UNIVERSIDAD DEL AZUAY

Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Ingeniería de Sistemas y Telemática

EXPLORACIÓN DE REDES BAYESIANAS PARA DETECTAR RELACIONES ENTRE LA PRECIPITACIÓN DEL ECUADOR Y TELECONEXIONES CLIMÁTICAS

Trabajo de graduación previo a la obtención del título de Ingeniero de Sistemas y Telemática

Autor:

Paúl Renato Ávila Andrade

Directora:

Agrim. Daniela Ballari, PhD

Cuenca – Ecuador

2019

DEDICATORIA

"Las monjas nos enseñaron que hay dos sendas en la vida: la de la naturaleza y la de la gracia. Debemos elegir cuál se seguirá. La gracia no trata de complacerse a sí misma. Acepta que la desaíren, que la olviden, que le tengan aversión. Acepta insultos y agravios. La naturaleza es interesada. Hace que otros la complazcan. Le gusta dominarlos. Para salirse con la suya. Encuentra razones para ser infeliz cuando el mundo brilla a su alrededor y el amor sonríe a través de todo." - *El árbol de la vida*.

A mi madre, padre y hermano por ser la gracia y naturaleza que han forjado mi vida.

AGRADECIMIENTO

A la doctora Daniela Ballari, a quien guardo mucho respeto y admiración, por todo el apoyo que me ha brindado. Estoy convencido que sin su ayuda este trabajo no hubiera sido posible. Mi eterna gratitud a ella.

Resumen

Las predicciones fiables de precipitación requieren comprender las teleconexiones climáticas sobre la precipitación. En el Ecuador, estas teleconexiones fueron estudiadas con métodos de correlación, careciéndose, sin embargo, de estudios multivariados con varios índices climáticos simultáneamente. El objetivo de este trabajo es explorar con una red bayesiana las relaciones entre la precipitación en el Ecuador continental e índices climáticos. Se emplean datos de precipitación y regiones de estacionalidad de precipitación para el aprendizaje de la red en R. Así se caracteriza la estructura y fuerza de la relación entre los índices climáticos y la precipitación.

Palabras claves: Teleconexiones climáticas, Redes bayesianas, propagación de probabilidades.

EXPLORATION OF BAYESIAN NETWORKS TO DETECT RELATIONSHIPS BETWEEN THE PRECIPITATION AND CLIMATIC TELECONNECTIONS IN ECUADOR

Abstract

Reliable precipitation predictions require understanding of climate teleconnections on precipitation. In Ecuador, these teleconnections were studied with correlation methods but there have been no multivariate studies with several simultaneous climatic indices. The objective of this work is to explore the relationships between precipitation and climatic indices in continental Ecuador with a Bayesian network. Precipitation data and precipitation seasonality regions are used for the learning of the network in R. This characterizes the structure and strength of the relationship between climatic indices and precipitation.

Keywords: Climatic teleconnections, Bayesian networks, propagation of probabilities.

Renato Ávila
Author

Agrim. Daniela Ballari, PhD
Thesis Director

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by
Ing. Paúl Arpi

ÍNDICE

Contenido

Introducción

Las precipitaciones juegan un papel fundamental en la economía de un país ya que influyen en los sectores productivos como por ejemplo: la agricultura (Ali et al., 2018; Sudha, Venugopal, & Ambujam, 2008), ganadería (Engler, Wehrden, & Baumgärtner, 2019), pesca (Pratiwi & Sukardjo, 2018), turismo (Abd-elhamid, Fathy, & Zelen, 2018) y generación de energía eléctrica (Hamududu, Killingtveit, & Engineering, 2012). En particular, este último sector es de especial interés para el Ecuador ya que produce gran parte de la energía eléctrica que se consume en el país. Además, existen nuevos proyectos hidroeléctricos que entrarán en funcionamiento en los próximos años, y para los cuales se requieren un conocimiento más profundo de la variable precipitación. En este contexto, se realizó un estudio en la cuenca del río Jubones al sur del Ecuador, donde se evaluó el impacto del cambio climático en la generación de energía hidroeléctrica a través de una herramienta de modelado (SWAT). Los resultados mostraron que durante las temporadas de sequías la planta experimentará una caída del 13% en la producción de energía debido a una reducción del 17% del flujo de agua (Hasan & Wyseure, 2018). La carencia de estudios sobre datos climáticos de precipitación ha dificultado la comprensión y predicción del clima, de ahí surge la importancia de las teleconexiones climáticas para permitir realizar estudios de pronósticos del clima de manera más fiables. Esto permitirá que los sectores económicos que son influenciados directamente por precipitaciones puedan planificar de mejor manera sus actividades, así como adoptar políticas orientadas a una gestión sostenible de los recursos hídricos.

Las teleconexiones climáticas se definen como una correlación entre variables climáticas separadas por grandes distancias. En la actualidad para el estudio de teleconexiones climáticas se usan diferentes métodos tales como: regresión lineal (Liu, Di, Chen, & DaMassa, 2018), análisis de componentes principales (PCA) (Fierro, 2014), análisis wavelet (Konapala, Valiya, & Ashok, 2017) y *Point-by-point correlation* (Fierro, 2014). En la literatura existen trabajos que hacen uso de estos métodos. Sin embargo, estos métodos presentan la limitación, según Torre-Gea, Soto-Zarazua, Guevara-Gonzalez, and Rico-Garcia (2011), que técnicas como regresión lineal, PCA y métodos de agrupamiento fragmentan la información y asumen independencia espacial para reducir la dimensionalidad. Por otra parte, existen técnicas de análisis multivariado como las teleconexiones ortogonales empíricas y las funciones ortogonales empíricas (EOF) las cuales se ajustan mucho mejor a este tipo de problemas. Una revisión de estos métodos se encuentra en (Lee, Lee, & Julien, 2018).

En los últimos años el uso de métodos bayesianos para análisis de teleconexiones climáticas se ha popularizado (Torre-Gea et al., 2011; Duc, Rivett, MacSween, & Le-Anh, 2017; Ebert-Uphoff & Deng, 2012). Una red bayesiana se define como "*un grafo acíclico dirigido que representa un conjunto de variables (nodos) y sus independencias condicionales probabilísticas (codificada en los arcos)*" (Correa, Bielza, & Pamies-teixeira, 2009) . Se trata de un modelo probabilístico que permite inferir variables desconocidas a partir de las variables que si son conocidas u observadas. Al ser un tipo de sistema experto, tiene aplicación en casi cualquier ámbito de las ciencias, por ejemplo: medicina, biología, astrofísica, informática, química, economía y clima.

En los últimos tiempos, la climatología se ha convertido en campo fértil para el empleo de técnicas con enfoque bayesiano, debido al creciente número de estaciones meteorológicas, así como a la mayor disponibilidad de datos climáticos a partir de imágenes satelitales. En la actualidad ya se han aplicado estas técnicas tanto para la predicción del clima (Das & Ghosh, 2015; Zeng, Hsieh, Shabbar, & Burrows, 2011), así como para el análisis de teleconexiones en donde se han aplicado las técnicas bayesianas tanto a variables climáticas (Torre-Gea et al., 2011) como a índices climáticos (Duc, Rivett, MacSween, & Le-Anh, 2017; Ebert-Uphoff & Deng, 2012). En un estudio realizado en Vietnam sobre precipitación, Duc, Bang, and Quang (2018) usaron el método de promediado bayesiano de modelos. En esta investigación se usaron 3 índices climáticos: Indian Ocean Dipole (IOD), Interdecadal Pacific Oscillation (IPO) y El Niño–Southern Oscillation (ENSO). Los resultados mostraron que estos índices tienen incidencia sobre las precipitaciones en Vietnam, y la fuerza con la que estos índices afectan al clima varía según sea el índice, la región y la estacionalidad. Además, se encontró que la interacción entre los índices también influye sobre las precipitaciones.

En el Ecuador existen investigaciones sobre las teleconexiones climáticas (Vicente-Serrano et al., 2017), sin embargo, hasta la fecha no se conocen estudios en los que se haya realizado análisis multivariado considerando además la interacción entre los mismos. Así, el objetivo de este trabajo es realizar un estudio multivariado utilizando redes bayesianas para determinar la influencia de Índices Climáticos sobre las precipitaciones en el país. El aplicar un enfoque bayesiano resulta útil para determinar si la influencia de un índice climático es homogéneo en todo el país o varía por región, así como identificar interacciones entre diferentes índices. Se espera que los resultados de este estudio sirvan para comprender de mejor manera las precipitaciones en nuestro país, y ayude a tomar decisiones sobre el recurso hídrico basadas en evidencias.

Metodología

Área de estudio

La zona de estudio es el Ecuador continental, que cuenta con una superficie aproximada de 250 000 km^2 y tres regiones claramente definidas: las planicies de la costa, los andes y la amazonía. El clima de Ecuador está influenciado por una gran cantidad de factores cuya influencia cambia según la región, y lo que resulta en una gran variabilidad espacio-temporal de la precipitación.

Materiales

Modos de variabilidad climática e índices climáticos

La variabilidad climática es el resultado de la combinación de patrones espaciales. Los patrones espaciales más destacados son conocidos como modos de variabilidad climática y tienen la capacidad de incidir en el clima a escala espacial y temporal, y también pueden afectar a otros modos de variabilidad. Los modos de variabilidad climática son identificados a través del estudio de las teleconexiones espaciales (Blunden et al., 2011).

Entre los modos de variabilidad climática más comunes tenemos: El Niño-Southern Oscillation (ENSO), Central Pacific El Niño (Modoki), Pacific Decadal and Interdecadal Varibility (PDO), North Atlatic Oscillation (NAO), Pacific/North America atmospheric teleconnection (PNA), etc. Los modos de variabilidad son definidos a través de índices climáticos, por ejemplo, ENSO está definido por varios índices como son: NIÑO3, NIÑO3.4, SOI entre otros. Los cambios en los índices están asociados a cambios a gran escala en los campos climáticos. Debido a que ningún índice puede aislar un fenómeno de todos los demás efectos, múltiples índices definen el mismo fenómeno climático, y por tanto cada índice se ve afectado por diversos fenómenos climáticos (Blunden et al., 2011).

Los índices climáticos a utilizar en este trabajo son:

Índice	Nombre	Acceso
Relacionados con ENSO		
Niño 1+2	El Niño región 1+2	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Niño 3	El Niño región 3	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Niño 4	El Niño región 4	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Niño 3.4	El Niño región 3,4	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EMI	El Niño Modoki	http://www.jamstec.go.jp/frsgc/research/d1/iod/e/elnmodoki/about_elnm.html
MEI	ENSO multivariado	https://www.esrl.noaa.gov/psd/data/climateindices/list/
ESPI	Monthly ENSO Precipitation Index	https://www.esrl.noaa.gov/psd/enso/dashboard.html
Relacionados con el Océano Pacífico		

TNI	Trans-Niño	https://www.esrl.noaa.gov/psd/data/climateindices/list/
ONI	Niño oceánico	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Relacionados con Océano Atlántico		
TNA	Tropical Northern Atlantic	https://www.esrl.noaa.gov/psd/data/climateindices/list/
TSA	Tropical Southern Atlantic	https://www.esrl.noaa.gov/psd/data/climateindices/list/
CAR_ersst	Caribbean SST Index	https://www.esrl.noaa.gov/psd/data/climateindices/list/
NTA_ersst	North Tropical Atlantic Index	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Amon	Atlantic multidecadal Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Ammsst	Atlantic Meridional Mode	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Relacionados con índices atmosféricos		
AO	Artic Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
AAO	Antarctic Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
QBO	Quasi-Biennial Oscillation	https://climatedataguide.ucar.edu/climate-data/qbo-quasi-biennial-oscillation
SOI	Southern Oscillation Index	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Glaam	Globally Integrated Angular Momentum	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Relacionados con teleconexiones		
PDO	Pacific Decadal Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EA	East Atlantic	https://www.esrl.noaa.gov/psd/data/climateindices/list/
WP	Western Pacific	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EP/NP	East Pacific North Pacific	https://www.esrl.noaa.gov/psd/data/climateindices/list/
PNA	Pacific North American	https://www.esrl.noaa.gov/psd/data/climateindices/list/
NAO	North Atlantic Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EPO	Eastern Pacific Oscillation	https://www.esrl.noaa.gov/psd/data/timeseries/daily/EPO/
Otros		
gmsst	Global Mean Lan/Ocean Temperature Index	https://www.esrl.noaa.gov/psd/data/climateindices/list/

Datos satelitales con reducción de escala

Se utilizaron imágenes satelitales mensuales de precipitación TRMM con reducción de escala a 5km (2001 - 2011). Para la reducción de escala se utilizaron covariables de imágenes satelitales. Para más información consultar (Ulloa, Ballari, Campozano, & Samaniego, 2017).

Métodos

Este trabajo se centra en la construcción e interpretación de una red bayesiana para identificar la relación entre regiones de estacionalidad de precipitación e índices climáticos. Los procedimientos de preparación de datos reportados a continuación fueron generados en el marco del proyecto CEPRAXII "Representación espacial de las teleconexiones climáticas en la precipitación del Ecuador", en el cual también se enmarca este trabajo de titulación. La figura 1 resume los pasos del proceso.

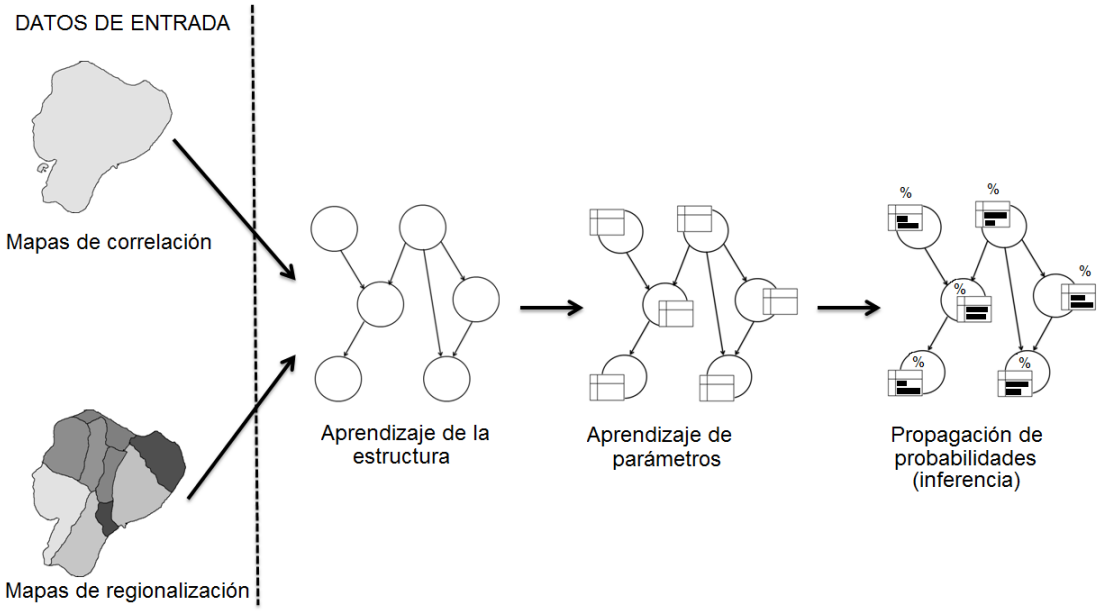


Figura 1: Pasos para la creación de una red bayesiana. Basado en (Rodríguez & Dolado, 2007).

1. Preparación de datos

a. Mapas de correlación

Ante la falta de normalidad de series temporales, se utilizó correlación no paramétrica de Spearman pixel a pixel entre la serie temporal de precipitación, correspondiente al pixel en cuestión, y los diferentes índices climáticos. La figura 2 muestra los mapas resultados de dichas correlaciones, con azul las correlaciones positivas y con rojo las correlaciones negativas. Los valores altos de correlación tanto positiva como negativa evidencian la existencia de teleconexiones climáticas.

b. Mapa de regionalización de estacionalidades de precipitación a 5km

Con las imágenes de precipitación a 5km se ejecutó el procedimiento de regionalización funcional reportado en el artículo (Ballari, Giraldo, Campozano, & Samaniego, 2018). Dicho

artículo se basó en imágenes TRMM de 27km y se encontraron 5 regiones de estacionalidad de precipitación. En este caso, aplicando la regionalización sobre las imágenes de 5km, se encontraron 9 regiones principales de estacionalidad. La figura 3 muestra las 9 regiones, de las cuales las regiones 1, 2 y 5 se localizan en la costa, las regiones 3, 7 y 8 en la sierra, y las 4, 6 y 9 en la amazonía. Es de destacar que la región 1 se caracteriza por estar compuesta de dos regiones, una en la estribación occidental de la sierra y la otra en la estribación oriental. Ambas se categorizaron como la misma región ya que presentan similitudes en la estacionalidad de la precipitación.

Las 9 regiones se pueden clasificar dentro de las regiones naturales del Ecuador (costa, sierra y amazonía) según su ubicación ya que abarcan superficies predominantes de alguna de estas tres. Sin embargo, existen dos casos especiales que por abarcar áreas de más de una región natural en proporciones casi iguales, su clasificación se torna difícil; este es el caso de la región 1 (que abarca costa, sierra y una pequeña parte de la amazonía) y de la región 9 (que abarca sierra y amazonía). Para efectos prácticos se decidió clasificar a la región 1 como costa y a la 9 como amazonía, sin embargo debemos recordar que se trata de regiones especiales y por tanto su categorización puede resultar ambigua.

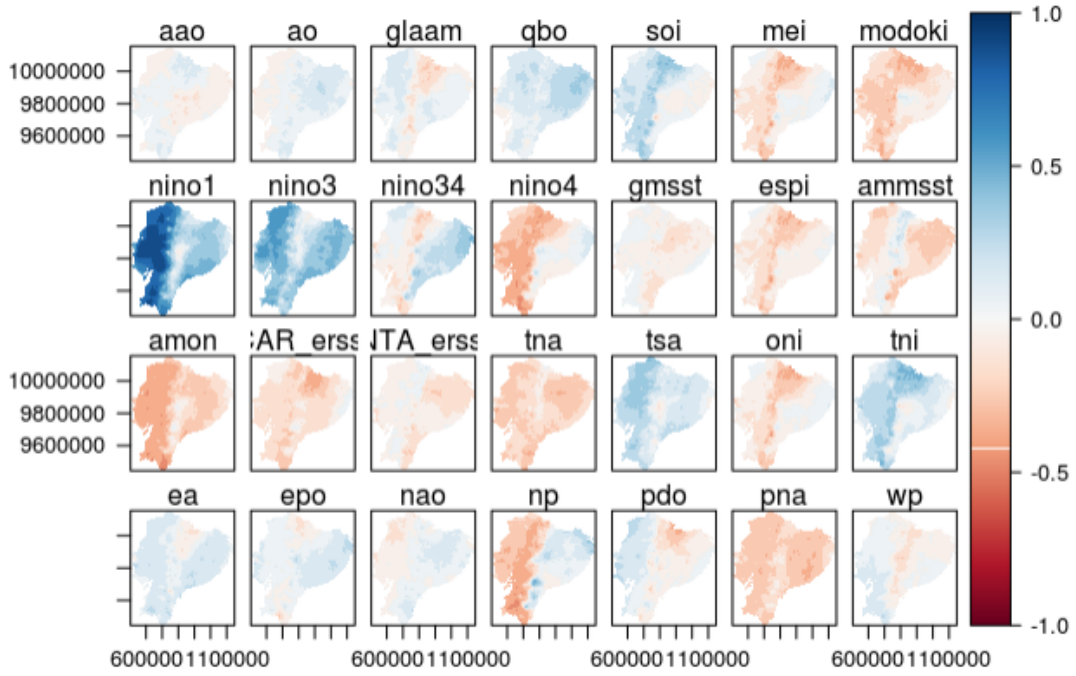


Figura 2: Mapas de correlación.

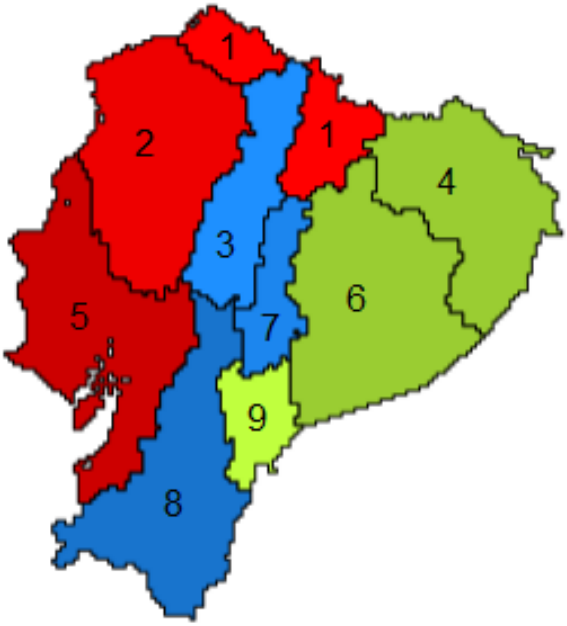


Figura 3: Mapa de regionalización a 5km.

Región	Descripción
1	Costa norte
2	Costa central
3	Sierra norte
4	Amazonía norte
5	Costa sur
6	Amazonía central
7	Sierra centro este
8	Sierra sur
9	Amazonía sur

Tabla 1: Descripción de las regiones de estacionalidad

c. Relación de regionalización y mapas de correlación

Cada pixel del mapa de regionalización se relaciona con los valores de correlación de los índices climáticos. Así se construye una tabla que será utilizada para el aprendizaje de la estructura y de los parámetros de la red bayesiana. El 80% de los registros de dicha tabla se seleccionan aleatoriamente para conformar el grupo de entrenamiento (muestreo aleatorio estratificado por región de estacionalidad), mientras que el 20% restante se utilizará como validación. La figura 4 muestra un extracto de la tabla, donde la primera columna de *layer* corresponde a la región (en este caso 1), y el resto de columnas son los valores de correlación de cada índice climático.

	layer <dbl>	aao <dbl>	ao <dbl>	glaam <dbl>	qbo <dbl>	soi <dbl>	mei <dbl>	modoki <dbl>	nino1 <dbl>
1	1	-0.040	-0.040	0.060	0.144	0.107	-0.044	-0.183	0.634
2	1	-0.059	-0.053	0.044	0.128	0.095	-0.031	-0.161	0.631
3	1	-0.052	-0.032	0.063	0.115	0.091	-0.027	-0.150	0.658
4	1	-0.044	-0.039	0.040	0.127	0.115	-0.056	-0.175	0.647
5	1	-0.052	-0.047	0.045	0.131	0.112	-0.054	-0.178	0.631
6	1	-0.065	-0.054	0.034	0.125	0.117	-0.065	-0.184	0.624
7	1	-0.038	-0.060	0.019	0.130	0.133	-0.079	-0.207	0.620
8	1	-0.051	-0.025	0.077	0.104	0.101	-0.023	-0.154	0.671
9	1	-0.044	-0.026	0.072	0.122	0.103	-0.028	-0.159	0.672
10	1	-0.041	-0.030	0.053	0.130	0.123	-0.049	-0.177	0.665

Figura 4: Extracto de datos para aprendizaje de la red bayesiana

2. Red Bayesiana

Conceptos

Una red bayesiana es un tipo de modelo gráfico probabilístico que se usa para representar el conocimiento sobre un dominio incierto. Se trata de un grafo acíclico dirigido (DAG), en la que los nodos representan una variable aleatoria y los arcos representan dependencias probabilísticas entre esas variables. Para determinar las dependencias se utilizan métodos estadísticos o computacionales (Ben-Gal, 2009).

Estas redes probabilísticas son construidas a partir de la distribución de probabilidad y usan las leyes de probabilidad para la predicción y detección de anomalías, razonamiento y diagnóstico, toma de decisiones bajo incertidumbre, predicción de series de tiempo y otros escenarios en donde no se conozca la probabilidad a priori (Bayesserver, 2018).

La red de la figura 5 es conocida como “red bayesiana de Asia”, es un ejemplo desarrollado por Lauritzen y Spiegelhalter en el año 1988 y es usando comúnmente para explicar de manera simple las redes bayesianas. Se trata de un modelo que podría ser usado para diagnosticar ciertas patologías en un paciente, cada nodo corresponde a alguna condición que el paciente puede o no presentar. Los arcos corresponden con la existencia de relación de probabilidad entre 2 nodos, es decir, si una persona es fumadora, esto incrementará las probabilidades de padecer cáncer o bronquitis (Norsys, 2018). En este ejemplo se observa tanto relaciones de dependencia e independencia condicional, por ejemplo los nodos “Visit to Asia” y “Smoker” son marginalmente independientes, pero cuando se establece el nodo “Tuberculosis or Cancer” se vuelven condicionalmente dependientes (ya que la ocurrencia de uno de los dos eventos afecta a la probabilidad de que ocurra el otro).

Cuando se estable el nodo “Smoker” como evidencia, los nodos “Has Lung Cancer” y “Has Bronchitis” se vuelve condicionalmente independientes (la ocurrencia de uno de los dos eventos no afecta a la probabilidad de que ocurra el otro).

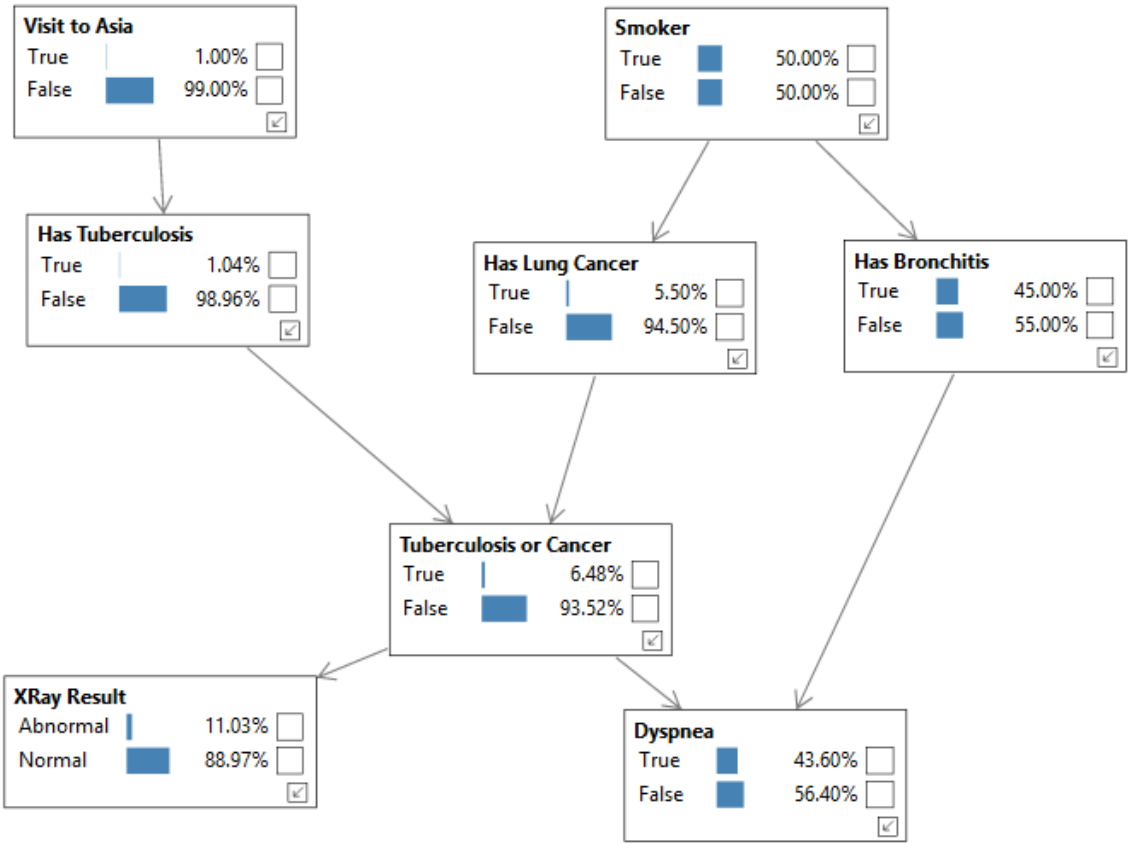


Figura 5: Ejemplo de una red bayesiana. Tomado de (Bayesserver, 2018)

En este trabajo, los nodos serán conformados por las regiones de estacionalidad y las correlaciones de la precipitación con cada uno de los índices climáticos (i.e. teleconexiones climáticas). Los arcos que conformen la estructura de la red serán dependencias probabilísticas entre las regiones y las teleconexiones climáticas.

Paso 1. Discretizar datos

Debido a que la mayoría de herramientas y algoritmos de aprendizaje se basan en nodos discretos, antes de realizar el aprendizaje de la red bayesiana se deben discretizar los datos. En este caso los datos de correlación ente la precipitación y los índices climáticos son de tipo continuo y, por ello, se discretizan. Las 9 regiones de estacionalidad, sin embargo, son discretas.

La discretización o agrupación de datos es dividir el rango de variables continuas en un número finito de intervalos exclusivos. Una consecuencia de la discretización es la pérdida de información que depende del dominio y número de intervalos seleccionados (Rodríguez & Dolado, 2007). La discretización también se puede aplicar cuando una o más variables presentan desviaciones o sesgos muy grandes.

Existen varios métodos de discretización (Das & Vyas, 2010), sin embargo, el método empleado para el presente trabajo se categoriza dentro de los “métodos no supervisados”. Estos utilizan la distribución de valores de la variable continua como única fuente de información (López, 2007). Dentro de esta clasificación tenemos al método EWD (Equal Width Discretization) que divide el rango de un atributo en un número k de intervalos de igual

tamaño, siendo el número de intervalos determinado por el usuario. Para determinar el intervalo se debe aplicar la siguiente formula:

$$w = \frac{max - min}{k}$$

En donde:
w = valor de intervalo
max = valor máximo
min = valor mínimo
k = número de intervalos

El presente trabajo utiliza valores de correlación, por tanto, estos varían desde -1 hasta 1. Los valores desde -1 hasta 0 corresponden a correlaciones negativas y aquellos desde 0 hasta 1 corresponden a correlaciones positivas. Se desea obtener 3 intervalos para cada tipo de correlación (bajo, medio y alto), debido a que el rango es el mismo para la correlación positiva y negativa, será suficiente aplicar la fórmula a uno de estos 2 rangos y luego usar este valor para la otra parte.

$$w = \frac{1 - 0}{3}$$

$$w = 0.333$$

De manera que los rangos quedan de la siguiente forma:

[0 – 0,3): Bajo
[0,3 – 0,6): Medio
[0,6 – 1]: Alto

De igual manera para los valores negativos:

[-1 - -0,6]: Alto
(-0,6 - -0,3]: Medio
(-0,3 - 0): Bajo

Paso 2. Aprendizaje de la red

El aprendizaje de una red bayesiana se trata de obtener un modelo tanto de estructura así como parámetros partiendo de datos. Esto se divide en: aprendizaje estructural que consiste en obtener la estructura o topología de la red (identificación de nodos, arcos y direcciones de los arcos), y aprendizaje paramétrico en el cual dada la estructura, se debe obtener las probabilidades asociadas a cada nodo (Sucar, 2013).

El aprendizaje de la estructura y parámetros de la red bayesiana se realizó con la librería bnlearn de R.

Aprendizaje de la estructura

Consiste en identificar la estructura gráfica de la red. Idealmente debería ser el mapa mínimo de la estructura, o en su defecto, una distribución apropiada en el espacio de probabilidad. Los

algoritmos existentes para el aprendizaje de la estructura se clasifican en 3 categorías (Nagarajan et al., 2013):

- Los algoritmos basados en restricciones se basan los trabajos de Pearl en los mapas y su aplicación a modelos gráficos causales. A partir del algoritmo de Pearl se ha desarrollado otros algoritmos mejorados tales como: PC, growk-shrink, asociación incremental, asociación incremental rápida y asociación incremental entrelazada. Todos estos algoritmos (excepto PC) aprenden primero el manto de Markov de cada nodo en la red, y como consecuencia, se reduce la complejidad computacional.
- Los algoritmos basados en puntajes son el resultado de una aplicación de técnicas de optimización heurística, a cada red candidata se le asigna una puntuación que refleje su calidad de ajuste. Ejemplos de estos algoritmos son: algoritmos de búsqueda codiciosos los cuales comienzan explorando desde una estructura de red generalmente vacía y desde aquí agrega, elimina o invierte un arco hasta que la puntación ya no pueda mejorarse; los algoritmos genéticos que emulan la evolución natural a través de la selección iterativa de los modelos más aptos y la combinación de sus características; y el algoritmo de recocido simulado que realiza una búsqueda local estocástica al aceptar cambios que aumentan la puntación de la red.
- Los algoritmos híbridos combinan lo mejor de los algoritmos basados en restricciones y basados en puntajes para compensar debilidades. Los más conocidos dentro de esta categoría son el algoritmo *Sparse Candidate* (SC) y el algoritmo *Max-Min- Hill-Climbing* (MMHC). Ambos algoritmos se basan en dos pasos denominados restringir y maximizar, en el primero el conjunto de candidatos para los padres de cada nodo Xi se reduce de la forma en que está relacionado de alguna manera con el de Xi, el segundo paso busca la red que maximiza una función de puntaje determinada, sujeto a las restricciones impuestas por los conjuntos de Ci. Estos pasos se aplican de manera iterativa en el algoritmo *Sparse Candidate* hasta que no haya cambios en la red o no pueda ser mejorada, en cambio, en el algoritmo MMHC estos pasos se realizan una sola vez.

Para este trabajo se seleccionó el método basados en puntajes *Hill-Climbing* (hc) por ser uno de los métodos frecuentemente utilizados (Sachs, 2005). *Hill-climbing* es un algoritmo de búsqueda local (esto significa que no siempre encontrará la solución óptima, si es que la hubiera), que consiste en una iteración que se mueve en la dirección que incremente su valor. Termina cuando ha alcanzado un pico y ningún vecino tiene un valor más alto. *Hill-climbing* no mira más allá de sus vecinos inmediatos (Russell & Norvig, 2010). En esta búsqueda se pueden presentar diferentes situaciones, que se presentan introducirá través de las diferentes regiones que podemos encontrar en el diagrama de estado-espacio (figura 6):

- *Global maximum* (Máximo global): Es el mejor estado posible dentro del diagrama.
- *Local maximum* (Máximo local): Es un estado que es mejor que sus vecinos, pero existe al menos 1 que es mejor que este (máximo global).
- *Flat local maximum* (Máximo local plano): Es una región plana, por tanto, los vecinos tienen el mismo estado.
- *Current state* (Estado actual): El estado desde el cual se da inicio la búsqueda.

- *Shoulder*: Es una región plana que presenta una pendiente positiva en uno de sus extremos.

El ejemplo escogido (figura 6) es unidimensional, de esta forma se simplificará la explicación. Como es lógico inicia la búsqueda en el estado actual, debido a que existe una pendiente positiva, el algoritmo avanza en esta dirección; continuará avanzando hasta que se encuentre en el máximo local. En este punto la búsqueda se detendrá y lanzará este estado como solución. En este caso no fue capaz de encontrar el máximo global debido a no era posible llegar desde el estado inicial.

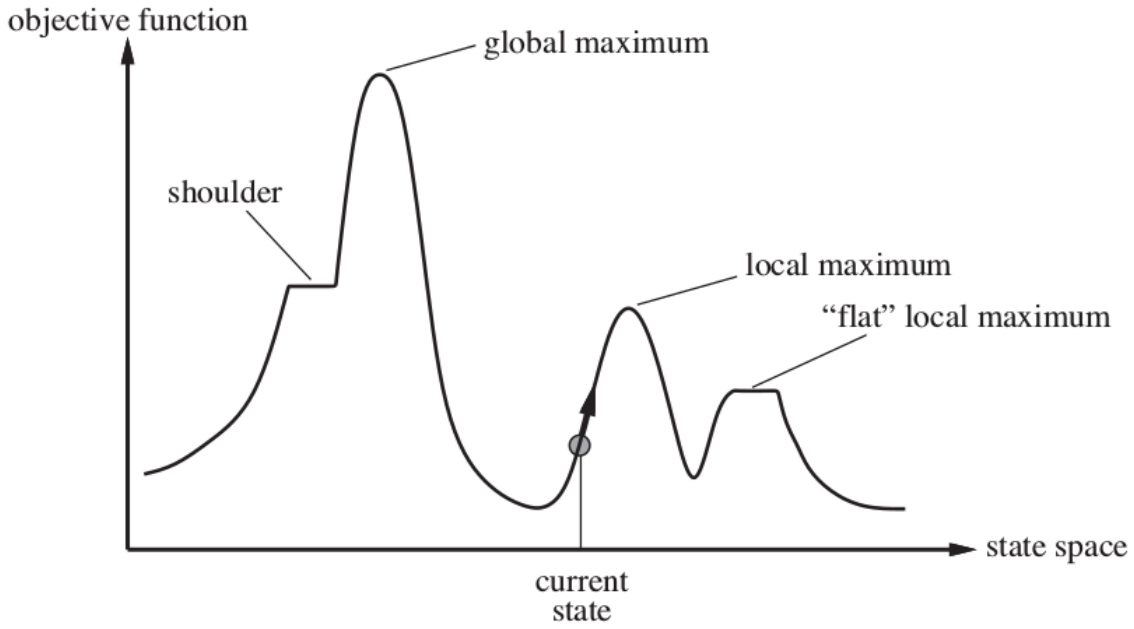


Figura 6: Ejemplo de búsqueda del algoritmo *hill-climbing*. Tomado de (Russell & Norvig, 2010, pp. 121)

El algoritmo *hill-climbing* seleccionará una estructura de red entre varias que fueron generadas (específicamente 100 redes). Para generar esta gran cantidad de redes fue necesario aplicar un método de remuestreo estadístico, el método empleado fue *bootstrapping*. El *bootstrapping* es una técnica estadística que consiste en obtener un elevado número de muestras a partir de la muestra inicial (*resampling*) con el fin de estudiar el comportamiento de determinados estadísticos (Flores, 2005).

Una vez seleccionada la estructura se emplea la función "averaged.network" para seleccionar únicamente aquellos arcos cuyo valor de fuerza es mayor a un valor de umbral (*threshold*) definido por el usuario. Para este trabajo el valor de *threshold* fue de 0.6

Criterios de selección de estructura

A pesar de que en los últimos años las redes bayesianas han ganado mucho protagonismo, la identificación de la estructura a partir de los datos sigue siendo un problema. Esto es debido a que el número de estructuras posibles se incrementa exponencialmente con el número de

nodos. Una forma de abordar este problema es a través de la búsqueda heurística basado en la optimización de las métricas de puntuación, que se realiza en algún espacio de búsqueda. Existen muchos algoritmos como por ejemplo *hill-climbing*, algoritmos genéticos, búsqueda tabú, etc. En todos estos algoritmos la búsqueda está basada en una función de puntuación que evalúa el grado de adecuación de la red (Carvalho, 2009). Existen muchas funciones de puntajes para redes bayesianas, entre las más comunes tenemos: *Akaike's information criterion* (AIC), *Bayesian information criterion* (BIC), *Bayesian Dirichlet equivalence score* (BDe). Para este trabajo se seleccionó la función BDe debido a que con esta función se obtuvo el mejor puntaje de red.

Aprendizaje de parámetros

Cuanto se cuentan con datos completos y suficientes, y asumiendo que la estructura ya ha sido obtenida, obtener los parámetros se realiza a través del método denominado: "estimador de máxima verosimilitud" (MLE), con esto se obtiene un aproximado de las probabilidades en base a las frecuencias de los datos (Sucar, 2013).

En muchas ocasiones cuando los datos no están completos. Existen dos tipos básicos de información incompleta: valores faltantes y nodos ocultos, las alternativas para el primer caso pueden ser (Sucar, 2013):

- Tomar un nuevo valor adicional para la variable, como "desconocido".
- Excluir registros con valores faltantes.
- Tomar el valor más probable basado en las otras variables.
- Considerar el promedio de la variable.
- Tomar la probabilidad de los diferentes valores en base a las otras variables.

Un nodo oculto es aquel en el que le faltan todos los valores de la variable, sin embargo todavía es posible estimar las tablas de probabilidad condicional usando el algoritmo de "maximización de la expectación". Este algoritmo consiste de dos pasos que se repiten de forma iterativa (Sucar, 2013):

- Paso E: Aproximar los datos que faltan, basado en los parámetros actuales.
- Paso M: Aproximar las probabilidades (parámetros) teniendo en cuenta los datos estimados.

Cuando se trata de nodos ocultos en redes bayesianas, se emplea el algoritmo EM de la siguiente manera (Sucar, 2013):

1. Iniciar los parámetros desconocidos con valores establecidos al azar.
2. Emplear datos conocidos con los parámetros actuales para calcular los valores de variables ocultas.
3. Usar los valores calculados para llenar la tabla de datos.
4. Volver a estimar los parámetros con los nuevos datos obtenidos.
5. Repetir los pasos 2, 3 y 4 hasta que no exista cambios importantes en las probabilidades.

Para este trabajo se seleccionó el estimador de máxima verosimilitud (MLE) que la librería bnlearn implementa para datos completos y discretos. Además, no se encontraron nodos ocultos.

Como resultado del aprendizaje de parámetros se obtienen tablas de probabilidad condicional y marginal. La probabilidad marginal se define como aquella probabilidad de un evento, es decir, que no depende de otro, mientras que la probabilidad condicional es la probabilidad de que suceda un evento "A" dado que ya sucedió un evento "B".

Para ejemplificar esto, consideramos nuevamente la red bayesiana de Asia. Luego de realizar el aprendizaje paramétrico se obtuvo que la probabilidad de que una persona haya contraído bronquitis es de 1,04% (probabilidad marginal), mientras que la probabilidad de que una persona haya contraído bronquitis dado que visitó Asia es de 5% (probabilidad marginal). Esto se puede apreciar en la figura 7.

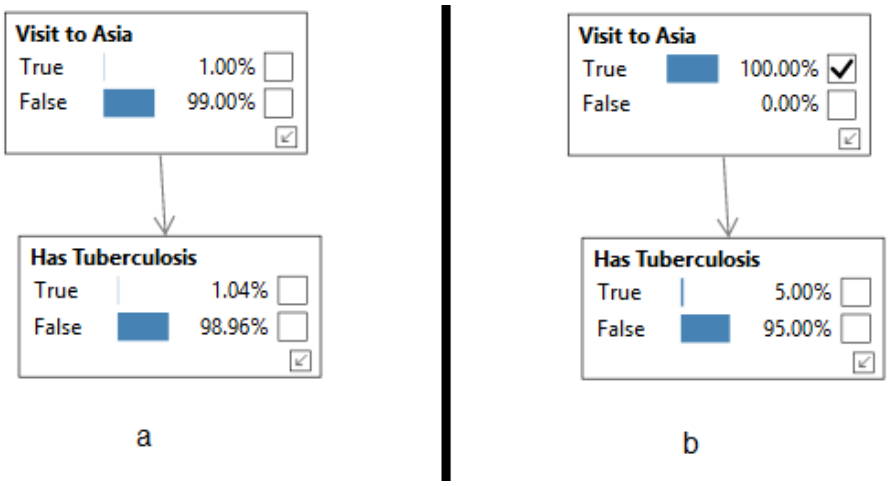


Figura 7: Tablas de probabilidad. A) Probabilidad marginal. B) Probabilidad marginal luego de propagar un nodo.

Propagación de probabilidades

Uno de los puntos fuertes de las redes bayesianas es que proporcionan un sistema de inferencia. Con ello se modifican las tablas de probabilidad una vez que se observa o selecciona nueva evidencia sobre el estado de ciertos nodos, y a su vez, estas nuevas probabilidades se propagan hacia el resto de nodos. Esto se conoce con el nombre de inferencia probabilística. Las probabilidades antes de establecer evidencia son llamadas probabilidades a priori; aquellas en las que ya se ha ingresado evidencias, se llaman probabilidades a posteriori (Rodríguez & Dolado, 2007).

El razonamiento probabilístico se trata de propagar los efectos causados por la evidencia por toda la red y determinar la probabilidad a posteriori en las variables. Por tanto, se establecen valores en determinadas variables, y se observa la probabilidad a posterior de las variables que no han sido establecidas como evidencia (Sucar, 2013). Existen 2 tipos de propagación:

forward y *backward*. Se explicará estos conceptos usando el ejemplo de la red bayasiana de Asia.

Sabemos que la probabilidad *a priori* de padecer disnea es de un 43,6%, ¿qué sucedería con esta probabilidad si conocemos que el paciente tiene bronquitis? Al establecer el nodo "Has Bronchitis" como verdadero (evidencia) observamos que la probabilidad de padecer disnea aumentó de 43.6% hasta un 80% (Figura 8), esto sucede porque una vez que se conoce nueva información, esta cambia las probabilidades no solo del nodo en el que se estableció sino de todos aquellos con los que tiene conexión (propagación de probabilidades). En este caso particular se realizó una propagación *forward* ya que el nodo padre se estableció como evidencia (Tiene bronquitis) y esta se propagó hacia los hijos (disnea).

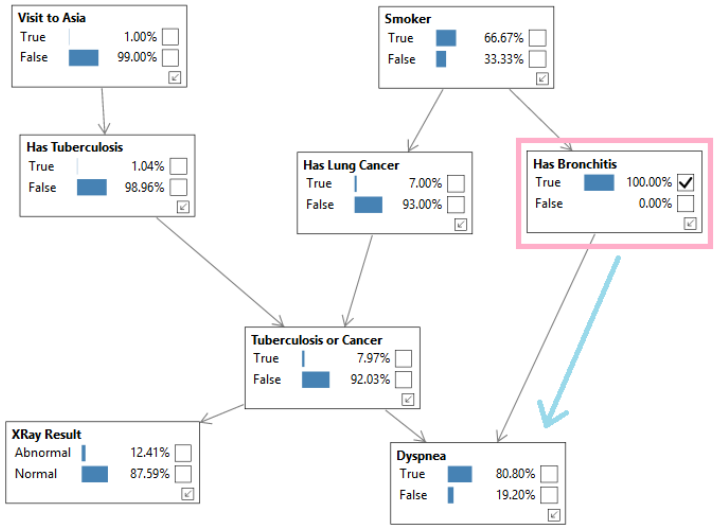


Figura 8: Propagación forward. Tomado de (Bayesserver, 2018)

¿Qué sucedería si lo que se tiene como evidencia es un nodo hijo? En este caso se produciría una propagación *backward* ya que las probabilidades se propagan hacia el o los nodos padres. Un ejemplo de esto podemos observar en la figura 9 en donde el nodo "Has Lung Cancer" es la evidencia y se observa que las probabilidades del nodo padre "Smoker" pasó de un 50% a un 90%.

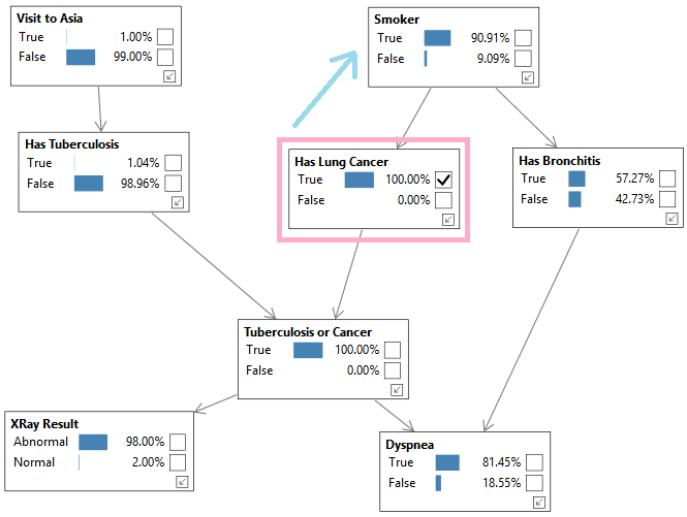


Figura 9: Propagación backward. Tomado de (Bayesserver, 2018)

Un problema que surge de la propagación es su coste computacional cuando existen muchos nodos y dependencias. Por ejemplo, una red con n variables booleanas contiene 2^n valores siendo un problema NP-complejo. Gracias al desarrollo de nuevos algoritmos de propagación ha sido posible usar redes capaces de modelar problemas reales. Los algoritmos de propagación se dividen en dos grandes grupos (Rodríguez & Dolado, 2007):

- Los algoritmos de propagación exactos son aquellos en los que no existe error en las probabilidades calculadas, ejemplos de estos son: inferencia mediante enumeración y eliminación de variables, estos funcionan bien con redes de hasta aproximadamente 30 nodos, no son apropiados para redes con mayor cantidad de nodos o altamente conexas.
- Los algoritmos aproximados son aquellos en los que existe cierto margen de error en las probabilidades estimadas, se clasifican en métodos de simulación estocástica (también llamados Montecarlo) y los métodos de búsqueda determinista, los primeros realizan la inferencia por medio de muestreos de números aleatorios que dependen de las probabilidades de la red. Las probabilidades condicionales se calculan dependiendo de las frecuencias generadas en el muestreo.

Para este trabajo se usó propagación exacta ya que esta se realiza a través de la librería gRain (Højsgaard, 2016), la cual emplea el algoritmo de algoritmo de Lauritzen y Spiegelhalter (Højsgaard, 2012). Para más información sobre el algoritmo consultar (Lauritzen & Spiegelhalter, 1988).

La librería gRain dispone de herramientas para leer la estructura y parámetros previamente desarrollos con bnlearn (Scutari, 2019). Además, permite realizar inferencias hacia delante o forward (es decir hacia los hijos) y hacia atrás o backward (es decir hacia los padres).

Se realizó la propagación de probabilidades usando tres enfoques diferentes:

1) De regiones a índices (forward)

2) De índices a regiones (backward)
3) Interacción entre índices

La propagación de regiones a índices permite conocer los índices que influyen en la precipitación de una determinada región, la propagación de índices a regiones permite determinar las regiones en las que un determinado índice tiene influencia, y por último, la interacción entre índices permite analizar "relaciones" entre índices, es decir, como los niveles de correlación con un índice puede afectar a otro.

Para facilitar la visualización de la red bayesiana, la red obtenida en R se exportó hacia un software que permita representar gráficamente el aprendizaje realizado, el software utilizado fue UnBBayes (UnBBayes, 2019).

Validación

Un modelo es una representación abstracta y simplificada de un fenómeno del mundo real, con el que se pretende dar explicaciones sobre dicho fenómeno. Una simplificación implica que pueden existir ciertas variables que influyen sobre el fenómeno pero que por algún motivo, no fueron consideradas a la hora elaborar el modelo. Debido a que todo modelo generado puede contener errores (o ha sido muy simplificado) es necesario realizar una validación del mismo. La validación es el proceso por el cual se determina si el modelo obtenido es apto para el propósito indicado. Es importante demostrar que el modelo es idóneo ya que por lo general se toman decisiones basadas en los resultados que estos proporcionan. Naturalmente si los resultados son erróneos, también lo serán las decisiones.

Existen muchas herramientas para realizar validación, la matriz de confusión es una de ellas y es bastante usada en el campo de "machine learning". Esta consiste en una matriz cuadrada en las que las filas toman los nombres de acuerdo a la cantidad de clases reales y las columnas de acuerdo a las clases pronosticadas en el modelo. Es útil para poder contrastar cuando una clase es confundida con otra (Recuero de los Santos, 2018). Para realizar la validación en este trabajo, se usó la función "confusionMatrix" de la librería "caret" (Kuhn, 2019). La figura 10 ilustra un ejemplo de matriz de confusión cuyos valores de clase se corresponden a valores binarios (positivo y negativo) pudiendo existir más de 2 tipos de valores. Los valores coloreados en la diagonal principal corresponden a aquellos que han sido clasificados correctamente, todos los demás han sido clasificados erróneamente.

Matriz de confusion		Valores predichos	
		Negativo	Positivo
Valores reales	Negativo	a	b
	Positivo	c	d

Figura 10: Matriz de confusión

La matriz de confusión permite diferenciar los distintos tipos de error que se pueden presentar, esto gracias a la variedad de métricas que ofrece tales como sensibilidad, especificidad, exactitud, precisión, porcentaje de acuerdo, coeficiente kappa, entre otros. Algunas de ellas se detallan a continuación

Precisión (*Accuracy* en inglés): Se define como el cociente entre el total de predicciones correctas y el total de predicciones, se usa para medir el grado de dispersión de los datos. Su fórmula es:

$$Accuracy = \frac{\sum(registros\ correctamente\ clasificados)}{\sum(total\ de\ registros)}$$

Porcentaje de acuerdo (PA): "Es la ratio entre el total de elementos correctamente clasificados (celdas de la diagonal principal), y el total de elementos en la matriz" (Ariza-López, Rodríguez-Avi, & Alba-Fernandez, 2018).

$$PA = \frac{1}{N}\sum_{i=1}^{M} n_{i,i} = \sum_{i=1}^{M} p_{i,i}$$

Donde,

M representa el número de clases
N expresa el número total de muestras (número de datos).
n_{ii} representa el número de casos en la diagonal.

Coeficiente de acuerdo aleatorio a posteriori: "Representa el porcentaje de acuerdo que cabe esperar al azar teniendo en cuenta que unas clases contienen un mayor número de celdillas que otras y que por lo tanto son más probables de estar bien clasificada" (Sánchez Muñoz, 2016).

$$Ca_{ps} = \frac{1}{N^2}\sum_{i=1}^{M} n_{i+}\, n_{+i}$$

Donde,
M representa el número de clases.
N expresa el número total de muestras (número de datos).
n_{i+} representa la suma de todos los elementos de la fila i de la matriz N.
n_{+i} la suma de todos los elementos de la columna i de la matriz N.

Coeficiente Kappa: es una medida basada en la diferencia entre el porcentaje de acuerdo indicado por los valores de la diagonal principal y el acuerdo aleatorio a posteriori, estimado a partir de los valores marginales (totales de las filas y columnas) (Ariza-López et al., 2018).

$$k = \frac{Pa - Ca_{ps}}{1 - Ca_{ps}}$$

Donde,

Pa representa el porcentaje de acuerdo
Ca_{ps} representa el coeficiente de acuerdo a posteriori

Para el presente trabajo se utilizó 2 métricas: precisión y coeficiente kappa.

Todo el código empleado tanto para el aprendizaje de la red así como propagación, predicción y generación de mapas, se encuentra en la sección de anexos.

Resultados

1. Estructura de la red bayesiana

El resultado del aprendizaje estructural se muestra en la figura 11. La estructura obtenida es muy similar a un clasificador bayesiano simple aumentado con una red (BAN, Naive Bayesian Network) en donde el nodo región es la variable clase. Sin embargo, debe notarse que aunque "región" es ancestro de la mayoría de los nodos, esto no significa que sea padre de estos nodos. Es decir, esto no implica una relación de causalidad entre región y los nodos de las teleconexiones climáticas.

Los nodos hijos en color beige son teleconexiones con el Océano Pacífico, y en color azul con el Océano Atlántico. Los nodos en blanco corresponden con teleconexiones globales y de los polos. El grosor de los nodos se relaciona con la fuerza de la relación entre nodos.

Los nodos hijos (inmediatos) de región son:

- Pacífico: niño 3, niño 1+2, np, niño 34, qbo
- Atlántico: ammsst, CAR_ersst, tsa, epo, nao
- Otros: glaam

Para conocer los nodos descendientes de "región" en un segundo nivel, se debe añadir a los nodos anteriores (hijos) los siguientes:

- Pacifico: mei, niño4, oni, espi, pdo, wp.
- Atlántico: ao, ea, NTA_ersst, amon.

Existen, además, 2 nodos que se encuentran totalmente aislados, estos son: tna y pna. Esto indica que no se ha encontrado relación alguna con las regiones o con otros nodos.

El nodo soi no se encuentra conectado directamente con "región", pero se encuentra conectado a través de los indices mei y tni, sin embargo hay que resaltar que si existe una relación considerable entre los nodos soi y región pero que no está representada a través de un arco debido a que el valor de fuerza coincide con el valor de *threshold* (solo se representa aquellos valores mayores al *threshold*). Para futuros trabajos, en caso que se reduzca el valor del *threshold*, nuevas relaciones pueden ser visibles.

Finalmente en la figura 11 también se puede apreciar que los arcos tienen diferente grosor, mostrando la fuerza de la relación entre nodos. Por tanto se puede concluir que la relación es fuerte entre "región" y los nodos: niño1+2, niño34, epo, qbo y ammsst; menos fuerte con los nodos: nao, np, CAR_ersst y niño3; y muy débil con los nodos: glaam y tsa.

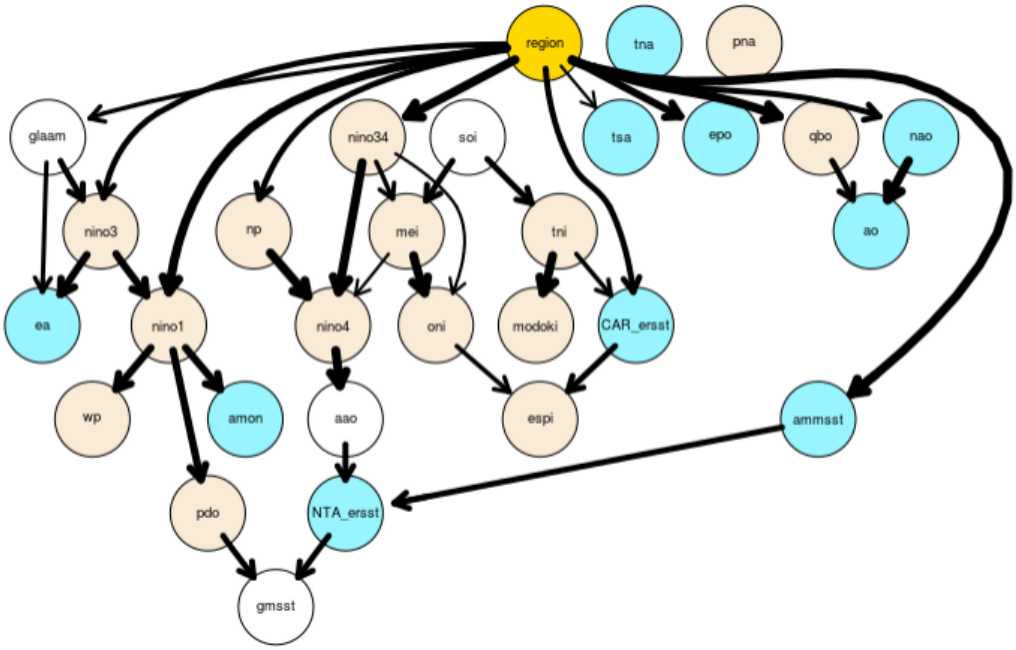


Figura 11. Estructura de una red bayesiana

1.2 Aprendizaje paramétrico

La figura 12 muestra la estructura y parámetros de la red bayesiana obtenida de R y exportada al software UnBBayes. En cada nodo se observa la probabilidad marginal, o también llamada *a priori*. En el caso de región, que consta de 9 clases, se observa la distribución de probabilidad de dichas regiones, o en otras palabras su distribución de frecuencias. En el caso de los nodos de teleconexiones se muestra las clases de correlación medias, bajas y altas, tanto positivas como negativas, y su distribución de probabilidad para todo el país.

Se puede apreciar que hay ciertos índices como modoki, CAR_ersst, espi, ea, NTA_ersst, y qbo cuya probabilidad es muy alta en un determinado rango (cercana al 90% y en algunos casos mayor). Existen también otros índices cuya probabilidad se encuentra distribuida de manera más uniforme entre sus rangos como es el caso de aao, pdo, wp y niño1+2.

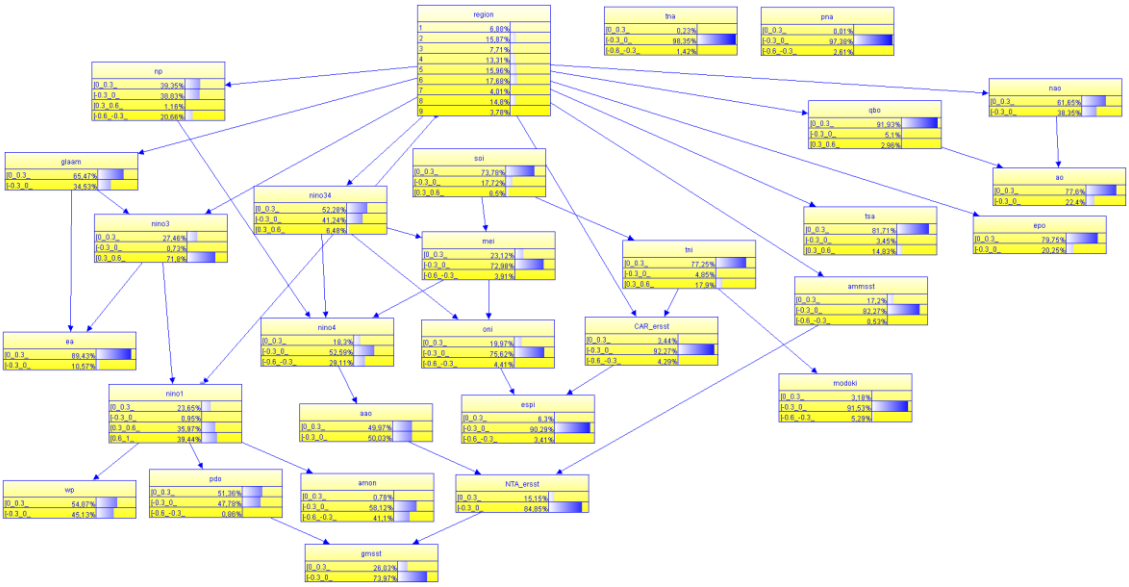


Figura 12. Aprendizaje paramétrico

En la figura 13 se puede apreciar la probabilidad marginal de un nodo de la red (nino34). En la figura 14 se observa la probabilidad marginal del mismo nodo cuando se establece la región 8 como evidencia.

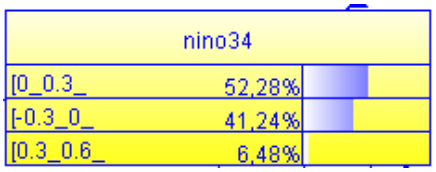


Figura 13: Probabilidad marginal

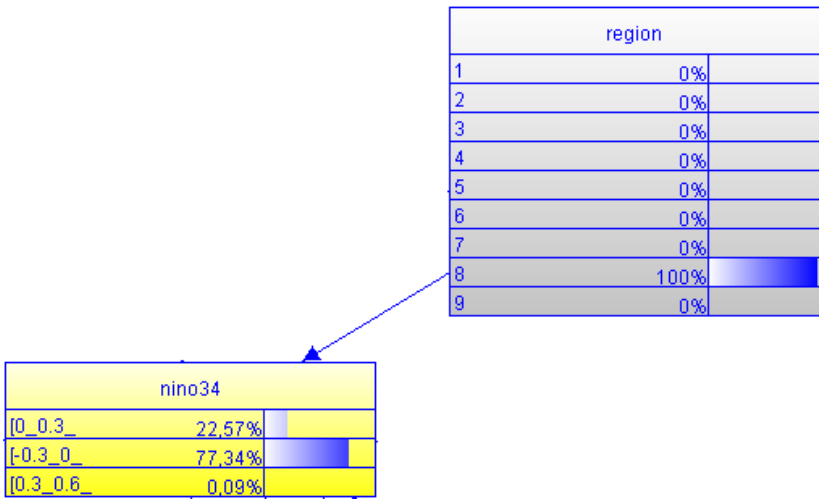


Figura 14: Probabilidad marginal con propagación de probabilidades

2. Propagación

Se realizó los tres tipos de propagación detallados en la sección de métodos. La propagación de regiones a índices permite conocer los índices que influyen en la precipitación de una determinada región, la propagación de índices a regiones permite determinar las regiones en las que un determinado índice tiene influencia, y por último, la interacción entre índices permite analizar "relaciones" entre índices, es decir, como los niveles de correlación con un índice puede afectar a otro.

Los resultados de la propagación de probabilidades se muestran en tablas las cuales constan de seis campos: región, descripción región, índice, probabilidad a priori, probabilidad a posteriori y rango. En el caso de la interacción entre índices los campos son: código, índice padre, rango del índice padre, índice hijo, rango del índice hijo, probabilidad a priori y probabilidad a posteriori, en cada sección se explicará cómo se debe interpretar las tablas.

Debe notarse que en todos los casos únicamente se muestra aquellos índices o regiones cuya probabilidad aumenta más allá de un 1% por considerar a este como un valor mínimo de representatividad. Además, no se tiene en cuenta a los rangos de los índices entre -0.3 y 0.3 ya que la correlación en estos rangos es débil.

Para mayor facilidad de interpretación, los índices fueron clasificados por su origen, es decir, aquellos que se relacionan con el Océano Atlántico fueron coloreados en marrón y aquellos del Océano Pacífico en beige. Existen ciertos índices que por ser globales o porque su zona de influencia no localizada, no fueron clasificados en ninguna de estas categorías. De igual manera la variable "Rango" fue clasificada usando también códigos de colores: rojo para correlación negativa y el azul para positiva.

a. De regiones a índices

En este tipo de propagación, el campo "región" es la variable a la que se le asignó la evidencia, y la variable "índice" son todos aquellos índices cuya probabilidad aumentó como consecuencia de la evidencia establecida. El campo "rango" es el rango de la correlación del índice para el cual existió dicho cambio de probabilidad, y finalmente los campos "probabilidad a priori" y "probabilidad a posteriori" muestran la probabilidad sin evidencia (a priori) y con la evidencia (a posteriori). En este caso, la probabilidad a priori se refiere a la influencia del índice a nivel del país (i.e. contabilización de frecuencia para todo el Ecuador), mientras que la probabilidad a posteriori se refiere a la influencia por región en particular (i.e. contabilización de frecuencia por regiones). La tabla 2 muestra los resultados de la propagación y resalta en negrita los resultados más relevantes cuando la probabilidad a posteriori aumentó más del 10%.

Tabla 2: Propagación de regiones a índices

Región	Descripción región	Índice	Probabilidad a priori	Probabilidad a posteriori	Rango
1	Costa norte	**tsa**	**14,83%**	**26,92%**	[0.3,0.6)
		mei	3,91%	5,59%	[-0.6,-0.3)
		espi	3,41%	4,65%	
		CAR_ersst	**4,29%**	**16,20%**	
		oni	4,41%	6,53%	
		pdo	0,86%	2,08%	
2	Costa central	**nino1+2**	**39,44%**	**94,68%**	[0.6,1)
		nino3	**71,80%**	**93,80%**	[0.3,0.6)
		tsa	**14,83%**	**67,59%**	
		nino4	29,11%	34,11%	[-0.6,-0.3)
		amon	**41,10%**	**88,33%**	
		np	**20,66%**	**30,82%**	
5	Costa sur	**nino1+2**	**39,44%**	**94,31%**	[0.6,1)
		nino3	**71,80%**	**92,42%**	[0.3,0.6)
		mei	3,91%	7,40%	[-0.6,-0.3)
		nino4	**29,11%**	**61,83%**	
		espi	3,41%	6,06%	
		amon	**41,10%**	**88,04%**	
		oni	4,41%	8,65%	
		np	**20,66%**	**47,79%**	
3	Sierra norte	**nino1+2**	**35,97%**	**53,43%**	[0.3,0.6)
		mei	3,91%	7,94%	[-0.6,-0.3)
		nino4	**29,11%**	**51,43%**	
		espi	3,41%	6,62%	
		oni	4,41%	9,27%	
7	Sierra centro este	pdo	0,86%	2,87%	[-0.6,-0.3)
8	Sierra sur	nino1+2	35,97%	42,93%	[0.3,0.6)
		nino1+2	39,44%	45,66%	[0.6,1)
		mei	3,91%	6,98%	[-0.6,-0.3)
		nino4	**29,11%**	**52,62%**	
		espi	3,41%	5,88%	
		amon	41,10%	47,72%	
		oni	4,41%	8,16%	
		np	**20,66%**	**44,21%**	
4	Amazonía norte	**qbo**	**2,96%**	**19,60%**	[0.3,0.6)
		nino1+2	**35,97%**	**49,81%**	
		nino3	**71,80%**	**95,64%**	
		nino34	**6,48%**	**45,08%**	
		np	1,16%	2,94%	
6	Amazonía	**nino1+2**	**35,97%**	**82,75%**	[0.3,0.6)

	central	**nino3**	**71,80%**	**85,39%**	
		CAR_ersst	4,29%	10,57%	[-0.6,-0.3)
9	Amazonía sur	nino1+2	35,97%	39,33%	[0.3,0.6)
		nino3	71,80%	80,33%	
		nino34	6,48%	11,00%	
		np	**1,16%**	**17,00%**	
		ammsst	0,53%	7,67%	[-0.6,-0.3)
		CAR_ersst	4,29%	5,97%	
		pdo	0,86%	2,13%	

En la región 1 (costa norte) existen 6 índices que tiene influencia: 4 del Pacífico (mei, espi, oni y pdo) y 2 del Atlántico (CAR_ersst y tsa), todos en un rango moderado negativo excepto tsa que lo hace en un rango moderado positivo. Para el caso de las teleconexiones con el Atlántico la probabilidad posteriori aumenta más del 10%, mientras que con el Pacífico aumenta menos del 6%.

En la región 2 (costa central), 4 de los índices con influencia pertenecen al Pacífico y 2 al Atlántico. Los del Pacífico son el niño1+2 mostrando el mayor aumento de probabilidad a poseriori (95%) y además con rango de correlación positiva de 0.6 a 1. Los índices amon, np y niño4 influyen en un rango moderado negativo, niño3 y tsa lo hacen en un rango moderado positivo. Todos los índices registran un cambio de probabilidad mayor al 10% excepto niño4.

En la región 5 (costa sur) se observa que todos los índices son del Pacífico con excepción de amon. Mei, niño4, espi, amon, oni y np tienen influencia en un rango moderado negativo, niño3 en un rango moderado positivo y niño1+2 en un rango alto positivo. Los índices niño1+2, niño3, niño4, amon y np registran un cambio de probabilidad a posteriori superior al 20% (siendo del 54,8% en el caso del niño1+2) mientras que los índices mei, espi y oni es inferior al 5%.

De los índices que afectan a la región 3 (sierra norte) todos guardan relación con el Pacífico. Todos influyen en un rango moderado negativo con excepción del índice niño1+2 que lo hace en un rango moderado positivo. Los índices que más aumentaron su probabilidad a posteriori son niño1+2 y niño4 (superior al 17% en ambos casos) mientras que en los demás índices es menor al 5%.

En la región 7 (sierra centro este) el único índice que tiene influencia es pdo, en un rango moderado negativo, sin embargo, el aumento de probabilidad es muy bajo, alcanzando apenas un 2%.

Todos los índices que afectan a la región 8 (sierra sur) pertenecen al pacífico (excepto amon), todos influyen en un rango moderado negativo excepto niño1+2 que registra una influencia tanto moderada positiva y alta positiva. Niño4 y np son los que mayor cambio de probabilidad registran: 23,5% en ambos casos. Para el caso de los demás índices el aumento de probabilidad no supera el 7%.

De los 5 índices que afectan a la región 4 (amazonía norte), todos son del pacífico (niño1+2, niño3, niño34). Todos los índices influyen en un rango moderado positivo y registran aumentos de probabilidad mayores al 13%, excepto np cuyo aumento apenas es del 1,7%.

Para la región 6 (amazonía central), los índices que afectan son niño1+2, niño3 (pacífico) y CAR_ersst (atlántico), los dos primeros tienen influencia moderada positiva mientras que el tercero influencia moderada negativa, el que mayor aumento de probabilidad registró fue niño1+2 con 46,7% mientras que el que menor cambio registró fue CAR_ersst con un 6,2%.

Para el caso de la región 9 (amazonía sur) existen 7 índices que influencian, de los cuales 5 pertenecen al pacífico y 2 al atlántico. Los índices niño1+2, niño3, niño34 y np tienen influencia moderada positiva mientras que ammsst, CAR_ersst y pdo influencia moderada negativa. Np es el índice que mayor aumento de probabilidad registra (15,8%), mientras que en el caso de los demás índices es inferior al 9%.

Existen 3 índices cuyas probabilidades no cambiaron con ninguna región que se estableció como evidencia, estos son: soi, tni y modoki.

Si bien es cierto que un aumento de probabilidad es un buen indicador de la presencia de un índice, se debe notar que es más importante la probabilidad *a posteriori* de dicho índice, independientemente de la magnitud del aumento. Un ejemplo de esto lo tenemos en la región 9 con los índices niño1+2 y niño3, cuyo aumento de probabilidad es de 3,36% y 8,53% respectivamente, inferior al 15,84% del índice np. Sin embargo, la probabilidad *a posteriori* del niño1+2 es de 39,33% y del niño3 es de 80,33% mientras que en el caso de np es del 17%, por tanto, la probabilidad de que tengan presencia los índices niño1+2 y niño3 es mayor a la probabilidad de presencia de np.

b. De índices a regiones

Para la propagación de índices a regiones, la variable que fue establecida como evidencia fue "índice", así como la variable "rango" indica el rango establecido como evidencia. Por lo tanto, en la variable "región" se propagan las probabilidades y se recogen la "probabilidad a priori" y "probabilidad a posteriori". La tabla 3 muestra los resultados de la propagación, y las figuras 15 y 16 los respectivos mapas de propagación por regiones. Del total de registros encontrados en la propagación de probabilidades de índices a regiones (50), 10 corresponden a índices del Atlántico (ammsst, amon, CAR_ersst, tsa), 40 corresponden a índices del Pacífico (np, pdo, qbo, niño1+2, niño3, niño34, niño4, mei, espi, oni). De estos 40, 29 registros son de índices usualmente utilizados para el monitoreo del fenómeno de El Niño (niño1+2, niño3, niño34, niño4, mei, espi, oni).

Tabla 3: Propagación de índices a regiones

Índice	Rango	Región	Descripción región	Probabilidad a priori	Probabilidad a posteriori
qbo	[0.3,0.6)	4	Amazonía N	13,31%	88,09%
mei	[-0.6,-0.3)	1	Costa N	6,88%	9,84%
		3	Sierra N	7,71%	15,67%

		5	Costa S	15,96%	30,23%
		8	Sierra S	14,80%	26,46%
nino1+2	[0.3,0.6)	3	Sierra N	7,71%	11,46%
		4	Amazonía N	13,31%	18,43%
		6	Amazonía C	17,68%	40,68%
		8	Sierra S	14,80%	17,66%
	[0.6,1)	2	Costa C	15,87%	38,10%
		5	Costa S	15,96%	38,16%
		8	Sierra S	14,80%	17,13%
nino3	[0.3,0.6)	2	Costa C	15,87%	20,73%
		4	Amazonía N	13,31%	17,73%
		5	Costa S	15,96%	20,54%
		6	Amazonía C	17,68%	21,03%
nino34	[0.3,0.6)	4	Amazonía N	13,31%	92,61%
		9	Amazonía S	3,78%	6,42%
nino4	[-0.6,-0.3)	2	Costa C	15,87%	18,59%
		3	Sierra N	7,71%	13,63%
		5	Costa S	15,96%	33,89%
		8	Sierra S	14,80%	26,74%
espi	[-0.6,-0.3)	1	Costa N	6,88%	9,41%
		3	Sierra N	7,71%	15,01%
		5	Costa S	15,96%	28,38%
		8	Sierra S	14,80%	25,55%
ammsst	[-0.6,-0.3)	8	Sierra S	14,80%	40,48%
		9	Amazonía S	3,78%	54,76%
amon	[-0.6,-0.3)	2	Costa C	15,87%	34,10%
		5	Costa S	15,96%	34,18%
		8	Sierra S	14,80%	17,18%
CAR_ersst	[-0.6,-0.3)	1	Costa N	6,88%	25,99%
		6	Amazonía C	17,68%	43,57%
		9	Amazonía S	3,78%	5,26%
tsa	[0.3,0.6)	1	Costa N	6,88%	12,49%
		2	Costa C	15,87%	72,30%
oni	[-0.6,-0.3)	1	Costa N	6,88%	10,19%
		3	Sierra N	7,71%	16,21%
		5	Costa S	15,96%	31,28%
		8	Sierra S	14,80%	27,37%
np	[-0.6,-0.3)	2	Costa C	15,87%	23,67%
		5	Costa S	15,96%	36,91%
		8	Sierra S	14,80%	31,67%
	[0.3,0.6)	4	Amazonía N	13,31%	33,70%
		9	Amazonía S	3,78%	55,43%
pdo	[-0.6,-0.3)	1	Costa N	6,88%	16,68%
		3	Sierra N	7,71%	12,21%
		4	Amazonía N	13,31%	28,25%

	7	Sierra CE	4,01%	13,43%
	9	Amazonía S	3,78%	9,38%

Respecto de los índices del Pacífico:

- El índice qbo tiene una influencia moderada positiva en la región 4, con el 88% de probabilidad.
- El índice mei tiene influencia moderada negativa sobre las regiones 1, 3, 5 y 8 siendo estas 2 últimas las que mayor aumento de probabilidad registran.
- La influencia del índice niño1+2 es moderada positiva sobre las regiones 3, 4, 6 y 8; y alta positiva sobre las regiones 2, 5 y 8.
- La influencia del índice niño3 se presenta como moderada positiva sobre las regiones 2, 4, 5 y 6.
- La influencia del índice niño34 es moderada positiva en las regiones 4 y 9, en la primera registra un cambio de probabilidad muy alto.
- El índice niño4 tiene una incidencia moderada negativa sobre las regiones 2, 3, 5 y 8.
- La influencia de espi sobre las regiones 1, 3, 5 y 8 se ubica en un rango moderado negativo.
- El índice oni tiene una influencia moderada negativa en las regiones 1, 3, 5 y 8.
- El índice np tiene influencia moderada negativa sobre las regiones 2, 5 y 8 e influencia moderada positiva sobre las regiones 4 y 9, además en esta última región registra un cambio de probabilidad considerable.
- La influencia que presenta el índice pdo sobre las regiones 1, 3, 4, 7 y 9 se ubica en un rango moderado negativo.

Respecto de los índices del Atlántico:

- El índice ammsst tiene incidencia sobre las regiones 8 y 9, en ambos casos se ubica en un rango moderada negativo, además registra un cambio de probabilidad considerable en la región 9.
- El índice amon tiene una influencia moderada negativa sobre las regiones 2, 5 y 8.
- La influencia del índice CAR_ersst es moderada negativa sobre las regiones 1, 6 y 9; sin embargo, en las 2 primeras regiones registra un aumento de probabilidad considerable mientras que en la última el aumento es mínimo.
- El índice tsa tiene una influencia moderada positiva en las regiones 1 y 2, en el segundo caso registra un aumento de probabilidad significativo.

Siete de los índices se utilizan usualmente para monitoreo del fenómeno de El Niño, 4 presentan una correlación negativa moderada (mei, niño4, espi y oni) y los restantes (niño1+2, niño3 y niño34) una correlación positiva. En el primer caso, las regiones a las que afectan son 1, 2, 3, 5 y 8, siendo 5 y 8 las que mayor aumento de probabilidad a posteriori registran (superiores al 10% en todos los casos). Mientras que en el segundo caso, las regiones afectadas son: 2, 3, 4, 5, 6, 8 y 9; siendo 2, 4, 5 y 6 las que mayor aumento de probabilidad registran (superiores al 20% en todos los casos).

El índice np es el único índice que tiene influencia tanto positiva (regiones 4 y 9) como negativa (regiones 2, 5 y 8), sin embargo el aumento de probabilidad a posteriori fue mayor en la correlación positiva.

De los 4 índices del atlántico, 3 tienen influencia moderada negativa: ammsst, amon y CAR_ersst (8 registros en total). Tsa tiene influencia moderada positiva (2 registros), sin embargo, este índice es el que mayor cambio de probabilidad registra (56,4% en uno de sus registros).

Para este caso, además de la tabla de propagación de probabilidad, se elaboraron mapas para facilitar el análisis de los resultados. Los mapas coloreados en azul representan correlación positiva y los de color naranja correlación negativa. Todos los mapas representan gráficamente el aumento de probabilidad una vez se ha establecido la evidencia (a posteriori). Los rangos de probabilidad son de 0 a 0.3, de 0.3 a 0.6 y de 0.6 a 1, y se muestran en los mapas con variación del tono del color. La figura 15 muestra para los índices del Pacifico y la 16 para los del Atlántico.

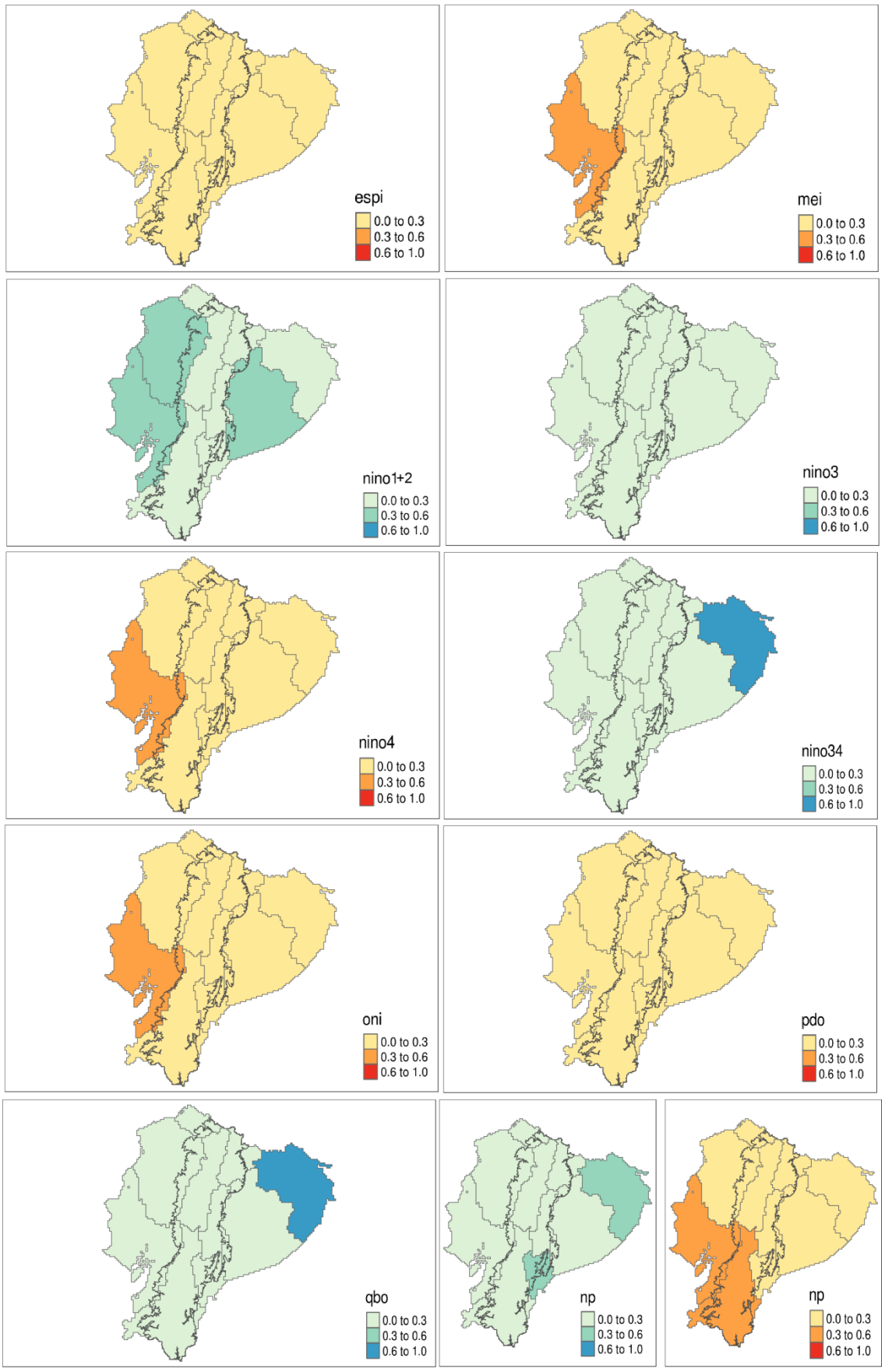


Figura 15: Índices del Pacífico: mapas de probabilidad a posteriori de propagación de índices a regiones.

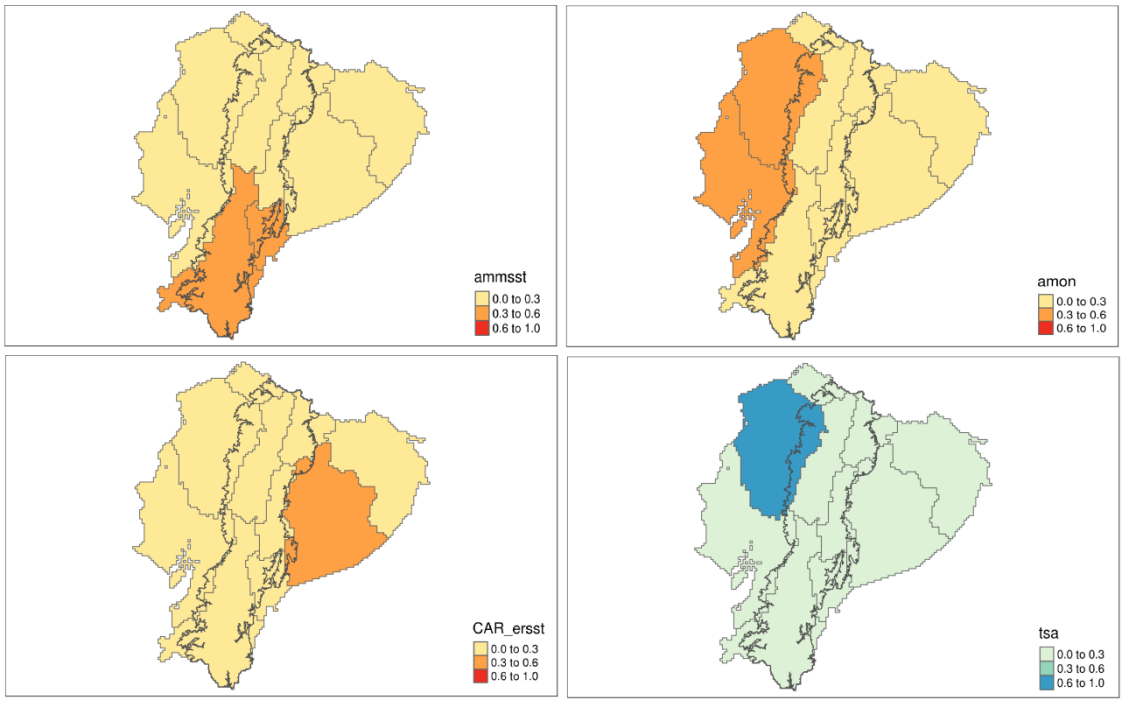


Figura 16: Índices del Atlántico: mapas de probabilidad a posteriori de propagación de índices a regiones.

c. Interacción entre índices

El tercer tipo de propagación fue desde un índice para explorar su interacción con otros índices En la tabla 4, propagación entre índices se muestra con los campos "índice Padre" y "rango índice padre", mientras los campos "índice hijo" y "rango índice hijo" muestran los índices afectados como resultados de la propagación y sus respectivos rangos. Las variables "Probabilidad a priori" y "Probabilidad a posteriori" muestran los cambios de probabilidad en los índices hijos.

La tabla 4 muestra los resultados de la propagación entre índices.

Tabla 4: Propagación de probabilidades e interacción entre índices.

Índice padre	Rango índice padre	Índice hijo	Rango índice hijo	Probabilidad a priori	Probabilidad a posteriori
qbo	[0.3,0.6)	nino3	[0.3,0.6)	71,80%	94,42%
		nino34		6,48%	39,74%
mei	[-0.6,-0.3)	soi	[0.3,0.6)	8,50%	37,34%
		tni		17,90%	34,61%
		nino1+2	[0.6,1)	39,44%	56,42%
		nino4	[-0.6,-0.3)	29,11%	87,67%
		espi		3,41%	46,74%
		amon		41,10%	55,52%
		CAR_ersst		4,29%	5,86%

		oni		4,41%	70,14%
		modoki		5,29%	10,21%
		np		20,66%	33,03%
nino1+2	[0.3,0.6)	qbo	[0.3,0.6)	2,96%	4,42%
		nino34		6,48%	8,89%
		nino3		71,80%	95,43%
	[0.6,1)	tsa		14,83%	31,90%
		mei	[-0.6,-0.3)	3,91%	5,59%
		nino4		29,11%	47,49%
		espi		3,41%	4,72%
		amon		41,10%	92,59%
		CAR_ersst		4,29%	5,81%
		oni		4,41%	6,53%
		np		20,66%	38,03%
nino3	[0.3,0.6)	nino1+2	[0.6,1)	39,44%	52,41%
		nino34	[0.3,0.6)	6,48%	8,53%
		amon		14,83%	17,44%
		tsa	[-0.6,-0.3)	41,10%	52,83%
		np		20,66%	21,70%
nino34	[0.3,0.6)	qbo	[0.3,0.6)	2,96%	18,17%
		nino3		71,80%	94,51%
nino4	[-0.6,-0.3)	soi	[0.3,0.6)	8,50%	10,12%
		tsa		14,83%	18,60%
		tni		17,90%	19,27%
		nino1+2	[0.6,1)	39,44%	64,33%
		mei	[-0.6,-0.3)	3,91%	11,76%
		espi		3,41%	7,70%
		amon		41,10%	62,54%
		oni		4,41%	11,06%
		np		20,66%	51,49%
espi	[-0.6,-0.3)	soi	[0.3,0.6)	8,50%	21,84%
		tni		17,90%	27,09%
		nino1+2	[0.6,1)	39,44%	54,68%
		mei	[-0.6,-0.3)	3,91%	53,62%
		modoki		5,29%	8,00%
		nino4		29,11%	65,79%
		amon		41,10%	54,13%
		CAR_ersst		4,29%	9,65%
		oni		4,41%	86,33%
		np		20,66%	31,81%
ammsst	[-0.6,-0.3)	nino34	[0.3,0.6)	6,48%	8,20%
amon	[-0.6,-0.3)	nino3	[0.3,0.6)	71,80%	92,29%
		tsa		14,83%	28,69%
		nino1+2	[0.6,1)	39,44%	88,84%
		mei	[-0.6,-0.3)	3,91%	5,28%

		nino4		29,11%	44,30%
		espi		3,41%	4,48%
		oni		4,41%	6,14%
		np		20,66%	35,09%
CAR_ersst	[-0.6,-0.3)	soi	[0.3,0.6)	8,50%	35,04%
		nino3		71,80%	72,84%
		tni		17,90%	79,62%
		mei	[-0.6,-0.3)	3,91%	5,34%
		modoki		5,29%	23,44%
		espi		3,41%	7,66%
tsa	[0.3,0.6)	nino1+2	[0.6,1)	39,44%	84,80%
		nino3	[0.3,0.6)	71,80%	84,43%
		nino4	[-0.6,-0.3)	29,11%	36,50%
		amon		41,10%	79,49%
		np		20,66%	29,96%
oni	[-0.6,-0.3)	soi	[0.3,0.6)	8,50%	23,43%
		tni		17,90%	24,01%
		nino1+2	[0.6,1)	39,44%	58,38%
		mei	[-0.6,-0.3)	3,91%	62,13%
		modoki		5,29%	7,09%
		nino4		29,11%	73,02%
		espi		3,41%	66,67%
		amon		41,10%	57,23%
		np		20,66%	34,17%
np	[-0.6,-0.3)	nino1+2	[0.6,1)	39,44%	72,60%
		nino3	[0.3,0.6)	71,80%	75,44%
		tsa		14,83%	21,51%
		qbo		2,96%	6,61%
		nino3		71,80%	82,44%
		nino34		6,48%	21,29%
	[0.3,0.6)	mei	[-0.6,-0.3)	3,91%	6,25%
		nino4		29,11%	72,56%
		espi		3,41%	5,24%
		amon		41,10%	69,82%
		oni		4,41%	7,30%
		ammsst		0,53%	4,46%
		pdo		0,86%	1,86%
pdo	[-0.6,-0.3)	qbo	[0.3,0.6)	2,96%	5,80%
		nino34		6,48%	13,81%
		CAR_ersst	[-0.6,-0.3)	4,29%	5,77%

Respecto de los índices del Pacífico:

- De los 94 registros encontrados, 74 corresponde a índices del Pacífico.

- Los índices qbo, niño1+2, niño3, niño34 tienen interacción positiva.
- Los índices pdo, oni, espi, niño4, mei tienen interacción negativa.
- El índice np tiene interacción tanto positiva como negativa.
- El índice np es el que más influencia tiene sobre los demás índices, con un total de 13 registros.
- Los índices qbo y niño34 son los que menos influencia tienen sobre los demás índices, con apenas 2 registros cada uno.

Respecto de los índices del Atlántico:

- De los 94 registros encontrados, 20 corresponde a índices del Atlántico.
- Los índices ammsst, amon, CAR_ersst tienen interacción negativa
- El índice tsa tiene influencia positiva
- El índice amon es el que más influencia tiene sobre los demás índices, con un total de 8 registros.
- El índice ammst es el que menos influencia tiene sobre los demás índices, con apenas 1 registro.

De la tabla anterior se extrajeron los 10 registros con mayor cambio en la probabilidad, con la finalidad de mostrar aquellos índices que tienen una interacción muy fuerte. Esta tabla es muy similar a la anterior, pero con una columna extra llamada “diferencia” la cual contiene el valor de la resta entre las 2 columnas de probabilidades.

Tabla 5: Registros con el más alto cambio de probabilidad

Índice padre	Rango índice padre	Índice hijo	Probabilidad a priori	Probabilidad a posteriori	Rango índice hijo	Diferencia
espi	[-0.6,-0.3)	oni	4,41%	86,33%	[-0.6,-0.3)	81,92%
modoki		tni	17,90%	99,52%	[0.3,0.6)	81,63%
soi	[0.3,0.6)	tni	17,90%	92,58%		74,68%
mei	[-0.6,-0.3)	oni	4,41%	70,14%	[-0.6,-0.3)	65,73%
oni		espi	3,41%	66,67%		63,26%
CAR_ersst		tni	17,90%	79,62%	[0.3,0.6)	61,72%
mei		nino4	29,11%	87,67%	[-0.6,-0.3)	58,56%
oni		mei	3,91%	62,13%		58,23%
nino1+2	[0.6,1)	amon	41,10%	92,59%		51,48%
espi	[-0.6,-0.3)	mei	3,91%	53,62%		49,71%

Es este caso, la interacción entre espi y oni es muy fuerte, ya que cuando se establece como evidencia a espi, la probabilidad de oni aumenta un 81,9%, de igual manera cuando se establece oni como evidencia la probabilidad de espi aumenta un 63,3%.

El caso es el mismo con los índices: mei y oni, cuando se establece mei como evidencia, la probabilidad de oni aumenta un 65,7% y cuando se establece oni como evidencia, la probabilidad de mei aumenta un 58,2%.

En los dos casos anteriores la relación fue directamente proporcional, sin embargo, también se encuentran situaciones en la que la relación es inversamente proporcional. Por ejemplo si establecemos como evidencia el índice modoki en un rango moderado negativo, la probabilidad de tni aumenta un 81,6% pero en un rango moderado positivo.

Si bien debe notarse que la correlación no indica causalidad, si podemos afirmar que existe una alta probabilidad de que estos índices interactúen entre sí.

Validación

La validación consiste en usar la función "confusionMatriz", a esta le pasamos como parámetro el nodo sobre el que queremos hacer la predicción y los datos de entrenamiento. Se realizó la validación usando 22 nodos, 10 de estos se encuentran representados en la tabla 6. En total 6 índices del Pacífico y 4 del Atlántico.

Tabla 6: Resultados de la matriz de confusión

Nodo	Precisión	Kappa
oni	0,916	0,787
espi	0,943	0,591
wp	0,794	0,595
nino34	0,787	0,593
nino1	0,809	0,705
nino3	0,894	0,743
amon	0,929	0,854
nao	0,888	0,754
NTA_ersst	0,886	0,521
tsa	0,868	0,534

Tenemos para el índice de precisión:

- 19 nodos registraron un valor superior a 0.8 que podríamos considerar como alto.
- 3 nodos registraron un valor entre 0.5 y 0.8 que podríamos considerar como moderado.
- No existen registros con un valor de precisión menor a 0.5

Tenemos que para el coeficiente kappa:

- 1 nodo registró un valor superior a 0.8 (alto).
- 17 nodos registraron un valor entre 0.5 y 0.8 (moderado).
- 4 nodos registraron un valor inferior a 0.5 (bajo).

Discusión

Resumen de resultados

Como resumen de resultados se muestran los registros más sobresalientes. Es decir, aquellos en los que los índices presentan un cambio de probabilidad es grande o cuya probabilidad a posteriori es muy alta. Esto se extrajo para los tres tipos de propagación de probabilidades empleadas.

Para la propagación forward: de regiones a índices

- En el caso de las regiones de la costa (1, 2 y 5) los índices que mostraron una mayor incidencia son: nino1+2, nino3, nino4, np, tsa y amon, siendo los 4 primeros índices del Pacífico y los 2 últimos del Atlántico.
- En el caso de las regiones de la sierra (3, 7 y 8) se tienen los índices: nino1+2, nino4 y np; siendo todos índices del Pacífico.
- Para el caso de la amazonía (4, 6 y 9) se tienen los índices: nino1+2, nino3, nino34 y qbo; siendo todos índices del Pacífico.

Para la propagación backward: de índices a regiones

- Se encontró que en el caso de que se establezca como evidencia a los índices qbo, nino1+2, nino34, np, ammsst, CAR_ersst y tsa, las regiones más afectadas son: 4, 6, 9, 8 y 2. Siendo las 3 primeras de la amazonía y las dos últimas de la sierra y costa respectivamente.

Para la interacción entre índices

- Los indices np, mei, amon, espi y oni son los que mayor influencia tiene sobres los demás índices.

Usos de la red bayesiana

La red bayesiana es una herramienta útil para extraer conclusiones sobre el comportamiento de las precipitaciones en las regiones del Ecuador. Es posible establecer índices como evidencia, y propagar sus probabilidades sobre las regiones. Además, es posible establecer evidencias en más de un índice que presenten interacción, y a su vez, es posible diferenciar la influencia de varios índices que afectan sobre una misma región, para conocer cuál de estos lo hace con una mayor fuerza.

También resulta útil determinar los índices que intervienen en las precipitaciones de una determinada región. Tanto en este caso como en el anterior (dada la naturaleza de esta red) se puede conocer, no solamente la probabilidad a posteriori, sino además el rango en el que se encuentra dicha probabilidad. Esto permitirá saber si la relación entre los índices y las regiones es directa o inversa.

Por último, también ayudaría a entender cómo interactúan los índices entre sí, es decir, si la presencia o ausencia de un índice influye en otro y que tan fuerte es esa relación.

Escenario

Las redes bayesianas son utilizadas para mostrar escenarios plausibles y observar sus efectos. En el ámbito de la climatología esto puede resultar una tarea de una complejidad muy alta debido a la gran cantidad de índices climáticos que interactúan. Sin embargo, debido a la alta capacidad de procesamiento de los ordenadores en la actualidad, una red bayesiana puede resolver este tipo de problemas, por lo que usarlas en estos escenarios ayudaría a entregar respuestas de manera ágil. Si a esto le añadimos que esta información está siendo usada por alguna entidad de prevención de riesgos, en donde el tiempo de respuesta es crítico, la obtención de respuestas ágiles toma aún más relevancia.

Se utilizará un escenario hipotético. Supongamos que el Instituto Nacional de Meteorología e Hidrología (INAMHI) informa que es probable la llegada del fenómeno de El Niño a las costas de Ecuador, y sabemos que este evento climático tiene influencia (muchas veces catastróficas) sobre el clima de nuestro país. Ante esta situación la Secretaría Nacional de Gestión de Riesgos (SNGR) debe elaborar planes de contingencia y de mitigación de riesgos, pero ¿Cómo saber sobre qué regiones debe actuar de acuerdo a las anomalías observadas en los índices climáticos? ¿Cuáles son las regiones más vulnerables y a cuáles debe dar prioridad?

Para puntualizar de manera más concreta este escenario, supongamos que el INAMHI informa que los índices niño1+2 y niño3 presentan valores “anormalmente altos” (podríamos considerar esto dentro de los rangos de correlación positivo moderado y alto). Al establecer estos índices como evidencia dentro de la red obtenemos que las regiones más afectadas son la número 6 (amazonía central) con un 29,2% de probabilidad de lluvia, la número 2 (costa central) con un 19,8% de probabilidad y la número 5 (costa sur) con un 19,6%.

En otro escenario supongamos que se detecta la presencia de los índices niño4 y np con valores “anormalmente altos” (consideramos esto dentro de los rangos de correlación negativo moderado y alto). En esta situación, al hablar de correlación negativa se refiere a que los valores altos de estos índices producirán reducción de las precipitaciones. Al establecer la evidencia se obtiene que las regiones más afectadas son el número 5 (costa sur) con un 40,2% de probabilidad, la número 8 (sierra sur) con un 33,5% de probabilidad y la número 2 (costa central) con un 17,9%.

Se debe notar que en este trabajo se utilizaron regiones de precipitación. Sin embargo, el método se puede aplicar con otro tipo de regionalización o división administrativa como por ejemplo provincias, cantones o parroquias.

Limitaciones del trabajo

A pesar de las ventajas que mostraron las redes bayesianas para la comprensión, de una manera visual e interactiva, qué índices que afectan una región y para la generación de escenarios y toma de decisiones, también se encontraron una serie de limitaciones que se detallan a continuación.

Discretización. Como se mencionó en capítulos anteriores, es necesario discretizar los valores ya que el algoritmo de aprendizaje empleado en este trabajo utiliza valores discretos. Sin embargo, esto también puede convertirse en una desventaja ya que, si no se eligen adecuadamente los intervalos, es posible extraer conclusiones sesgadas o erróneas. Además,

el método aquí usado para discretizar (EWD) puede producir intervalos con muchos datos y otros intervalos con pocos datos, lo que podría afectar la red resultante. A futuro sería importante utilizar métodos de aprendizaje que garanticen una mejor distribución de datos o que trabaje sobre la distribución continua de los datos.

Interpretación de causalidad. A menos que se disponga de evidencias externas, a partir del aprendizaje de la red bayesiana no se puede concluir causalidad entre nodos. Por ejemplo, entre 2 nodos que se encuentren conectados no existe relación de causalidad, sino que a través del aprendizaje se encuentra una "relación" de probabilidad entre esos nodos. Es importante recordar esto para evitar obtener conclusiones erróneas.

Valores de correlación de los índices. Los datos usados para obtener la red bayesiana de este trabajo fueron datos de correlación y no los valores originales de los índices climáticos. Esto significa que para crear esta red primero se tuvo que obtener los mapas de correlación de los índices con las regiones, previo a la creación de la red. Sería interesante a futuro conformar una red a partir de los valores originales de los índices.

Uso de datos. Del total de registros, el 80% se usó para el aprendizaje de la red, y un 20% para su validación. Aunque es importante realizar la validación, esto limita el número de registros usados para el aprendizaje.

Conclusiones

Este trabajo permitió, mediante el lenguaje de programación R, la creación, entrenamiento y propagación de probabilidades de una red bayesiana. Gracias a esto fue posible estudiar diferentes índices climáticos que tienen influencia sobre las precipitaciones en las diferentes regiones de estacionalidad del Ecuador, medir que tan fuerte es esa influencia en términos de correlación, así como determinar si existe interacción entre índices climáticos.

El uso de esta herramienta demostró que existe una relación de probabilidad entre la mayoría de los índices climáticos y las regiones de precipitación. Esta probabilidad varía de acuerdo al índice así como al rango en que es establecida la evidencia pudiendo encontrar índices que tienen una relación de probabilidad muy fuerte (como es el caso del índice nino34 sobre la región 4), hasta índices con una relación de probabilidad muy débil (el caso del índice espi sobre la región 1), e incluso se encontraron índices aislados que no interactúan con ningún nodo de la red (tna y pna).

De manera general, los resultados encontraron fueron que en las regiones de la costa los índices con mayor incidencia son nino1+2, nino3, nino4, np, tsa y amon; en las regiones de la sierra nino1+2, nino4 y np; y en la amazonía nino1+2, nino3, nino34 y qbo.

Sobre las métricas usadas en la validación, la precisión ofrece muy buenos resultados, sin embargo, para el caso del coeficiente kappa los resultados obtenidos fueron moderados.

Los sistemas expertos son herramientas útiles para ayudar en la toma de decisiones. Las redes bayesianas cuentan con la ventaja de que al ser un sistema experto basado en probabilidades, pueden ser usados en problemas en donde exista incertidumbre y el conocimiento no sea determinista. Estos abarcan la mayor parte de problemas del mundo real y en áreas tan diversas como medicina, economía, Inteligencia Artificial o en este caso, la climatología.

Esta herramienta puede ser utilizada conjuntamente por organismo como el INAMHI para el análisis y estudio de índices climáticos y como estos afectan a las precipitaciones en el país. Así como también por organismos de control de riegos como la Secretaría Nacional de Gestión de Riesgos para tomar decisiones orientadas a minimizar el impacto ante posibles anomalías climáticas como sequías o inundaciones (por ejemplo usando la información que oportunamente habrá brindado el INAMHI). También puede ser usada como una herramienta de apoyo para estudiar teleconexiones climáticas ya que utiliza un enfoque diferente (inferencia bayesiana).

Debido a que este trabajo emplea valores de correlación de los índices, a futuro se propone crear una nueva red bayesiana que utilice los valores reales de precipitación e índices climáticos, en lugar de su correlación. Además, se propone usar un clasificador bayesiano ingenuo ya que por lo general estos son mucho más rápidos y sus resultados igualmente buenos. Lo interesante de estos clasificadores es que asumen independencia condicional entre los nodos hijos (en este caso serían los índices), por lo que se convierte en un nuevo enfoque para abordar estos problemas.

Finalmente, se podría considerar otro tipo de regionalización como por ejemplo por provincias, cantones o incluso parroquias, esto podría ofrecer información mucho más precisa para los tomadores de decisiones.

Bibliografía

Abd-elhamid, H. F., Fathy, I., & Zelen, M. (2018). Flood prediction and mitigation in coastal tourism areas, a case study: Hurghada, Egypt. https://doi.org/10.1007/s11069-018-3316-x

Ali, S., Jan, A., Manzoor, Sohail, A., Khan, A., Khan, M. I., … Daur. (2018). Soil amendments strategies to improve water-use efficiency and productivity of maize under different irrigation conditions. Agricultural Water Management, 210(March), 88–95. https://doi.org/10.1016/j.agwat.2018.08.009

Ariza-López, F. J., Rodríguez-Avi, J., & Alba-Fernandez, V. (2018). CONTROL ESTRICTO DE MATRICES DE CONFUSIÓN POR MEDIO DE DISTRIBUCIONES MULTINOMIALES, (21), 215–226. Retrieved from https://dialnet.unirioja.es/descarga/articulo/6522539.pdf%0A%0A

Ballari, D., Giraldo, R., Campozano, L., & Samaniego, E. (2018). Spatial functional data analysis for regionalizing precipitation seasonality and intensity in a sparsely monitored region: Unveiling the spatio-temporal dependencies of precipitation in Ecuador. International Journal of Climatology, 38(8), 3337–3354. https://doi.org/10.1002/joc.5504

Bayesserver. (2018). Asia [Figura]. Retrieved 11 June 2019, from https://www.bayesserver.com/examples/networks/asia.

Bayesserver. (2018). Bayesian networks - an introduction. Retrieved 10 June 2019, from https://www.bayesserver.com/docs/introduction/bayesian-networks.

Ben-Gal, I. (2009). Bayesian networks. Wiley Interdisciplinary Reviews: Computational Statistics, 1(3), 307–315. https://doi.org/10.1002/wics.48

Blunden, J., Arndt, D. S., Baringer, M. O., Achberger, C., Ackerman, S. A., Ahlstrom, A., … Zimmermann, S. (2011). State of the climate in 2010. Bulletin of the American Meteorological Society, 92(6), 92 (6), pp. S1–S236. https://doi.org/10.1175/1520-0477-92.6.S1

Carvalho, A. (2009). Scoring functions for learning Bayesian networks. Inesc-Id Tec. Rep, 54, 1–27. Retrieved from https://pdfs.semanticscholar.org/6efe/f4bacfb14cfe4c1abababae751904431b75cc9.pdf

Correa, M., Bielza, C., & Pamies-teixeira, J. (2009). Expert Systems with Applications Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. Expert Systems With Applications, 36(3), 7270–7279. https://doi.org/10.1016/j.eswa.2008.09.024

Das, K., & Vyas, O. P. (2010). A Suitability Study of Discretization Methods for Associative Classifiers. International Journal of Computer Applications, 5(10), 46–51. https://doi.org/10.5120/944-1322

Das, M., & Ghosh, S. K. (2015). A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables. 9th International Conference on Industrial and Information Systems, ICIIS 2014. https://doi.org/10.1109/ICIINFS.2014.7036528

Duc, H. N., Bang, H. Q., & Quang, N. X. (2018). Influence of the Pacific and Indian Ocean climate drivers on the rainfall in Vietnam. International Journal of Climatology, (June), 1–16. https://doi.org/10.1002/joc.5774

Duc, H. N., Rivett, K., MacSween, K., & Le-Anh, L. (2017). Association of climate drivers with rainfall in New South Wales, Australia, using Bayesian Model Averaging. Theoretical and Applied Climatology, 127(1–2), 169–185. https://doi.org/10.1007/s00704-015-1622-8

Ebert-Uphoff, I., & Deng, Y. (2012). Causal discovery for climate research using graphical models. Journal of Climate, 25(17), 5648–5665. https://doi.org/10.1175/JCLI-D-11-00387.1

Engler, J., Wehrden, H. Von, & Baumgärtner, S. (2019). Land Use Policy Determinants of farm size and stocking rate in Namibian commercial cattle farming. Land Use Policy, 81(February 2018), 232–246. https://doi.org/10.1016/j.landusepol.2018.10.009

Fierro, A. O. (2014). Relationships between California rainfall variability and large-scale climate drivers. International Journal of Climatology, 34(13), 3626–3640. https://doi.org/10.1002/joc.4112

Flores, J. G. I. L. (2005). Aplicacion del metodo Bootstrap al contraste de Hipotesis en la Investigacion Educativa. Revista de Educacion, 336(1979), 251–265.

Hamududu, B., Killingtveit, A., & Engineering, E. (2012). Assessing Climate Change Impacts on Global Hydropower, (2005), 305–322. https://doi.org/10.3390/en5020305

Hasan, M. M., & Wyseure, G. (2018). Impact of climate change on hydropower generation in Rio Jubones Basin, Ecuador. Water Science and Engineering. https://doi.org/10.1016/j.wse.2018.07.002

Højsgaard, S. (2012). Graphical Independence Networks with the gRain Package for R. Journal of Statistical Software, 46(10), 37–44. https://doi.org/10.4324/9780429468872-4

Højsgaard, S. (2016). Graphical Independence Networks “gRain.” Retrieved from https://cran.r-project.org/web/packages/gRain/gRain.pdf

Konapala, G., Valiya, A., & Ashok, V. (2017). Teleconnection between low flows and large-scale climate indices in Texas River basins. Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-017-1460-6

Kuhn, M. (2019). Classification and Regression Training, 216. Retrieved from https://cran.r-project.org/web/packages/caret/caret.pdf

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Statistical Society, 50(2), 353–366. https://doi.org/10.1007/BF01531016

Lee, J. H., Lee, J., & Julien, P. Y. (2018). Catena Global climate teleconnection with rainfall erosivity in South Korea. Catena, 167(June 2017), 28–43. https://doi.org/10.1016/j.catena.2018.03.008

Liu, Y. C., Di, P., Chen, S. H., & DaMassa, J. (2018). Relationships of rainy season precipitation and temperature to climate indices in California: Long-Term variability and extreme events. Journal of Climate, 31(5), 1921–1942. https://doi.org/10.1175/JCLI-D-17-0376.1

López, D. A. G. (2007). Algoritmo de Discretización de Series de Tiempo Basado en Entropía y su Aplicación en Datos Colposcópicos. Retrieved from http://cdigital.uv.mx/bitstream/123456789/32352/1/garcialopezdaniel.pdf

Norsys. (2018). Introduction to Bayes Nets. Retrieved 20 May 2019, from https://www.norsys.com/tutorials/netica/secA/tut_A1.htm.

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). Bayesian Networks in R. https://doi.org/10.1007/978-1-4614-6446-4

Pratiwi, R., & Sukardjo, S. (2018). EFFECTS OF RAINFALL ON THE POPULATION OF SHRIMPS Penaeus monodon Fabricius IN SEGARA ANAKAN LAGOON , CENTRAL JAVA , INDONESIA, 2(3), 156–169. https://doi.org/10.11598/btb.2018.25.3.830

Recuero de los Santos, P. (2018). Machine Learning a tu alcance: La matriz de confusión. España: Think Big/Empresas. Recuperado de https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/

Rodríguez, D., & Dolado, J. (2007). Redes bayesianas en la ingeniería del software. Cc.Uah.Es, 1–21. https://doi.org/10.2196/jmir.7.3.e31

Russell, S., & Norvig, P. (2010). Artificial Intelligence A Modern Approach Third Edition. Pearson. https://doi.org/10.1017/S0269888900007724

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721), 523-529.

Sánchez Muñoz, J. M. (2016). Análisis de Calidad Cartográfica mediante el estudio de la Matriz de Confusión. Pensamiento Matemático, 6(2), 9–26. Retrieved from https://dialnet.unirioja.es/descarga/articulo/5998855.pdf

Scutari, M. (2019). Package ‘ bnlearn .’ Retrieved from https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf

Sudha, V., Venugopal, K., & Ambujam, N. K. (2008). Reservoir operation management through optimization and deficit irrigation, 93–102. https://doi.org/10.1007/s10795-007-9041-3

Sucar, L. E. (2013). Redes Bayesianas. Test, (X), 18. Retrieved from http://ccc.inaoep.mx/~esucar/Clases-mgp/caprb.pdf

Torre-Gea, G. D. la, Soto-Zarazua, G. M., Guevara-Gonzalez, R. G., & Rico-Garcia, E. (2011). Bayesian networks for defining relationships among climate factors. International Journal of Physical Sciences, 6(18), 4412–4418. https://doi.org/10.1016/j.jmaa.2015.01.055

Ulloa, J., Ballari, D., Campozano, L., & Samaniego, E. (2017). Two-step downscaling of TRMM 3b43 V7 precipitation in contrasting climatic regions with sparse monitoring: The case of

Ecuador in tropical South America. Remote Sensing, 9(7), 1–23. https://doi.org/10.3390/rs9070758

UnBBayes. (2019). UnBBayes: Framework & GUI for Bayes Nets and other probabilistic models. Brasilia, Brasil.: SOURCEFORGE. Retrieved 10 May 2019, from https://sourceforge.net/projects/unbbayes/

Vicente-Serrano, S. M., Aguilar, E., Martínez, R., Martín-Hernández, N., Azorin-Molina, C., Sanchez-Lorenzo, A., … Nieto, R. (2017). The complex influence of ENSO on droughts in Ecuador. Climate Dynamics, 48(1–2), 405–427. https://doi.org/10.1007/s00382-016-3082-y

Zeng, Z., Hsieh, W. W., Shabbar, A., & Burrows, W. R. (2011). Seasonal prediction of winter extreme precipitation over Canada by support vector regression. Hydrology and Earth System Sciences, 15(1), 65–74. https://doi.org/10.5194/hess-15-65-2011

Anexos

Código generado en r

#Cargar en memoria todas las librerias necesarias para trabajar, estas contienen funciones que
#nos ayudan entre otras cosas a plotear (como gglpot2, gridExtra), realizar el aprendizaje de la
#red (Bnlearn, gRain) o la manipulacion de datos (dplry)

```
library(rgdal)
library(rasterVis)
library(latticeExtra)
library(sf)
library(gridExtra)
library(ggplot2)
library(bnlearn)
library(caret)
library(gRain)
library(dplyr)
```

#Cargar el archivo con las regiones de precipitación en la variable p9, este archivo es de tipo
#*.shp (shapefile)

```
p9<- st_read("/home/primrose/teleconexiones/Regionalizacion5km/renato/p9.shp")
```

#Cargar el archivo con los mapas de correlaciones entre los índices climáticos y la precipitación

```
load("/home/primrose/teleconexiones/Regionalizacion5km/renato/corr.indexes.map.RData
")
```

#Generar a partir de los mapas de correlaciones los gráficos de nivel y de contorno, y los
#almacena en la variable p, primero se debe convertir los mapas de correlaciones en un objeto
#de tipo RasterBrick usando la función “brick”

```
p <- levelplot(brick(corr.indexes.map), at=seq(-1, 1, 0.1))
```

#Se establece la semilla para la generación de números pseudo-aleatorios, esto es importante
#ya que así garantizamos que los resultados sean reproducibles

```
set.seed(123)
```

#Asociar los mapas de correlación con las regiones de precipitación y convertirlos a un objeto
#de tipo raster

```
p9.r <- raster::rasterize(p9, brick(corr.indexes.map)[[1]],"cluster9")
```

#Agregar una nueva capa al objeto p9.r, como resultado se obtiene un objeto de tipo
#rasterStack

```
datos.GD.9.r<- addLayer(p9.r,brick(corr.indexes.map))
```

#Convertir el objeto de tipo rasterStack a tipo Data frame de puntos espaciales

```
datos.GD.9<- as(datos.GD.9.r, "SpatialPointsDataFrame")@data
```

#Verificar que los datos del data frame generado se encuentren completos, es decir que no
#exista valores faltantes

```
datos.GD.9<- datos.GD.9[complete.cases(datos.GD.9),]
```

#Del data frame anterior, se excluyen los indices “hurr”, “solar”, “trend”, “sahelrain”

datos.GD.9 <- select(datos.GD.9, -hurr,-solar, -trend, -sahelrain)

#Tomar una muestra del data frame “datos.GD.9” usando la columna de las regiones (layer)
#como variable de agrupamiento, se toma el 80% del total de datos como muestra (esto está
#dado por el tercer parámetro de la función). El resultado se almacena en un nuevo data
frame #llamado “data_train.9” (que lo usaremos como datos de entrenamiento)
data_train.9<-fifer::stratified(datos.GD.9, "layer", 0.8)

#Con el 20% de datos restantes de la variable “datos.GD.9” se genera otro data frame que
#servirá para realizar la validación
data_test.9<-datos.GD.9[-as.numeric(rownames(data_train.9)),]

#Generar un data frame discretizando los valores de correlación de los índices, en los rangos
#definidos en la variable “break” con intervalo abierto hacia la derecha
breaks = c(-1, -0.6,-0.3 ,0 , 0.3,0.6,1)
data_train.9.discr = apply(as.data.frame(data_train.9[,-1]), 2, cut, breaks, right=FALSE)

#Al data frame generado se le agrega la columna de regiones (que previamente se quitó para
#discretizar los valores) y luego a esta misma variable se la convierte en factor (factor es una
#variable categórica con un numero finito de valores o niveles)
data_train.9.discr<- data.frame(region=as.factor(data_train.9[,1]) ,data_train.9.discr)
data_train.9.discr[] <- lapply(data_train.9.discr, factor)

#Genera arcos entre cada nodo y calcula la fuerza así como la dirección, esto lo hace a través
#de un proceso estadístico conocido como bootstrapping que consiste en remuestrear los
#datos para crear muchas muestras simuladas. Los parámetros pasados a este procedimiento
#son:
#algoritmo de aprendizaje (algorithm): hill-climbing (hc)
#Tamaño de la muestra (m): 500
#Número de réplicas de bootstrap (R): 100
#cpdag: TRUE (usa una clase equivalente en lugar de la estructura, esto debería hacer más fácil
#identificar arcos equivalentes en puntajes)
#Score: Criterio de selección de modelos
bootstrength<-boot.strength(data_train.9.discr, algorithm = "hc",R = 100, m=500,
cpdag=TRUE, algorithm.args = list(score = "bde"))

#Construye una red usando solamente los nodos significativos (aquellos cuyo valor de fuerza
#sea mayor al threshold, en este caso, 0.6)
av_net<- averaged.network(bootstrength, threshold = 0.6)

#Convierte la estructura en un grafo acíclico no dirigido, ya que hasta este momento la
#estructura podría contener arcos no dirigidos, con esta función garantizamos que estos arcos
#desaparezcan
res = pdag2dag(av_net, ordering = nodes(av_net))

#Calcula el puntaje de la red, los argumentos son: la estructura de la red, los datos de
#entrenamiento y el criterio de selección de modelos
score(res, data = data_train.9.discr, type="aic")

#Funciones para verificar si efectivamente la red es acíclica y dirigida
acyclic(res)
directed(res)

#Crear una red bayesiana y genera los arcos de acuerdo a la fuerza de las dependencias que #representan.
gR <- strength.plot(res, bootstrength)

#Establecer el atributo “fill” para clasificar a los nodos, este atributo permite colorear los #mismos con la finalidad de poder diferenciarlos, los índices del atlántico fueron coloreados en #azul, los del pacifico en un tono beige y el nodo región de color dorado, los que no fueron #coloreados son índices de influencia global
nodeRenderInfo(gR)$fill[c("tna", "nao", "tsa", "epo",
"ao","ea","amon","ammsst","CAR_ersst","NTA_ersst")] = "cadetblue1"
nodeRenderInfo(gR)$fill[c("pna", "nino3", "qbo", "np",
"nino4","nino1","tni","modoki","mei","wp","pdo","oni","espi","nino34")] = "antiquewhite"
nodeRenderInfo(gR)$fill[c("region")] = "gold"

#Plotear la red bayesiana
renderGraph(gR)

#Generar las tablas de probabilidad condicional a través de la funcion “bn.fit” (aprendizaje de #parámetros)
fitted <- bn.fit(res, data = data_train.9.discr, method = "mle")

#Graficar la probabilidad marginal de un nodo (en este caso oni)
bn.fit.barchart(fitted$oni, data = data_train.9.discr, xlab = "Probabilities",ylab = "Levels",
main = "Marginal Probabilities")

#Exportar las tablas de probabilidad condicional hacia el formato “*.net” para poder ser #representado gráficamente con la ayuda de otro software
write.net(file="fitted-deseada.net", fitted)

#Similiar a lo que se hizo con los datos de entrenamiento, se debe generar un data frame #discretizando los valores de correlación de los índices, esta vez para los datos de validación, #luego a esta variable generada se le agrega la columna de regiones y se la convierte en factor.
data_test.9.discr = apply(as.data.frame(data_test.9[,-1]), 2, cut, breaks, right=FALSE)
data_test.9.discr<- data.frame(region=as.factor(data_test.9[,1]) ,data_test.9.discr)
data_test.9.discr[] <- lapply(data_test.9.discr, factor)

#Graficar las tablas de probabilidad condicional de cada nodo
bn.fit.barchart(fitted$region, data = data_train.9.discr, xlab = "Probabilities",ylab = "Levels",
main = "Marginal Probabilities")

#Realizar una prediccion de un nodo dado la red bayesiana y los datos de prueba
pred = predict(fitted, node = "nino1", data_test.9.discr)
#Generar una matriz de confusion para el nodo con el que se realizó la predicción
confusionMatrix(pred, data_test.9.discr[, "nino1"])

#Convierte las tablas de probabilidad en un objeto de tipo grain y luego lo compila
bngrain<- compile(as.grain(fitted))

#Realizar una propagación forward estableciendo como evidencia la región 1 del nodo “region”
e.reg1<- setFinding(bngrain, nodes="region", states = "1")

```
#Realiza un query de la probabilidad del nodo “nino1”, el objeto usado es la variable “bngrain”
#por lo que en este caso realiza una consulta de la probabilidad a priori, ya que no se ha
#establecido ningún nodo como evidencia
prior<- cbind(as.data.frame(querygrain(bngrain, nodes =
"nino1")$nino1),"without_evidence")
#Crea una columna de “rangos” que se añade al data frame anterior para identificar a que
#rango pertenece una probabilidad
prior$factor<- rownames(prior)
#Cambiar los nombres de las columnas del data frame
names(prior)<- c("probabilities", "evidence", "type")

#Realiza un query de la probabilidad del nodo “nino1”, el objeto usado es la variable “e.reg1”
#por lo que en este caso realiza una consulta de la probabilidad a posteriori, ya que dentro de
#esta variable se realizó una propagación estableciendo la región 1 como evidencia
posterior<- cbind(as.data.frame(querygrain(e.reg1, nodes =
"nino1")$nino1),"with_evidence")
#Crea una columna de “rangos” que se añade al data frame anterior para identificar a qué
#rango pertenece una probabilidad
posterior$factor<- rownames(posterior)
#Cambiar los nombres de las columnas del data frame
names(posterior)<- c("probabilities", "evidence", "type")

#Combina los data frame anteriormente creados en uno solo (por filas)
df1 <- rbind(prior, posterior)

#Plotear el data frame “df1”
ggplot(df1, aes(x=type, y=probabilities, fill=evidence)) +  geom_bar(stat='identity',
position='dodge')+
  ggtitle("P(nino1|region=1)")


##************************************************
#Generación de mapas
##************************************************

#Importar las librerías necesarias
library(sf)
library(tmap)
library(sp)

#Cargar los archivos que contiene un joint entre las regiones de precipitación y los resultados
#de la propagación
load("/home/primrose/teleconexiones/renato2/mapas.RData")
load("/home/primrose/teleconexiones/renato2/indices.prior.post.Rdata")

#Cargar el archivo vectorial con el polígono del Ecuador y transformación a sistema de
#referencia UTM
load("/home/primrose/teleconexiones/renato2/Ecuador_shape_4326.RData")#ecuador
ecuador<- spTransform(ecuador, CRS("+init=epsg:32717"))
load("/home/primrose/teleconexiones/renato2/curva_nivel_1000m_4326.RData")
cn_32717<- spTransform(cn, CRS("+init=epsg:32717"))
```

```
#Cargar datos de regionalización
Seasonality_9<- st_read("/home/primrose/teleconexiones/renato2/Region9_poly.shp")

#Bucle para generación de mapas, se deben realizar 28 iteraciones ya que esta es la cantidad
#de índices que existe
for (i in 1:28) {
#Combina en un solo data frame los mapas de estacionalidad con los valores de los índices
#climáticos
 Seasonality_9.prior<-cbind(Seasonality_9,p9.prior.indices[[i]])
 Seasonality_9.posterior<- cbind(Seasonality_9,p9.post.indices[[i]])
#Banderas que se usaran para verificar si un rango es positivo o negativo
 bandera.positivo=FALSE
 bandera.negativo=FALSE

#De los data frame generados previamente, se realizan 4 copias:
#*Probabilidad a priori positiva
#*Probabilidad a priori negativa
#*Probabilidad a posteriori positiva
#*Probabilidad a posteriori negativa
#Se hace esto con el fin de facilitar la labor de graficar los mapas
 copia.Seasonality.prior.p<-Seasonality_9.prior
 copia.Seasonality.prior.n<-Seasonality_9.prior
 copia.Seasonality.posterior.p<-Seasonality_9.posterior
 copia.Seasonality.posterior.n<-Seasonality_9.posterior

#Bucle para recorrer todas las regiones de precipitacion
 for(j in 1:9) {

#Si el rango del indice es positivo, entonces los data frame con valores negativos se setean a 0,
#y la bandera “bandera.positivo” se setea a TRUE.
#Si el rango del indice es negativo, entonces los data frame con valores positivos se setean a 0,
#y la bandera “bandera.negativo” se setea a TRUE.
  if(Seasonality_9.posterior$rango[j]=="[0.3,0.6)" |
Seasonality_9.posterior$rango[j]=="[0.6,1)"){
    copia.Seasonality.posterior.n[[6]][j]=0
    copia.Seasonality.prior.n[[6]][j]=0
    bandera.positivo=TRUE
  }else if(Seasonality_9.posterior$rango[j]=="[-0.6,-0.3)" |
Seasonality_9.posterior$rango[j]=="[-1,-0.6)"){
    copia.Seasonality.posterior.p[[6]][j]=0
    copia.Seasonality.prior.p[[6]][j]=0

    bandera.negativo=TRUE
  }else{
  }
 }

#Si un índice tiene la bandera “bandera.positivo” seteado como verdadera y la bandera
#“bandera.negativo” seteado como falsa significa que solamente tiene rangos positivos, si
#tiene la bandera “bandera.negativo” como verdadera y “bandera.positivo” como falso
#significa que #solamente tiene rangos negativos; si tiene las 2 banderas seteadas como
```

```
#verdadero significa que tiene rangos tanto negativos como positivos. Es necesario realizar
#esta distinción ya que dependiendo de esto se graficara con una paleta de colores u otra, o en
#el caso de que tenga los dos tipos de rangos, se graficara con las 2 paletas de colores

  if(bandera.positivo & bandera.negativo){
 #Graficar los mapas cuyos rangos son positivos
   indiceX_post.p<- tm_shape(copia.Seasonality.posterior.p[6]) + tm_polygons(col
=colnames(p9.prior.indices[[i]])[3],breaks = c(0, 0.3, 0.6, 1),palette="GnBu")+
tm_shape(cn_32717) + tm_borders('gray30')

 #Graficar los mapas cuyos rangos son negativos
   indiceX_post.n<- tm_shape(copia.Seasonality.posterior.n[6]) + tm_polygons(col
=colnames(p9.prior.indices[[i]])[3],breaks = c(0, 0.3, 0.6, 1),palette="YlOrRd")+
tm_shape(cn_32717) + tm_borders('gray30')

#Guardar los mapas generados
   imagen<-tmap_arrange(indiceX_post.p,indiceX_post.n)
   tmap_save(tm = imagen,filename =
paste("/home/primrose/teleconexiones/Mapas/",colnames(p9.prior.indices[[i]])[3],".png",s
ep = ""))

  }else if(bandera.positivo){
#Graficar los mapas cuyos rangos son positivos
   indiceX_post<- tm_shape(copia.Seasonality.posterior.p[6]) + tm_polygons(col
=colnames(p9.prior.indices[[i]])[3],breaks = c(0, 0.3, 0.6,1),palette="GnBu")+
tm_shape(cn_32717) + tm_borders('gray30')

#Guardar los mapas generados
   imagen<-tmap_arrange(indiceX_post)
   tmap_save(tm = imagen,filename =
paste("/home/primrose/teleconexiones/Mapas/",colnames(p9.prior.indices[[i]])[3],".png",s
ep = ""))

  }else if(bandera.negativo){
#Graficar los mapas cuyos rangos son negativos
   indiceX_post<- tm_shape(copia.Seasonality.posterior.n[6]) + tm_polygons(col
=colnames(p9.prior.indices[[i]])[3],breaks = c(0, 0.3, 0.6, 1),palette="YlOrRd")+
tm_shape(cn_32717) + tm_borders('gray30')

#Guardar los mapas generados
   imagen<-tmap_arrange(indiceX_post)
   tmap_save(tm = imagen,filename =
paste("/home/primrose/teleconexiones/Mapas/",colnames(p9.prior.indices[[i]])[3],".png",s
ep = ""))

  }else{
  }
}
```

Scripts

```
# <---------------------------------------------------------------------------------->
#Script 1
# El siguiente script genera una gran cantidad de redes bayesianas utilizando diferentes
#parámetros como son: el algoritmo, score, número de bootstrap y tamaño del bootstrap. El
#objetivo de esto es determinar cuál es la red con el mejor "score" y esta se usara para
#realizar el aprendizaje paramétrico así como la propagación
# <---------------------------------------------------------------------------------->

#Vectores de parámetros usados en la generación de las redes bayesianas, los algoritmos son
#de tipo scored-base (basados en puntajes)
bAlgoritmo <-c("hc","tabu")
bThreshold <-c(0.4,0.5,0.6,0.7)
bM <-c(400,450,500)
bScore <-c("aic","bic","k2","bde")
bScoreType <-c("aic","bic","k2","bde")

#Vector en donde se almacenara los score de las redes generadas
vectorPuntajesScore <-c(1:384)
#Identificador de fila
indiceVector <- 0

#Bucles para generar todas las permutaciones posibles de las redes bayesianas con los
#parámetros que se establecieron
for (vAlgoritmo in bAlgoritmo) {
 for (vThreshold in bThreshold) {
  for (vM in bM) {
   for (vScore in bScore) {
    for (vScoreType in bScoreType) {
     bootstrength<-boot.strength(data_train.9.discr, algorithm = vAlgoritmo,R = 100,
m=vM, cpdag=TRUE, algorithm.args = list(score = vScore))
     av_net<- averaged.network(bootstrength, threshold = vThreshold)
     res = pdag2dag(av_net, ordering = nodes(av_net))
#Una vez que se realiza el aprendizaje, se incrementa en 1 el contador y se almacena en el
#vector de puntajes
     indiceVector=indiceVector+1
     vectorPuntajesScore [indiceVector]<- score(res, data = data_train.9.discr,
type=vScoreType)
    }
   }
  }
 }
}
}



#Vectores de parámetros usados en la generación de las redes bayesianas, los algoritmos son
#de tipo constraint-base (basados en restricciones)
bAlgoritmo <-c("gs","iamb","fast.iamb","inter.iamb")
```

```
bThreshold <-c(0.4,0.5,0.6,0.7)
bM <-c(400,450,500)
bScoreType ←c("aic","bic","k2","bde")
#Vector en donde se almacenara los score de las redes generadas
vectorPuntajesRestricciones <-c(1:192)
#Identificador de fila
indiceVector <- 0

#Bucles para generar todas las permutaciones posibles de las redes bayesianas con los
#parametros que se establecieron
for (vAlgoritmo in bAlgoritmo) {
 for (vThreshold in bThreshold) {
  for (vM in bM) {
    for (vScoreType in bScoreType) {
      bootstrength<-boot.strength(data_train.9.discr, algorithm = vAlgoritmo,R = 100,
m=vM, cpdag=TRUE)
      av_net<- averaged.network(bootstrength, threshold = vThreshold)
      res = pdag2dag(av_net, ordering = nodes(av_net))
#Una vez que se realiza el aprendizaje, se incrementa en 1 el contador y se almacena en el
#vector de puntajes
      indiceVector=indiceVector+1
      vectorPuntajesRestricciones[indiceVector]<- score(res, data = data_train.9.discr,
type=vScoreType)
    }
  }
 }
}
}
```

```
#Vectores de parámetros usados en la generación de las redes bayesianas, los algoritmos son
#de tipo hibrido
bAlgoritmo <-c("mmhc","rsmax2")
bThreshold <-c(0.4,0.5,0.6,0.7)
bM <-c(400,450,500)
bScoreType ←c("aic","bic","k2","bde")
#Vector en donde se almacenara los score de las redes generadas
vectorPuntajesHibridos <-c(1:96)
indiceVector <- 0

#Bucles para generar todas las permutaciones posibles de las redes bayesianas con los
#parámetros que se establecieron
for (vAlgoritmo in bAlgoritmo) {
 for (vThreshold in bThreshold) {
  for (vM in bM) {
    for (vScoreType in bScoreType) {
      bootstrength<-boot.strength(data_train.9.discr, algorithm = vAlgoritmo,R = 100,
m=vM, cpdag=TRUE)
      av_net<- averaged.network(bootstrength, threshold = vThreshold)
      res = pdag2dag(av_net, ordering = nodes(av_net))
```

```
#Una vez que se realiza el aprendizaje, se incrementa en 1 el contador y se almacena en el
#vector de puntajes
        indiceVector=indiceVector+1
        vectorPuntajesHibridos[indiceVector]<- score(res, data = data_train.9.discr,
type=vScoreType)
      }
    }
  }
}
}

# <---------------------------------------------------------------------------->
#Fin script
# <---------------------------------------------------------------------------->



# <---------------------------------------------------------------------------->
#Script 2
#El siguiente script realiza propagaciones de manera automática, estableciendo como
#evidencia las regiones de precipitación, luego mide la probabilidad de cada uno de los índices
#climáticos, en cada uno de sus respectivos rangos, en aquellos cuya probabilidad aumente en
#1% o más y cuyo rango no se encuentre entre -0.3 y 0.3, quedara registrado en un data frame
#que almacenara la información no solo de las probabilidades del índice, sino además de la
#región así como el rango del índice
# <---------------------------------------------------------------------------->

#Crear un vector con los números de las regiones, (1,2,3,4,5,6,7,8,9)
vectorRegiones <- as.vector(unique(data_train.9.discr$region))

#Crear un vector con los índices climáticos, y excluye el primer elemento y que corresponde a
#"region"
vectorIndices <- colnames(data_train.9.discr)
vectorIndices<-vectorIndices[-1]

#Creación de un data frame (df) vacío que almacenara los datos posteriormente
df_indicesVsRegiones<-data.frame()

#Identificador para el df
indiceCodigo=1

#Bucle para recorrer cada una de las regiones
for (variableRegion in vectorRegiones) {
#Realiza una propagación estableciendo cada una de las regiones como evidencia
  e.regx<- setFinding(bngrain, nodes="region", states = variableRegion)
#Bucle para recorrer todos los índices en busca de cambios de probabilidad
  for (variableNodos in vectorIndices) {

#Probabilidad a priori de un determinado indice
    prior<- cbind(as.data.frame(querygrain(bngrain, nodes =
variableNodos)),"without_evidence")
    prior$factor<- rownames(prior)
```

```
   names(prior)<- c("probabilities", "evidence", "type")

#Probabilidad a posteriori de un determinado indice
   posterior<- cbind(as.data.frame(querygrain(e.regx, nodes =
variableNodos)),"with_evidence")
   posterior$factor<- rownames(posterior)
   names(posterior)<- c("probabilities", "evidence", "type")

#Se obtiene el total de registros encontrados
   numFilas=length(prior[,1])

#Con el número de registros previamente obtenido, se realiza una nueva iteración, esta
#consiste en comparar cada registro obtenido y determinar la diferencia de sus probabilidades
#(posteriori – priori), si dicha diferencia es mayor a 1% (0.01) y su rango se encuentre fuera de
#los valores entre -0.3 y 0.3 se almacena en el data frame final
   for (contadorNumFilas in 1:numFilas) {
    if(posterior[contadorNumFilas,1]-prior[contadorNumFilas,1]>=0.01 &
prior$type[contadorNumFilas]!="[-0.3,0)" & prior$type[contadorNumFilas]!="[0,0.3)"){

#Se almacenan los valores en variables temporales
      reg_temp=variableRegion
      nod_temp=variableNodos
      prob_prior_temp=as.vector(prior$probabilities[contadorNumFilas])
      prob_posterior_temp=as.vector(posterior$probabilities[contadorNumFilas])
      type_temp=prior$type[contadorNumFilas]

#Las variables temporales se guardan en el data frame final
      df_temp<-
data.frame(indiceCodigo,reg_temp,nod_temp,prob_prior_temp,prob_posterior_temp,type_
temp)
      df_indicesVsRegiones<-rbind(df_indicesVsRegiones,df_temp)
      indiceCodigo=indiceCodigo+1
    }
   }
  }
}

#Renombrar las columnas del data frame
names(df_indicesVsRegiones)<-c("codigo","region","indice","prob prior","prob
posterior","rango")

#Guardar el data frame, la primera sentencia lo guarda en un archivo de tipo *.Rdata y la
#segunda en un archivo *.txt
save(df_indicesVsRegiones, file= "df_regiones.RData")
write.table(df_indicesVsRegiones, file="df_regiones.txt")

# <---------------------------------------------------------------------------->
#Fin script
# <---------------------------------------------------------------------------->
```

```
# <---------------------------------------------------------------------------->
#Script 3
#De manera similar al anterior, el siguiente script realiza propagaciones de manera automática
#y almacena la información de aquellos registros cuyo cambio de probabilidad es mayor a
#0.01 y su rango no se encuentre entre -0.3 y 0.3, la diferencia radica en que la evidencia en
#este script se  setea en los índices, y se observa cómo se comportan las regiones.
# <---------------------------------------------------------------------------->

#Obtener el vector de regiones (de 1 a 9)
vectorRegiones <- as.vector(unique(data_train.9.discr$region))

#Obtener un vector de indices climaticos y luego se excluye el primer elemento (que es
"region")
vectorIndices <- colnames(data_train.9.discr)
vectorIndices<-vectorIndices[-1]

#Inicializar el data frame
df_indicesVsRegiones<-data.frame()

#Contador para el número de registros
contador=1

#Bucle para recorrer todos los índices climáticos
for (variableIndice in vectorIndices) {
#en la variable "rangoIndices" se almacenan todos los rangos de un determinado índice (no
#todos los índices tienen los mismos rangos) y luego son usados para una nueva iteración
  rangoIndices<-as.vector(unique(data_train.9.discr[,variableIndice]))

#Bucle para recorrer todos los rangos de un determinado índices
  for (variableRango in rangoIndices) {
#Realizar una propagación estableciendo como evidencia un índice y un rango
    indiceX<- setFinding(bngrain, nodes=variableIndice, states = variableRango)

#Probabilidad a priori del nodo región
    prior<- cbind(as.data.frame(querygrain(bngrain, nodes = "region")),"without_evidence")
    prior$factor<- rownames(prior)
    names(prior)<- c("probabilities", "evidence", "type")

#Probabilidad a posteriori del nodo región
    posterior<- cbind(as.data.frame(querygrain(indiceX, nodes = "region")),"with_evidence")
    posterior$factor<- rownames(posterior)
    names(posterior)<- c("probabilities", "evidence", "type")

#Bucle para recorrer cada uno de los registros, debido a que son 9 regiones se debe iterar 9
#veces para recorrer todos los registros
    for (variableFilaIndices in 1:9) {

#Si el cambio de probabilidad es mayor a 0.01 y el rango se encuentre fuera de los valores
#entre -0.3 y 0.3 entonces se almacena en el data frame
```

```
    if(posterior$probabilities[variableFilaIndices]-
prior$probabilities[variableFilaIndices]>=0.01 & variableRango!="[-0.3,0)" &
variableRango!="[0,0.3)"){

#Se almacenan los valores en variables temporales
      reg_temp=prior$type[variableFilaIndices]
      nod_temp=variableIndice
      prob_prior_temp=as.vector(prior$probabilities[variableFilaIndices])
      prob_posterior_temp=as.vector(posterior$probabilities[variableFilaIndices])
      rango_temp=variableRango

#Las variables temporales se guardan en el data frame final
      df_temp<-
data.frame(contador,reg_temp,nod_temp,prob_prior_temp,prob_posterior_temp,rango_te
mp)
      df_indicesVsRegiones<-rbind(df_indicesVsRegiones,df_temp)
      contador=contador+1
    }
   }
  }
}

#Renombrar las columnas del data frame
names(df_indicesVsRegiones)<-c("codigo","region","indice","prob prior","prob
posterior","rango")

#Guardar el data frame, la primera sentencia lo guarda en un archivo de tipo *.Rdata y la
#segunda en un archivo *.txt
save(df_indicesVsRegiones, file= "df_indices.RData")
write.table(df_indicesVsRegiones, file="df_indices.txt")

# <---------------------------------------------------------------------------->
#Fin script
# <---------------------------------------------------------------------------->




# <---------------------------------------------------------------------------->
#Script 4
# El siguiente script también realiza propagaciones automáticas y guarda la información de
#aquellos cuyo cambio de probabilidad es mayor a 0.01, la evidencia es seteada en los índices
#y se observa el comportamiento de los demás índices, mas no de las regiones. No se setea
#los índices entre los rangos [-0.3, 0) y [0, 0.3).
# <---------------------------------------------------------------------------->

#Generar el vector de índices, excluyendo el nodo “region” y los nodos aislados “pna” y “tna”
vectorIndices <- colnames(data_train.9.discr)
vectorIndices<-vectorIndices[c(-1,-19,-28)]

#Inicializar el data frame
df_indicesVsRegiones<-data.frame()
```

```
#Contador para el número de registros
contador=1

#Bucle para iterar sobre cada uno de los índices climáticos
for (variableIndice in vectorIndices) {
#Variable en la que se almacena los rangos de un determinado índice
 rangoIndices<-as.vector(unique(data_train.9.discr[,variableIndice]))

#Bucle para iterar sobre los rangos de un determinado índice
 for (variableRango in rangoIndices) {
 #Si el rango es diferente de [-0.3, 0) y [0, 0.3) entonces realiza la propagación, caso contrario
#realiza una nueva iteración
  if(variableRango!= "[-0.3,0)" & variableRango!="[0,0.3)"){
#Propagacion estableciendo como evidencia un índice en un determinado rango
    indicePadre<- setFinding(bngrain, nodes=variableIndice, states = variableRango)
#Bucle para iterar sobre los todos los demás índices, excepto aquel que haya sido seteado
#como evidencia
    for (variableIndiceHijo in vectorIndices) {
#Condición para verificar que no se compare el cambio de probabilidad en un índice, si este
#mismo fue seteado como evidencia
     if(variableIndice!=variableIndiceHijo){
#Probabilidad a priori de un índice seteado como evidencia otro índice
       prior<- cbind(as.data.frame(querygrain(bngrain, nodes =
variableIndiceHijo)),"without_evidence")
       prior$factor<- rownames(prior)
       names(prior)<- c("probabilities", "evidence", "type")

#Probabilidad a posteriori de un índice seteado como evidencia otro índice
       posterior<- cbind(as.data.frame(querygrain(indicePadre, nodes =
variableIndiceHijo)),"with_evidence")
       posterior$factor<- rownames(posterior)
       names(posterior)<- c("probabilities", "evidence", "type")

#Se obtiene el total de registros encontrados
       numFilas=length(prior[,1])

#Bucle para iterar sobre el total de registros encontrados
       for (variableFilaIndices in numFilas) {
#Si el cambio de probabilidad es mayor a 0.01 y el rango no es [-0.3, 0) ni [0, 0.3) entonces
#almacena los registros en el data frame
        if(posterior$probabilities[variableFilaIndices]-
prior$probabilities[variableFilaIndices]>=0.01 & prior$type[variableFilaIndices]!="[-0.3,0)" &
prior$type[variableFilaIndices]!="[0,0.3)"){


#Guardar los valores de probabilidad en variables temporales
         prob_prior_temp=as.vector(prior$probabilities[variableFilaIndices])
         prob_posterior_temp=as.vector(posterior$probabilities[variableFilaIndices])
         rango_hijo_temp=prior$type[variableFilaIndices]

#Almacena el registro en el data frame
```

```
          df_temp<-
data.frame(contador,variableIndice,variableRango,variableIndiceHijo,prob_prior_temp,prob
_posterior_temp,rango_hijo_temp)
          df_indicesVsRegiones<-rbind(df_indicesVsRegiones,df_temp)
          contador=contador+1
        }
       }
      }
     }
   }
  }
}
```

#Renombrar las columnas del data frame
names(df_indicesVsRegiones)<-c("codigo","Indice Padre","Rango Indice padre","Indice hijo","prob prior","prob posterior","Rango Indice hijo")

#Guardar el data frame, la primera sentencia lo guarda en un archivo de tipo *.Rdata y la #segunda en un archivo *.txt
save(df_indicesVsRegiones, file= "df_indicesVSindices.RData")
write.table(df_indicesVsRegiones, file="df_indicesVSindices.txt")

<---------->
#Fin script
<---------->

UNIVERSIDAD DEL AZUAY

Doctora María Elena Ramírez Aguilar, Secretaria de la Facultad de Ciencias de la Administración de la Universidad del Azuay

C E R T I F I C A:

Que, el Consejo de Facultad en sesión del 23 de enero de 2019, conoció y aprobó la solicitud para realización del trabajo de titulación, presentada por:

Estudiante:	Paúl Renato Ávila Andrade (cód. 73072)
Tema:	**"Exploración de Redes Bayesianas para detectar relaciones entre la precipitación del Ecuador y teleconexiones climáticas ".** Previo a la obtención del título de Ingeniero de Sistemas y Telemática
Director:	Agrim. Daniela Ballari
Tribunal:	Ing. Marcos Orellana Cordero e Ing. Chester Sellers

Plazo de presentación del trabajo de titulación: Se fijó como plazo para la entrega del trabajo de titulación, conforme a la Disposición Tercera del Reglamento de Régimen Académico, un período académico contado desde la fecha de aprobación del diseño del trabajo, esto es hasta el 23 de julio de 2019.

E INFORMA:

Que, en aplicación de la Disposición General Cuarta del Reglamento de Régimen Académico vigente, en caso de que el estudiante no culmine y apruebe el trabajo de titulación luego de dos períodos académicos contados a partir de su fecha de culminación de estudios, deberá realizar la actualización de conocimientos previa a su titulación.

Cuenca, 24 de enero de 2019

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad de
Ciencias de la Administración

UNIVERSIDAD DEL AZUAY
Facultad de Ciencias de la Administración
SECRETARIA

CONVOCATORIA

Por disposición de la Junta Académica de la escuela de Ingeniería de Sistemas y Telemática se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: **Exploración de redes bayesianas para detectar relaciones entre la precipitación del Ecuador y Teleconexiones Climáticas,** presentado por el estudiante Paúl Renato Ávila Andrade con código 73072, previa a la obtención del título de Ingeniero de Sistemas y Telemática, para el día **Jueves, 17 de enero de 2019 a las 10:00**

Tomar en cuenta que posterior a la sustentación del Diseño del Trabajo de Titulación, por ningún concepto se puede realizar modificaciones ni cambios en los documentos; únicamente, en caso de diseño aprobado con modificación, el Director adjuntará al esquema un oficio indicando que se procede con los cambios sugeridos.

Cuenca, 15 de enero de 2019

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad

Agrim. Daniela Ballari .....

Ing. Marcos Orellana Cordero .....

Ing. Chester Sellers .....

Oficio Nro. 003-2019-DIST-UDA

Cuenca, 4 de enero de 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De nuestras consideraciones,

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 4 de enero del 2019, revisó la documentación del trabajo de titulación denominado **"EXPLORACIÓN DE REDES BAYESIANAS PARA DETECTAR RELACIONES ENTRE LA PRECIPITACIÓN DEL ECUADOR Y TELECONEXIONES CLIMÁTICAS"**, por la/el estudiante **PAÚL RENATO ÁVILA ANDRADE,** con código estudiantil UA073072, estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por **DANIELA BALLARI**, previo a la obtención del título de Ingeniero de Sistemas y Telemática.

La Junta Académica considera que la documentación cumple con las normas legales y reglamentarias de la Universidad y de la Facultad de Ciencias de la Administración y designa como miembros del tribunal a MARCOS ORELLANA y CHESTER SELLERS, así por su digno intermedio, el conocimiento y aprobación por parte del Consejo de Facultad.

Atentamente,

Marcos Orellana Cordero
Coordinador de la Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

UNIVERSIDAD DEL AZUAY

ACTA
SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

Fecha de sustentación: **Jueves, 17 de enero de 2019 a las 10:00**

1.9. Nombre del estudiante: Paúl Renato Ávila Andrade
1.10. Código: 73072
1.11. Director sugerido: Agrim. Daniela Ballari
1.12. Codirector (opcional): _____
1.12.1. Tribunal: Ing. Marcos Orellana Cordero e Ing. Chester Sellers
1.12.2. Título propuesto: **Exploración de redes bayesianas para detectar relaciones entre la precipitación del Ecuador y Teleconexiones Climáticas**

1.12.3. Aceptado sin modificaciones : ✓

1.12.4. Aceptado con las siguientes modificaciones:

_____

_____

_____

1.12.5. No aceptado
1.12.6. Justificación:

_____

_____

_____

Tribunal

Agrim. Daniela Ballari

Ing. Marcos Orellana Cordero

Ing. Chester Sellers

Sr. Paúl Renato Ávila Andrade

Dra. María Elena Ramírez Aguilar
Secretaria de la Facultad

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN
(Tribunal)

1.11. Nombre del estudiante: Paúl Renato Ávila Andrade
1.12. Código : 73072
1.13. Director sugerido: Agrim. Daniela Ballari
1.13.1. Codirector (opcional):
1.14. Título propuesto: **Exploración de redes bayesianas para detectar relaciones entre la precipitación del Ecuador y Teleconexiones Climáticas**
1.15. Revisores tribunal: Ing. Marcos Orellana Cordero e Ing. Chester Sellers
1.16. Recomendaciones generales de la revisión:

	Cumple	No cumple
Problemática y/o pregunta de investigación	✓	
29. ¿Presenta una descripción precisa y clara?	✓	
30. ¿Tiene relevancia profesional y social?	✓	
Objetivo general		
31. ¿Concuerda con el problema formulado?	✓	
32. ¿Se encuentra redactado en tiempo verbal infinitivo?	✓	
Objetivos específicos		
33. ¿Permiten cumplir con el objetivo general?	✓	
34. ¿Son comprobables cualitativa o cuantitativamente?	✓	
Metodología		
35. ¿Se encuentran disponibles los datos y materiales mencionados?	✓	
36. ¿Las actividades se presentan siguiendo una secuencia lógica?	✓	
37. ¿Las actividades permitirán la consecución de los objetivos específicos planteados?	✓	
38. ¿Las técnicas planteadas están de acuerdo con el tipo de investigación?	✓	
Resultados esperados		
39. ¿Son relevantes para resolver o contribuir con el problema formulado?	✓	
40. ¿Concuerdan con los objetivos específicos?	✓	
41. ¿Se detalla la forma de presentación de los resultados?	✓	
42. ¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	✓	

Agrim. Daniela Ballari

Ing. Marcos Orellana Cordero

Ing. Chester Sellers

UNIVERSIDAD
DEL AZUAY

DOCTORA LARIZA ROBLES SERRANO, SECRETARIA (E) DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN DE LA UNIVERSIDAD DEL AZUAY

C E R T I F I C A:

Que, el señor **AVILA ANDRADE PAUL RENATO** con código de estudiante Nro. **73072,** alumno de la carrera de **INGENIERIA DE SISTEMAS Y TELEMATICA,** tiene aprobado el **91,89%** de créditos de su malla curricular.

Cuenca, 19 de diciembre de 2018

Dra. Lariza Robles Serrano
SECRETARIA (E) DE LA FACULTAD
DE CIENCIAS DE LA ADMINISTRACIÓN

UNIVERSIDAD DEL AZUAY
Facultad de Ciencias de la Administración
SECRETARÍA

Derecho No. **001-002-000076762**

mjmr.-

0880546

Cuenca, 20 de diciembre de 2018

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Yo, **Daniela Elisabet Ballari** informo que he revisado el protocolo de trabajo de titulación, elaborado previo a la obtención del título de Ingeniero de Sistemas y Telemática, **"EXPLORACIÓN DE REDES BAYESIANAS PARA DETECTAR RELACIONES ENTRE LA PRECIPITACIÓN DEL ECUADOR Y TELECONEXIONES CLIMÁTICAS"**, realizado por el estudiante **Paúl Renato Ávila Andrade,** con código estudiantil **73072** protocolo que a mi criterio, cumple con los lineamientos y requerimientos establecidos por la carrera.

Por lo expuesto, me permito sugerir que sea considerado para la revisión y sustentación del mismo,

Sin otro particular, me suscribo.

Atentamente

Daniela Elisabet Ballari

Universidad del Azuay

Cuenca, 4 de enero de 2019

Ingeniero,
Oswaldo Merchán Manzano
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN
UNIVERSIDAD DEL AZUAY

De mi consideración,

Estimado Señor Decano, yo **Paúl Renato Ávila Andrade** con C.I. **1400918288,** código estudiantil 73072, estudiante de la Carrera de Ingeniería de Sistemas y Telemática, solicito encarecidamente a usted, la aprobación del protocolo de trabajo de titulación con el tema **"EXPLORACIÓN DE REDES BAYESIANAS PARA DETECTAR RELACIONES ENTRE LA PRECIPITACIÓN DEL ECUADOR Y TELECONEXIONES CLIMÁTICAS."** previo a la obtención del título de Ingeniero de Sistemas y Telemática para lo cual adjunto la documentación respectiva.

Por la favorable acogida que brinde a la presente, anticipo mi agradecimiento.

Atentamente

Paúl Renato Ávila Andrade
Estudiante de la Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

0880864

UNIVERSIDAD DEL AZUAY

GUÍA PARA LA ELABORACIÓN Y PRESENTACIÓN DE LA DENUNCIA/PROTOCOLO DE TRABAJO DE TITULACIÓN

1. DATOS GENERALES

1.1 Nombre del estudiante: Ávila Andrade Paúl Renato

1.1.1 Código: ua073072

1.1.2 Contacto: Telf.: 07 2780 337 Cel.: 09 8888 5901 correo: r4avila@hotmail.com

1.2 Director sugerido: Agrim. Ballari Daniela Elisabet, PhD

1.2.1 Contacto: 098 120 1458; dballari@uazuay.edu.ec

1.3 Co-director sugerido:

1.3.1 Contacto:

1.4 Asesor metodológico: Gabriel Barros, PhD

1.5 Tribunal designado:

1.6 Aprobación:

1.7 Línea de investigación de la carrera:

1.7.1 Código UNESCO: 1203.04 Ciencia de los Ordenadores, Inteligencia Artificial

1.7.2 Tipo de Trabajo: Estudios comparados

1.8 Área de estudio: Sistemas expertos, estadística, sistemas de información geográfica

1.9 Título propuesto:

Exploración de Redes Bayesianas para detectar relaciones entre la precipitación del Ecuador y teleconexiones climáticas.

1.10 Subtítulo:

1.11 Estado del proyecto: Nuevo

2. CONTENIDO

2.1 Motivación de la investigación:

Las precipitaciones juegan un papel fundamental en la economía de un país ya que influyen en muchos sectores productivos, por citar algunos ejemplos tenemos: la agricultura (Ali et al., 2018; Sudha, Venugopal, & Ambujam, 2008), ganadería (Engler, Wehrden, & Baumgärtner, 2019), pesca (Pratiwi & Sukardjo, 2018), turismo (Abd-elhamid, Fathy, & Zelen, 2018) y generación de energía eléctrica. En un estudio realizado en la cuenca del río Jubones en Ecuador, se evaluó el impacto del cambio climático en la generación de energía hidroeléctrica a través de una herramienta de modelado (SWAT), los resultados demostraron que durante las temporadas de sequías la planta experimentará una caída del 13% en la producción de energía debido a una reducción del 17% del flujo de agua (Hasan & Wyseure, 2018). La carencia de estudios sobre datos climáticos de precipitación ha dificultado la comprensión y predicción del clima, de ahí surge la importancia del estudio de las teleconexiones climáticas ya que ayudarán a realizar pronósticos del clima mucho más fiables, permitiendo que todos los sectores económicos que son influenciados directamente por precipitaciones puedan planificar de mejor manera sus actividades así como adoptar políticas orientadas a una gestión sostenible de los recursos hídricos.

2.2 Problemática:

En el Ecuador existen investigaciones sobre las teleconexiones climáticas (Vicente-Serrano et al., 2017), sin embargo hasta la fecha no se conocen estudios en los que se haya realizado análisis multivariado más allá de los índices de ENSO (*El Niño Southern Oscillation*). Realizar este tipo de investigación es importante, ya que los índices climáticos no se encuentran aislados sino que interactúan entre sí, por lo que representan una mejor aproximación de la realidad. De ahí que el objetivo de este trabajo es realizar un estudio multivariado utilizando redes bayesianas para determinar la influencia de índices climáticos tales como ENSO, *Trans-Niño Index* (TNI), *Southern Oscillation Index* (SOI), *Pacific Decadal Oscillation* (PDO) sobre las precipitaciones en el país. Además al aplicar un enfoque bayesiano puede ser útil para determinar si la influencia del índice climático es dependiente de otros índices o de diferentes regiones. Se espera que la información obtenida a través de este estudio pueda servir para comprender de mejor manera el comportamiento de las precipitaciones en nuestro país.

2.3 Pregunta de investigación:

¿Cuáles son las teleconexiones que tienen una influencia directa sobre la precipitación? ¿Y cuáles una influencia indirecta? ¿Existen diferencias en las relaciones de acuerdo a la localización de la precipitación (Costa, Sierra, Oriente)?

2.4 Resumen

Este trabajo explora, a través del aprendizaje de una red bayesiana, la estructura y fuerza de las relaciones entre la precipitación a escala del Ecuador continental y teleconexiones climáticas. Los datos a utilizar consistirán en imágenes satelitales de precipitación e índices climáticos relacionados con ENSO, TNI, SOI, PDO entre otros. La red bayesiana es un modelo probabilístico que se compone de nodos que representan variables, conectados a través de arcos directos que representan influencia y, bajo ciertos supuestos, causalidad entre los nodos. Así, la red permite propagar probabilidades a través de sus nodos y arcos para explorar preguntas como: ¿Cuáles son las teleconexiones que tienen una influencia directa sobre la precipitación? ¿Y cuáles una influencia indirecta? ¿Existen diferencias en las relaciones de acuerdo a la localización de la precipitación (Costa, Sierra, Oriente)?

2.5 Estado del arte y marco teórico

Las teleconexiones se definen como una correlación entre variables climáticas separadas por grandes distancias. En la actualidad para el estudio de teleconexiones climáticas se usan diferentes métodos tales como: regresión lineal (Liu, Di, Chen, & DaMassa, 2018), análisis de componentes principales (PCA) (Fierro, 2014), análisis wavelet (Konapala, Valiya, & Ashok, 2017) y correlación punto a punto (Fierro, 2014). En la literatura podemos encontrar trabajos que hacen uso de estos métodos. Sin embargo, esto métodos presentan la limitación, según Torre-Gea, Soto-Zarazua, Guevara-Gonzalez, & Rico-Garcia. (2011), que técnicas como regresión lineal, PCA y métodos de agrupamiento fragmentan la información y asumen independencia espacial para reducir dimensionalidad. Por otra parte existen técnicas de análisis multivariado como las teleconexiones ortogonales empíricas y las funciones ortogonales empíricas (EOF) las cuales se ajustan mucho mejor a este tipo de problemas, una revisión de estos métodos se puede encontrar en (Lee, Lee, & Julien, 2018).

En los últimos años el uso de métodos bayesianos para análisis de teleconexiones climáticas se ha popularizado. Una red bayesiana se define como "un grafo acíclico dirigido que representa un conjunto de variables (nodos) y sus independencias condicionales probabilísticas (codificada en los arcos)" (Correa, Bielza, & Pamies-teixeira, 2009). Se trata de un modelo probabilístico que permite inferir sobre variables desconocidas a partir de las variables que si son conocidas u observadas. Al ser un tipo

de sistemas expertos tiene aplicación en casi cualquier ámbito de las ciencias, por ejemplo: medicina, biología, astrofísica, informática, química y economía.

Debido al creciente número de estaciones meteorológicas, así como a la mayor disponibilidad de datos climáticos a partir de imágenes satelitales, la climatología se ha convertido en campo fértil para el empleo de técnicas con enfoque bayesiano. En la actualidad ya se han aplicado estas técnicas tanto para la predicción del clima (Das & Ghosh, 2015; Zeng, Hsieh, Shabbar, & Burrows, 2011), así como para el análisis de teleconexiones en donde se han aplicado las técnicas bayesianas tanto a variables climáticas (Torre-Gea et al., 2011) como a índices climáticos (Hiep Nguyen Duc, Rivett, MacSween, & Le-Anh, 2017; Ebert-Uphoff & Deng, 2012). En un estudio realizado en Vietnam sobre precipitación, Hiep. Duc, Bang & Quang (2018) usaron el método de promediado bayesiano de modelos, en esta investigación se usaron 3 índices climáticos: Indian Ocean Dipole (IOD), Interdecadal Pacific Oscillation (IPO) y El Niño–Southern Oscillation (ENSO). Los resultados mostraron que estos índices tienen incidencia sobre las precipitaciones en Vietnam, la fuerza con la que estos índices afectan al clima varía según sea el índice, la región y la estacionalidad. Además, se encontró que la interacción entre los índices también influye sobre las precipitaciones.

2.6 Hipótesis

Las redes bayesianas pueden ser empleadas para el análisis de teleconexiones climáticas y los resultados obtenidos tendrán relación con otras técnicas empleadas en el análisis climatológico.

2.7 Objetivo general

Explorar, a través del aprendizaje de una red bayesiana, la estructura y fuerza de las relaciones entre la precipitación a escala del Ecuador continental y teleconexiones climáticas.

2.8 Objetivos específicos

- Preparar datos de precipitación en regionalización del Ecuador e índices climáticos a utilizar.
- Crear y entrenar una red bayesiana usando el lenguaje de programación R y los datos de precipitación obtenidos.
- Determinar e interpretar la relación entre los índices climáticos y las precipitaciones en el Ecuador.

2.9 Metodología

2.9.1 Área de estudio

El lugar de estudio seleccionado fue el Ecuador continental, que cuenta con una superficie de 250 000 km^2 y 3 regiones claramente definidas: las planicies de la costa, los andes y la amazonía. El clima de Ecuador está influencia por una gran cantidad de

factores cuya influencia cambia según la región, lo que resulta en una gran variedad de precipitación espacio-temporal.

2.9.2 Materiales

Modos de variabilidad climática e índices climáticos

La variabilidad climática es el resultado de la combinación de patrones espaciales, los patrones espaciales más destacados son conocidos como modos de variabilidad climática y tiene la capacidad de incidir en el clima a escala espacial y temporal, y también pueden afectar a otros modos de variabilidad. Los modos de variabilidad climática son identificados a través del estudio de las teleconexiones espaciales (Blunden et al., 2011).

Entre los modos de variabilidad climática más comunes tenemos: El Niño-Southern Oscillation (ENSO), Central Pacific El Niño (Modoki), Pacific Decadal and Interdecadal Varibility (PDO), North Atlatic Oscillation (NAO), Pacific/North America atmospheric teleconnection (PNA), etc. Los modos de variabilidad son definidos a través de índices climáticos, por ejemplo, ENSO está definido por varios índices como son: NINO3, NINO3.4, SOI entre otros. Los cambios en los índices están asociados a cambios a gran escala en los campos climáticos. Debido a que ningún índice puede aislar un fenómeno de todos los demás efectos, múltiples índices definen el mismo fenómeno climático, por tanto cada índice se ve afectado por muchos fenómenos climáticos (Blunden et al., 2011).

Los índices climáticos a utilizar en este trabajo son:

Índice	Nombre	Acceso
Relacionados con ENSO		
Niño 1+2	El Niño región 1+2	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Niño 3	El Niño región 3	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Niño 4	El Niño región 4	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Niño 3.4	El Niño región 3,4	https://www.esrl.noaa.gov/psd/data/climateindices/list/
SOI	Southern Oscillation Index	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EMI	El Niño Modoki	http://www.jamstec.go.jp/frsgc/research/d1/iod/e/elnmodoki/about_elnm.html
TNI	Trans-Niño	https://www.esrl.noaa.gov/psd/data/climateindices/list/
MEI	ENSO multivariado	https://www.esrl.noaa.gov/psd/data/climateindices/list/
ONI	Niño oceánico	https://www.esrl.noaa.gov/psd/data/climateindices/list/

Relacionados con el Océano Pacífico		
PDO	Pacific Decadal Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Relacionados con Océano Atlántico		
AMO	Atlantic Multidecadal Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
TNA	Tropical Northern Atlantic	https://www.esrl.noaa.gov/psd/data/climateindices/list/
TSA	Tropical Southern Atlantic	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Relacionados con índices atmosféricos		
AO	Artic Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
Relacionados con teleconexiones		
EA	East Atlantic	https://www.esrl.noaa.gov/psd/data/climateindices/list/
WP	Western Pacific	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EP/NP	East Pacific North Pacific	https://www.esrl.noaa.gov/psd/data/climateindices/list/
PNA	Pacific North American	https://www.esrl.noaa.gov/psd/data/climateindices/list/
NAO	North Atlantic Oscillation	https://www.esrl.noaa.gov/psd/data/climateindices/list/
EA/WR	East Atlantic Western Russian	https://www.esrl.noaa.gov/psd/data/climateindices/list/
DMI	Dipole Mode Index	http://www.jamstec.go.jp/frsgc/research/d1/iod/e/iod/about_iod.html

Datos satelitales con reducción de escala

Imágenes satelitales TRMM con reducción de escala a 5km (2001 - 2011), procesado por el director de este proyecto en investigaciones previas. Es el resultado del análisis de reducción de escala con covariables. Para más información consultar: Ulloa, Jacinto; Ballari, Daniela; Campozano, Lenin; Samaniego, Esteban. 2017. "Two-Step Downscaling of TRMM 3b43 V7 Precipitation in Contrasting Climatic Regions with Sparse Monitoring: The Case of Ecuador in Tropical South America." Remote Sensing. 9, no. 7: 758.

2.9.3 Métodos

Este trabajo se centra en la construcción e interpretación de una red bayesiana para para identificar la relación entre regiones de estacionalidad de precipitación e índices climáticos. Los procedimientos de preparación de datos reportados a continuación fueron generados en el marco del proyecto CEPRAXII "Representación espacial de las teleconexiones climáticas en la precipitación del Ecuador", en el cual también se enmarca este proyecto de Trabajo de Titulación.

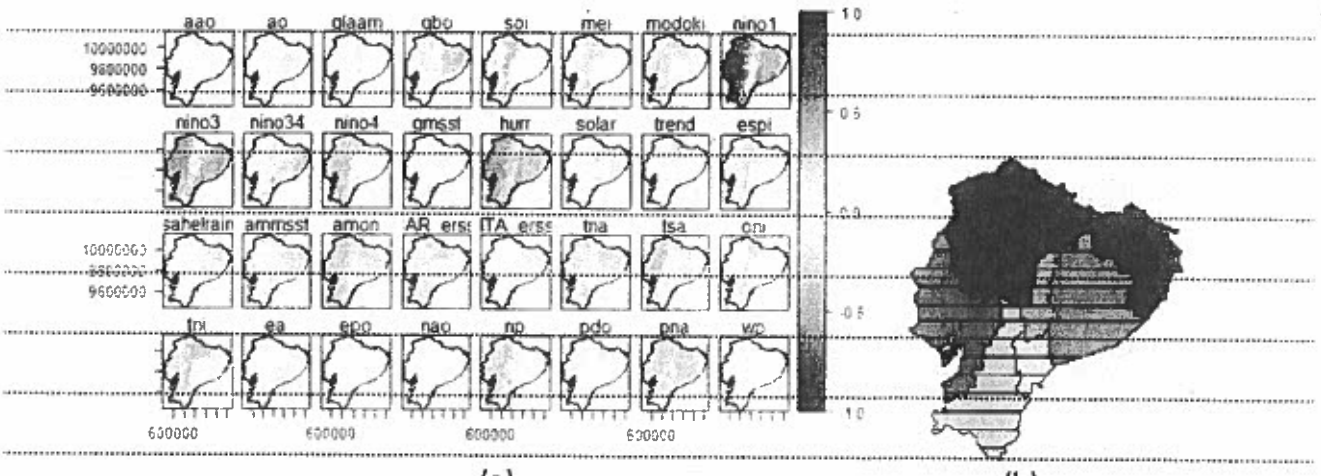
1. Preparación de datos

a. Correlación pixel a pixel de precipitación e índices climáticos

Se utilizó correlación no paramétrica pixel a pixel entre la serie temporal de precipitación, correspondiente al pixel en cuestión, y los diferentes índices climáticos. La figura 1.a muestra los mapas de dichas correlaciones, con azul las correlaciones positivas y con rojo las correlaciones negativas. Valores altos de correlación tanto positiva como negativa representan la existencia de teleconexiones climáticas.

b. Regionalización de estacionalidades de precipitación a 5km

Con las imágenes de precipitación a 5km se ejecutó el procedimiento de regionalización funcional reportado en el artículo "Ballari D, Giraldo R, Campozano L, Samaniego E. *Spatial functional data analysis for regionalizing precipitation seasonality and intensity in a sparsely monitored region: Unveiling the spatio-temporal dependencies of precipitation in Ecuador. Int J Climatol.* 2018; 38:3337–3354." Dicho artículo se basó en imágenes TRMM de 27km y se encontraron 5 regiones de estacionalidad de precipitación. En este caso, aplicando la regionalización sobre las imágenes de 5km, se encontraron 9 regiones principales de estacionalidad. La figura 1.b muestra las 9 regiones.



(a) (b)

Figura 1. Materiales. a. Mapas de correlación; b. Mapa de regionalización a 5km.

c. **Relación de regionalización y mapas de correlación**

Finalmente, para cada pixel de los mapas se relacionan los valores de correlación de cada índice con la región a la que pertenece. Así que construye una tabla que será utilizada para el aprendizaje de la estructura y de los parámetros de la red bayesiana. El 80% de los registros de dicha tabla se seleccionan aleatoriamente para conformar el grupo de entrenamiento, mientras que el 20% restante se utilizará como validación. La figura 2 muestra un extracto de la tabla, donde la primera columna de *layer* corresponde a la región (en este caso 1), y el resto de columnas son las correlaciones con cada índice climático.

	layer	aao	ao	glaam	qbo	soi
1	1	-0.0396265106	-0.0399369170	0.0598420267	0.14379467	0.10688980
2	1	-0.0591332206	-0.0532709634	0.0441620956	0.12800822	0.09515583
3	1	-0.0524648843	-0.0316902097	0.0625031797	0.11465588	0.09090696
4	1	-0.0436155114	-0.0389611914	0.0396224815	0.12743948	0.11517875
5	1	-0.0518074427	-0.0466495958	0.0453752683	0.13055973	0.11225830
6	1	-0.0649093148	-0.0537666529	0.0335566180	0.12482276	0.11685428
7	1	-0.0376254879	-0.0597592775	0.0193194493	0.13018665	0.13316335
8	1	-0.0507560579	-0.0248575219	0.0768342715	0.10404288	0.10091581
9	1	-0.0441868595	-0.0256088828	0.0724094524	0.12237822	0.10337170
10	1	-0.0410770563	-0.0300387812	0.0526412596	0.13015274	0.12250987

Figura 2. Extracto de datos para aprendizaje de la red bayesiana

2. Red Bayesiana

Conceptos

Una red bayesiana es un tipo de modelo gráfico probabilístico que se usa para representar el conocimiento sobre un dominio incierto, se trata de un grafo acíclico dirigido (DAG), en la que los nodos representan una variable aleatoria y las aristas representan dependencias probabilísticas entre esas variables. Para determinar las dependencias se utilizan métodos estadísticos o computacionales (Ben-Gal, 2009).

Estas redes probabilísticas son construidas a partir de la distribución de probabilidad y usan las leyes de probabilidad para la predicción y detección de anomalías, razonamiento y diagnóstico, toma de decisiones bajo incertidumbre, predicción de series de tiempo y otros escenarios en donde no se conozca la probabilidad a priori (Bayesserver, 2018).

El siguiente ejemplo (Figura 3) ilustra de manera gráfica una red bayesiana. Considere que una persona puede padecer fiebre, dicho síntoma puede ser causado ya sea por intoxicación o por salmonela. Cualquiera de estas dos enfermedades puede ser causadas por comida en mal estado, pero además la intoxicación puede ser causado por consumo excesivo de alcohol. En este ejemplo se puede observar tanto relaciones de dependencia e independencia condicional, por ejemplo las variables alcohol y comida son marginalmente independientes, pero cuando se produce el evento

intoxicación se vuelven condicionalmente dependientes; cuando se produce el evento salmonela las variables dolor muscular y fiebre se vuelven condicionalmente independientes, cuando se produce el evento intoxicación, la variable indigestión se vuelve condicionalmente independiente de sus ancestro comida y alcohol.

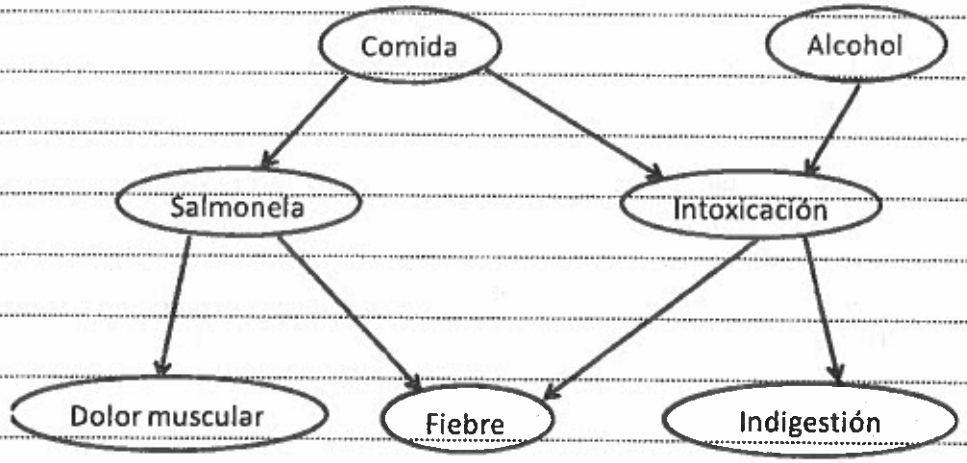


Figura 3: Ejemplo de una red bayesiana. Tomado de (Sucar, 2013).

Paso 1. Discretizar datos

Debido a que la mayoría de herramientas y algoritmos de aprendizaje se basan en nodos discretos, antes de realizar el aprendizaje de la red bayesiana se deben discretizar los datos. En este caso los datos de correlación ente la precipitación y los índices climáticos son de tipo continuo y se discretizarán. Sin embargo, las 9 regiones de estacionalidad son discretas.

La discretización o agrupación de datos es dividir el rango de variables continuas en un número finito de intervalos exclusivos. Una consecuencia de la discretización es la pérdida de información la cual depende del dominio y número de intervalos seleccionados (Rodríguez & Dolado, 2007). La discretización también se puede aplicar cuando una o más variables presentan desviaciones o sesgos muy grandes.

Existen diferentes formas de discretizar (Nagarajan, Scutari, & Lèbre, 2013). Estas estrategias suponen una compensación entre la precisión de la representación de los datos originales y la eficiencia computacional que resulta de la transformación, estas son:

- Uso de conocimientos previos sobre los datos. Los límites de los intervalos se definen, para cada variable, para corresponder a escenarios del mundo real significativamente diferentes, como la concentración de un contaminante particular (ausente, peligroso, letal) o clases de edad (niños, adultos, ancianos).

- Uso de heurísticas antes de conocer la estructura de la red. Algunos ejemplos son las reglas de Sturges, Freedman-Diaconis o Scott.

- Elegir el número de intervalos y sus límites para equilibrar la precisión y la pérdida de información, nuevamente una variable a la vez y antes de que se haya aprendido la estructura de la red.

- Realizar el aprendizaje y la discretización de manera iterativa hasta que no se realice ninguna mejora.

- Medidas estadísticas de división de datos, como es el caso de los cuartiles, que divide a todo el conjunto de datos en 4 grupos ordenados de datos, cada uno conteniendo el 25% del total. Así el primer cuartil contendrá el 25% de los datos con valores más bajos, mientras que el cuarto cuartil en 25% de valores más altos. Este último método es el elegido para aplicar en este trabajo.

Paso 2. Aprendizaje de la red

El aprendizaje de redes bayesianas consiste en estimar un modelo de estructura y parámetros asociados a partir de datos. Este puede dividirse naturalmente en dos partes: aprendizaje estructural que consiste en obtener la estructura o topología de la red (identificación de nodos, arcos y direcciones de los arcos) y aprendizaje paramétrico en el cual dada la estructura, se debe obtener las probabilidades asociadas (Sucar, 2013).

El aprendizaje de la estructura y parámetros de la red bayesiana se realizará con la librería bnlearn de R.

Aprendizaje de la estructura

Consiste en identificar la estructura gráfica de la red, idealmente debería ser el mapa mínimo de la estructura, o en su defecto, una distribución cercana a la correcta en el espacio de probabilidad, los algoritmos existentes para el aprendizaje de la estructura se clasifican en 3 categorías (Nagarajan et al., 2013):

- Los algoritmos basados en restricciones se basan los trabajos de Pearl en los mapas y su aplicación a modelos gráficos causales. A partir del algoritmo de Pearl se ha desarrollado otros algoritmos mejorados tales como: PC, growk-shrink, asociación incremental, asociación incremental rápida y asociación incremental entrelazada. Todos estos algoritmos (excepto PC) aprenden primero el manto de Markov de cada nodo en la red, como consecuencia, se reduce la complejidad computacional.

- Los algoritmos basados en puntajes son el resultado de una aplicación de técnicas de optimización heurística, a cada red candidata se le asigna una puntuación que refleje su calidad de ajuste. Ejemplos de estos algoritmos son:

algoritmos de búsqueda codiciosos los cuales comienzan explorando desde una estructura de red generalmente vacía y desde aquí agrega, elimina o invierte un arco hasta que la puntación ya no pueda mejorarse; los algoritmos genéticos que emulan la evolución natural a través de la selección iterativa de los modelos más aptos y la combinación de sus características; y el algoritmo de recocido simulado que realiza una búsqueda local estocástica al aceptar cambios que aumentan la puntación de la red.

- Los algoritmos híbridos combinan lo mejor de los algoritmos basados en restricciones y basados en puntajes para compensar debilidades. Los más conocidos dentro de esta categoría son el algoritmo Sparse Candidate (SC) y el algoritmo Max-Min- Hill-Climbing (MMHC). Ambos algoritmos se basan en dos pasos denominados restringir y maximizar, en el primero el conjunto de candidatos para los padres de cada nodo Xi se reduce de la forma en que está relacionado de alguna manera con el de Xi, el segundo paso busca la red que maximiza una función de puntaje determinada, sujeto a las restricciones impuestas por los conjuntos de Ci. Estos pasos se aplican de manera iterativa en el algoritmo Sparse Candidate hasta que no haya cambios en la red o no pueda ser mejorada, en cambio, en el algoritmo MMHC estos pasos se realizan una sola vez.

Para este trabajo se seleccionó el método basados en puntajes Hill-Climbing (hc) por ser uno de los métodos frecuentemente utilizados (Sachs, 2005).

Aprendizaje de parámetros

Cuanto se cuentan con datos completos y suficientes, y asumiendo que la estructura ya ha sido obtenida, obtener los parámetros se realiza a través del método llamado: "estimador de máxima verosimilitud" (MLE), bajo el cual se estiman las probabilidades en base a las frecuencias de los datos. Debido a que por lo general no se cuentan con suficientes datos, existe incertidumbre en las probabilidades estimadas, esto puede ser representado mediante una distribución de probabilidad para considerar de forma explícita la incertidumbre sobre las probabilidades (Sucar, 2013).

En muchas ocasiones cuando los datos no están completos, existen dos tipos básicos de información incompleta: valores faltantes y nodos ocultos, las alternativas para el primer caso pueden ser:

- Eliminar los registros donde aparecen valores faltantes.
- Considerar un nuevo valor adicional para la variable, como "desconocido".
- Tomar el promedio de la variable.
- Considerar el valor más probable en base a las otras variables.
- Considerar la probabilidad de los diferentes valores en base a las otras variables.

Para este trabajo se seleccionó el estimador de máxima verosimilitud (MLE) que la librería bnlearn implementa para datos completos y discretos.

Propagación de probabilidades

Uno de los puntos fuertes de las redes bayesianas es que proporcionan un sistema de inferencia. Con ello se modifican las tablas de probabilidad una vez que se encuentra o selecciona evidencia sobre el estado de ciertos nodos, y a su vez, estas nuevas probabilidades se propagan hacia el resto de nodos, esto se conoce con el nombre de inferencia probabilística. Las probabilidades antes de introducir evidencias se conocen como probabilidades a priori; una vez introducidas evidencias, las nuevas evidencias propagadas se llaman probabilidades a posteriori (Rodríguez & Dolado, 2007).

Un problema que surge de la propagación es su coste computacional cuando existen muchos nodos y dependencias, por ejemplo, una red con n variables booleanas contiene 2n valores. Gracias al desarrollo de nuevos algoritmos de propagación hoy en día es posible usar redes bayesianas capaces de modelar problemas reales. Los algoritmos de propagación se dividen en dos grandes grupos (Rodríguez & Dolado, 2007):

- Los algoritmos de propagación exactos son aquellos en los que no existe error en las probabilidades calculadas, ejemplos de estos son: inferencia mediante enumeración y eliminación de variables, estos funcionan bien con redes de hasta aproximadamente 30 nodos, no son apropiados para redes con mayor cantidad de nodos o altamente conexas.

- Los algoritmos aproximados son aquellos en los que existe cierto margen de error en las probabilidades estimadas, se clasifican en métodos de simulación estocástica (también llamados Montecarlo) y los métodos de búsqueda determinista, los primeros realizan la inferencia por medio de muestreos de números aleatorios que dependen de las probabilidades de la red. Las probabilidades condicionales se calculan dependiendo de las frecuencias generadas en el muestreo.

La propagación de inferencias en este trabajo ser realizará con la librería gRain. Esta librería dispone de herramientas para leer la estructura y parámetros previamente desarrollos con bnlearn. Además, permite realizar inferencias hacia delante o forward (es decir hacia los hijos) y hacia atrás o backward (es decir hacia los padres).

2.10 Alcances y resultados esperados

Al culminar el proyecto se espera contar con una red bayesiana capaz de determinar el grado de influencia que tiene cada uno de los índices estudiados sobre las

precipitaciones en Ecuador así como conocer si existe variabilidad entre las diferentes regiones del país.

2.11 Supuestos y riesgos

Riesgo	Probabilidad	Solución
Seleccionar una localidad sin suficientes datos representativos	Baja	Cambiar por otro sitio que cuente con una mayor cantidad de datos
Baja capacidad de procesamiento de los datos	Baja	Adquirir componentes que ayuden al procesamiento de los datos, tales como CPU o tarjetas gráficas
Datos anómalos que puedan alterar los resultados de la investigación	Baja	Realizar un filtrado de los datos para asegurar que se trabajará con datos representativos

2.12 Presupuesto

Rubro-Denominación	Justificación	Costo (Detalle)	Cantidad	Costo Total
Computadora	Recurso para el desarrollo del proyecto	1000	1	1000
Salario investigador	Necesario para el desarrollo de la investigación, en este caso se trata del propio investigador.	800	6	4800
Internet	Recursos de apoyo para el desarrollo de la tesis.	35	6	210
Documentos varios	Derechos, solicitudes y cualquier otro documento que se necesite adquirir	10	5	50
Total				6060

2.13 Financiamiento

El proyecto será financiado por el propio investigador.

2.14 Esquema tentativo

I) Introducción
II) Estado del arte
III) Método
   a. Materiales
   b. Aprendizaje de la estructura de la red

c. Aprendizaje de los parámetros de la red
d. Propagación de probabilidades

IV) Resultados

V) Discusiones y conclusiones

2.15 Cronograma

			Planificación por semanas															
			Febrero				Marzo					Abril				Mayo		
Objetivo Específico	Actividad	Resultados esperados	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Revisión de la literatura	Búsqueda de documentos	Contar con documentos relevantes para la investigación																
	Selección de los mejores documentos	Contar con documentos relevantes para la investigación																
	Lectura y extracción de información	Adquirir conocimiento sobre el tema a tratar																
	Síntesis de la información	Contar con un documento que sintetise la información extraída																
Red bayesiana	Aprendizaje de la estructura	Identificación la estructura gráfica de la red																
	Aprendizaje de los parámetros	Estimación de las probabilidades en base a la frecuencia de los datos																
	Propagación de probabilidades	Contar con un sistema de inferencia que permita la modificación de las tablas de probabilidades cuando se cuenta coó información a posteriori																
Resultados	Análisis e interpretación de los resultados	Tener un documento en el que se detalle las conclusiones extraídas de los estudios realizados																
Documento Final	Plasmar los resultados obtenidos en el documento final de titulación	Documento final de titulación terminado																

2.16 Referencias

Abd-elhamid, H. F., Fathy, I., & Zelen, M. (2018). Flood prediction and mitigation in coastal tourism areas, a case study: Hurghada, Egypt. https://doi.org/10.1007/s11069-018-3316-x

Ali, S., Jan, A., Manzoor, Sohail, A., Khan, A., Khan, M. I., ... Daur. (2018). Soil amendments strategies to improve water-use efficiency and productivity of maize under different irrigation conditions. *Agricultural Water Management*, *210*(March), 88–95. https://doi.org/10.1016/j.agwat.2018.08.009

Bayesserver. (2018). Bayesian networks - an introduction. Retrieved 20 December 2018, from https://www.bayesserver.com/docs/introduction/bayesian-networks.

Ben-Gal, I. (2009). Bayesian networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(3), 307–315. https://doi.org/10.1002/wics.48

Blunden, J., Arndt, D. S., Baringer, M. O., Achberger, C., Ackerman, S. A., Ahlstrom, A., ... Zimmermann, S. (2011). State of the climate in 2010. *Bulletin of the American Meteorological Society*, *92*(6), 92 (6), pp. S1–S236. https://doi.org/10.1175/1520-0477-92.6.S1

Castillo, E., & Hadi, A. S. (2010). Sistemas Expertos y Modelos de Redes Probabilísticas, 639.

Correa, M., Bielza, C., & Pamies-teixeira, J. (2009). Expert Systems with Applications Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. *Expert Systems With Applications*, *36*(3), 7270–7279. https://doi.org/10.1016/j.eswa.2008.09.024

Das, M., & Ghosh, S. K. (2015). A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables. *9th International Conference on Industrial and Information Systems, ICIIS 2014*. https://doi.org/10.1109/ICIINFS.2014.7036528

Duc, H. N., Bang, H. Q., & Quang, N. X. (2018). Influence of the Pacific and Indian Ocean climate drivers on the rainfall in Vietnam. *International Journal of Climatology*, (June), 1–16. https://doi.org/10.1002/joc.5774

Duc, H. N., Rivett, K., MacSween, K., & Le-Anh, L. (2017). Association of climate drivers with rainfall in New South Wales, Australia, using Bayesian Model Averaging. *Theoretical and Applied Climatology*, *127*(1–2), 169–185. https://doi.org/10.1007/s00704-015-1622-8

Ebert-Uphoff, I., & Deng, Y. (2012). Causal discovery for climate research using graphical models. *Journal of Climate*, *25*(17), 5648–5665. https://doi.org/10.1175/JCLI-D-11-00387.1

Engler, J., Wehrden, H. Von, & Baumgärtner, S. (2019). Land Use Policy Determinants of farm size and stocking rate in Namibian commercial cattle farming. *Land Use Policy*, *81*(February 2018), 232–246. https://doi.org/10.1016/j.landusepol.2018.10.009

Fierro, A. O. (2014). Relationships between California rainfall variability and large-scale climate drivers. *International Journal of Climatology*, *34*(13), 3626–3640. https://doi.org/10.1002/joc.4112

Hasan, M. M., & Wyseure, G. (2018). Impact of climate change on hydropower generation in Rio Jubones Basin, Ecuador. *Water Science and Engineering*. https://doi.org/10.1016/j.wse.2018.07.002

Konapala, G., Valiya, A., & Ashok, V. (2017). Teleconnection between low flows and large-scale climate indices in Texas River basins. *Stochastic Environmental Research and Risk Assessment*. https://doi.org/10.1007/s00477-017-1460-6

Lee, J. H., Lee, J., & Julien, P. Y. (2018). Catena Global climate teleconnection with rainfall erosivity in South Korea. *Catena*, *167*(June 2017), 28–43. https://doi.org/10.1016/j.catena.2018.03.008

Liu, Y. C., Di, P., Chen, S. H., & DaMassa, J. (2018). Relationships of rainy season precipitation and temperature to climate indices in California: Long-Term variability and extreme events. *Journal of Climate*, *31*(5), 1921–1942. https://doi.org/10.1175/JCLI-D-17-0376.1

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian Networks in R*. https://doi.org/10.1007/978-1-4614-6446-4

Pratiwi, R., & Sukardjo, S. (2018). EFFECTS OF RAINFALL ON THE POPULATION OF SHRIMPS Penaeus monodon Fabricius IN SEGARA ANAKAN LAGOON , CENTRAL JAVA , INDONESIA, *2*(3), 156–169. https://doi.org/10.11598/btb.2018.25.3.830

Rodríguez, D., & Dolado, J. (2007). Redes bayesianas en la ingeniería del software. *Cc.Uah.Es*, 1–21. https://doi.org/10.2196/jmir.7.3.e31

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 523-529.

Sucar, L. E. (2013). Redes Bayesianas. *Test*, (X), 18. Retrieved from http://ccc.inaoep.mx/~esucar/Clases-mgp/caprb.pdf

Sudha, V., Venugopal, K., & Ambujam, N. K. (2008). Reservoir operation management through optimization and deficit irrigation, 93–102. https://doi.org/10.1007/s10795-007-9041-3

Torre-Gea, G. D. la, Soto-Zarazua, G. M., Guevara-Gonzalez, R. G., & Rico-Garcia, E. (2011). Bayesian networks for defining relationships among climate factors. *International Journal of Physical Sciences*, *6*(18), 4412–4418. https://doi.org/10.1016/j.jmaa.2015.01.055

Vicente-Serrano, S. M., Aguilar, E., Martínez, R., Martín-Hernández, N., Azorin-Molina, C., Sanchez-Lorenzo, A., … Nieto, R. (2017). The complex influence of ENSO on droughts in Ecuador. *Climate Dynamics*, *48*(1–2), 405–427. https://doi.org/10.1007/s00382-016-3082-y

Zeng, Z., Hsieh, W. W., Shabbar, A., & Burrows, W. R. (2011). Seasonal prediction of winter extreme precipitation over Canada by support vector regression. *Hydrology and Earth System Sciences*, *15*(1), 65–74. https://doi.org/10.5194/hess-15-65-2011

2.17 Anexos

2.18 Firma de responsabilidad (estudiante)

Paúl Renato Avila Andrade
Estudiante

2.19 Firma de responsabilidad (director sugerido)

Agrim. Daniela Elisabet Ballari
Directora del proyecto

2.20 Fecha de entrega: Miércoles 2 de enero de 2019