustos y flexibles para muchas aplicaciones de regresión, usando estos para evaluar la relación entre los factores de sequía y las condiciones de sequía.

Con base en esto se aplicó el SPI, NDVI, y LST obteniendo que el NDVI era la variable más importante en las regiones húmedas y el LST en las regiones áridas al momento de caracterizar la sequía[24].

Otros estudios realizan un análisis de los comportamientos globales de la sequía, y las perspectivas a futuro del fenómeno, teniendo en cuenta las proyecciones climáticas sustentadas en el cambio climático [1] . Sin embargo, también aplican el algoritmo de Random Forest en su estudio para determinar la importancia de los diferentes predictores; concluyendo que a raíz del cambio climático, es altamente probable que los tiempos de recuperación después de un evento de sequía sean insuficientes para los ecosistemas, aumentando la exposición y vulnerabilidad de estos mismos ante los fenómenos de sequía[1].

Parte VIII

Metodología

10 Zona de Estudio

El departamento de La Guajira, es la zona continental más septentrional de Colombia, limita al oriente con Venezuela, al sur con los departamentos del Cesar, y el Magdalena, al occidente y al norte con el Mar Caribe como se puede detallar en la figura 8.

Geográficamente el departamento está constituido por la península de La Guajira, una planicie en su centro, y partes de la sierra Nevada de Santa Marta y de la Serranía del Perijá, extendiéndose en una superficie de 20 848 (km²). La presencia de estas formaciones montañosas, hace que en este departamento se encuentren todos los pisos térmicos de la zona intertropical donde el ecosistema predominante son las zonas desérticas y semidesértica.

Por último en cuanto a recursos hídricos sus principales fuentes son el río Ranchería y el río Cesar, además del Jerez, Ancho y Palomino; además de arroyos, pozos acuíferos o jagüeyes [7]; sin embargo esta hidrografía presenta una muy baja regulación hídrica (es decir muchos caudales son temporales) lo que limita la recuperación que puede llegar a tener tanto los ecosistemas como las personas a lo fenómenos de sequía[7][8].



Figura 8: Zona de estudio del proyecto

11 Conjunto de Datos

El conjunto de datos utilizado para realizar el presente proyecto, se describe a continuación.

Variables derivadas de datos *in situ*:

Precipitación.

Esta variable climática se deriva de los datos diarios de precipitación acumulada, de la red de estación hidrometeorológicas y climatológicas del IDEAM en la zona, con mediciones entre 1980 y 2012, localizadas en el departamento de La Guajira ¹.

Variables derivadas del sensor MODIS y su resolución espacial:

Tabla 2: Variables derivadas del sensor MODIS

Variable	Resolución Espacial	Resolución Temporal	Referencia
Temperatura superficial (LST)	1000 (m)	8 días	[28]
Índice de vegetación Normalizado (NDVI)	250 (m)	16 días	[29]
Índice de vegetación Mejorado (EVI)	250 (m)	16 días	[29]
Índice de área foliar (LAI)	500 (m)	8 días	[30]
Evapotranspiración	500 (m)	8 días	[31]

Para las anteriores variables se utilizaron los datos entre los años 2000 y 2014, para las variables de 8 días se procesaron 192 raster y para las de 16 días 96 rasters.

Además, esta información fue descargada del portal ftp de distribución de datos *e4ftl01*² del

¹<http://dhime.ideam.gov.co/atencionciudadano/>

²<https://e4ftl01.cr.usgs.gov/>

servicio geológico de los estados unidos conocido por sus siglas en inglés USGS.

Variables derivadas del productos CHIRPS :

Precipitación. (5000 m)

Este producto satelital, tiene una resolución diaria, y es generado por el *Group Infrared Precipitation with station data* de la universidad de California - Santa Monica ³.

³<https://www.chc.ucsb.edu/data/chirps>

12 Métodos

Como se puede ver en la figura 9 donde se puede observar el flujo de trabajo diseñado.

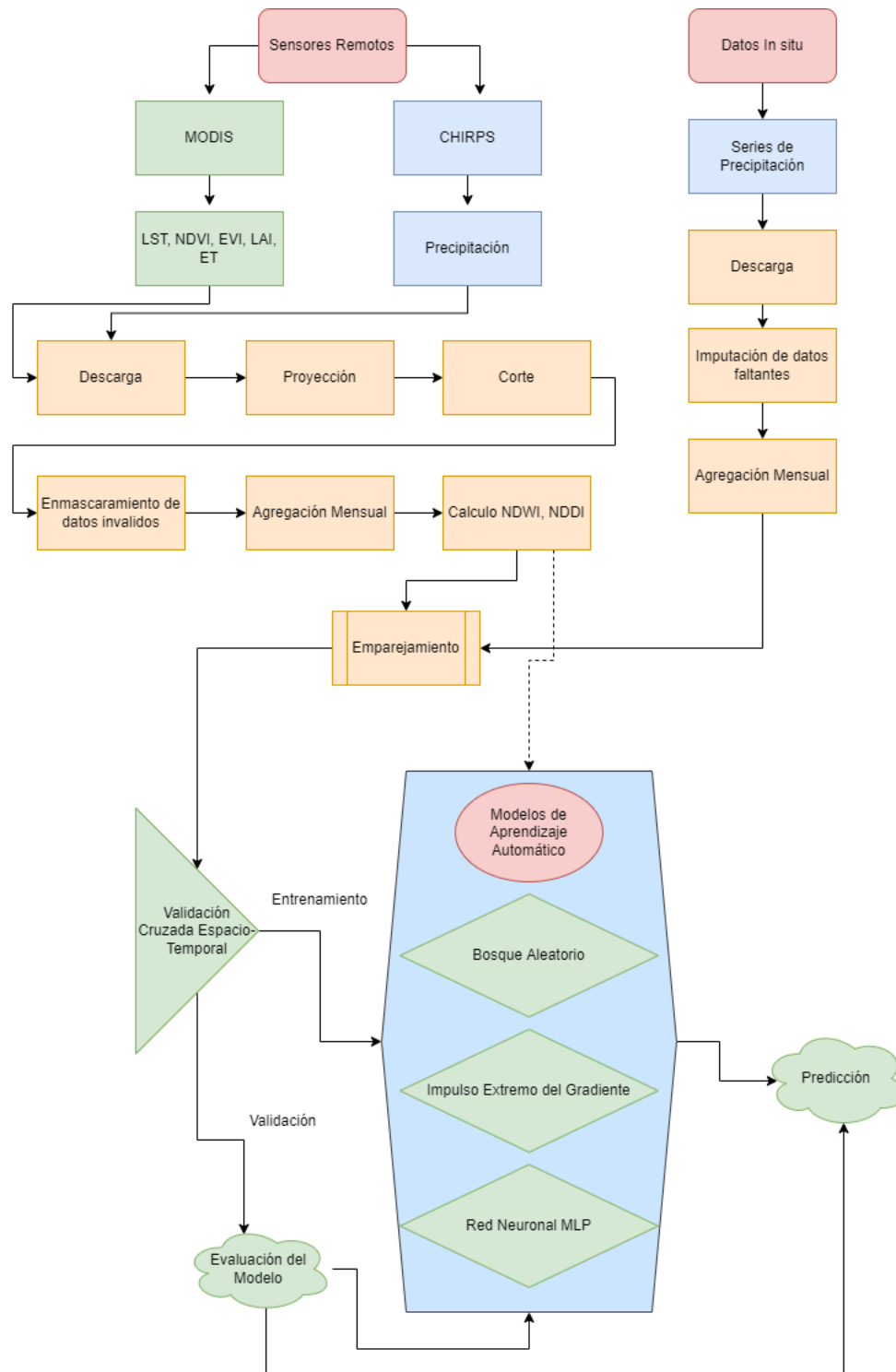


Figura 9: Flujograma General del Proyecto

12.1. Pre-procesamiento de datos *in situ*

Selección de estaciones

En la zona de estudio se encuentra 160 estaciones meteorológicas o hidrológicas con mediciones de precipitación diaria. A paso seguido se aplicó un doble filtro de calidad sobre la información, el primero fue la selección de aquellas estaciones con registros faltantes o nulos no mayores al 10 % de todos los registros, y el segundo una continuidad de los registros no menos a 32 años, entre 1980 y 2012.

Dado que el SPI requiere que la entrada de datos tenga el registro completo de la serie de tiempo, se procedió a imputar los valores faltantes.

Imputación de datos faltantes

Se estudiaron distintos métodos de imputación de datos, se seleccionaron los siguientes:

- *Last observation carried forward*
- Suavizado de Kalman
- *Seasonally Decomposed Missing Value Imputation*
- *Seasonally Splitted Missing Value Imputation*

De los anteriores se seleccionó el método de *Seasonally Decomposed Missing Value Imputation*, pues obtuvo los errores de imputación más bajos sobre una muestra de 10 estaciones con alrededor del 10 % de datos faltantes.

De manera práctica se utilizó la librería de R *ImputeTS* para realizar este procesamiento.

Una vez obtenidas las series completas, se procedió a acumular los datos de precipitación a escala mensual.

Cálculo SPI no paramétrico

Seguidamente se realizó la estimación del SPI no paramétrico para las ventanas temporales de 3, 6 y 9 meses. Para esto se utilizó una biblioteca originalmente diseñada para *MATLAB* llamada *Standardized Drought Analysis Toolbox (SDAT)*.

Se implementaron unas pequeñas modificaciones sobre la biblioteca, la primera es que se ejecuto sobre un ambiente de software libre llamado *Octave* para asegurar que se cumplieran los postulados de la ciencia reproducible ¹. Y el segundo que se realizo una mínima modificación sobre el código fuente para que la biblioteca calculara el SPI no paramétrico sobre todas las estaciones del estudio.

12.2. Pre-procesamiento de datos raster

En esta sección de la metodología se constituye de todos los pasos necesarios para transformar los datos descargados de los servidores de las agencias espaciales y de las Unidades de Investigación de distintas universidades del mundo, transformándolos para ser consumidos por los modelos seleccionados para realizar la fusión.

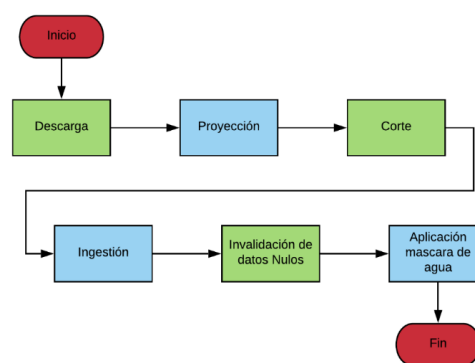


Figura 10: Flujograma de preprocesamiento raster

Descarga de Productos

El primer paso consiste en la descarga automática de los productos raster utilizados, para esta descarga se utilizaron dos estrategias, según el proveedor de datos.

El primer proveedor es la *Universidad de California Santa Bárbara*, en la página web de su *Centro de Riesgos Climáticos*, que mediante el uso de la herramienta de línea de comandos *WGET* se creo una pequeña rutina que pudiese descargar los datos diarios de precipitación para el periodo de tiempo de 1980 - 2016.

¹Ver sección 18

En segundo lugar y por medio de la utilidad de línea de comandos de la librería de *Python* *PyModis* se ejecuto, la descarga de los productos derivados del sensor MODIS.

Proyección de datos

El siguiente paso de la metodología consiste en la asignación de una proyección cartográfica a los datos MODIS, pues el formato nativo de esta es *.hdf carece de dicho sistema de referencia. Así que se procedió a utilizar el módulo *modis_mosaic.py* para construir el mosaico virtual de cada uno de los productos creándose un archivo en formato *.vrt. A continuación se realiza la conversión la información a formato *.TIF haciendo uso del módulo *modis_convert.py*.

Recorte

Luego se realiza el recorte de la información raster a la zona de estudio, dibujada a mano alzada, de manera que incluye la totalidad del Departamento de La Guajira, para este procesamiento se utilizó el modulo *gdal_warp* de la librería GDAL (Librería de Abstracción de Datos Geográficos). La figura muestra el resultado del recorte de las imágenes a la zona de estudio.

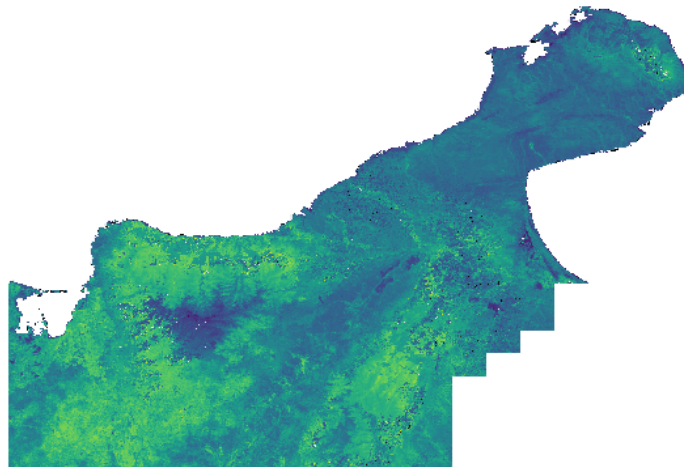


Figura 11: Recorte a la zona de estudio

Ingestión en GRASS GIS

A paso seguido, se realizó la ingestión de la información en el software GRASS GIS, de manera que se aprovechará todas las potencialidades de este software que incluyen, pero no se limitan al soporte de un marco para series de tiempo de datos raster y vectoriales, manejo eficiente de la memoria RAM, orden en el conjunto de datos y finalmente un almacenamiento eficiente en disco duro.

Además, se escalan los valores de los productos, para que estos queden almacenados en sus valores reales físicos, para esto se usa la información interna de los productos para enmascarar los píxeles no válidos.

Aplicación Máscara de agua

Dado que la zona de estudio se encuentra a orillas del Mar Caribe, es necesario tener en cuenta que la zona de modelación está limitada por el contorno de la superficie de agua, por lo tanto se utiliza el producto de máscara de agua de MODIS, para invalidar los valores pertenecientes a las superficies de agua.

Homogeneización de la serie raster

Acto seguido y teniendo en cuenta que los diferentes productos tienen resoluciones temporales diferentes es necesario homogeneizar las series, dado que el indicador objetivo el SPI no paramétrico se estima a escala mensual, es necesario agrupar la información de los productos a esta escala.

Según la procedencia del producto se procede de distintas maneras. Por un lado están el producto de precipitación y por el otro tenemos los productos MODIS que se distribuyen en dos resoluciones temporales, los productos de 8 días y los de 16 días.

- CHIRPS: Para la precipitación los datos de CHIRPS pueden ser descargados a distintas escalas, en particular para este estudio se descargo el producto diario de precipitación y este se acumulo a escala mensual.
- MODIS 8 días : Estos productos que incluyen ET, LAI, LST. Estos se promediaron a escala mensual.
- MODIS 16 días : El producto distribuido con esta resolución temporal es el MOD13, el producto de Índices de Vegetación, del que se obtiene el NDVI, el EVI y las bandas infrarroja cercana e infrarroja media. A estas series se imputo sobre la serie el valor promedio en el tiempo, para obtener una serie de 8 días, posteriormente se promediaron los valores para obtener la serie mensual.

Cálculo NDWI y NDDI

Adicional a lo anterior, se generaron dos variables explicativas adicionales de los productos descargados, estos se describen a continuación.

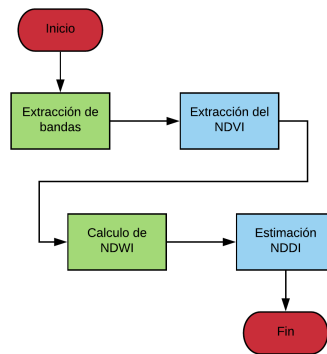


Figura 12: Cálculo NDDI

Se extraen las bandas de infrarrojo cercano e infrarrojo medio del producto MOD13 y se aplica la ecuación 25.

A continuación, se calcula el NDDI con la ecuación (26), a través de los productos de NDVI (23) y NDWI (25).

Relleno de datos faltantes

Dada la persistente nubosidad de la zona de estudio, y a pesar del uso de ventanas temporales de 8 días para minimizar este efecto, aún se pueden encontrar amplias zonas con píxeles no válidos, debido en particular a la presencia de nubes.

Se aplicó un método de relleno de datos faltantes para series de tiempo raster homogéneos en sus resoluciones espaciales y temporales, basado en una modificación del análisis de componentes principales e implementado mediante el software R, en particular el paquete o librería *sinker* donde se encuentra implementado el algoritmo *DINEOF* para su formalización teórica ver la sección .

Una vez surtido lo anterior se da por finalizado el pre-procesamiento raster.

12.3. Procesamiento

Una vez los datos de las fuentes raster e *in situ* están homogenizadas a una muestreo temporal común, y dadas las diferencias en la ventana temporal de disponibilidad de datos para ambas fuentes de datos se selecciona la intersección de ambas para generar un conjunto el conjunto de datos de entrenamiento.

Esto es tomar los datos *in situ* con una ventana temporal entre 1982 - 2014, los datos raster con una ventana temporal de 2000 - 2012, de modo que el periodo de predicción comprende entre 2000 y 2012.

Una vez se completa lo anterior, se realiza la extracción de valores de las diferentes variables explicativas en la localización de las estaciones. Y se construye lo que se conoce como una **Matriz de entrenamiento**; matriz que sera usada para ajustar los algoritmos seleccionados.

12.4. Entrenamiento y ajuste de hiperparamétros

Una vez constituida la matriz de entrenamiento se procede a inicializar los hiper-parámetros para ajustar los algoritmos seleccionados, estos son:

12.4.1. Random Forest

- Número de estimadores : 100
- Criterio : error medio cuadrático
- Número mínimo de muestras a dividir : 2
- Número mínimo de muestras por hoja : 1
- Mínimo de decrecimiento de la impureza : 0.0

12.4.2. Extrem Gradient Boostings

- Número de estimadores : 10

- Criterio : error medio cuadrático
- Número de columnas de muestra por árbol : 0.3
- Tasa de aprendizaje : 0.1
- Profundidad máxima : 5
- Alfa: 10

12.4.3. MLP

- Arquitectura :
 - Número de entradas: 8
 - Capas ocultas : 1
 - Numero de neuronas en la capa oculta : 11
 - Capa de salida : 1
- Función de Activación : ReLu y linear
- Dropout: 0.05
- Tamaño del batch : 25
- Número de épocas : 1000
- Tasa de aprendizaje : 0.0001

12.5. Validación Cruzada

Una vez inicializados los hiper-parametros, durante el entrenamiento de los algoritmos se aplico la estrategia de **validación cruzada espacio-temporal (LLTO)**. Esta estrategia (introducida en 9.3.5) tiene como objeto evaluar la robustez de los métodos aplicados, para ello se aplico una validación cruzada con 20 subconjuntos de datos, divididos a su vez en 70 % para entrenamiento y 30 % para validación. A cada uno de estos subconjuntos se ajusto los modelos seleccionados de manera independiente, y se evaluaron las métricas de Error Medio Absoluto (MAE) y Error Medio Cuadrático (MSE).

La implementación de esta sección se realizó haciendo uso de la librería CAST implementada en R, adicionalmente se aplicó una interfaz con Python a través de R2py2.

12.6. Importancia de las Variables

Finalmente los modelos basados en árboles, en particular el Random Forest permite evaluar la importancia de las variables predictoras sobre la estimación final, se realizó la extracción de esta información para mejorar la interpretabilidad de los modelos.

Parte IX

Resultados

A continuación se expondrán los resultados parciales y finales de la metodología expuesta en la sección VIII.

13 Pre-procesamiento de datos *in situ*

Primeramente se presenta los resultados del preprocesamiento de los datos *in situ*, de las 162 estaciones de la zona de estudio seleccionada, solo 51 estaciones pasaron los dos filtros, el primero de una suficiente longitud de los registros, y de una incompletitud de los registros menor al 10%.

A partir de estas 51 estaciones se aplico el método de imputación de datos *Seasonally Decomposed Missing Value Imputation*. En la figura 13 se puede visualizar un ejemplo de imputación de datos sobre una estación de muestra, con los valores imputados en rojo.

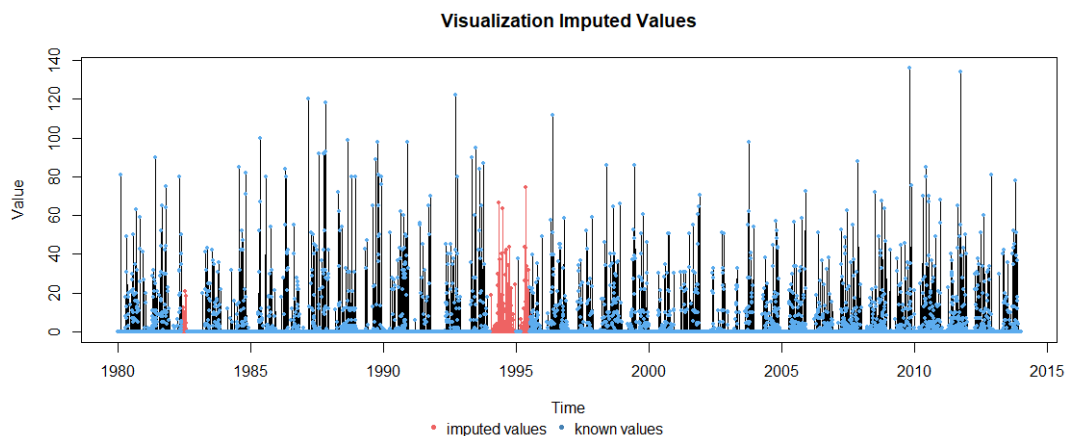


Figura 13: Imputación de datos faltantes en una estación de muestra

Gráfico de una Estación de muestra

Seguidamente se acumula la precipitación en ventanas temporales de un mes y se estima el SPI no paramétrico, podemos visualizar un ejemplo de lo anterior en la 69, y 69.

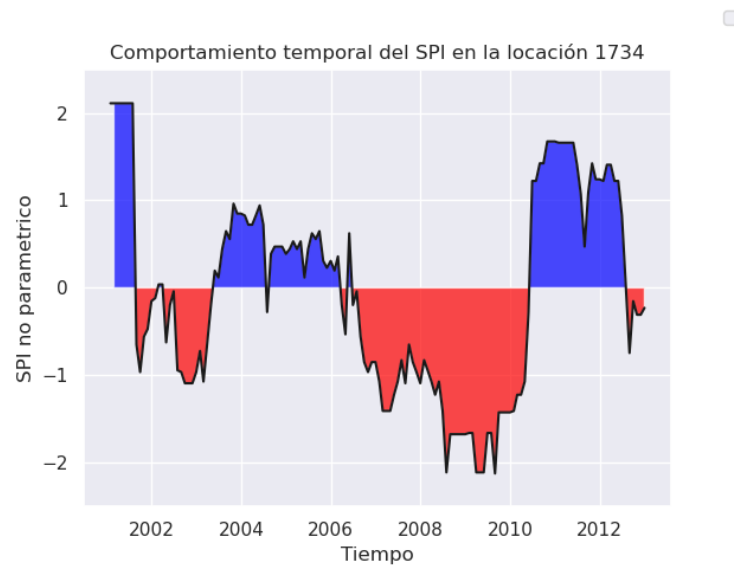


Figura 14: Comportamiento del SPI en una estación de muestra entre el año 2000 y el año 2014.

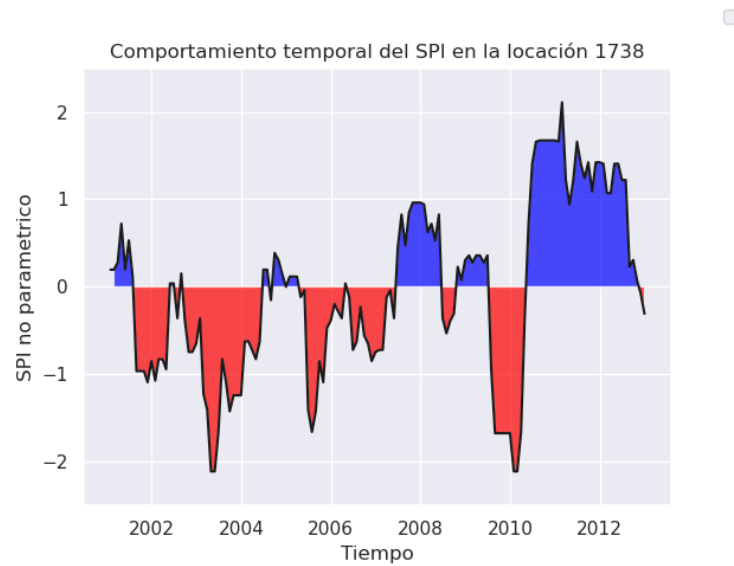


Figura 15: Comportamiento del SPI en una estación de muestra entre el año 2000 y el año 2014.

14 Pre-procesamiento de datos raster

A continuación se presenta un resumen de la fase de pre-procesamiento de los datos raster, estos gráficos corresponden a los valores promedios mensuales a lo largo del periodo de observación de los datos raster, comprendido entre el año 2000 y 2014.

Y en general las variables de estudio el impacto de la composición oreo-gráfica del terreno, con una zona nororiental plana y con condiciones semi-deserticas, una zona sur occidental montañosa conformada por la Sierra Nevada de Santa Marta y la serranía del Perijá al sur-oriental; y finalmente las zonas de transición y la zona transición con alturas moderadas.

14.1. Precipitación

Primeramente la precipitación, esta variable dada su resolución más baja presenta un comportamiento más suave entre las regiones descritas anteriormente, sin embargo es posible diferenciarlas marcadas por sus diferencias en las precipitaciones, con unas precipitaciones más altas sobre la zona de la Sierra Nevada de Santa Marta, seguida por una zona de transición y finalmente una zona semi desértica al norte con unas precipitaciones cercanas a cero a lo largo del periodo de datos de MODIS disponibles.

En el comportamiento en frecuencia se ve claramente la concentración de los valores de precipitaciones nulas y concentradas a inicio del año.

Que se complementa con el gráfico de caja y bigotes donde se observa un comportamiento unimodal en la distribución de la precipitación a lo largo del año, en el promedio de la ventana temporal del año 2000 al 2014.

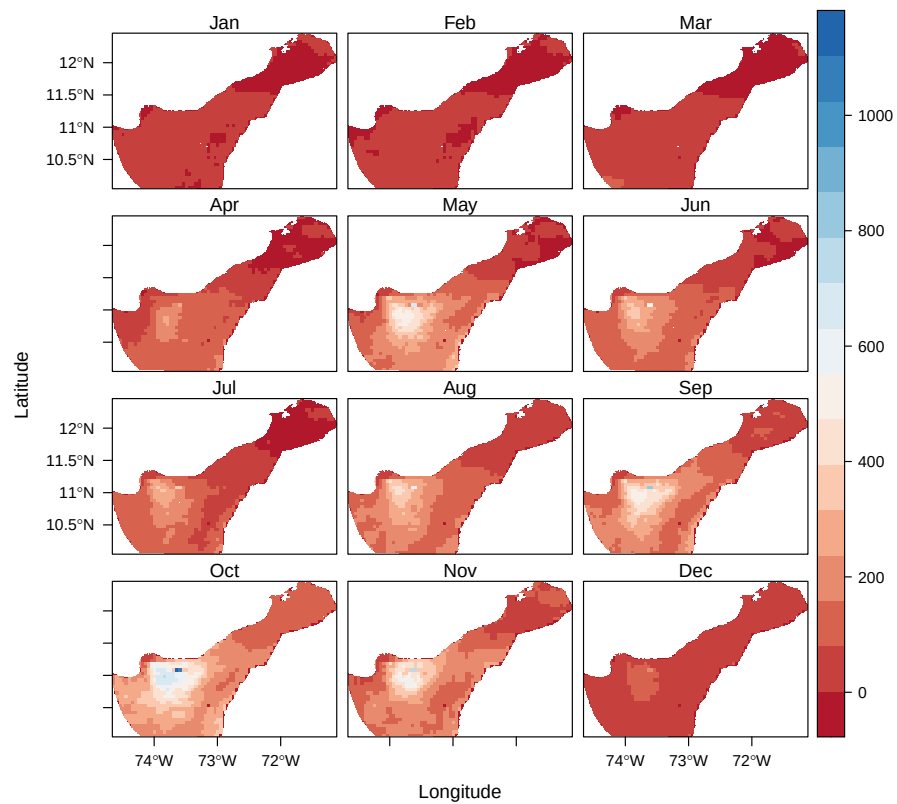


Figura 16: Comportamiento promedio mensual de la precipitación

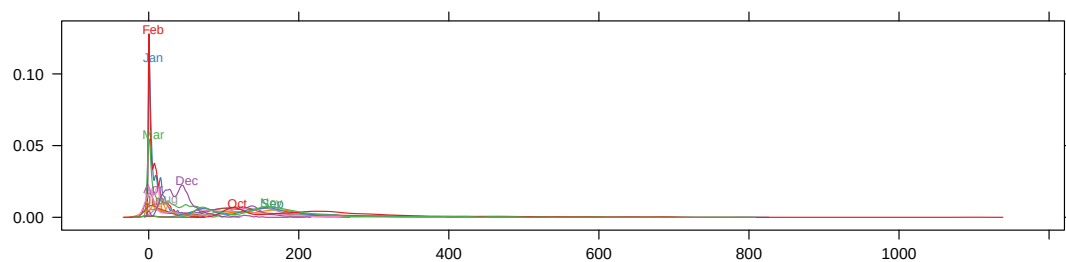


Figura 17: Gráfico de densidad promedio mensual de la precipitación

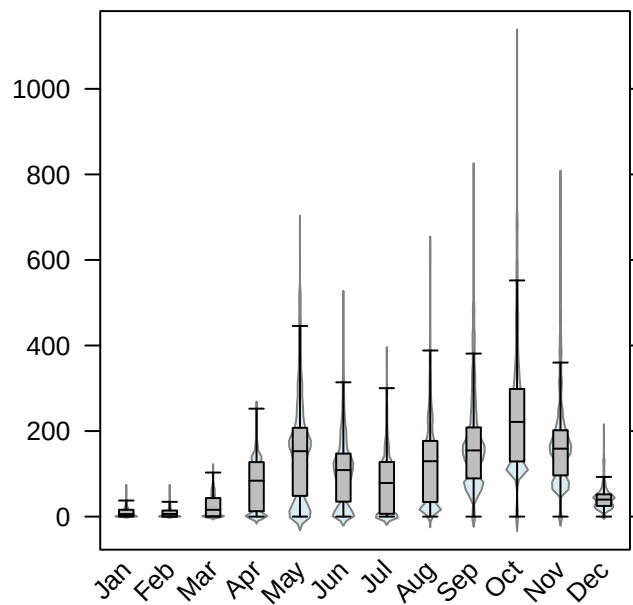


Figura 18: Gráfico de caja-violín promedio mensual de la precipitación

14.2. Evapotranspiración

Seguidamente la evapotranspiración, tiene un comportamiento bastante homogéneo en el espacio, con valores de evapotranspiración ahí donde se espera que haya presencia de vegetación y valores bajos donde esta es escasa.

Esto se corrobora en el gráfico de densidad donde el rango de valores es constante a lo largo del año y una alta variabilidad en sus valores inter-mensuales.

Sin embargo podemos ver como los valores de evapotranspiración son más bajo a inicio del año, en especial en febrero, marzo y abril, siendo más constante el resto del año, esto es esperable dada la disponibilidad de agua a lo largo del año, vista en el régimen de precipitaciones.

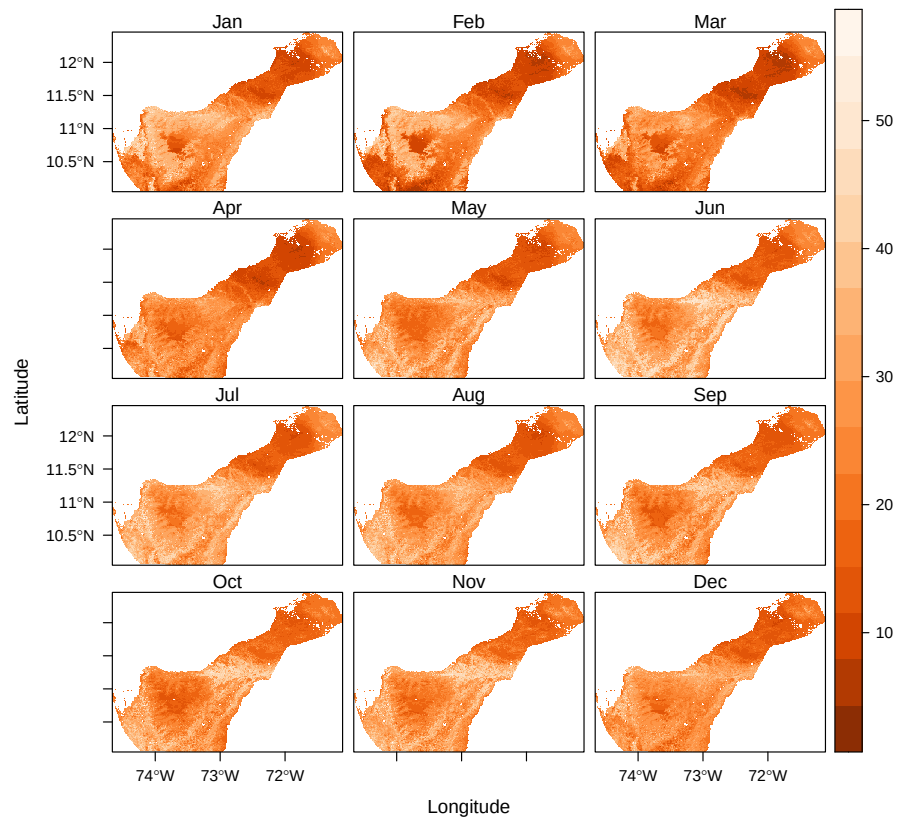


Figura 19: Comportamiento promedio mensual de la evapotranspiración

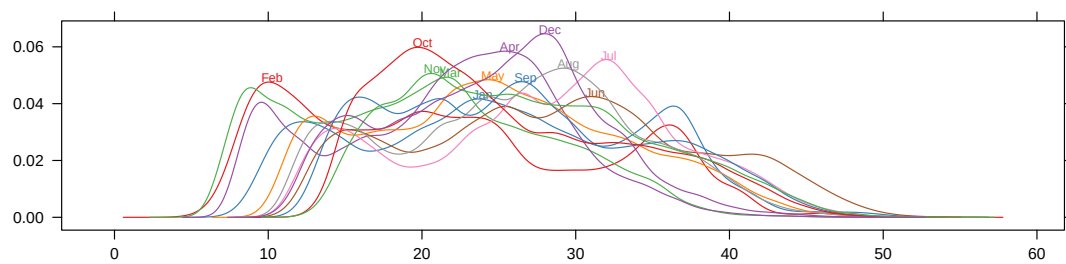


Figura 20: Gráfico de densidad promedio mensual de la evapotranspiración

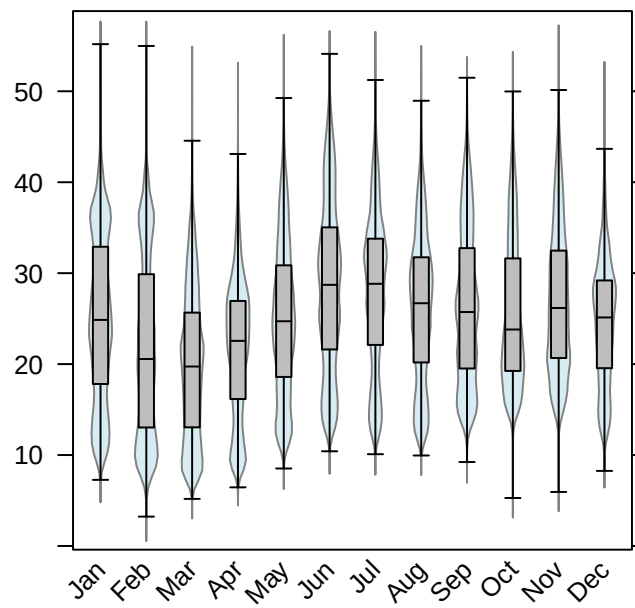


Figura 21: Gráfico de caja-violín promedio mensual de la evapotranspiración

14.3. Temperatura Superficial

Por su parte, la temperatura superficial representa un comportamiento relativamente homogéneo a lo largo del año, marcándose diferencia por oreografía. Es de resaltar las altas temperaturas promedio a las que se ve expuesta esta zona de estudio.

Los valores se distribuyen a lo largo del año sin una tendencia clara.

En el gráfico de cajas y bigotes se observa como la temperatura mantiene valores muy altos a lo largo del año y estos bajan menos de unos pocos grados a final de año, época en la que se concentran las lluvias.

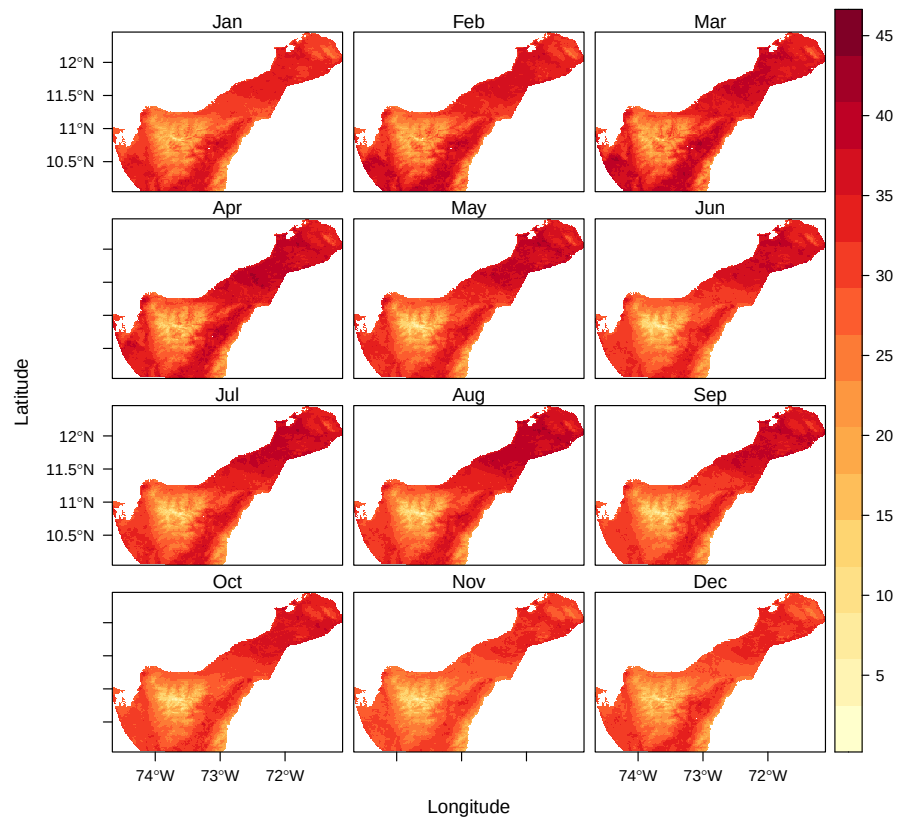


Figura 22: Comportamiento promedio mensual de la temperatura Superficial

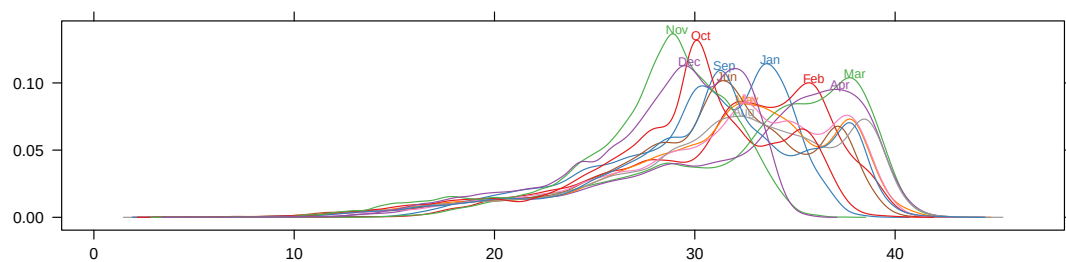


Figura 23: Gráfico de densidad promedio mensual de la temperatura Superficial

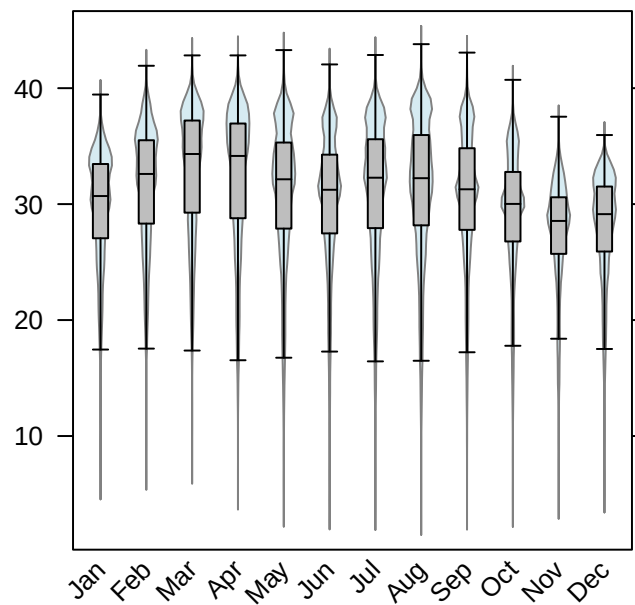


Figura 24: Gráfico de caja-violín promedio mensual de la temperatura Superficial

14.4. NDWI

El comportamiento de la humedad estimada por el NDWI es más alta en las zonas montañosas y de transición, a excepción de las picos más altos dentro de la sierra nevada, por el contrario los valores bajos de las zona semi-desértica, esta en particular aumenta sus valores, y por tanto su humedad hacia el final del año.

En el gráfico de densidad, se visualiza como los meses húmedos concentran los valores más altos del indicador y una ligera disminución al inicio del año en la temporada seca.

Esto se corrobora en el gráfico de cajas y bigotes. Donde se visualiza claramente valores bajos a inicio de año en los meses de febrero, marzo, abril.

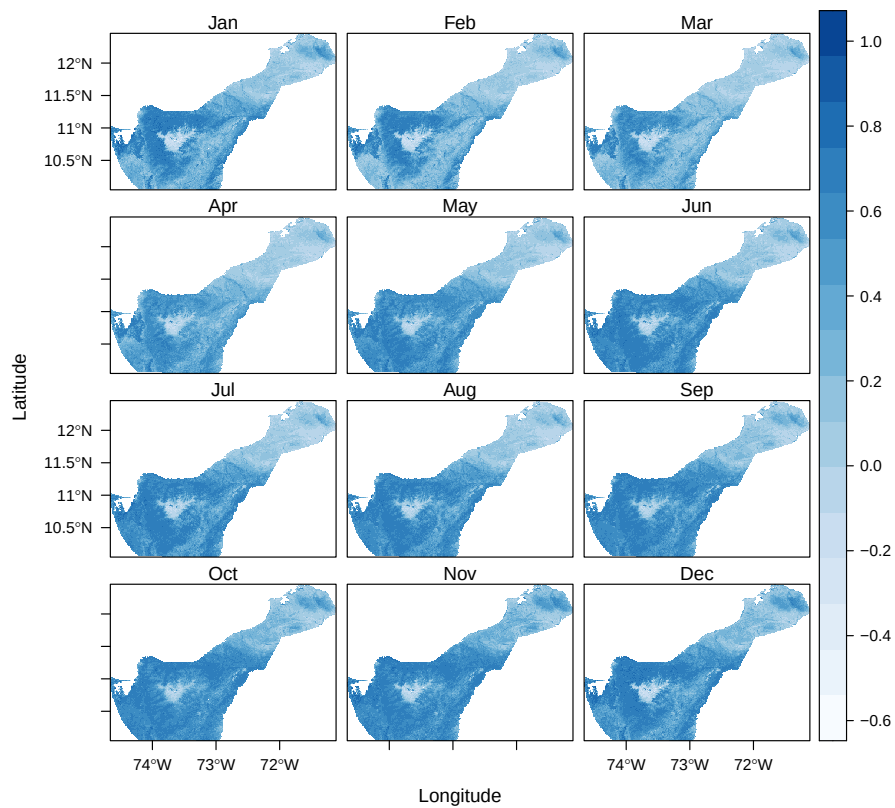


Figura 25: Comportamiento promedio mensual del NDWI

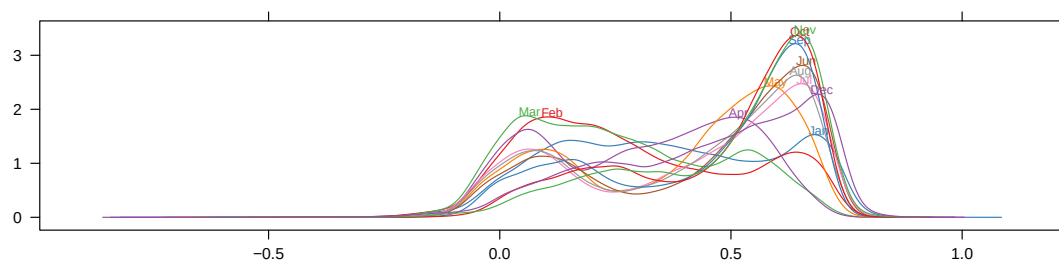


Figura 26: Gráfico de densidad promedio mensual del NDWI

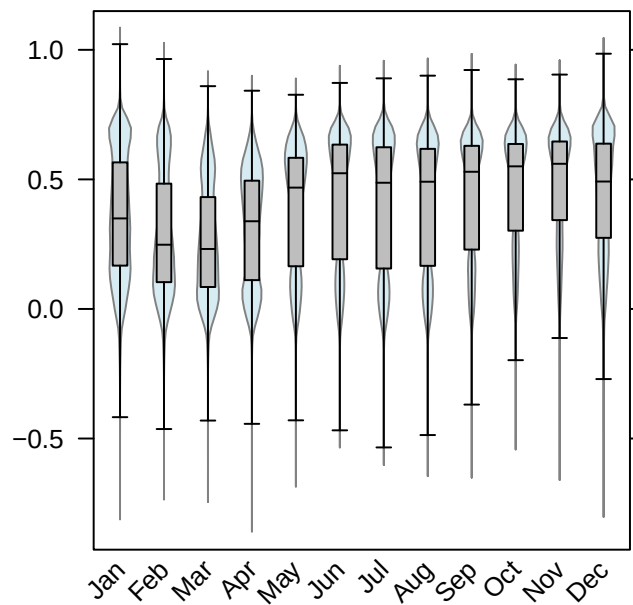


Figura 27: Gráfico de caja-violín promedio mensual del NDWI

14.5. NDVI

El comportamiento del NDVI permite realizar una estimación visual de la distribución de la vegetación. Donde podemos ver como la vegetación se concentra en la zona montañosa, a excepción de los grandes picos del centro de la sierra nevada.

En cuanto a su comportamiento en el gráfico de densidad, se visualiza la concentración de los valores altos del NDVI a final del año, y con valores un poco más bajos concentrados en los meses iniciales del año.

Distribución que se corrobora con el gráfico de cajas y bigotes.

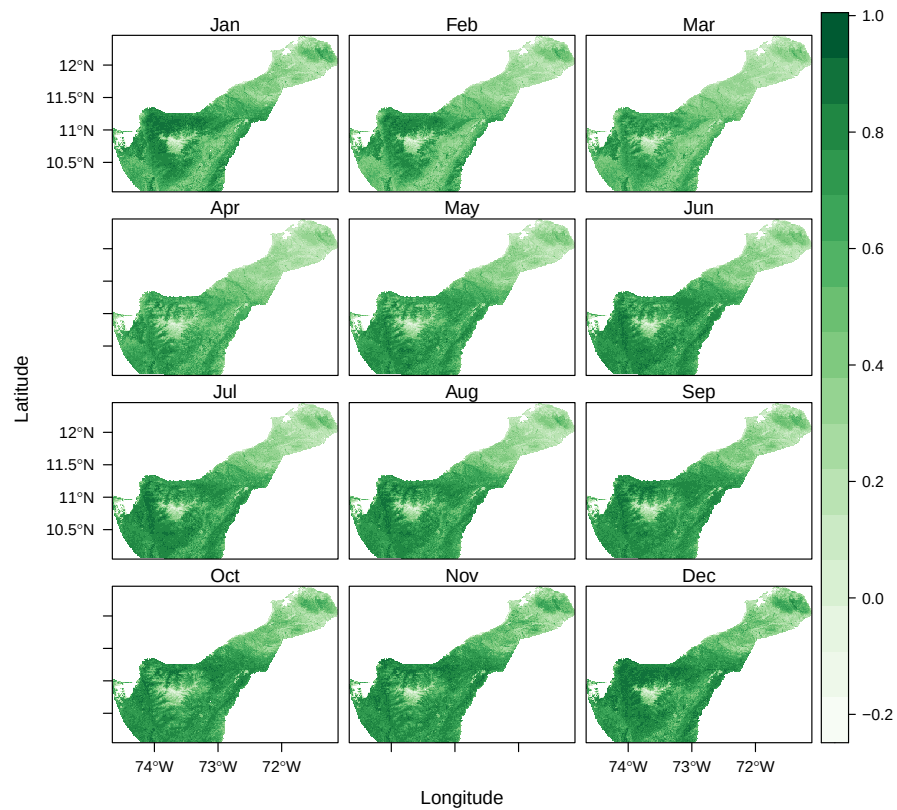


Figura 28: Comportamiento promedio mensual del NDVI

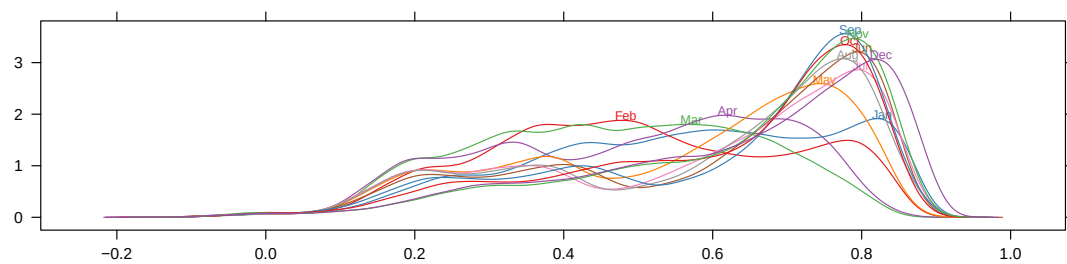


Figura 29: Gráfico de densidad promedio mensual del NDVI

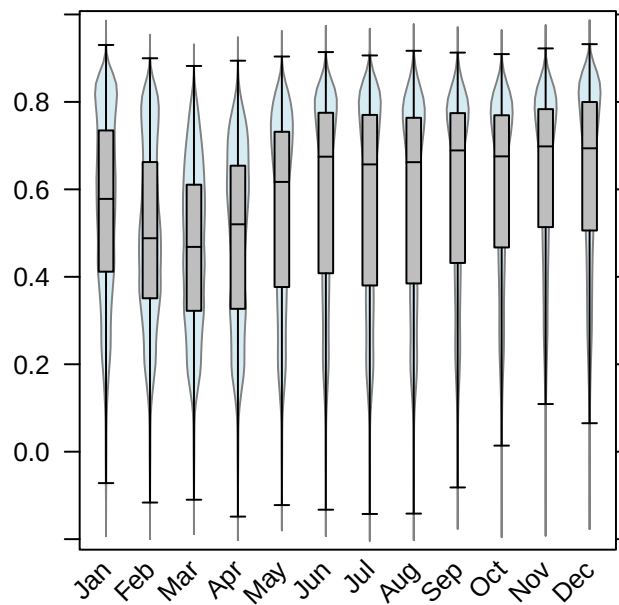


Figura 30: Gráfico de caja-violín promedio mensual del NDVI

14.6. EVI

Los valores del EVI o índice de vegetación mejorada es muy similar al del NDVI, pero dada la corrección atmosférica explicada anteriormente la distribución de los valores es menor.

En cuanto al comportamiento en densidad se ve claramente la distribución de los valores, y una distinción más marcada a lo largo del año con los valores bajo concentrados en el inicio del año y los valores altos al final, en coincidencia con los patrones de precipitación.

Patrón que se repite en el gráfico de caja y bigotes.

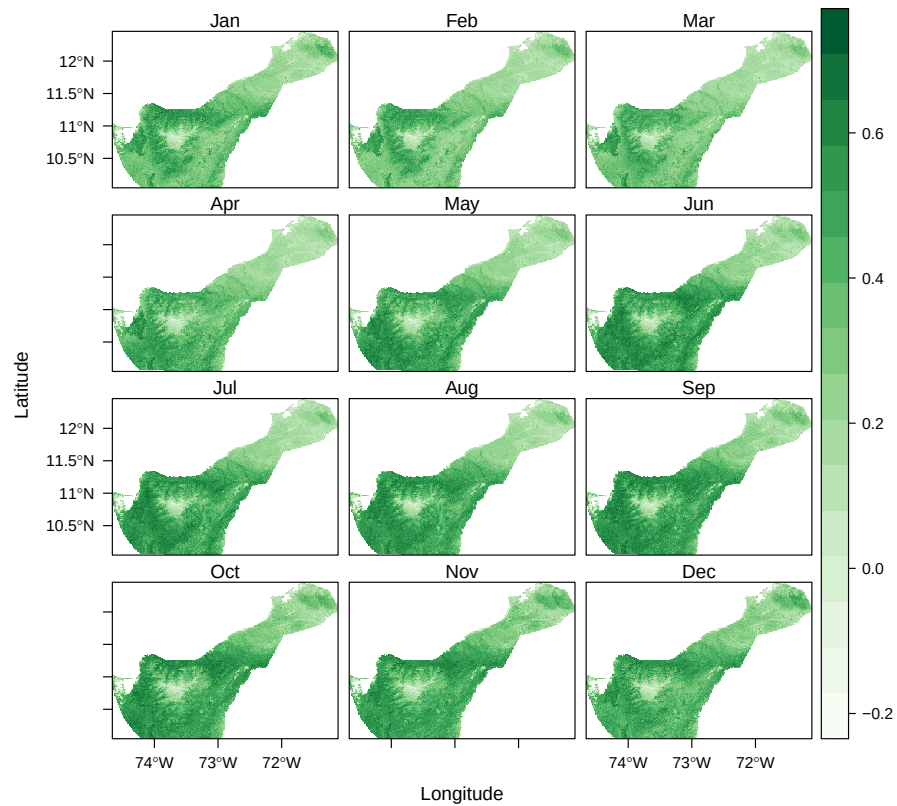


Figura 31: Comportamiento promedio mensual del EVI

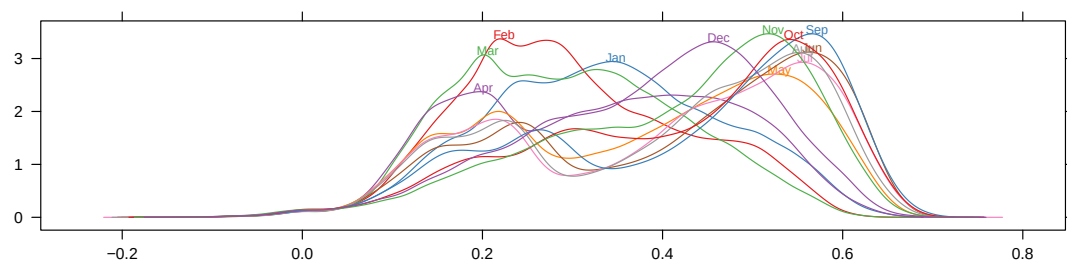


Figura 32: Gráfico de densidad promedio mensual del EVI

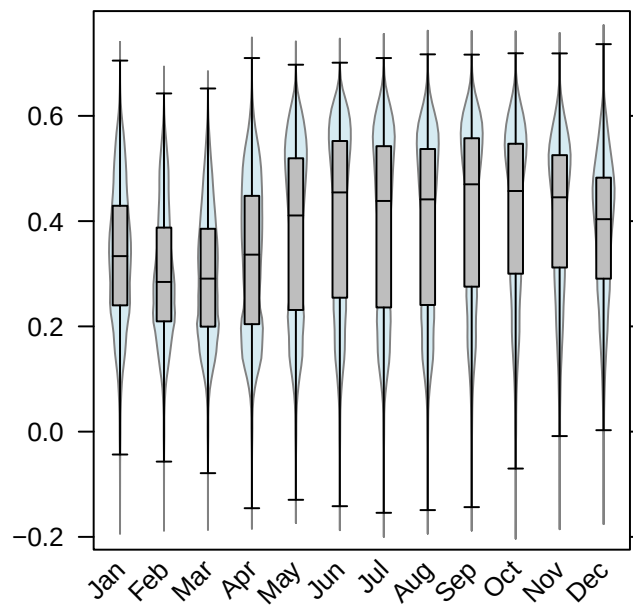


Figura 33: Gráfico de caja-violín promedio mensual del EVI

14.7. LAI

Por su parte la variable LAI o índice de área foliar, es posible identificar muy claramente las zonas ore-gráficas de la zona de estudio. En particular unos valores bajos del índice en la zona seca y en las partes más altas de la zona montañosa y con valores mucho más altos en la estribaciones de la sierra nevada de Santa Marta.

En el gráfico de densidad, se visualiza como los valores bajos se concentran al inicio del año y los valores altos en la parte final del año.

Esta distribución se repite en el gráfico de caja-violín.

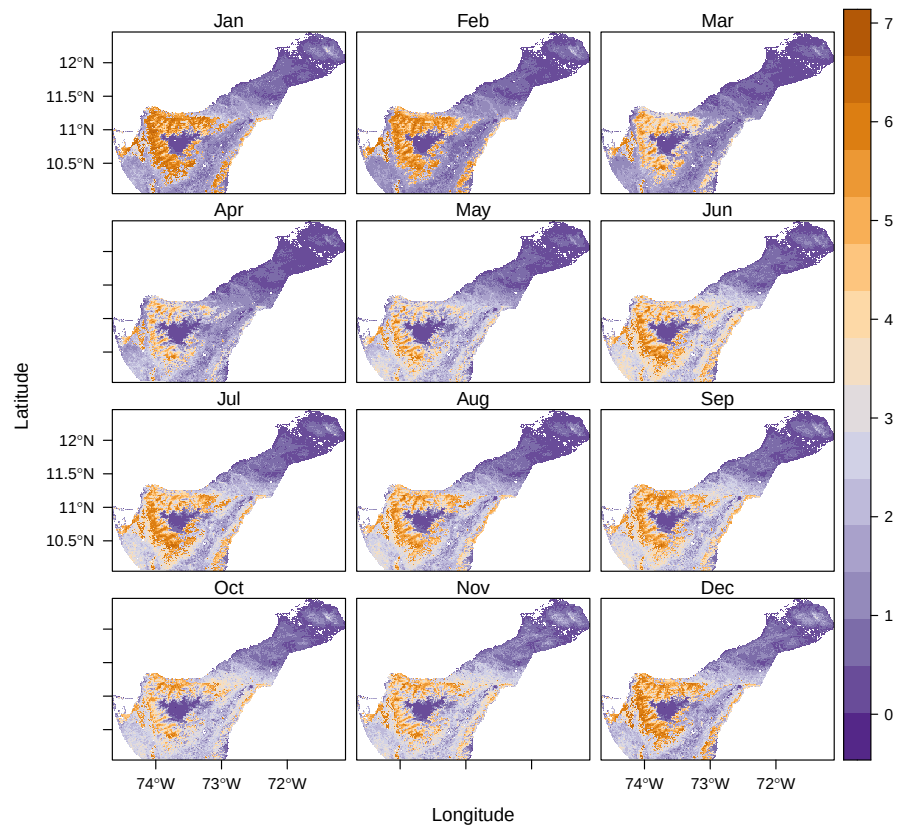


Figura 34: Comportamiento promedio mensual del LAI

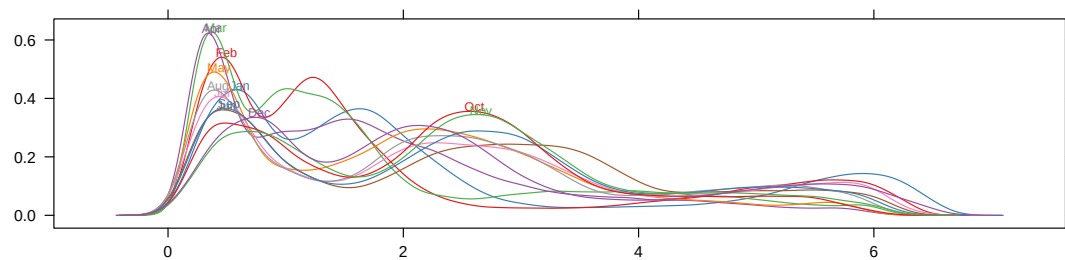


Figura 35: Gráfico de densidad promedio mensual del LAI

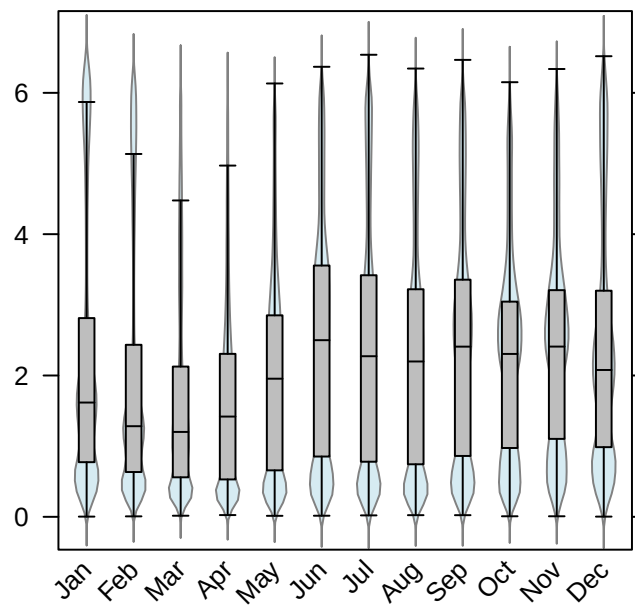


Figura 36: Gráfico de caja-violín promedio mensual del LAI

14.8. NDDI

Finalmente para el índice de sequía normalizado (NDDI), se visualiza una clara distinción entre las zonas orográficas con valores altos del índice en las zonas montañosa húmeda. Esto nos dice que la sequía impacta más severamente las zonas donde la humedad es más alta.

En el gráfico de frecuencia se comporta similar a los indicadores estudiados, donde al inicio de año se concentran los valores bajos y al final del año se concentran los valores altos.

Situación que se repite en el gráfico de caja-violín.

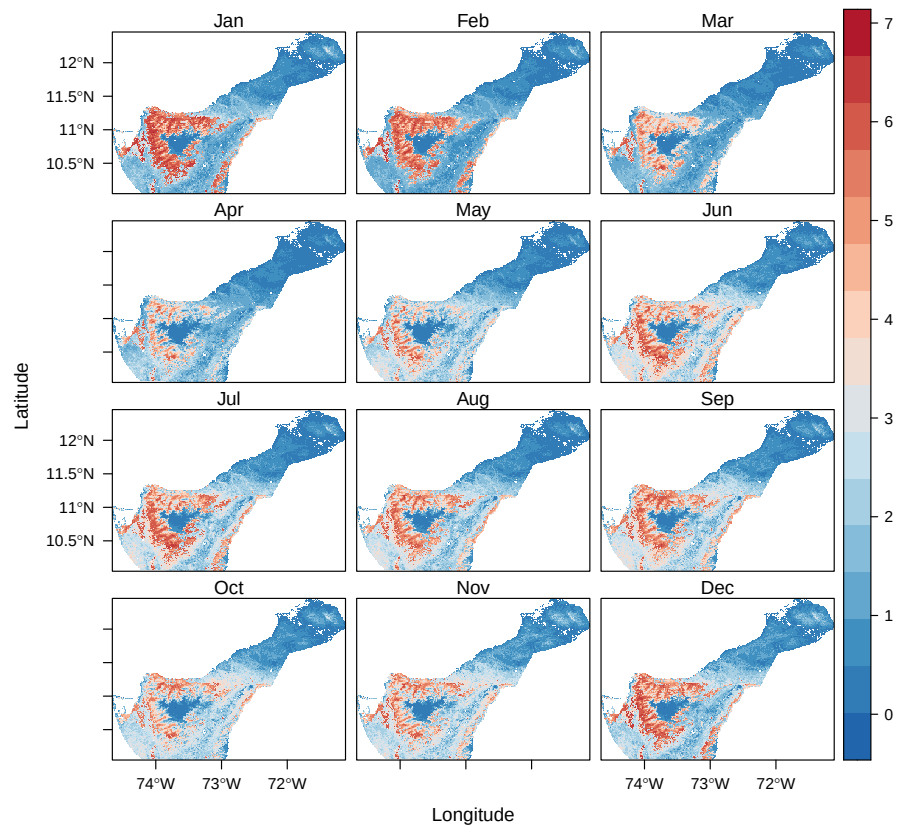


Figura 37: Comportamiento promedio mensual del NDDI

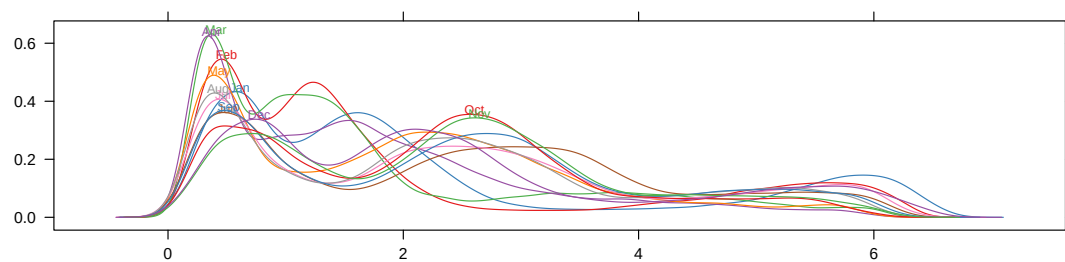


Figura 38: Gráfico de densidad promedio mensual del NDDI

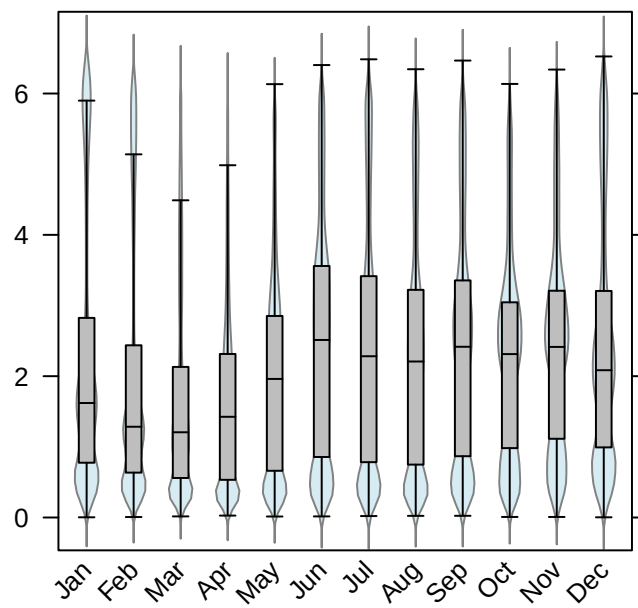


Figura 39: Gráfico de caja-violín promedio mensual del NDDI

15 SPI no paramétrico estimado

15.1. Ajuste de los Valores

15.1.1. Redes Neuronales

El primero modelo evaluado gráficamente es el de las redes neuronales tipo MLP con la configuración especificada en la sección 12.4. Un gráfico de dispersión para una estación de muestra es mostrada en la figura 40.

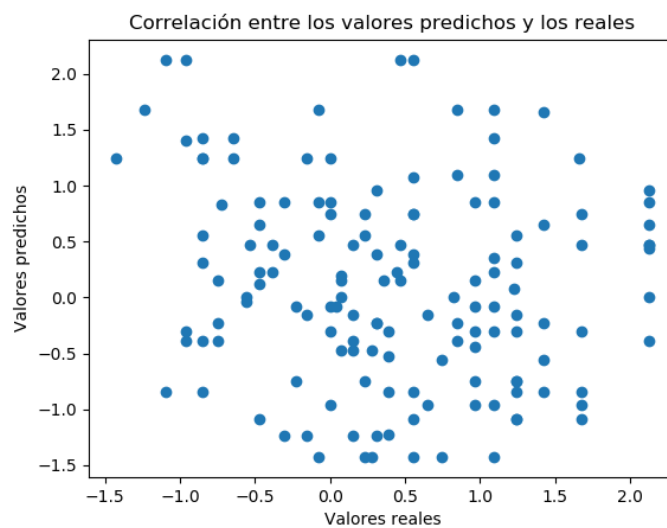


Figura 40: Correlacion entre los valores reales y predichos Red Neuronal 3 Meses

Dada la imposibilidad de la red neuronal tipo Perceptrón Multi-capas implementado de converger, este modelo es descartado de los análisis posteriores.

15.1.2. Bosque Aleatorio (Random Forest)

Para el siguiente modelo, se genera un gráfico de dispersión en base a la información de las 51 estaciones seleccionadas, y se ajusto una regresión lineal, para cada una de las ventanas temporales seleccionadas de 3, 6 y 9 meses.

Es posible detallar a través de los tres gráficos como la dispersión de los datos mejora para ventanas temporales mayores, lo que implica que el modelo tiene más problemas para capturar el comportamiento en periodos cortos de tiempo.

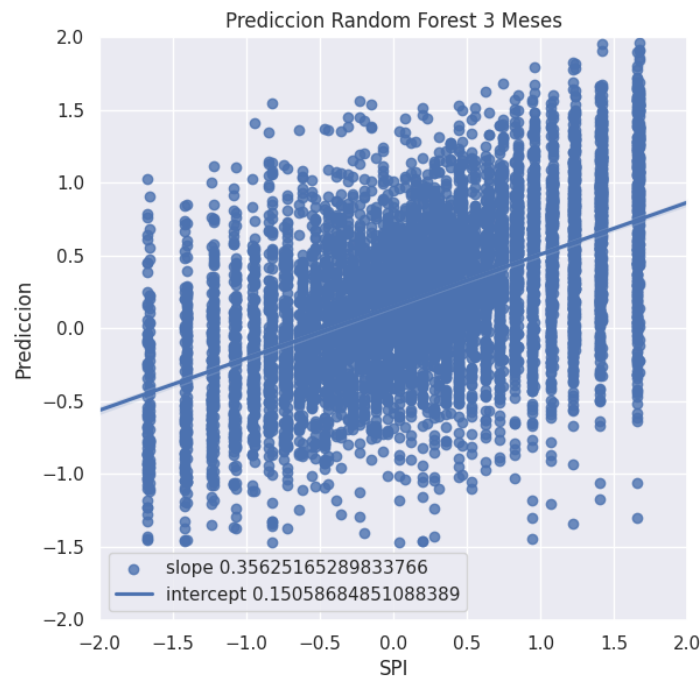


Figura 41: Correlacion entre los valores reales y predichos Random Forest 3 meses

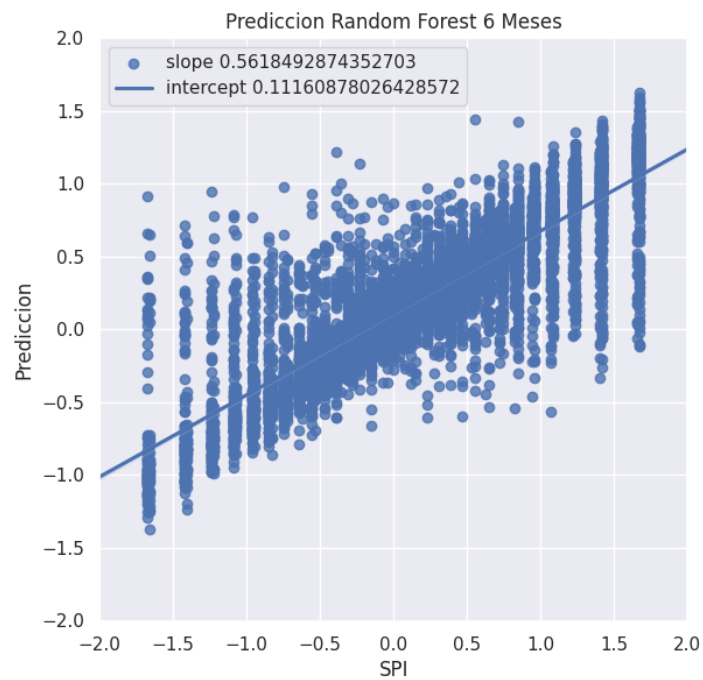


Figura 42: Correlacion entre los valores reales y predichos Random Forest 6 meses

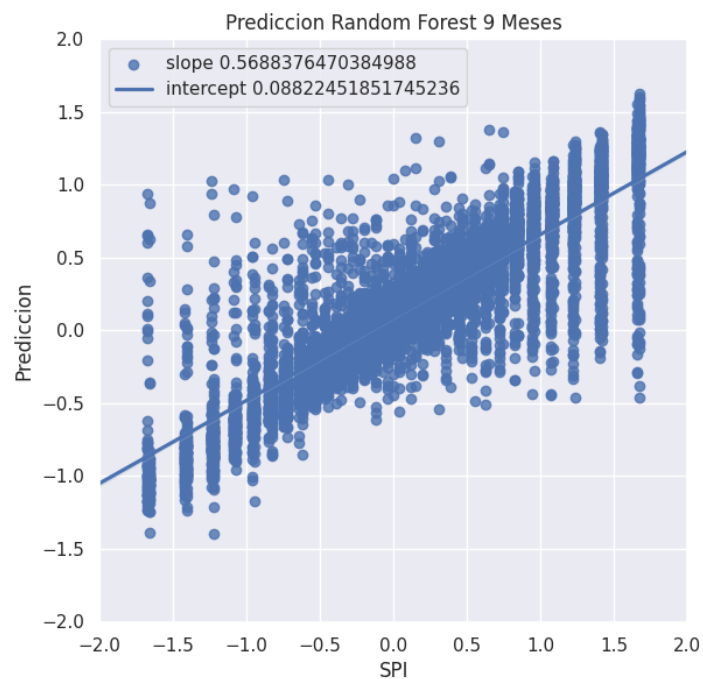


Figura 43: Correlacion entre los valores reales y predichos Random Forest 9 meses

15.1.3. Aumento Extremo del Gradiente (Extrem Gradient Boosting)

De forma similar para el siguiente modelo se el Extrem Gradient Boosting se genera una gráfica de dispersión para todas las ventanas temporales. En este podemos visualizar como de manera similar al modelo Random Forest, la dispersión de la predicción aumenta conforme la ventana temporal aumenta. Esto sumado al hecho de tener unas correlaciones más altas que el modelo random forest.

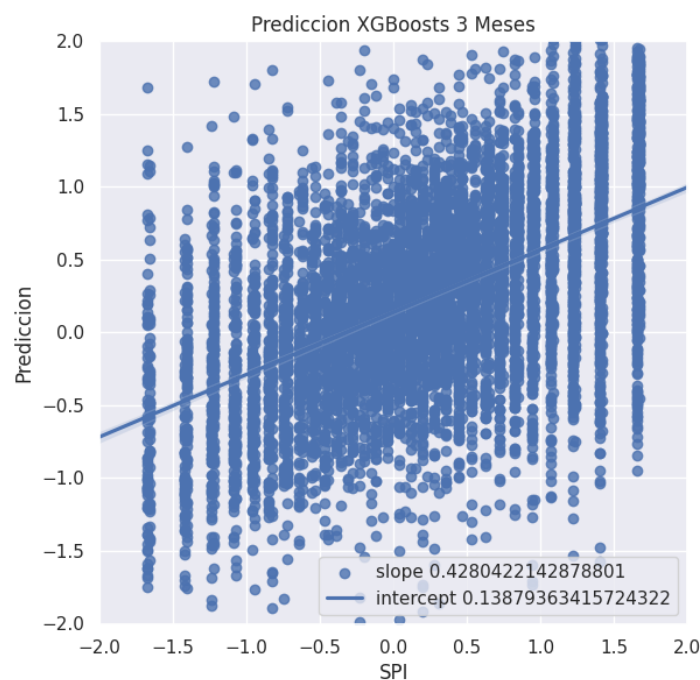


Figura 44: Correlacion entre los valores reales y predichos Extrem Gradient Boosting

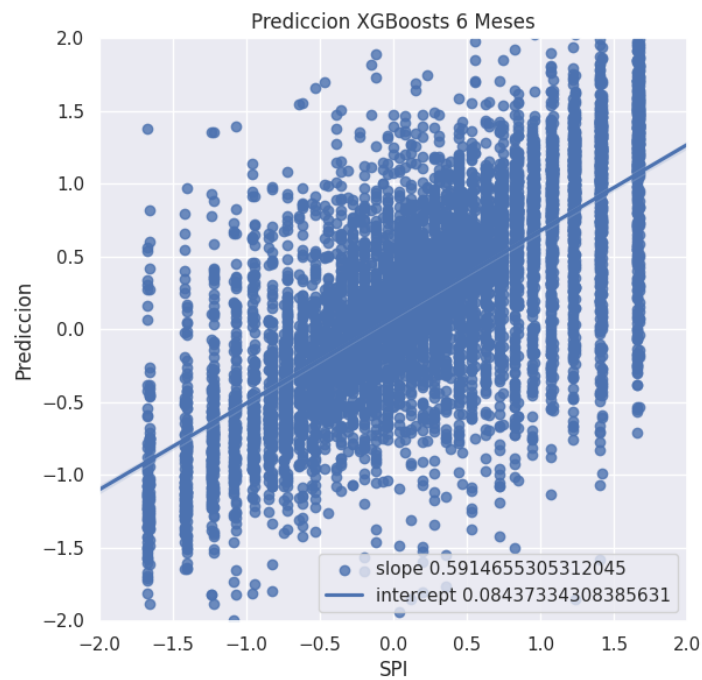


Figura 45: Correlacion entre los valores reales y predichos Extrem Gradient Boosting

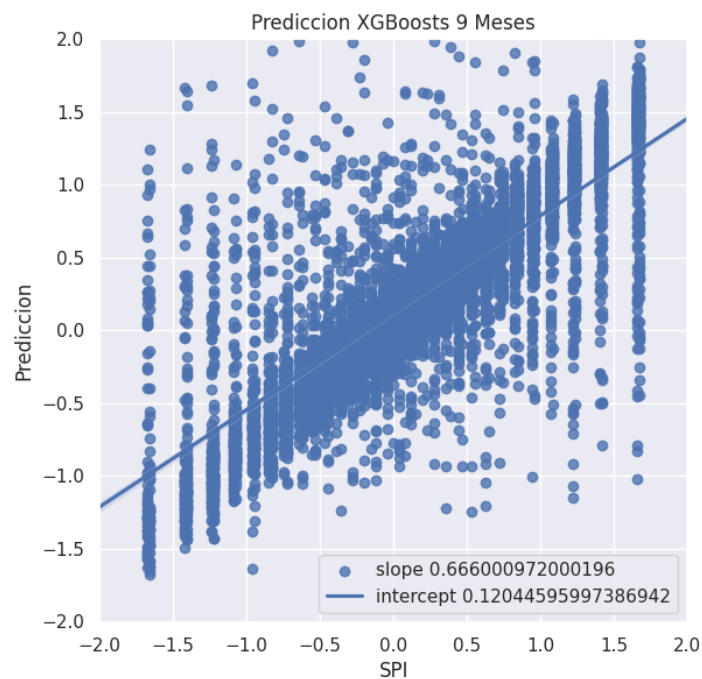


Figura 46: Correlacion entre los valores reales y predichos Extrem Gradient Boosting

15.2. Métricas de Error

Una vez más en detalle, se visualiza el comportamiento de las métricas de error del entrenamiento de los modelos, en particular sobre los datos no vistos por el modelo, o los datos de validación.

Como se mencionó antes, la estrategia escogida fue la de validación cruzada usando el criterio espacio-temporal LLTO, el resultado de dicha estrategia puede ser visualizado en la figura 47. Donde la columna izquierda representa las rondas del modelo random forest durante el entrenamiento, a su vez la columna derecha representa las rondas del modelo Extrem Gradient Boosting. Por otro lado las filas representan las ventanas temporales de los modelo entrenados, 3, 6 y 9 meses para la primera, segunda y tercera fila. Las 20 barras de cada gráfico representan individualmente las rondas de entrenamiento computadas, y el color de la barra representa la métrica monitoreada, azul para error medio absoluta (MAE) y naranja para el error medio cuadrático (MSE).

Se puede visualizar un comportamiento bastante homogeno en la distribución de los errores para los modelos, con resultados ligeramente superiores para el modelo Extrem Gradient Boostings. Adicionalmente en los cuadros 3 y 4 tenemos los valores promedio de las métricas para cada ventana temporal.



Figura 47: Métricas de Error Validación Cruzada el eje x representa 20 subconjuntos de datos y el eje y representa los errores en cada ronda de entrenamiento

Tabla 3: Tabla de errores MAE promedio.

Ventana Temporal	Random Forest MAE	Extrem Gradient Boosting MAE
3 meses	0.736	0.785
6 meses	0.783	0.863
9 meses	0.814	0.875

Tabla 4: Tabla de errores MSE promedio

Ventana Temporal	Random Forest MSE	Extrem Gradient Boosting MSE
3 meses	0.856	0.883
6 meses	0.883	0.927
9 meses	0.901	0.934

15.3. Comportamiento Temporal

Por otro lado, el ajuste de los modelos al comportamiento temporal de una estación de muestra se visualiza en la presente sección donde se selecciona una estación aleatoria y comparamos el comportamiento real del indicador SPI y los valores predichos por el modelo.

15.3.1. Random Forest

Para el modelo Random Forests tenemos podemos visualizar el comportamiento en las figuras 48 49 50 en estos gráficos tenemos en negro el valor de SPI estimado y en verde la predicción del modelo. El ajuste temporal de los modelos es sigue el comportamiento general del fenómeno, con dificultades para la los valores más altos y más bajos del indicador. Parece existir una tendencia hacia los valores altos del indicador que indican condiciones húmedas, más que las condiciones secas .Además al incrementarse la ventana temporal el comportamiento tiene una tendencia más clara y menos aleatorio por lo que el modelo puede describir más fidedignamente el comportamiento.

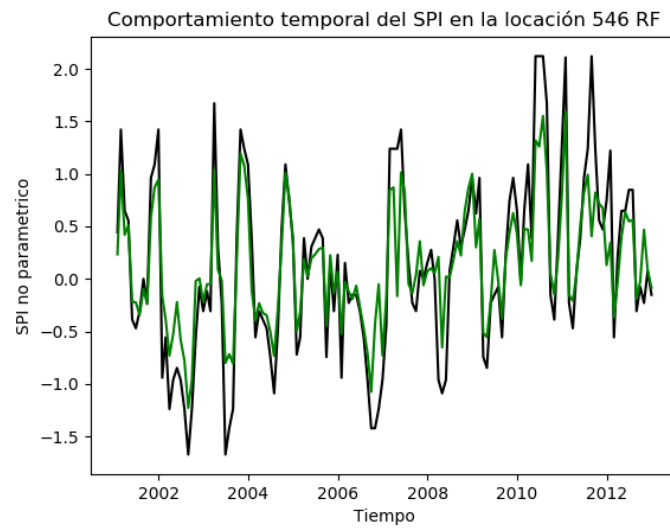


Figura 48: Perfil temporal entre los valores reales y predichos Random Forest 3 Meses (En negro el SPI en verde la estimación del modelo)

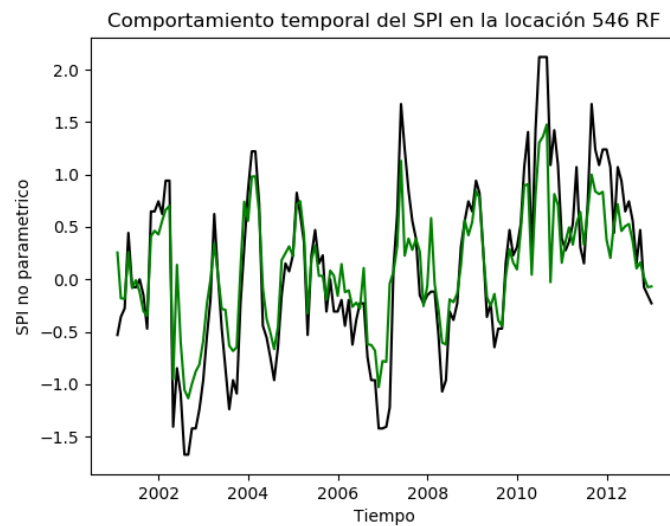


Figura 49: Perfil temporal entre los valores reales y predichos Random Forest 6 Meses (En negro el SPI en verde la estimación del modelo)

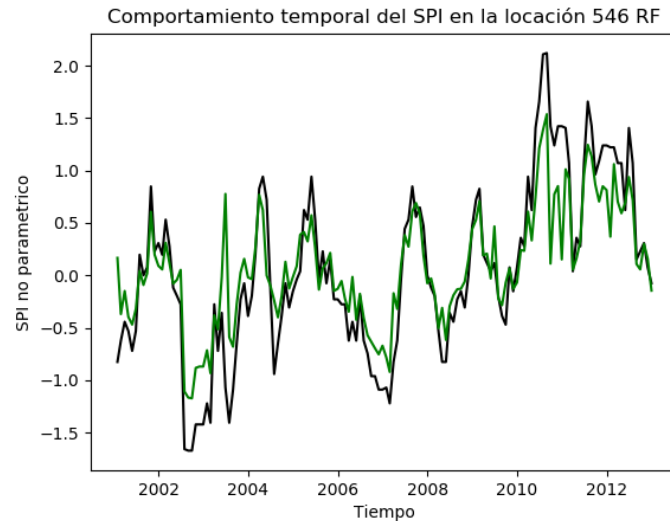


Figura 50: Perfil temporal entre los valores reales y predichos Random Forest 9 Meses (En negro el SPI en verde la estimación del modelo)

15.3.2. Extrem Gradient Boostings

De manera similar el modelo Extrem Gradient Boosting, se construyen una serie de visualizaciones presentadas en las figuras 51 52 53 aquí se puede observar lo siguiente. El modelo reproduce el comportamiento temporal de una manera más cercana en comparación con el modelo Random Forests con menos problemás para estimar los valores extremos aunque se observa cierto desplazamiento temporal en la curva descrita. Se observa igualmente una tendencia hacia valores de SPI positivos. Finalmente parece existir un menor impacto en el tamaño de la ventana temporal de estudio sobre la capacidad predictiva del modelo.

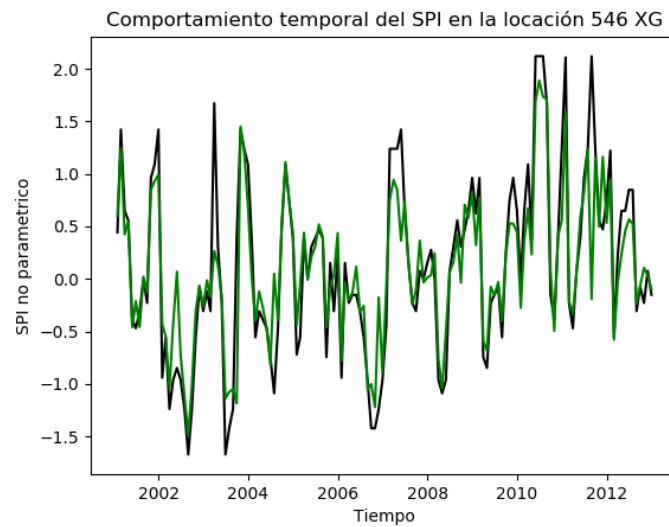


Figura 51: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses (En negro el SPI en verde la estimación del modelo)

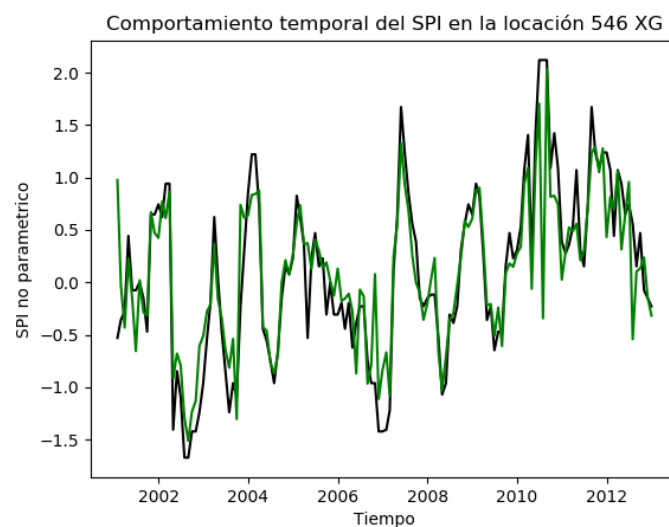


Figura 52: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 6 Meses (En negro el SPI en verde la estimación del modelo)

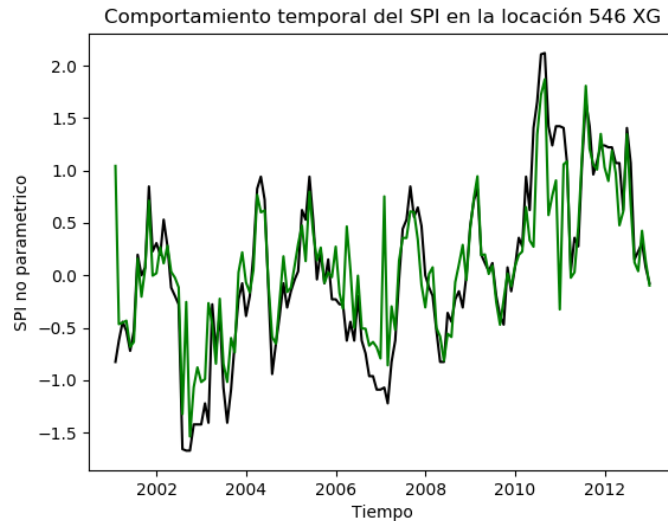


Figura 53: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 9 Meses (En negro el SPI en verde la estimación del modelo)

15.4. Comportamiento Espacial

Para visualizar el comportamiento espacial, se selecciono el uso de la ventana temporal de 3 meses, y computar los valores promedios del indicador a nivel mensual, con el objetivo de tratar de visualizar tendencias regidas por el régimen de lluvias anual de la zona de estudio.

15.4.1. Random Forests

Para el modelo de Random Forest un comportamiento muy homogéneo a lo largo del tiempo, y con valores normales del índice, y con una zona en sequía alrededor de la ciénaga grande de Santa Marta. Es posible que sea un efecto de borde.

Para los valores en densidad se visualiza una diferencia más pronunciada en la distribución de los valores a lo largo del año, con el periodo seco concentrado en los meses de enero, febrero, una periodo de transición en diciembre y mayo; y condiciones húmedas al final de año.

Finalmente, en el gráfico de cajas-bigotes vemos que el algoritmo tiene dificultades que para capturar la variabilidad de los datos, lo que es consistente con lo visto, en los resultados

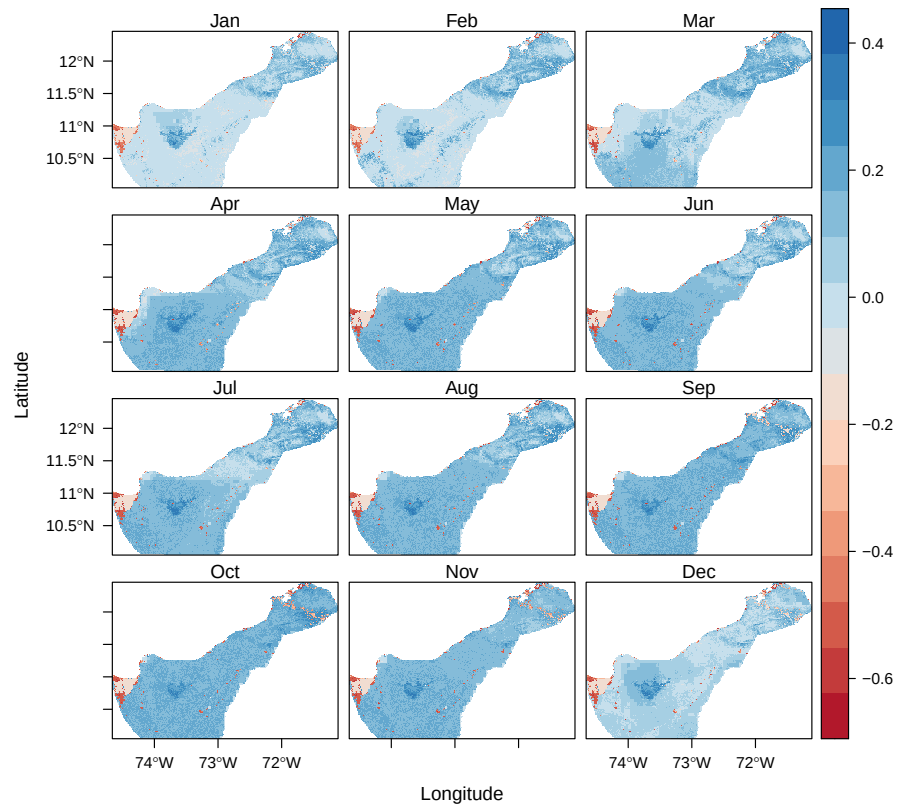


Figura 54: Comportamiento promedio mensual de la predicción SPI por Random Forest

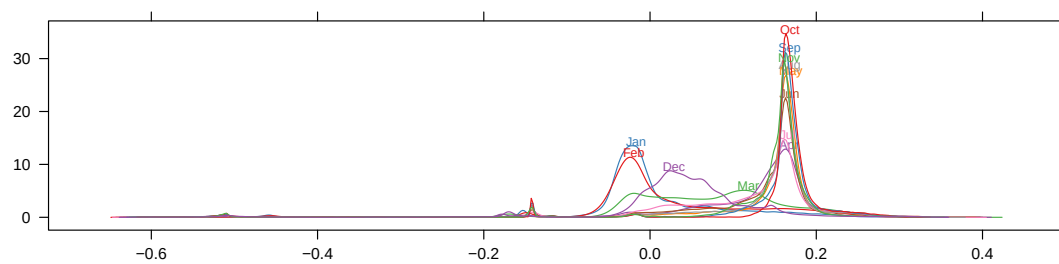


Figura 55: Gráfico de densidad promedio mensual de la predicción SPI por Random Forest

anteriores.

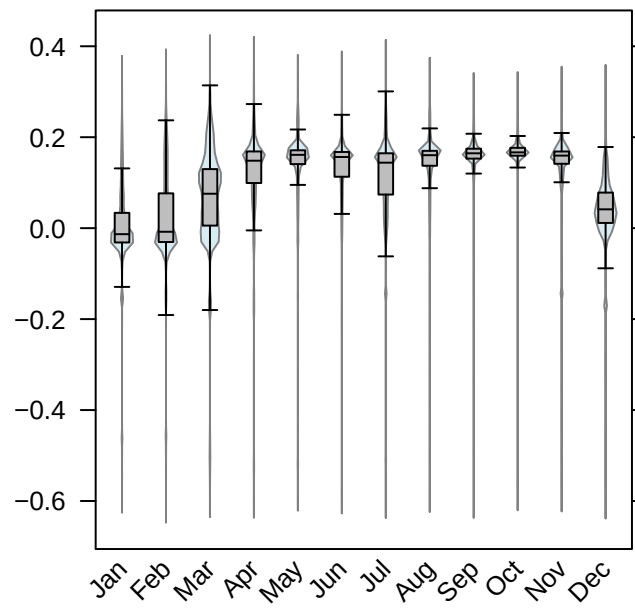


Figura 56: Gráfico de caja-violín promedio mensual de la predicción SPI por Random Forest

15.4.2. Extrem Gradient Boostings

Por su parte el algoritmo de Extrem Gradient Boosting, se detalla una demarcación más precisa de las zonas semi-áridas en los valores del indicador. Donde además vemos como los valores bajos del indicador se concentran al inicio del año en los meses de enero, febrero y marzo, con un mes de transición en diciembre. Y los valores altos del indicador se concentran en las zonas de transición de la zona montañosa.

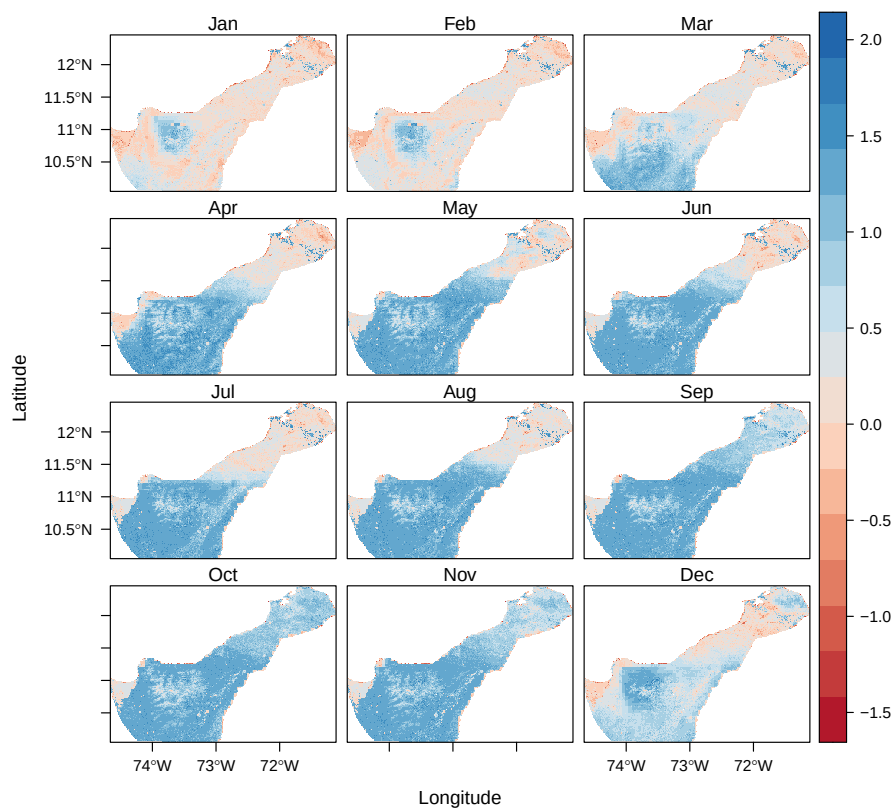


Figura 57: Comportamiento promedio mensual de la predicción SPI por Extrem Gradient Boosting

Este comportamiento se replica en los valores en densidad, con los valores húmedos altamente concentrados al final del año, a excepción de diciembre.

Esto finalmente se corrobora en el gráfico de cajas y bigotes.

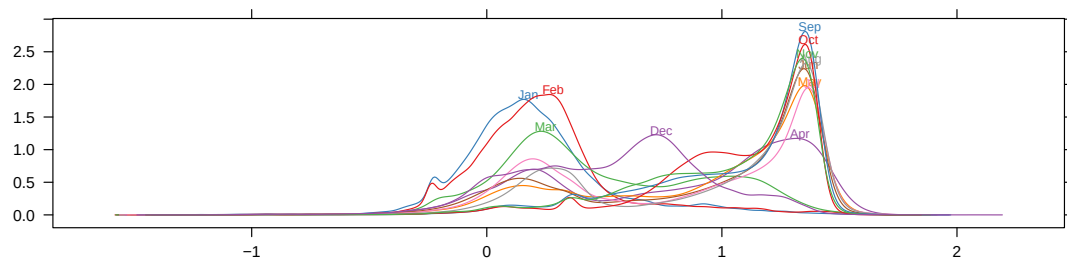


Figura 58: Gráfico de densidad promedio mensual de la predicción SPI por Extrem Gradient Boosting

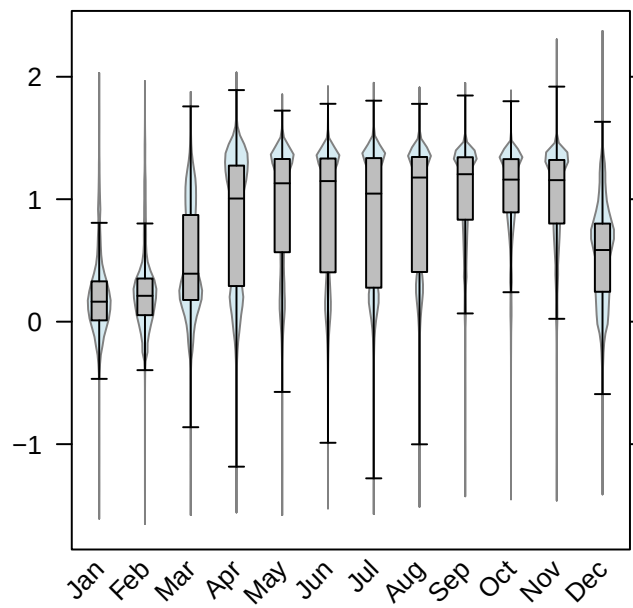


Figura 59: Gráfico de caja-violín promedio mensual de la predicción SPI por Extrem Gradient Boosting

15.5. Importancia de los predictores

Por último, un aspecto positivo de los modelos basados en árboles es la posibilidad de pesar la importancia de las variables predictoras sobre la predicción final, a continuación se visualiza la importancia de las variables según el modelo Random Forest.

Es posible visualizar en este gráfico que la importancia de las variables en los modelos aunque cambia sus valores relativos su orden se conserva a través de las distintas ventanas temporales del estudio. con Un primer grupo de variables como la precipitación y la temperatura superficial relacionadas con la climatología de la zona y seguida de el NDDI que sería una especie de conglomerado entre las variables relacionadas con la vegetación y la humedad. A continuación las variables cercanas o directamente relacionadas con el comportamiento de la vegetación y finalmente la humedad detectada por el satélite a través del NDWI.

La importancia de la precipitación, la temperatura superficial eran esperados, es relevante detectar la importancia del NDDI lo que puede inducir a interpretar que el uso de las variables vegetativas y del NDWI es muy redundante. Finalmente es posible que dada la climatología seca de la zona se obtengan unos pesos tan bajos de NDWI en la predicción.

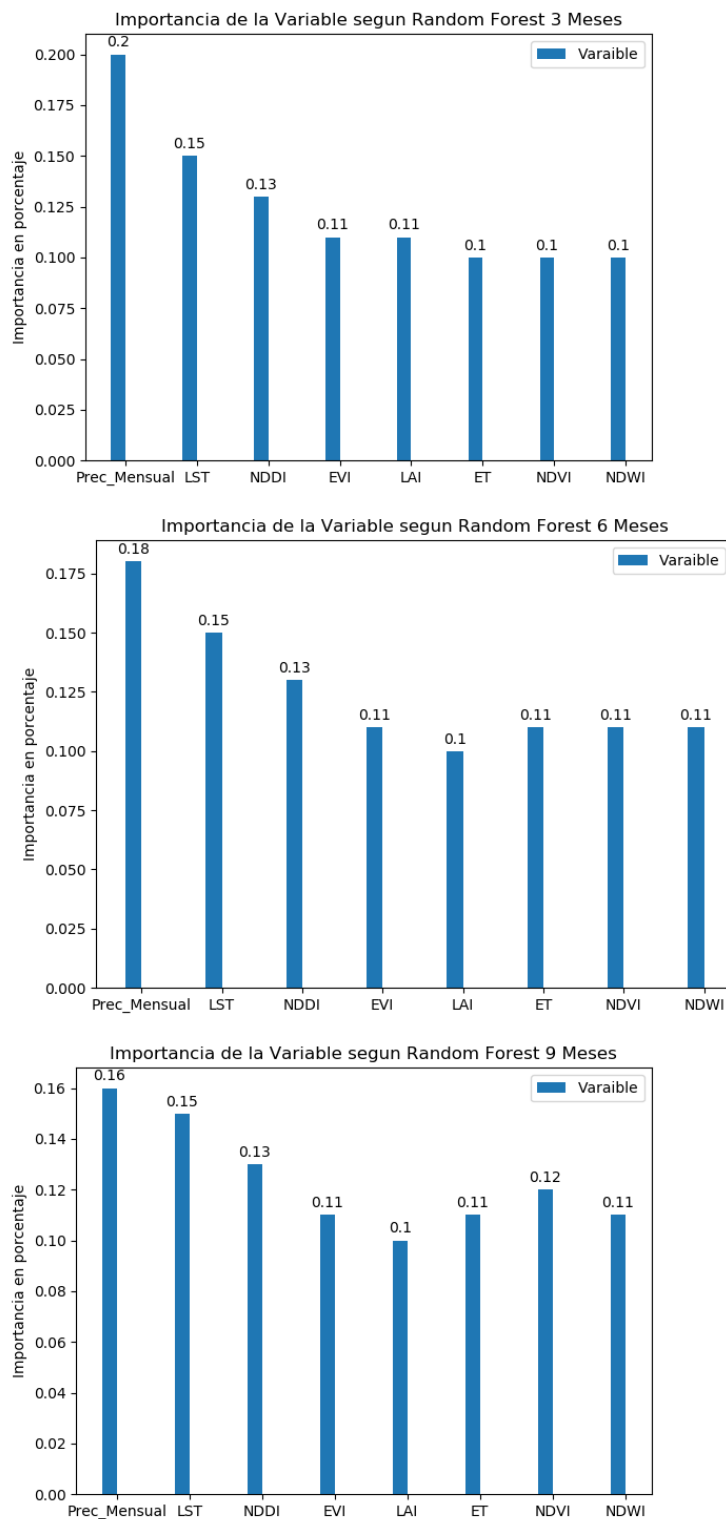


Figura 60: Importancia de las variables predictoras Random Forest 3 Meses

Parte X

Discusión

Distintos autores han investigado el modelamiento y monitoreo de sequías, cada uno proponiendo distintas metodologías que aún no convergen a un único tratamiento del problema, hoy en día varias investigaciones se centran en el uso de técnicas de *machine learning*, como en el presente trabajo, donde se propende por fusionar datos de diferentes sensores, para aprovechar al máximo las virtudes de cada fuente.

Por ejemplo en Colombia, el organismo encargado de monitorear el estado de los recursos hídricos a nivel nacional es el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Que al igual que lo aquí expuesto realiza un monitoreo mensual del indicador de sequía meteorológica SPI, sin embargo su monitoreo es realizado en ventanas temporales de 1 y 12 meses. Además los métodos usados en su estimación difieren puesto que el IDEAM basa su metodología en un monitoreo clásico desde el punto de vista meramente hidrológico, por lo que sus fuentes de datos se limitan a las estaciones meteorológicas, realizando un exhaustivo análisis hidrológico que incluye estimación de periodos de retorno, análisis tendenciales, entre otros. Finalmente el estudio del IDEAM espacializa el comportamiento de la sequía usando interpolaciones basadas en métodos geoestadísticos, en contraposición a los modelos impulsados por datos expuestos en este trabajo. Dicho lo anterior, no es claro cual es el método utilizado por el IDEAM para espacializar sus registros de SPI calculados sobre las estaciones meteorológicas, tampoco es claro el error o incertidumbre asociada a dicha estimación.[8] [32].

Por otro lado, desde el punto de vista de la academia se han presentado una serie de propuestas de mejoras sobre la modelación de sequías, mejoras que dadas las limitaciones de datos y recursos de computo se intento implementar, por ejemplo la estimación estándar el **SPI** se realiza mediante la estandarización de una función de probabilidad que se ajuste a un racha de precipitación medida en una estación, meteorológica dada; sin embargo esta aproximación supone cierta homogeneidad en la zona de estudio, puesto que asume que el comportamiento de la precipitación sera similar en un área determinada.

Este no siempre es el caso, sobre todo cuando los estudios cubre áreas extensas, debido a lo anterior[16], introduce en al literatura el **SPI No Paramétrico**, que proponen dada la falta de flexibilidad del método actual, este consiste en ajustar una función de probabilidad empírica a las precipitaciones de cada estación meteorológica de la zona de estudio, y posteriormente realizar la estandarización para estimar el SPI.

En su estudio [16] en comparación al presente estudio, trabajan su modelación de sequía a

una resolución menos detallada, en una extensión global y con unas fuentes de datos que se limitan a datos de precipitación calculados por diferentes centros especializados a nivel internacional, además por la amplitud de la zona de estudio usan el algoritmo k-vecinos (para identificar 5 regiones climáticas) como aproximación a la regionalización climática de su zona de estudio y proceden a calcular modelos individuales especializados para cada uno de los cinco regiones. Finalmente y probablemente debido a la mayor amplitud de las fuentes de información obtengan resultados relevantes como un R^2 mínimo de 0.95, en comparación al R^2 de 0.777 en promedio para el Random Forest y el R^2 de 0.769 para el Extrem Gradient Boostings obtenidos en este estudio.

Más recientemente la literatura en cuanto a la modelación de sequías a virado a el uso de técnicas que permitan una espacialización fidedigna de los índices de sequía calculados usando datos de estaciones meteorológicas, es el caso de [27], [24], [13].

Esta clase de aproximaciones a la sequía, al ser basados en modelos impulsados por datos, usualmente requieren de una gran cantidad de datos para ajustar los modelos, por lo que la baja densidad y en ocasiones los problemás de calidad de los datos en la zona estudiada pueden llegar a impactar la capacidad de los modelos para explicar el fenómeno de sequía estudiado.

Aunque los autores consultados realizaron estudios con un número de estaciones variable lo que puede impactar la comparación de los estudios, que aunque poseen condiciones similares, estas distan de ser idénticas. Por ejemplo [33] usa en su estudio dos redes de estaciones independientes, localizadas en el sur-oriente de Australia, uno correspondiente a 13 estaciones usadas para entrenar y validar el modelo, y otra red que asciende a 2242 estaciones, usadas unicamente con el propósito de testear los modelos, en un área que asciende a 360 000 km². Por su parte [27] centra su estudio sobre la península coreana, incluyendo 61 estaciones sobre un área de estudio de unos 100 284 km². En comparación el presente estudio se utilizaron 51 estaciones sobre un área de unos 30 000 km² aproximadamente.

En cuanto a los datos de precipitación, su uso no varía mucho entre los diferentes autores, se usan los datos en mm como principal predictor de los modelos aplicados. Por ejemplo [34] enfocó su estudio únicamente en precipitación, no uso predictores adicionales, para compensar esto combinó diferentes productos satelitales para tener distintas estimaciones de sequía para cada unidad mínima observada.

Sin embargo, debido a la gran cantidad de productos derivados del sensor MODIS, este provee de varios predictores, por ejemplo [27], utiliza variables como la temperatura superficial (LST), la Precipitación, Índices de vegetación (NDVI), además de algunas bandas espectrales 7 en total. Por su parte [13] en su estudio enfocado en la sequía agrícola, y la humedad del suelo, utilizó con predictores la temperatura superficial (LST), la evapotranspiración (ET), el Albedo, índice de área foliar (LAI), e índices de vegetación (NDVI, EVI). Basado en lo anterior, en el presente estudio se seleccionaron las variables predictoras, Precipitación, Temperatura Superficial (LST), variables del ciclo hidrológico Evapotranspiración (ET) comportamiento de la vegetación, índice de área foliar (LAI), e índices de vegetación (NDVI, ET), un índice de humedad (NDWI) y una aproximación a la sequía (NDDI).

Finalmente, [24] indica como las variables Temperatura Superficial (LST) y Evapotranspiración (ET) muestran una importancia alta para la estimación del SPI en el corto plazo, a diferencia de otros predictores como el los índices de vegetación (NDVI) y el Índice Normalizado de Sequía Multi-Banda (NDMI) que tienen implicaciones en el largo plazo. Esto contrasta con los resultados obtenidos en este estudio donde las ventanas temporales seleccionadas se conservo el orden de la importancia de las variables predictoras, encabezadas por la Precipitación, Temperatura Superficial (LST) y el Índice de Sequía Normalizado (NDI).

De manera similar los estudios consultados son consistentes con el presente en que los modelos derivados de los CART como el Bosque Aleatorio (Random Forest) o el Aumento Extremo del Gradiente (Extrem Gradient Boosting) obtienen mejores resultados en este tipo de problemás. Por ejemplo [13] al igual que [24] evalúa el uso de modelos como Bosque Aleatorio (Random Forest), Cubista, y *Bossted Regression Trees*; por su parte [27] opta por el uso de CART y Árboles Extremadamente Aleatorios.

Por el contrario otros autores consideran otros modelos como las Máquinas de Soporte Vectorial (SVM), y las Redes Neuronales en sus más simples arquitecturas [33]. Otros optan por modelos menos comunes como lo son *Adaptative Neuro-Fuzzy Inference System* (ANFIS), *M5P model tree*, *Group Method of Data Handling* (GMDH) [34].

La mayor parte de los autores anteriormente mencionados son bastante consistentes, que dado que se trata de resolver un problema de regresión, las métricas utilizadas para entrenar y evaluar los modelos se centraron principalmente en Raíz del Error Medio Cuadrático (RMSE por sus siglas en ingles) y el Error Absoluto Medio (MAE por sus siglas en ingles).

Sumado al uso de R^2 , como indicador adicional de que tan ajustada es la predicción de los valores por parte de los modelos. Obteniendo diversos resultados, por ejemplo [13] obtuvo un R^2 0.52 y un rmse DE 0.09.

Por su parte [24] obtuvo un R^2 de 0.93 y un RMSE de 0.3, lo que implica que a pesar de un buen ajuste en de los valores los errores son altos, lo que podría indicar un posible sesgo por parte del modelo. De igual modo [33] obtiene un R^2 de 0.28 y un RMSE de 0.9. En el ente estudio se obtuvo \ obtiene un R^2 de 0.28 y un RMSE de 0.9. En el presente estudio se obtuvo \ obtiene un R^2 de 0.28 y un RMSE de 0.9. En el presente estudio se obtuvo en promedio un R^2 de para Bosque Aleatorio y n R^2 para Aumento Extremo del Gradiente (Extrem Gradient Boosting) con un RMSE asociado de para Random Forest y de Extrem Gradient Boostings.

Existe sin embargo una dificultad para la comparación directa entre estos modelos, y esta es precisamente la evaluación de los mismos, dadas las particularidades de los datos espacio-temporales, descritas en este documento, hacen que el evaluar un modelo que pretende capturar el comportamiento de un fenómeno espacio-temporal como la sequía es algo no trivial, existen dos posibles estrategias de evaluación, la primera es el uso de dos conjuntos independientes de datos que permita un testeo del modelo como en el caso de [33]. Sin embargo, esto es difícil de realizar, sobre todo en países en vías de desarrollo; por lo que otro método popular para realizar la evaluación es la validación cruzada.

En el presente trabajo se implemento una rutina para aplicar el método de Validación Cruzada LLTO propuesto por [MEYER20181] (y descrito en 9.3.5) usando esta metodología la idea es realizar evaluación más realista de los modelos espacio-temporales, los autores de esta estrategia aplicaron este método para tratar de estimar la Temperatura Superficial (LST) en el continente Antártico [35].

Finalmente es de resaltar que la evaluación de los modelos espacio-temporales es un tópico aún en investigación, y en proceso de consolidación, por lo que el debate de aplicar o no estas estrategias aún esta abierto.

Parte XI

Conclusiones

En el presente trabajo se implementó, una metodología para el moldeamiento de una función de interpolación no lineal, para la espacialización del indicador SPI no paramétrico en la zona de estudio del departamento de La Guajira, Colombia entre el año 2000 y el año 2014. Como un estudio de caso en el modelamiento de sequías.

Para lo anterior, se sustentaron los pasos metodológicos en revisión de literatura, de diferentes disciplinas, que desde sus perspectivas apoyaron a diferentes partes del presente trabajo.

El proceso metodológico comprendió la identificación, evaluación y selección de una serie de estaciones meteorológicas entre 1982 y el 2012 aptas para el cálculo del SPI no paramétrico mediante la agregación de la precipitación a acumulados mensuales y su relleno en aquellas ventanas temporales donde existían datos faltantes mediante un método de imputación de datos; además de la descarga, proyección, recorte, aplicación de máscara y homogenización temporal con el acumulado o promedio de datos a nivel mensual.

Se seleccionaron variables explicativas y se optó por tres modelos de aprendizaje de máquina para realizar la interpolación, las redes neuronales, y clásico y robusto Random Forest, y la promesa Extrem Gradient Boostings.

De estos el que mejor resultado obtuvo fue en Extrem Gradient Boostings con una correlación o R^2 de 0.777 y de un MSE de 0.914 y un MAE de 0.841 frente a Random Forest con un R^2 de 0.768 y un MSE de 0.914 y un MAE de 0.841.

Al presentar por su origen los datos, autocorrelación espacial y temporal se optó por realizar el entrenamiento de los modelos usando la metodología de validación cruzada k-fold tipo LLTO (Leave Location and Time Out) con el fin de obtener una evaluación más fidedigna.

Esta metodología presenta fortalezas como la posibilidad de aplicarse sobre vastos territorios sin cambio significativos, su escalabilidad para realizar un monitoreo a nivel nacional, funciona con una precisión alta a pesar de la poca disponibilidad de información. Pero tiene debilidades como la necesidad de investigar más sobre los modelos espacio temporales y como evaluarlos.

El presente trabajo como caso de uso, puede ser un insumo para la formulación de monitoreo de sequías por parte de las autoridades del país, las metodologías expuestas aunque en

una área limitada, demuestran que es posible escalarlas.

Algunos temas dejados por fuera de este trabajo tienen el potencial de mejorar los resultados aquí expuestos, por ejemplo la incorporación de experimentos en métodos más sofisticados de selección de características de evaluación de los modelos con conjuntos de datos independientes, entre otros.

Parte XII

Observaciones y Recomendaciones

Una observación importante sobre la metodología presentada y es que el sensor *MODIS* (*Moderate-Resolution Imaging Spectroradiometer*) a bordo de los vehículos satelitales *Terra* y *Aqua*, y del que se derivan la mayoría de las variables predictoras de este proyecto se encuentra muy cerca del final de su vida útil, y su reemplazo el sensor VIIRS (*Visible Infrared Imaging Radiometer Suite*) a bordo del vehículo satelital *Suomi NPP* (*National Polar-Orbiting Partnership*), aún no tiene en fase operacional los mismos productos científicos que el actual *MODIS*, por lo que la continuidad de las observaciones no está garantizada.

Parte XIII

Anexos

16 AED Estaciones Metereológicas

A continuación se presentan unas muestras del comportamiento del SPI entre el 2000 y el 2014, aunque se ven claras diferencias entre las distintas estaciones es de destacar un inicio de década más bien húmedo, y con un periodo de sequía a inicios de la década de 2010, en algunas estaciones sin embargo es posible detallar el fenómeno de la niña del año 2011.

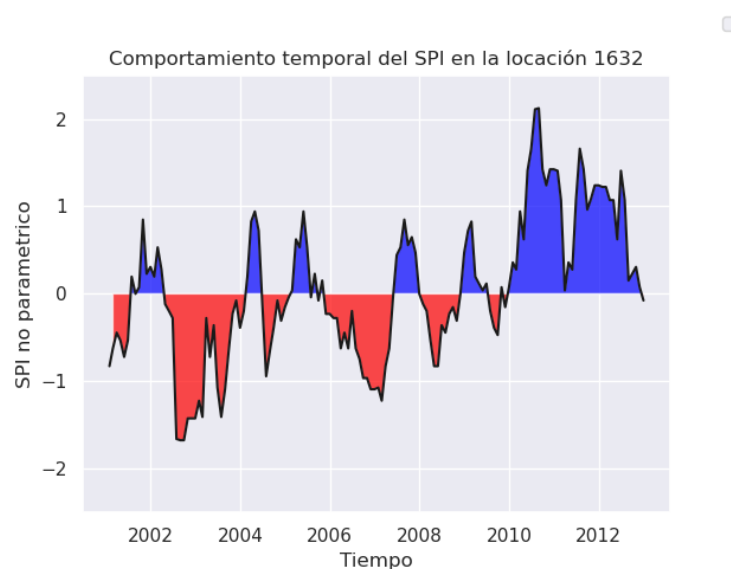


Figura 61: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

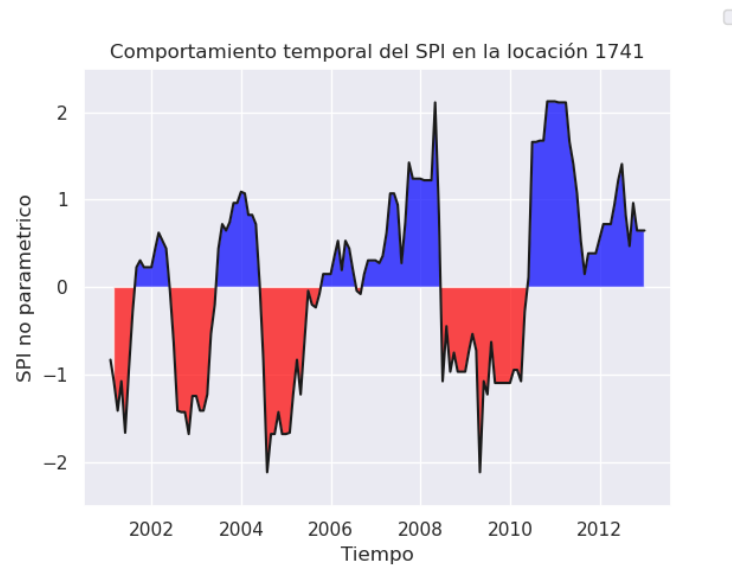


Figura 62: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

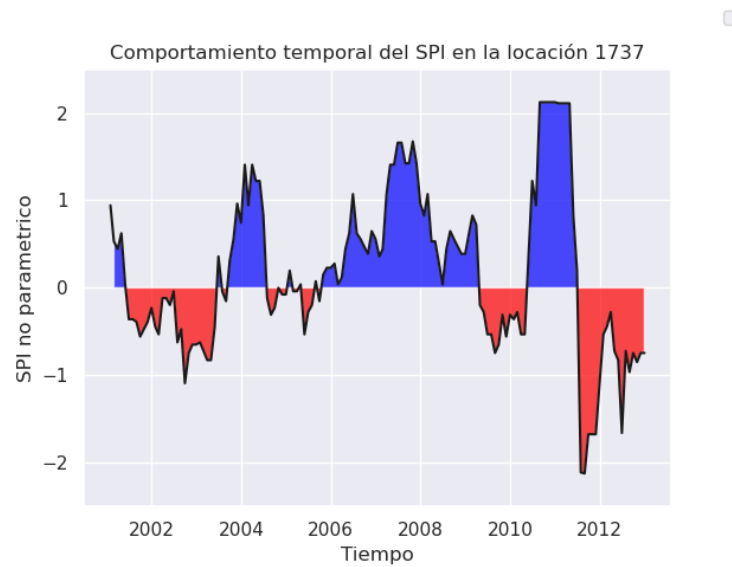


Figura 63: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

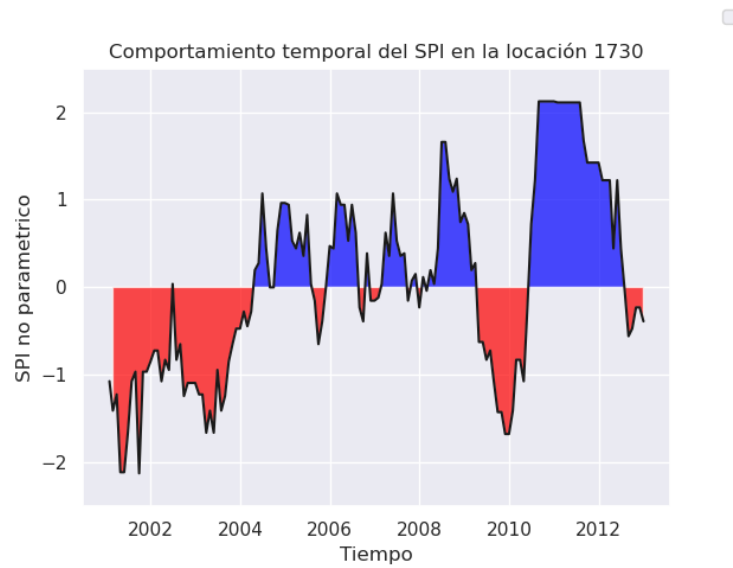


Figura 64: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

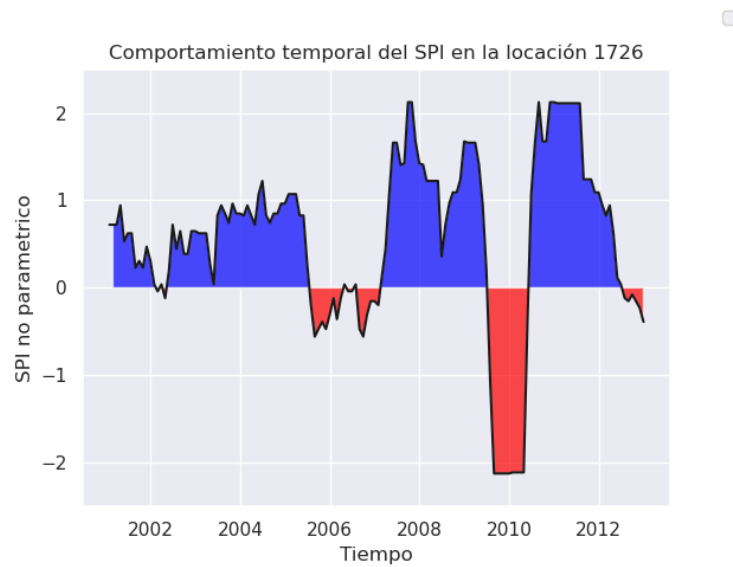


Figura 65: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

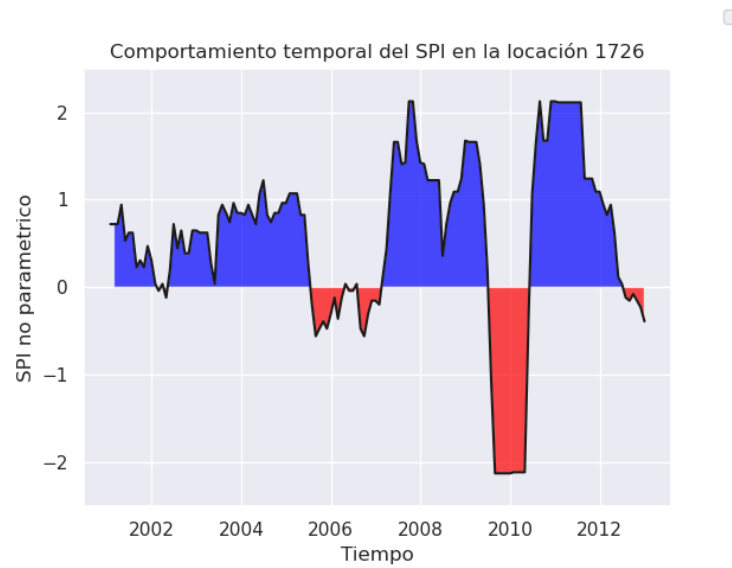


Figura 66: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

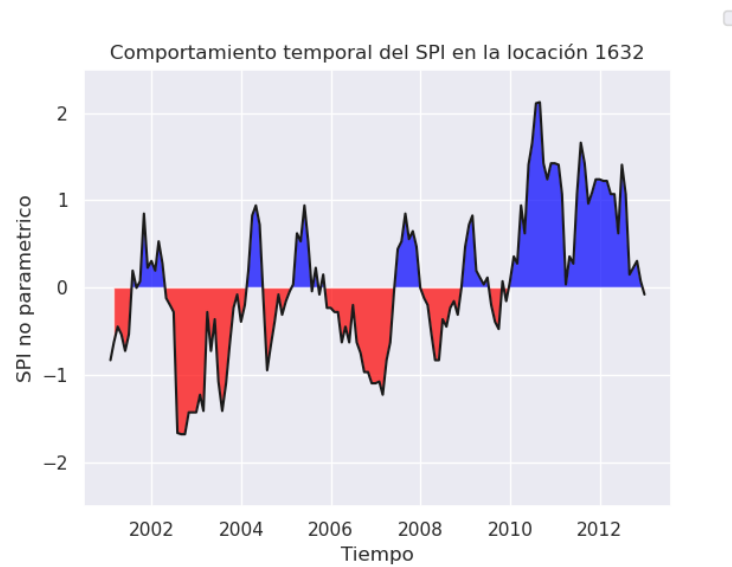


Figura 67: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

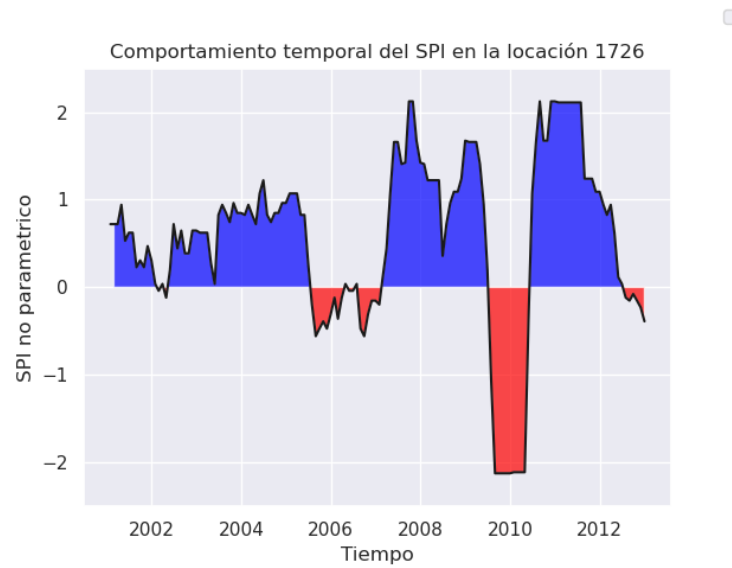


Figura 68: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

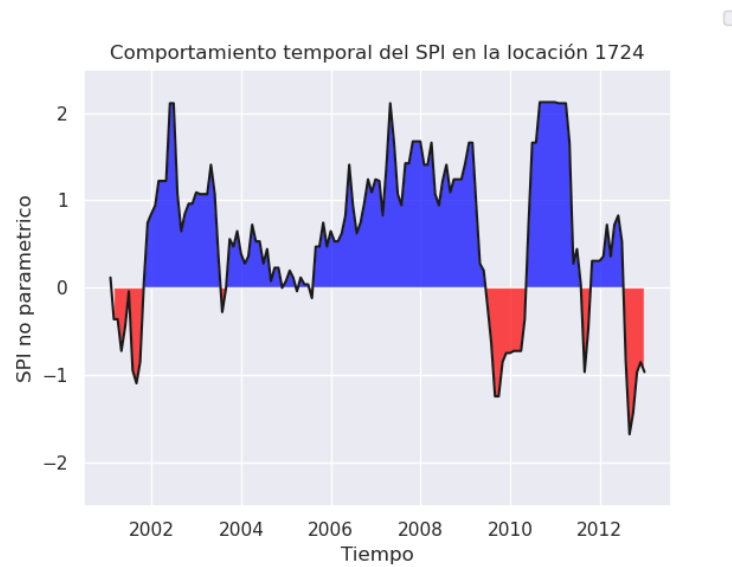


Figura 69: Perfil temporal entre los valores reales y predichos Extrem Gradient Boostings 3 Meses

17 Imputación de datos

El problema del relleno de datos faltantes es común en las ciencias ambientales, muchas veces se genera información que por diversas razones está contaminada y puede ser explotada solo parcialmente o no ser explotada en absoluto.

Por tal motivo desde distintas disciplinas se han desarrollado métodos para resolver este problema dependiendo del tipo de dato que se pierde y adaptando a dominios precisos.

Dado que los productos satelitales ópticos e infrarrojos son altamente afectados por la presencia de nubes, en este trabajo aplicamos el algoritmo **DINEOF** desarrollado para el relleno de datos en series remotamente sensadas de datos oceanográficos para el relleno de los datos, de las variables predictoras afectadas por la nubosidad (LST, NDVI, EVI, LAI, ET).

Dicho método se basa en las **EOF** o *Empirical Orthogonal Functions* por como son conocidos estos métodos en la comunidad oceanográfica, también conocida en otras comunidades como un tipo especial de los *Componentes Principales*.

18 Ciencia Reproducible

Con el animo de seguir los postulados de la *Ciencia Reproducible*, que abogan por compartir libremente no solo los resultados de una investigación sino los elementos necesarios para reproducirla, esto incluye una descripción clara del procedimiento seguido, los datos en orden de ejecutar los experimentos sumado a el código fuente y el ambiente de desarrollo usado, son compartidos, con el fin de detectar posibles errores.

En este orden de ideas es posible consultar esta información a través de internet los elementos anteriormente descritos, en las siguientes plataformas¹.

- Ambiente de desarrollo: **DockerHub**
- Código desarrollado: **Repositorio en GitHub**
- La base de datos generada para este proyecto: **Sequía**

¹Pulse sobre los textos en negrita para seguir el enlace.

Parte XIV

Bibliografía

-
- [1] Christopher R Schwalm y col. "Global patterns of drought recovery". En: *Nature* 548.7666 (2017), págs. 202-205.
- [2] UNCCD. "Global land outlook". En: Secretariat of the United Nations Convention to Combat Desertification Bonn ... 2017.
- [3] Jeffrey Gettleman. *Drought and War Heighten Threat of Not Just 1 Famine, but 4*. URL: <https://www.nytimes.com/2017/03/27/world/africa/famine-somalia-nigeria-south-sudan-yemen-water.html>.
- [4] WFP. *Colombia: la sequía afecta a miles de personas en La Guajira | WFP | Programa Mundial de Alimentos - Luchando contra el hambre en el mundo*. URL: <http://es.wfp.org/colombia-sequia-afecta-miles-personas-la-guajira..>
- [5] Deutsche Welle. *Sequía y hambre en La Guajira*. URL: <http://sostenibilidad.semana.com/impacto/articulo/sequia-y-hambre-en-la-guajira/35773>.
- [6] La Libertad. *La Guajira, el segundo departamento más pobre del país, según el DPS | Diario La Libertad*. URL: <http://lalibertad.com.co/wp/2016/10/17/la-guajira-el-segundo-departamento-mas-pobre-del-pais-segun-el-dps/>.
- [7] Gobernación de La Guajira. *Sequía y hambre en La Guajira*. URL: http://www.laguajira.gov.co/web/index.php?option=com_content&view=article&id=1182&Itemid=78.
- [8] M Garcia y col. *Estudio Nacional del Agua, 2014*. 2014.
- [9] WM WMO. *Vigilancia y Alerta Temprana de la sequia*. 2006.
- [10] Ciudad y Territorio Ministerio de Vivienda. *Portal Minvivienda Aprobados mas de 45 mil millones de pesos para proyectos ambientales de agua y saneamiento en La Guajira*. URL: <http://www.minvivienda.gov.co/sala-de-prensa/noticias/2017/octubre/aprobados-mas-de-45-mil-millones-de-pesos-para-proyectos-ambientales-de-agua-y-saneamiento-en-la-guajira>.
- [11] Javier Álvarez-Rodríguez y col. *Caracterización hidrológica de sequías. M-127. CEDEX; Centro de Publicaciones, Ministerio de Fomento*. Dic. de 2015. ISBN: 978-84-7790-563-9 ISBN DIGITAL (PDF)978-84-7790-564-6 NIPO 163-15-006-0 ; 007-6.
- [12] Mark Svoboda, Brian Fuchs y col. "Handbook of drought indicators and indices". En: (2016).
- [13] Seonyoung Park y col. "Drought monitoring using high resolution soil moisture thorough multi-sensor satellite data fusion over the Korean peninsula". En: *Agricultural and Forest Meteorology* 237 (2017), págs. 257-269.
- [14] Thomas B McKee. "Drought monitoring with multiple time scales". En: *Proceedings of 9th Conference on Applied Climatology, Boston, 1995*. 1995.
- [15] Diana Carolina Fonnegra Mora y col. "Desarrollo de un sistema operativo para el cálculo de índices de sequia basados en información espacial". Tesis de mtría. 2017.

-
- [16] Alireza Farahmand y Amir AghaKouchak. "A generalized framework for deriving non-parametric standardized drought indicators". En: *Advances in Water Resources* 76 (2015), págs. 140-145.
 - [17] Robert H Shumway, David S Stoffer y David S Stoffer. *Time series analysis and its applications*. Vol. 3. Springer, 2000.
 - [18] Jonathan D Cryer y Kung-Sik Chan. *Time series analysis: with applications in R*. Vol. 2. Springer, 2008.
 - [19] Emilio Chuvieco. *Fundamentos de teledetección espacial*. 526.982 C564 1996. Ediciones Rialp, 1996.
 - [20] Antonio Criminisi y Jamie Shotton. *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.
 - [21] Jordi Inglada y col. "Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series". En: *Remote Sensing* 9.1 (2017). ISSN: 2072-4292. DOI: 10.3390/rs9010095. URL: <https://www.mdpi.com/2072-4292/9/1/95>.
 - [22] Tianqi Chen y Carlos Guestrin. "Xgboost: A scalable tree boosting system". En: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, págs. 785-794.
 - [23] Christopher K Wikle, Andrew Zammit-Mangion y Noel Cressie. *Spatio-temporal Statistics with R*. CRC Press, 2019.
 - [24] Seonyoung Park y col. "Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions". En: *Agricultural and Forest Meteorology* 216 (2016), págs. 157-169. ISSN: 0168-1923. DOI: <https://doi.org/10.1016/j.agrformet.2015.10.011>. URL: <https://www.sciencedirect.com/science/article/pii/S0168192315007467>.
 - [25] Wilmar Loaiza Cerón, Yesid Carvajal Escobar y Olga L Baquero Montoya. *Sequias & adaptación: principios para su evaluación en sistemas productivos agrícolas del Valle del Cauca, Colombia*. Universidad del Valle, Programa Editorial, 2014.
 - [26] Xinhui Xu, Fei Xie y Xingyu Zhou. "Research on spatial and temporal characteristics of drought based on GIS using Remote Sensing Big Data". En: *Cluster Computing* 19.2 (2016), págs. 757-767.
 - [27] Jinyoung Rhee y Jungho Im. "Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data". En: *Agricultural and Forest Meteorology* 237 (2017), págs. 105-122.
 - [28] Zhengming Wan. "MODIS land-surface temperature algorithm theoretical basis document (LST ATBD)". En: *Institute for Computational Earth System Science, Santa Barbara* 75 (1999).

-
- [29] Alfredo Huete, Chris Justice y Wim Van Leeuwen. "MODIS vegetation index (MOD13)". En: *Algorithm theoretical basis document* 3.213 (1999).
- [30] Y Knyazikhin. "MODIS leaf area index (LAI) and fraction of photosynthetically active radiation absorbed by vegetation (FPAR) product (MOD 15) algorithm theoretical basis document". En: <http://eospsso.gsfc.nasa.gov/atbd/modistabls.heml> (1999).
- [31] Qiaozhen Mu, Maosheng Zhao y Steven W Running. "MODIS global terrestrial evapotranspiration (ET) product (NASA MOD16A2/A3)". En: *Algorithm Theoretical Basis Document, Collection 5* (2013).
- [32] I d IDEAM. *Estudio Nacional del Agua 2018*. 2018.
- [33] Puyu Feng y col. "Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia". En: *Agricultural Systems* 173 (2019), págs. 303-316.
- [34] Mohammad Reza Alizadeh y Mohammad Reza Nikoo. "A fusion-based methodology for meteorological drought estimation using remote sensing data". En: *Remote sensing of environment* 211 (2018), págs. 229-247.
- [35] Hanna Meyer y col. "Mapping daily air temperature for Antarctica based on MODIS LST". En: *Remote Sensing* 8.9 (2016), pág. 732.