Universidad de Cantabria

Facultad
de
Ciencias

Modelos gráficos probabilísticos para modelización de relaciones causales en clima

(Probabilistic networks for modeling causal relationships in clime)

Trabajo de Fin de Grado
para acceder al
GRADO EN MATEMÁTICAS

Autor: Celia Melendi Ortega
Director: Inés González
Co-director: José Manuel Gutiérrez
Julio – 2023

Resumen

Las redes probabilísticas son técnicas de aprendizaje automático basadas en probabilidad que permiten modelar problemas multivariados a partir de conjuntos de datos. Estos modelos utilizan grafos para extraer y representar de forma automática las relaciones relevantes entre las variables, y probabilidad para definir una probabilidad conjunta consistente con las dependencias. Estas técnicas han sido aplicadas exitosamente en distintas disciplinas, pero su aplicación en clima es reciente y limitada.

En este trabajo de fin de grado se ha explorado la aplicación de las redes probabilísticas para la modelización de datos climáticos (en particular datos globales de temperatura para un período de 30 años) y la identificación de causalidad en teleconexiones (correlaciones de largo alcance). Para ello se han estudiado los métodos más extendidos para detectar causalidad en series temporales, en particular causalidad de Granger y nuevas técnicas empíricas basadas en redes probabilísticas dinámicas. Para problemas de alta dimensionalidad (como el problema de clima tratado en este trabajo) se ha visto que las redes probabilísticas dinámicas ofrecen una metodología eficiente, que permite identificar teleconexiones que están de acuerdo con el conocimiento de causalidad que se tiene del problema.

Palabras clave: Redes probabilísticas, aprendizaje automático, causalidad, clima, temperatura global, teleconexiones.

Abstract

Probabilistic networks are probability-based machine learning techniques for modelling multivariate problems from data. These models use graphs to automatically extract and represent the relevant relationships between variables, and probability to define a joint probability consistent with the dependencies. These techniques have been successfully applied in different disciplines, but their application in climate is recent and limited.

This thesis has explored the application of probabilistic networks to model climate data (in particular global temperature data for a 30-year period) and to identify causality in teleconnections (long-range correlations). For this purpose, the most widespread methods for detecting causality in time series have been studied, in particular Granger causality and new empirical techniques based on dynamic probabilistic networks. For high-dimensional problems (such as the climate problem discussed in this work), dynamic probabilistic networks have been found to offer an

efficient methodology, which allows us to identify teleconnections that coincide with the knowledge of causality that we have of the problem.
Key words: Probabilistic networks, machine learning, causality, climate, global temperature, teleconnections.

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1. Motivación

Las redes probabilísticas son técnicas de inteligencia artificial (en concreto de aprendizaje automático, o de representación del conocimiento) introducidas hace cuatro décadas por J. Pearl (1988). Estas técnicas permiten extraer las relaciones relevantes entre un conjunto de variables a partir de una muestra de datos y representarlas cualitativamente en un grafo, y cuantitativamente por medio de un modelo probabilístico asociado (de ahí el nombre de redes probabilísticas), permitiendo el análisis y predicción del problema subyacente. Estas técnicas han sido aplicadas exitosamente en distintas disciplinas, pero su aplicación en clima es reciente y limitada (Graafland, 2022; Gutiérrez et al., 2004).

El sistema climático (en particular la atmósfera) es un sistema dinámico no lineal regido por las leyes fundamentales de la Física (dinámica de fluidos para la atmósfera, incluyendo conservación de masa, momento y energía). Estas leyes se expresan en forma de ecuaciones en derivadas parciales que describen la compleja evolución de la atmósfera sobre el globo terrestre, con fenómenos que se manifiestan en distintas escalas temporales y espaciales (Santos Burguete et al., 2018), y nos ayudan a modelar y entender el tiempo y el clima. Por ejemplo, la atmósfera muestra una gran sensibilidad a las condiciones iniciales y, por tanto, la predicción a partir de unas condiciones iniciales dadas es posible sólo para un tiempo finito (típicamente dos semanas). A escalas mayores de tiempo, la evolución del sistema climático responde a procesos de gran escala más lentos (con duración de meses), que modulan el comportamiento sincronizado de regiones distantes e introducen dependencias/correlaciones entre puntos alejados, denominadas “teleconexiones”. Un ejemplo muy conocido es El Niño-Southern Oscilation (ENSO), caracterizado por un calentamiento de la superficie del Pacífico central y oriental, que fuerza una va-

riación de la atmósfera a gran escala, en particular en zonas tropicales con efectos en todo el globo (Gutiérrez et al., 2004).

Identificar teleconexiones es un problema de gran interés en clima porque estas dependencias de larga escala coexisten con las fuertes dependencias locales propias de los campos meteorológicos (por ejemplo, la temperatura) y definen las propiedades globales del clima. Además, establecer la causalidad de las mismas (qué zona es la que influye a qué otra) permite entender los mecanismos físicos subyacentes.

En este trabajo de fin de grado se explora la aplicación de las redes probabilísticas para la modelización de datos climáticos, en concreto para la búsqueda de teleconexiones y la identificación de causalidad en las mismas. Para ello se estudian los métodos más extendidos para detectar causalidad en series temporales, en particular causalidad de Granger y también nuevas técnicas empíricas basadas en redes probabilísticas dinámicas, para estudiar su aplicabilidad a este problema. Más concretamente se trabaja con un ejemplo práctico introducido en Graafland (2022) que consiste en datos de anomalías mensuales de temperatura en una rejilla global de 36×18 puntos para un período de 30 años.En este trabajo se extienden los resultados de este trabajo previo para introducir causalidad en estos modelos teniendo en cuenta el tiempo.

1.2. Objetivos del TFG

El objetivo final de este trabajo de fin de grado es estudiar las posibilidades que ofrecen las redes probabilísticas para el estudio de causalidad en problemas de alta dimensión, analizando un ejemplo práctico en clima. Concretamente, se pretende estudiar la aplicación y limitaciones de las redes bayesianas dinámicas, que permiten introducir una componente temporal en las redes probabilísticas analizadas previamente para aplicaciones en clima.

1.3. Organización de la memoria

El resto de la memoria está organizado como sigue: En primer lugar, en el capítulo 2, se introducen distintos conceptos de probabilidad, y se explica la evolución de los modelos probabilísticos, desde los sistemas expertos hasta las redes probabilísticas, en concreto las redes bayesianas. En el capítulo 3 se expone el concepto de causalidad y se desarrolla sobre distintos métodos para detectar relaciones causales, centrándose

más en las redes bayesianas dinámicas. En el capítulo 4 se desarrolla la parte práctica del trabajo, el aprendizaje de redes bayesianas y redes bayesianas dinámicas y su posterior análisis. Finalmente, en el capítulo 5 se presentan las conclusiones del trabajo realizado.

Capítulo 2

Sistemas inteligentes y redes probabilísticas

En este capítulo se introducen los modelos utilizados en este trabajo (redes probabilísticas) en el contexto de la Inteligencia Artificial y, más concretamente, de los sistemas inteligentes. Se comienza introduciendo distintos conceptos de probabilidad, incluyendo algunos básicos, para asentar las definiciones y la notación; a continuación, se exponen los conceptos de dependencia e independencia condicional, que son de gran importancia para entender los sistemas expertos/inteligentes probabilísticos. Posteriormente, se expone la evolución de estos sistemas, desde los primeros sistemas expertos (con estructuras *ad-hoc* basadas en distintas hipótesis), hasta modelos basados en datos (*data-driven*), como las redes probabilísticas, que se aprenden a partir de una muestra de datos que caracterizan el problema en estudio.

2.1. Conceptos de probabilidad

A continuación, se presentan algunos conceptos de probabilidad que serán empleados posteriormente.

Sea p una probabilidad, $\{X_1, \ldots, X_n\}$ un conjunto de variables aleatorias discretas y $Val(X_i)$ el conjunto de las posibles realizaciones de la variable X_i. Denotaremos con $\{x_1, \ldots, x_n\}$ a una de las posibles realizaciones de $\{X_1, \ldots, X_n\}$.

Definición 1 (Función de probabilidad de las variables $\{X_1, \ldots, X_n\}$)**.** Una función de probabilidad de las variables $\{X_1, \ldots, X_n\}$ se define como

$$p(x_1, \ldots, x_n) = p(X_1 = x_1, \ldots, X_n = x_n) \tag{2.1}$$

Definición 2 (Función de probabilidad marginal)**.** La función de probabilidad mar-

ginal de la variable X_i viene dada por

$$p(x_i) = p(X_i = x_i) = \sum_{x_j \in Val(X_j); j \neq i} p(x_1, \ldots, x_n) \tag{2.2}$$

Definición 3 (Probabilidad condicional)**.** Sean X, Y dos conjuntos disjuntos de variables aleatorias de $\{X_1, \ldots, X_n\}$ y x e y posibles realizaciones de las variables de dichos conjuntos, tales que $p(y) > 0$. Entonces, la función de probabilidad condicionada, $p(x|y)$, se obtiene de la siguiente manera

$$p(x|y) = p(X = x|Y = y) = \frac{p(x, y)}{p(y)} \tag{2.3}$$

Por lo tanto, la función de probabilidad de X, Y se puede escribir como

$$p(x, y) = p(y)p(x|y) = p(x)p(y|x) \tag{2.4}$$

La probabilidad condicional puede emplearse para calcular la probabilidad marginal como sigue: dados X, Y dos conjuntos de variables aleatorias,

$$p(x) = \sum_{y \in Val(Y)} p(x|y)p(y) \tag{2.5}$$

Definición 4 (Independencia probabilística de dos variables)**.** Dados dos subconjuntos disjuntos X, Y del conjunto de variables aleatorias $\{X_1, \ldots, X_n\}$. Se dice que X es independiente de Y si y solo si:

$$p(x|y) = p(x) \tag{2.6}$$

para todas las realizaciones posibles x, y de X, Y. En caso contrario, se dice que X es dependiente de Y.

La condición $p(y) > 0$ se verifica, puesto que solo se consideran las realizaciones posibles de Y y por lo tanto, su probabilidad es mayor que cero. Por lo tanto, es posible combinar (2.4) y (2.6) para obtener otra definición de independencia, ya que, que la variable X sea independiente de la variable Y es equivalente a:

$$p(x, y) = p(x)p(y) \tag{2.7}$$

Por lo tanto, la relación de independencia es simétrica y se dirá simplemente que dos variables son independientes.

Esta noción se puede extender a más de dos variables de la siguiente manera:

Definición 5 (Independencia probabilística de un conjunto de variables)**.** Se dice que las variables aleatorias $X_1, \ldots, X_k$ son independientes si y solo si:

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i) \tag{2.8}$$

para todas las realizaciones posibles de $X_1, \ldots, X_k$. En otro caso, se dice que son dependientes.

Definición 6 (Independencia probabilística condicional)**.** Dados X, Y, Z, conjuntos disjuntos de variables aleatorias, se dice que X es condicionalmente independiente de Y dado Z si y solo si:

$$p(x|z, y) = p(x|z) \tag{2.9}$$

para todas las realizaciones posibles x, y, z de X, Y, Z y se denota $I(X, Y|Z)$. En otro caso, X e Y son condicionalmente dependientes dado Z, lo cual se escribe $D(X, Y|Z)$.

Otra definición alternativa pero equivalente es

$$p(x, y|z) = p(x|z)p(y|z) \tag{2.10}$$

La equivalencia entre (2.9) y (2.10) es análoga a la equivalencia entre (2.6) y (2.7)

Extendiendo esto al caso multivariado, se tiene que, si $X_1, \ldots, X_k$ son condicionalmente independientes dado $Y_1, \ldots, Y_r$, entonces

$$p(x_1, \ldots, x_k|y_1, \ldots, y_r) = \prod_{i=1}^{k} p(x_i|y_1, \ldots, y_r) \tag{2.11}$$

La independencia (no condicional) puede ser entendida como un caso concreto de independencia condicionada, escribiendo $I(X, Y|\emptyset)$ para denotar que X e Y son incondicionalmente independientes, con $\emptyset$ el conjunto vacío.

A partir de las ecuaciones (2.3), (2.4) y (2.5), se obtiene un resultado muy famoso de teoría de probabilidad, el Teorema de Bayes.

Teorema 7 (Teorema de Bayes)**.** *Sean X, Y dos conjuntos disjuntos de variables. Entonces*

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)} = \frac{p(x)p(y|x)}{\sum_x p(x)p(y|x)} \tag{2.12}$$

En el caso de que X es una única variable e Y es un subconjunto de variables, se tiene

$$p(x_i|x_1, \ldots, x_k) = \frac{p(x_i)p(x_1, \ldots, x_k|x_i)}{\sum_{x_i} p(x_i)p(x_1, \ldots, x_k|x_i)} \tag{2.13}$$

A continuación, se enuncian y demuestran algunas propiedades de independencia condicional.

Proposición 8 (Propiedades independencia condicional)**.** *Sean X, Y, Z, W conjuntos de variables aleatorias. Las siguientes propiedades las cumple todo modelo probabilístico p.*

1. ***Simetría**. Si X es condicionalmente independiente de Y dado Z, entonces Y es condicionalmente independiente de X dado Z.*

Demostración. Si X es condicionalmente independiente de Y dado Z, por (2.10), se tiene que

$$p(x, y|z) = p(x|z)p(y|z) = p(y|z)p(x|z) = p(y, x|z)$$

para todas las realizaciones posibles de X, Y, Z y por lo tanto Y es condicionalmente independiente de X dado Z □

2. ***Descomposición***. *Si X es condicionalmente independiente de $Y \cup W$ dado Z, entonces X es condicionalmente independiente de Y dado Z y X es condicionalmente independiente de W dado Z.*

Demostración. De acuerdo con Pearl (1988), la notación se interpreta como sigue: $I(X, Y \cup W|\emptyset)$ es equivalente a

$$p(X = x, Y = y, Z = z) = p(X = x)p(Y = y, Z = z)$$

Veamos que X es condicionalmente independiente de Y dado Z. Sean x, y, z tres realizaciones cualquiera de X, Y, Z. Entonces

$$p(x, y|z) = \sum_{w \in Val(W)} p(x, y, w|z) = \sum_{w \in Val(W)} p(x|z)p(x, y|z) = p(x|z)p(y|z)$$

y así, se tiene $I(X, Y|Z)$. De forma análoga se prueba para W. □

3. ***Unión débil***. $I(X, Y \cup W|Z) \Longrightarrow I(X, Y|Z \cup W)$ *y* $I(X, W|Z \cup Y)$

Demostración. Sean x, y, z, w cuatro realizaciones cualesquiera de X, Y, Z, W respectivamente. Por ser X independiente de $Y \cup W$ dado Z se tiene

$$p(x|y, w, z) = p(x|z)$$

Por otra parte, por la propiedad de descomposición, como $I(X, Y \cup W|Z)$, se obtiene $I(X, Y|Z)$ y $I(X, W|Z)$ y así

$$p(x|w, z) = p(x|z) = p(x|y, w, z)$$

Para cualquier realización de las variables, de forma que se concluye que $I(X, Y|Z \cup W)$. De forma análoga se demuestra que $I(X, W|Z \cup Y)$ □

4. ***Contracción***.$I(X, Y|Z)$ *y* $I(X, W|Z \cup W) \Longrightarrow I(X, Y \cup W|Z)$.

Demostración. Sean x, y, z, w cuatro realizaciones cualesquiera de X, Y, Z, W respectivamente. Por ser X independiente de W dado $Z \cup Y$ se tiene

$$p(x|w, y, z) = p(x|y, z)$$

y por ser X independiente de Y dado Z, se obtiene

$$p(x|w, y, z) = p(x|y, z) = p(x|z)$$

Se concluye que X es independiente de $Y \cup W$ dado Z. □

Por último, se introduce la regla de la cadena, que permite factorizar una función de probabilidad mediante probabilidades condicionales. Esta se obtiene aplicando (2.4) tantas veces como sea necesario.

Definición 9 (Regla de la cadena)**.** Dados $X = \{X_1, \dots X_n\}$ un conjunto de variables y $\{Y_1, \dots Y_r\}$ una partición de X, cualquier función de probabilidad de X puede ser factorizada como

$$p(x_1, \dots x_n) = \prod_{i=1}^{r} p(y_i|z_i) \tag{2.14}$$

donde $Z_i = \{Y_1, \dots Y_{i-1}\}$ son las variables previas a Y_i.

Ejemplo 10 (Factorización e independencia). Sean X_1, X_2, X_3, X_4 cuatro variables binarias, con $I(X_4, X_3|X_2)$. Por la regla de la cadena, su función de probabilidad se puede factorizar de la siguiente manera

$$p(x_1, \dots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$$

la cual depende de $2^4 - 1 = 15$ parámetros. Debido a que $I(X_4, X_3|X_2)$, se obtiene

$$p(x_1, \dots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2)$$

que depende de $1 + 2 + 4 + 4 = 11$ parámetros.

En esta idea se basan los modelos probabilísticos para reducir el número de parámetros de la función de probabilidad mediante las independencias e independencias condicionales que existen entre las variables en el conjunto de datos.

2.2. Sistemas expertos probabilísticos

El campo de la inteligencia artificial comienza a desarrollarse en los años 50 del siglo XX y con ello surgen los primeros sistemas expertos. Estos se pueden definir como sistemas informáticos (hardware y software) que simulan a los expertos humanos

en un área de especialización dada (ver Castillo et al., 1996) y se pueden clasificar en deterministas o estocásticos. Los primeros tratan problemas deterministas y se conocen como sistemas basados en reglas. En cambio, los segundos tratan problemas con incertidumbre y se basan en probabilidad.

Los sistemas expertos estocásticos que tratan la incertidumbre mediante la probabilidad se denominan sistemas expertos probabilísticos y su base de conocimiento consiste en un conjunto de variables, $\{X_1, \ldots, X_n\}$ y una función de probabilidad definida, $p(x_1, \ldots, x_n)$, que describe las relaciones entre ellas. El motor de inferencia de estos sistemas emplea la base de conocimiento y evidencia, es decir, datos, para obtener nuevas conclusiones realizando evaluaciones de probabilidades condicionadas.

Una de las principales críticas a los primeros sistemas expertos probabilísticos es que la función de probabilidad dependía de un número muy alto de parámetros (en el caso de n variables binarias, p depende de 2^n parámetros), por lo que resultaban inmanejables desde el punto de vista práctico. No obstante, en su aplicación, suele ocurrir que subconjuntos de variables son independientes o condicionalmente independientes entre ellos, por lo que es posible reducir el número de parámetros factorizando la función de probabilidad en funciones que involucran un menor número de parámetros, sin perder representatividad (ver ejemplo 10).

En un principio, se emplearon diversos modelos *ad hoc* en los que se imponían distintas hipótesis de dependencia e independencia para reducir el número de parámetros. Una de las áreas de mayor aplicación de los sistemas expertos fue el campo de la medicina; se introdujeron distintos modelos probabilísticos para el problema del diagnóstico médico (que también se generalizaron a otras áreas). En este problema, se tienen n síntomas binarios $S_1, \ldots, S_n$ y una variable aleatoria E, que solo puede tomar como valor una de las enfermedades $e_1, \ldots, e_r$. El problema del diagnóstico consiste en calcular la enfermedad que tiene mayor probabilidad de padecer un paciente que presenta un conjunto de síntomas dados $S_{i_1} = s_{i_1}, \ldots, S_{i_k} = s_{i_k}$. Para ello se puede modelar a partir de los datos disponibles una función de probabilidad conjunta de los síntomas y la enfermedad $p(e, s_1, \ldots, s_n)$; sin embargo, como se comentó anteriormente, en su forma más general esta probabilidad depende de muchos parámetros y no resulta de utilidad práctica. Por ello, se introdujeron distintos modelos simplificados, considerando distintas hipótesis de independencia e independencia condicional entre síntomas y/o enfermedad.

1. Modelo de síntomas dependientes (MSD)

Para este modelo se supone que las enfermedades son independientes entre ellas dados los síntomas, los cuales sí son dependientes entre ellos. Por lo tanto, por (2.3) y (2.4) se obtiene

$$p(e_i|s_1,\ldots,s_n) = \frac{p(e_i,s_1,\ldots,s_n)}{p(s_1,\ldots,s_n)} = \frac{p(e_i)p(s_1,\ldots,s_n|e_i)}{p(s_1,\ldots,s_n)}$$

Entonces, la información necesaria para el MSD es las probabilidades $p(e_i)$ para todos las realizaciones de E y la probabilidad condicional $p(s_1,\ldots,s_n|e_i)$ para todas las combinaciones posibles de enfermedades y síntomas, mientras que $p(s_1,\ldots,s_n)$ es una constante de normalización.

Así, para determinar la función de probabilidad de E, son necesarios $m-1$ parámetros en el caso de m enfermedades, y para determinar la probabilidad condicionada $p(s_1,\ldots,s_n|e_i)$, son necesarios $m(2^n-1)$ (hay n síntomas binarios), 2^n-1 para cada enfermedad. Por lo tanto, este modelo depende de $m(2^n-1)+m-1 = m2^n-1$ parámetros, que es un número demasiado alto como para ser manejable. Para dar solución a este problema, se imponen nuevas hipótesis: que los síntomas son independientes entre sí.

2. Modelo de síntomas independientes (MSI)
   Ahora se supone que los síntomas son independientes entre sí, dada una enfermedad, es decir, son condicionalmente independientes dada E. Por (2.11) se obtiene
   $$p(s_1\ldots,s_n|e_i) = \prod_{j=1}^{n} p(s_j|e_i)$$
   y así
   $$p(e_i|s_1,\ldots,s_n) = \frac{p(e_i)p(s_1,\ldots,s_n|e_i)}{p(s_1,\ldots,s_n)} = \frac{p(e_i)\prod_{j=1}^{n} p(s_j|e_i)}{p(s_1,\ldots,s_n)}$$
   Este modelo depende de $m-1$ parámetros, correspondientes a la probabilidad marginal de E y de mn, por $p(s_j|e_i)$ para cada síntoma y cada enfermedad. Por esto, el MSI depende de $mn+m-1 = m(n+1)-1$ parámetros. Comparado con el modelo anterior, para $m=50$, $n=100$ (50 enfermedades y 100 síntomas), el MSD depende de más de 10^{31} parámetros, mientras que el MSI lo hace de 5049.

3. Modelo de síntomas relevantes independientes (MSRI)
   En este modelo se supone que los síntomas son independientes entre sí y que cada enfermedad tiene un número reducido de síntomas relevantes, es decir, que para el resto de los síntomas, esa enfermedad e_i es independiente de estos. Para la enfermedad e_i, sus síntomas relevantes son $S_{1_i},\ldots,S_{r_i}$ y aquellos irrelevantes son $S_{(r+1)_i},\ldots,S_{n_i}$.

Por lo tanto, se tiene que

$$p(e_i|s_1,\ldots,s_n) = \frac{p(e_i)\prod_{j=1}^{n} p(s_j|e_i)}{p(s_1,\ldots,s_n)} = \frac{p(e_i)\prod_{j=1_i}^{r_i} p(s_j|e_i)\prod_{j=(r+1)_i}^{n_i} p(s_j|e_i)}{p(s_1,\ldots,s_n)} = \frac{p(e_i)\prod_{j=1_i}^{r_i} p(s_j|e_i)\prod_{j=(r+1)_i}^{n_i} k_j}{p(s_1,\ldots,s_n)}$$

donde dado j, $k_j = p(s_j|e_i)$ tiene el mismo valor para todos los síntomas s_j irrelevantes para e_i.

Entonces, se necesita conocer las probabilidades marginales de $p(e_i)$ $(m-1$ parámetros), las probabilidades condicionales $p(s_j|e_i)$ para cada enfermedad y cada valor posible de sus síntomas relevantes ($\sum_{i=1}^{m} r_i$ parámetros) y las probabilidades $p(s_j|e_i) = k_j$ para cada e_i que tiene algún síntoma irrelevante ($n-a$ parámetros, donde a es el número de síntomas que son relevantes para todas las enfermedades). Así, el MSRI depende de $m-1+n-a+\sum_{i=1}^{m} r_i$ parámetros.

4. Modelo de Síntomas Relevantes Dependientes (MSRD)
   El modelo anterior consigue reducir el número de parámetros de manera notable, pero no es realista suponer que los síntomas relevantes para una enfermedad son independientes entre sí. Este modelo es similar al anterior, con la diferencia de que no impone que los síntomas relevantes sean independientes entre sí dada la enfermedad, solamente lo serán aquellos que son irrelevantes para esta. Con la notación anterior, para la enfermedad e_i, sus síntomas relevantes son $S_{1_i},\ldots,S_{r_i}$ y los irrelevantes $S_{(r+1)_i},\ldots,S_{n_i}$. Entonces,

$$p(e_i|s_1,\ldots s_n) = \frac{p(e_i)p(s_1,\ldots,s_n|e_i)}{p(s_1,\ldots s_n)} = \frac{p(e_i)p(s_{1_i},\ldots,s_{r_i}|e_i)\prod_{j=(r+1)_i}^{n_i} p(s_j|e_i)}{p(s_1,\ldots,s_n)}$$

   Para el MSRD, es necesario conocer las probabilidades marginales de $p(e_i)$ $(m-1$ parámetros), las probabilidades condicionales $p(s_{1_i},\ldots,s_{r_i}|e_i)$ para las posibles enfermedades y sus síntomas relevantes ($2^{r_i}-1$ parámetros para cada enfermedad e_i) y las probabilidades $p(s_j|e_i)$ para las enfermedades que tengan algún síntoma irrelevante ($n-a$ parámetros, donde a es el número de síntomas que son relevantes para todas las enfermedades). Así, el número total de parámetros para el MSRD es $m-1+(\sum_{i=1}^{m} 2^{r_i}-1)+n-a = n-a-1+\sum_{i=1}^{m} 2^{r_i}$

Pese a conseguir reducir el número de parámetros, estos modelos siguen presentando un problema: solo se pueden aplicar en situaciones particulares, ya que la función de probabilidad se obtiene imponiendo ciertas hipótesis, que en ocasiones pueden no ser ciertas. Una solución a este problema son las redes probabilísticas, una alternativa dirigida por datos, que permite factorizar la función de probabilidad

en otras de probabilidad condicionada más sencillas, en función de las relaciones de independencia entre las variables que se obtengan a partir del conjunto de datos.

2.3. Redes probabilísticas basadas en grafos

Las definiciones de esta sección se basan en Castillo et al. (1996). Para comenzar, se introducen los conceptos de modelo de dependencia y modelo probabilístico.

Definición 11 (Modelo de dependencia)**.** Un modelo M de un conjunto de variables $\{X_1, \ldots, X_n\}$ es un modelo de dependencia si permite saber si la relación $I(X, Y|Z)$ es cierta o no, para cualesquiera X, Y, Z subconjuntos disjuntos.

Definición 12 (Modelo probabilístico)**.** Un modelo de dependencia M es un modelo de dependencia probabilístico si contiene todas las relaciones de independencia dadas por una función de probabilidad $p(x_1, \ldots, x_n)$.

Un enfoque para definir modelos de dependencia probabilísticos es hacerlo mediante grafos, ya que tienen la ventaja de que sus enlaces expresan de una manera directa y cualitativa relaciones de dependencia y las conservan para cualquier asignación de parámetros numéricos. Estos modelos se conocen como redes probabilísticas y se obtienen a partir de conjuntos de datos empleando grafos, que representan explícitamente las dependencias e independencias entre las variables (los nodos del grafo) y una factorización de la función de probabilidad conjunta, que proporciona una descripción completa de dichas relaciones.

Para describir las estructuras de independencia y dependencia que contiene un grafo como relaciones de independencia condicional son necesarios criterios de separación gráfica. En el caso de grafos no dirigidos, se emplea el criterio de u-separación.

Definición 13 (Criterio de u-separación)**.** Dados X, Y, Z conjuntos disjuntos de nodos de un grafo G no dirigido, se dice que Z separa X e Y si y solo si cada camino entre nodos de X y nodos de Y contiene un nodo de Z. Cuando Z separe a X e Y, se denotará $I(X, Y|Z)_G$ para indicar que la relación de independencia se obtiene a partir de un grafo y en caso contrario se denotará $D(X, Y|Z)_G$ para indicar que X e Y son condicionalmente dependientes dado Z en el grafo G. En los casos en los que la relación venga dada por el modelo, se escribirá $I(X, Y|Z)_M$.

Para grafos dirigidos se utiliza otro criterio de separación gráfica, el criterio de $d-$separación. Antes de definirlo, se introduce otro concepto.

Definición 14 (Nodo de aristas convergentes en un camino)**.** Dado un grafo dirigido y un camino no dirigido (. . . – C – A – B – . . .), el nodo A se conoce como un nodo

de aristas convergentes en este camino si el grafo dirigido contiene las artistas C $\longrightarrow$ A y A $\longleftarrow$ B

Definición 15 (Criterio de $d-$separación)**.** Dados X, Y, y Z tres subconjuntos disjuntos de nodos en un grafo acíclico dirigido G, se dice que X e Y son condicionalmente independientes dado Z en el grafo y se denota $I(X, Y|Z)_G$, si y solo si a lo largo de todo camino no dirigido entre cualquier nodo de X y cualquiera de Y existe un nodo W que satisface únicamente una de las siguientes condiciones:

1. W es un nodo de aristas convergentes en el camino y ni W ni ninguno de sus descendientes están en Z

2. W no es un nodo de aristas convergentes en el camino y W está en Z

En caso contrario, X e Y son condicionalmente dependientes dado Z en el grafo G y se denota $D(X, Y|Z)_G$.

Todo grafo da lugar a un modelo de dependencia, sin embargo, no para todos los modelos de dependencia existe un grafo G tal que, cada relación de independencia que contiene G, esta también está contenida en el modelo y viceversa. Si esto ocurre, el grafo debe ser un mapa de independencia minimal.

Definición 16 (Mapa de independencia)**.** Un grafo G es un mapa de independencia de un modelo de dependencia M si

$$I(X, Y|Z)_G \Longrightarrow I(X, Y|Z)_M$$

lo cual implica

$$D(X, Y|Z)_M \Longrightarrow D(X, Y|Z)_G$$

Definición 17 (Mapa de independencia minimal)**.** Un grafo G es un mapa de independencia minimal (I-mapa minimal) de un modelo de dependencia M si es un mapa de independencia de M, pero deja de serlo si se elimina cualquiera de sus aristas.

El siguiente teorema permite concluir que, dado un grafo dirigido acíclico , al construir su función de probabilidad asociada p como el producto de las probabilidades condicionadas de las variables dados sus padres, el grafo será un $I-$mapa para p.

Teorema 18. *Dado un grafo acíclico dirigido G que describe las dependencias entre las variables $\{X_1, \ldots, X_n\}$ y $p(x_1, \ldots, x_n)$ un modelo probabilístico de dichas variables, son equivalentes:*

1. G es un mapa de independencia de $p(x_1, \ldots, x_n)$

2. $p(x_1, \ldots, x_n)$ *se puede factorizar como*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) \tag{2.15}$$

donde π_i *son los padres de la variable* i *en el grafo* G

Demostración. 1. $\Longrightarrow$ 2. Supongamos sin pérdida de generalización que $X_1, \ldots X_n$ es un orden topológico de las variables consistente con G, tal que si existe la arista $X_i \to X_j$ en el grafo, entonces $i < j$. Por la regla de la cadena 2.14, se tiene

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_1, \ldots x_{i-1})$$

Dado $i \in \{1, \ldots, n\}$, aplicando el criterio de d-separación, se tiene $I(X_i, ND(X_i)|\pi_i)_G$ con $ND(X_i)$ el conjunto de variables que no son descendientes de X_i.

Como G es un mapa de independencia de $p(x_1, \ldots, x_n)$, todas las relaciones de independencia que contiene G, las contiene también p y por lo tanto se tiene $I(X_i, ND(X_i)|\pi_i)_p$.

Debido al orden topológico de las variables, $\pi_i \subseteq \{X_1, \ldots, X_i - 1\} = A$, mientras que los descendientes de X_i no pueden estar en dicho conjunto. Por lo tanto, $A = \pi_i \cup B$, con $B \subseteq ND(X_i)$, así, existe un conjunto $C \subseteq \{X_1, \ldots, X_n\}$ tal que $ND(X_i) = B \cup C$. Entonces, aplicando la propiedad de descomposición de independencia condicional, se tiene

$$I(X_i, ND(X_i)|\pi_i) \Longleftrightarrow I(X_i, B \cup C|\pi_i) \Longrightarrow I(X_i, B|\pi_i)$$

Esto es equivalente a $I(X_i, \{X_1, \ldots, X_{i-1}\} \setminus \pi_i \,|\pi_i)$ y se concluye

$$p(x_i|x_1, \ldots x_{i-1}) = p(x_i|\pi_i)$$

de forma que

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i)$$

2. $\Longrightarrow$ 1. Queremos ver que, si la función de probabilidad factoriza de la siguiente manera

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i)$$

entonces toda independencia que contiene el grafo, está también contenida en la probabilidad p, equivalente a demostrar que $I(X_i, ND(X_i)|\pi_i)_G \Longrightarrow I(X_i, ND(X_i)|\pi_i)_p$.

Debido a que $\{X_1, \ldots, X_{i-1}\} \subseteq \{X_i, ND(X_i)\}$, pero la igualdad no tiene por qué darse, se reordenan las variables de forma que $\{X_i, ND(X_i)\} = \{Z_1, \ldots, Z_m\}$, con $i \leq m$ y $Z_m = Z_i$ y siendo esta nueva ordenación un orden topológico. Sea α_i los padres de la variable Z_i en G,

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) = p(z_1, \ldots, z_n) = \prod_{i=1}^{n} p(z_i|\alpha_i)$$

Por lo tanto, aplicando el teorema de Bayes

$$p(x_i|ND(X_i)) = \frac{p((x_i, ND(X_i))}{p(ND(X_i))} = \frac{p(z_1, \ldots, z_m)}{p(z_1, \ldots, z_{m-1})} = \frac{\prod_{i=1}^{m} p(z_i|\alpha_i)}{\prod_{i=1}^{m-1} p(z_i|\alpha_i)}$$

$$= p(z_m|\alpha_m) = p(x_i|\pi_i)$$

Por lo tanto, se tiene que $I(X_i, ND(X_i)|\pi_i)_p$ y G es un mapa de independencia. □

Es decir, para construir modelos probabilísticos a partir de grafos, se aplican criterios de separación gráfica para inferir un modelo de dependencia que represente la estructura cualitativa del modelo y se obtiene una factorización de la función de probabilidad que define su estructura cuantitativa.

Los modelos de dependencia que emplean grafos no dirigidos se conocen como redes de Markov y los que emplean grafos acíclicos dirigidos, redes bayesianas. En este trabajo nos centraremos en redes bayesianas,

2.4. Redes bayesianas

La principal desventaja de las redes de Markov es que no es posible representar dependencias no transitivas. La propiedad transitiva establece que si una variable A es dependiente de B y B es dependiente de otra variable C, entonces A es dependiente de C. Aunque este tipo de dependencias son muy comunes, también existen otras en las que dos variables A y C no son marginalmente dependientes aunque tengan una dependencia común B, sino condicionalmente dependientes cuando se conoce B. Este tipo de relaciones no transitivas no es posible en redes de Markov (pues A - B - C implica que A y C son dependientes).

Las redes bayesianas permiten definir relaciones transitivas y no transitivas mediante la dirección de las aristas ($A \rightarrow B \rightarrow C$, frente a $A \rightarrow B \leftarrow C$) resultando así más expresivas para modelar distintos conjuntos de datos reales en los que coexisten este tipo de relaciones.

Definición 19 (Red bayesiana)**.** Una red bayesiana es un par $B(\theta) = (G, P(\theta))$, donde G es un grafo dirigido acíclico y $P(\theta) = \{p(x_1|\,\pi_1;\theta_1), \ldots,$
$p(x_n|\,\pi_n;\theta_n)\}$ es un conjunto de n funciones de probabilidad condicionada, una para cada variable. θ_i es un conjunto de parámetros correspondientes a la función de probabilidad condicionada $p(x_i|\pi_i)$ y π_i es el conjunto de padres del nodo X_i en G. El conjunto $P(\theta)$ define una función de probabilidad asociada mediante la factorización

$$p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i|\pi_i; \theta_i) \tag{2.16}$$

con $\theta = \{\theta_1, \ldots, \theta_n\}$, el conjunto de parámetros de la función de probabilidad conjunta. El grafo dirigido acíclico G es un I-mapa minimal de $p(x_1 \ldots, x_n)$. En los casos en los que la referencia al conjunto de parámetros no sea necesaria, se denotará $B = (G, P)$

Por el teorema 18, todas las relaciones de independencia que se infieran del grafo mediante el criterio de $d-$separación, también estarán incluidas en el modelo probabilístico asociado.

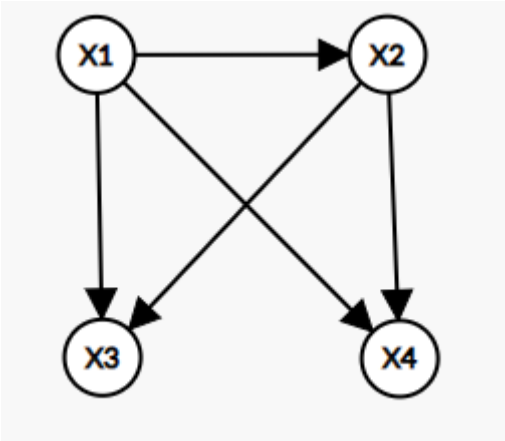


Figura 2.1: Grafo empleado para construir red bayesiana

Ejemplo 20 (Red bayesiana). Sean X_1, X_2, X_3, X_4 cuatro variables binarias y considérese el grafo de la figura 2.1. Este define una red bayesiana junto a una función de probabilidad, la cual es factorizada en funciones de probabilidad condicionadas de la siguiente manera:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2)$$

Esta factorización coincide con la que se obtuvo en el ejemplo 10 y depende de 11 parámetros. Este grafo es un mapa de independencia minimal de p. Las funciones de

probabilidad condicionadas se corresponden con tablas de probabilidad, que recogen las posibles combinaciones de valores que pueden tomar las variables. El grafo, junto a dichas tablas (los parámetros), determinan una red bayesiana.

Dependiendo de las variables del grafo, discretas, continuas o ambas, y de la distribución que se considere para ellas, existen distintos tipos de redes bayesianas. En este trabajo se emplearán redes bayesianas multinomiales, que tienen únicamente variables discretas y cuyas probabilidades condicionadas asociadas son multinomiales. Por otra parte, la dimensión de una red bayesiana $B = (G, P(\theta))$ es el número de parámetros necesarios para definir la función de probabilidad $P(\theta)$ y se denota como $dim(B)$.

2.5. Aprendizaje de redes bayesianas

En la práctica, es habitual que tanto la estructura cualitativa como la cuantitativa del modelo probabilístico no se conozcan y por lo tanto, se infieren a partir de un conjunto de datos la estructura de dependencia, las funciones de probabilidad condicionada asociadas y los parámetros de los que depende. Este proceso es conocido como aprendizaje y consta de dos fases: el aprendizaje estructural, que consiste en definir la estructura de dependencia gráfica (qué aristas incluir en el grafo) y el aprendizaje paramétrico, que consiste en determinar las probabilidades (parámetros) de P. La parte más compleja de este proceso es el aprendizaje estructural, que es un problema NP-completo (véase Koller y Friedman (2009)). Una vez se obtiene el grafo que represente las relaciones de dependencias entre las variables, se puede conocer la factorización de la función de probabilidad y es posible estimar sus valores.

2.5.1. Aprendizaje paramétrico

Los enfoques para obtener los parámetros de la función de probabilidad asociada al grafo G, a partir del conjunto de datos $D = \{D_1, \ldots, D_n\}$, se dividen, principalmente, en dos tipos: basados en estimadores de máxima verosimilitud y basados en métodos bayesianos. En este trabajo utilizaremos el estimador de máxima verosimilitud (MLE), que se basa en obtener los parámetros $\hat{\theta}$ que maximicen la función de verosimilitud, es decir,

$$L(\hat{\theta}|G, D) = max_\theta \, L(\theta|G, D) \tag{2.17}$$

donde

$$L(\theta|G, D) = p(D|G, \theta) \tag{2.18}$$

El conjunto de parámetros es $\theta = \{\theta_{ijk} : i \in \{1 \dots n\},\ j \in \{1, \dots r_i\},\ k \in \{1 \dots s_i\}\}$, donde n es el número de variables, r_i es el número de posibles valores que puede tomar la variable X_i, s_i es el número de posibles realizaciones de los padres de X_i (π_i) y θ_{ijk} es la probabilidad de que la variable i tome su valor j-ésimo y sus padres en el grafo G tomen su k-ésima realización, $\theta_{ijk} = p(X_i = x_{i_j} | \pi_i = w_k)$. Así, asumiendo que los datos son independientes e igualmente distribuidos, se tiene que

$$p(D|G,\theta) = \prod_l p(D_l|G,\theta) = \prod_l \prod_{i,j,k} \theta_{ijk} \cdot \chi(i,j,k : D_l) \tag{2.19}$$

$$= \prod_i^n \prod_j^{r_i} \prod_k^{s_i} \theta_{ijk}^{N_{ijk}} \tag{2.20}$$

con N_{ijk} el número de casos en los que la variable X_i toma el valor j-ésimo y sus padres toman los valores de la realización k-ésima y χ la siguiente función

$$\chi(i,j,k : D_l) = \begin{cases} 1 & \text{si } X_i = x_{i_j} \text{ y } \pi_i = w_k \text{ en } D_l \\ 0 & \text{en otro caso} \end{cases} \tag{2.21}$$

Habitualmente, se maximiza el logaritmo de la verosimilitud, puesto que es más sencillo que maximizar la verosimilitud (los productos se convierten en sumatorios) y puesto que el logaritmo es una función monótona, ambos procedimientos son equivalentes. En este caso, la fórmula del logaritmo la verosimilitud que se debe maximizar es la siguiente:

$$\ell(\theta|G,D) = \sum_i^n \sum_j^{r_i} \sum_k^{s_i} N_{ijk} \log \theta_{ijk} \tag{2.22}$$

En el caso de redes bayesianas multinomiales, la asignación de parámetros que la maximiza se corresponde con las propias frecuencias empíricas en el conjunto de datos, es decir, $\hat{\theta} = \{\hat{\theta}_{ijk} : i \in \{1 \dots n\},\ j \in \{1, \dots r_i\},\ k \in \{1 \dots s_i\}\}$, con

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ik}} \tag{2.23}$$

siendo N_{ik} el número de veces que los padres de la variable X_i toman su k-ésima realización en el conjunto de datos. Para más detalles consultar Koller y Friedman (2009).

Sin embargo, este método presenta un claro inconveniente. Si la muestra de datos no es suficientemente grande, o el tamaño de los dominios de las variables es muy extenso en relación con los datos, puede darse el caso de que algunos parámetros sean estimados como nulos debido a que no ocurran ciertas configuraciones en los datos.

2.5.2. Aprendizaje estructural

Existen tres enfoques principales para obtener la estructura gráfica: el primero es el aprendizaje basado en restricciones, que consiste en realizar tests de independencia y dependencia sobre los datos para después encontrar la red que mejor represente las relaciones obtenidas, el segundo es el aprendizaje basado en puntuación y búsqueda, sobre el que se desarrollará a continuación, puesto que es el que se utiliza en este trabajo, y por último, el aprendizaje híbrido, que es una combinación de los dos anteriores.

Aprendizaje basado en restricciones

El aprendizaje basado en restricciones tiene como objetivo encontrar el grafo que mejor represente las relaciones de independencia de los datos, es decir, encontrar el mejor mapa de independencia minimal. Para encontrar dichas relaciones emplean tests de independencia condicional, que determinan si, dadas dos variables A y B, son condicionalmente independientes dado un conjunto de variables $S = \{S_1, \dots S_d\}$, es decir, si se tiene $I(A, B|S)$ o $D(A, B|S)$. También se determina la dependencia o independencia no condicionada con $S = \emptyset$. El tamaño del conjunto S está limitado dependiendo del problema (habitualmente, un máximo de tres o cuatro variables, Koski y Noble (2012)), para que el algoritmo sea realizable.

Aprendizaje basado en puntuación y búsqueda

El aprendizaje basado en puntuación y búsqueda aplica técnicas de optimización para encontrar de manera iterativa la estructura de red que mejor se adapta a los datos y consta de dos elementos: un algoritmo de búsqueda, que permita recorrer el espacio de grafos candidatos, y una medida de calidad, que permita evaluar cada uno de ellos. Cabe mencionar que, de acuerdo con Koller y Friedman (2009), dado que este tipo de aprendizaje estructural considera la estructura entera de la red, es menos sensible a fallos individuales y es mejor en equilibrar las dependencias entre las variables y el coste de añadir nuevas aristas. Sin embargo, no siempre es posible encontrar la red de mayor calidad, pero sí máximos locales. Existen distintas estrategias para iterar en el espacio de grafos, siendo la opción más simple la que emplea el algoritmo K2 (Cooper y Herskovits (1992)), que comienza con un grafo de nodos ordenados, sin aristas, y las añade vorazmente (sin crear ciclos), hasta que la calidad de la red no aumenta añadiendo enlaces o el grafo es completo.

En este trabajo se emplea el algoritmo Hill-Climbing (Heckerman et al. (1995)), ya que estudios previos que han analizado el comportamiento de distintos algorit-

mos para aplicaciones en clima concluyen que es el más adecuado y eficaz (para más información consultar Graafland (2022)). Hill-Climbing puede iniciar la búsqueda a partir de cualquier grafo, sin embargo suele comenzar con uno vacío y en cada iteración considera todos los grafos vecinos del actual, aquellos que se pueden formar eliminando, añadiendo o cambiando la dirección de una arista en la red actual. Si no existe ningún vecino que aumente la calidad de la red, se ha llegado a un máximo local (en ocasiones global). Para evitar obtener un máximo local poco favorable, se realizan reinicios locales, en los que se aplican cambios aleatorios al grafo en el máximo local, para después reiniciar el algoritmo desde ahí. Para más información sobre distintos algoritmos de puntuación y búsqueda consultar Kitson et al. (2023)

Por otra parte, dada una red bayesiana $B = (G, p(\theta))$ y un conjunto de datos D, una medida de calidad $Q(B|D)$ permite evaluar las distintas redes y debe dar el mismo valor a redes que codifiquen la misma estructura de independencias. A continuación se describen distintos tipos de medidas de calidad para redes bayesianas multinomiales.

Medidas de calidad bayesianas Las medidas de calidad basadas en estadística bayesiana, consisten en designar a cada red un valor proporcional a la probabilidad del grafo dados los datos, $p(G|D)$. Por el teorema de Bayes (2.12),

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)} \propto p(G)p(D|G) \tag{2.24}$$

ya que $p(D)$ es igual para todas las estructuras gráficasypor lo tanto es una constante de normalización. Entonces, tomando logaritmos, se define una medida de calidad bayesiana como

$$Q_B(G, D) = \log p(D|G) + \log p(G) \tag{2.25}$$

donde el término $p(D|G)$ permite tener en cuenta que los parámetros también son variables aleatorias,

$$p(D|G) = \int_\theta p(D|\theta, G)p(\theta|D)d\theta \tag{2.26}$$

Las medidas que se obtienen mediante estadística bayesiana son equivalentes a otras más simples, basadas en teoría de la información. Para más información sobre medidas de calidad bayesianas, consultar Castillo et al. (1996).

Medidas de información Otro tipo de medidas de calidad son las medidas de información. Estas se basan en elegir la estructura de red que mejor representa los datos, con una penalización proporcional al número de parámetros que se necesitan

para definir la función de probabilidad asociada, $dim(B)$, para así evitar sobreajustes. Utilizando la notación empleada en 2.5.1, se introduce su fórmula general:

$$Q_I(B, D) = \log p(G) + \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{i=1}^{s_i} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} - dim(B) f(N) \tag{2.27}$$

donde N es el número de casos que tiene el conjunto de datos D, $f(N)$ es la función de penalización ($f(N) \geq 0$). Es habitual que $p(G)$ sea una distribución uniforme. Además,

$$dim(B) = \sum_{i=1}^{n} (r_i - 1) \cdot s_i \tag{2.28}$$

En este trabajo se emplean las siguientes medidas de calidad de información:

1. Criterio de máxima verosimilitud de información (LL). ($f(N) = 0$). Se basa en encontrar la estructura G que maximice la verosimilitud de los datos dada la red (2.17). Para conseguir esto se emplean los parámetros obtenidos mediante MLE para ese mismo grafo, ya que, como se ha visto en 2.5.1 maximizan la verosimilitud. Es decir, en el caso de redes multinomiales, se aplican los parámetros (2.23) en (2.19), y, al tomar logaritmos, se obtiene

$$LL(B, D) = \log p(G) + \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{i=1}^{s_i} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} \tag{2.29}$$

   Puesto que la penalización es nula, favorece las estructuras completas.

2. Criterio de información de Akaike (AIC). ($f(N) = 1$).

$$AIC(B, D) = LL(B, D) - dim(B) \tag{2.30}$$

3. Criterio de información bayesiano (BIC). ($f(N) = \frac{1}{2} \log N$).

$$AIC(B, D) = LL(B, D) - \frac{1}{2} dim(B) \log N \tag{2.31}$$

Una vez se obtiene la estructura gráfica, es posible estimar los parámetros de su función de probabilidad asociada.

Capítulo 3

Causalidad

3.1. Causalidad en términos probabilísticos

Durante la segunda mitad del siglo XX, las principales teorías de causalidad se basaron en la idea de que la causa debe aumentar la probabilidad de sus efectos. A continuación, se introducen distintas definiciones para la causalidad desde un punto de vista probabilístico.

El concepto probabilístico más básico de causalidad es que A es la causa de B si $p(B|A) > p(B|\bar{A})$ (Kleinberg (2013)). Sin embargo, esto no es útil para establecer relaciones causa-efecto, ya que puede existir una relación causal que no lo verifique o viceversa. Una alternativa es el principio de causa común. Este establece que, dados dos sucesos correlacionados, A y B, o bien uno es la causa del otro, o tienen una causa común C. Se dice que C es una causa común anterior de A y B si y solo si se verifica

1. $I(A, B|C)$
2. $p(A|C) > p(A|\bar{C})$
3. $p(B|C) > p(B|\bar{C})$

No obstante, puede darse el caso en el que dos sucesos A y B estén correlacionados pero ninguno sea causa del otro y que su correlación no pueda ser explicada por una causa común. Esto puede ser porque dicha causa no existe, o porque, habiendo una causa común C, se tiene que $D(A, B|C)$ (consultar Hitchcock y Rédei (2021) para contraejemplos).

Por otra parte, tomando en consideración el tiempo, Hans Reichenbach establece que, dados C^t y $E^{t'}$, dos sucesos que tienen lugar en t y t' respectivamente, C^t es causa de $E^{t'}$ si y solo si

1. $t' > t$
2. $p(E^{t'}|C^t) > p(E^{t'}|\bar{C}^t)$
3. No existe ningún suceso $D^{t''}$ con $t'' \leq t$ tal que $I(E^{t'}, C^t|D^{t''})$

En Hitchcock (2021) se detallan otras muchas definiciones de causalidad, las cuales permitieron su entendimiento desde un punto de vista probabilístico, pero no produjeron métodos para identificar relaciones causales. En este trabajo nos centramos en modelos causales para series temporales, como son la causalidad de Granger y las redes bayesianas dinámicas.

3.2. Causalidad de Granger

Uno de los métodos más extendidos para determinar relaciones causales en datos de series temporales, especialmente en economía, es la causalidad de Granger, que permite determinar si dadas dos series temporales, una puede predecir la otra. La definición, dada por Granger (1980) establece que X_1 es causa en el sentido de Granger de X_2 si

$$p(X_1^{t+1} \,|\, W^{0:t}) \neq p((X_1^{t+1} \,|\, W^{0:t} \setminus X_2^{0:t}) \tag{3.1}$$

es decir, si X_1^{t+1} y $X_2^{0:t}$ son condicionalmente dependientes dado $W^{0:t} \setminus X_2^{0:t}$,

$$D\left(X_1^{t+1}, X_2^{0:t} \,|\, W^{0:t} \setminus X_2^{0:t}\right) \tag{3.2}$$

donde X_i^t es el valor de la variable X_i en el tiempo t, $X_i^{0:t}$ es el conjunto de valores de la variable X_i para cualquier tiempo igual o anterior a t, es decir $X_i^{0:t} = \{X_i^t, X_i^{t-1}, \ldots, X_i^0\}$ y $W^{0:t}$ es el conjunto de toda la información disponible hasta el tiempo t (incluido).

Utilizar todas las variables posibles, durante un periodo infinitamente largo de tiempo, no es factible, por lo que existen distintas adaptaciones. Una de ellas es el test bivariado, que se puede aplicar mediante el uso de un modelo autoregresivo, el cual difiere de la definición original, ya que no utiliza toda la información disponible, sino solo aquella relativa a las variables que se corresponden con la causa y el efecto (C y E respectivamente). Si los coeficientes de los valores retardados de C son distintos de cero, se dice que C es causa Granger de E. De acuerdo con Kleinberg (2013), para calcular el valor de una variable en un tiempo t, un modelo de autoregresión de m retrasos utiliza valores retardados hasta el tiempo $t - m$, es decir, para los tiempos $t - 1, \ldots, t - m$, las variables X_1 y X_2 influyen sobre el valor de X_1^t, mientras que utilizando la definición original, m tendría que ser infinito. Mediante un modelo de

autoregresión lineal de m retrasos, se puede representar la variable X_1 de la siguiente manera:

$$X_1^t = \sum_{j=1}^{m} A_{11}^j X_1^{t-j} + \sum_{j=1}^{m} A_{12}^j X_2^{t-j} + \varepsilon_1^t \tag{3.3}$$

donde los coeficientes A_{ij} representan el efecto que tiene la variable X_j sobre X_i, y ε_1 es una variable aleatoria (habitualmente con media cero) que representa el error. Se comprueba si X_2 es causa, en el sentido Granger, de X_1 estudiando si, para valores distintos de cero en los coeficientes A_{12}, se produce una reducción estadísticamente significativa para la varianza del error frente a cuando sí son nulos.

Sin embargo, este test bivariado no tiene en cuenta causas comunes a ambas variables o relaciones indirectas para establecer relaciones causa-efecto. Una alternativa es utilizar un test multivariado, incluyendo el resto variables en el modelo. El modelo de autoregresión lineal aplicado a n variables es el siguiente:

$$X^t = \sum_{j=1}^{m} A^t X^{t-j} + \varepsilon^t \tag{3.4}$$

donde A^t son matrices de coeficientes de dimensión $n \times n$, $X^t = (X_1^t, \ldots, X_n^t)$ es un vector que contiene los valores de todas las variables para un tiempo t y $\varepsilon^t = (\varepsilon_1^t, \ldots, \varepsilon_n^t)$ es un vector que contiene el error para cada variable. De acuerdo con Shojaie y Fox (2022), las relaciones de causalidad en el sentido de Granger se corresponden con los elementos distintos de cero en los coeficientes de autoregresión.

A pesar de que el test multivariado permite considerar más variables, computacionalmente, es inviable aplicarlo para un número moderado de variables y series temporales, raramente se utiliza con más de seis. Por este motivo, cuando se aplica el test de Granger, habitualmente se corresponde con el bivariado (Kleinberg (2013)).

3.3. Redes bayesianas dinámicas

Las redes bayesianas dinámicas son una extensión de las redes bayesianas que introducen otra dimensión, el tiempo, que en este trabajo será discreto. Habitualmente, hay una primera red que muestra las relaciones entre las variables en un tiempo t, y otras en los tiempos $t+1$, $t+2$, ... con relaciones entre las redes de series temporales distintas, mostrando las dependencias entre nodos del tiempo inicial y aquellos de tiempos futuros.

Para la forma más general de una red bayesiana dinámica, su función de probabilidad factoriza como:

$$p(X^0, \ldots, X^T) = p(X^0) \prod_{t=1}^{T} p(X^t | X^{0:t-1}) \tag{3.5}$$

donde $X^t = \{X_1^t, \ldots, X_n^t\}$ es el conjunto de todas las variables en un tiempo t.

En este trabajo, se trabajan con redes bayesianas dinámicas que verifican las siguientes condiciones:

1. La suposición de Markov. Esta establece que, dado un proceso dinámico, un tiempo t solo está condicionado por el anterior, $t-1$, es decir, dado $t \in \mathbb{N}^+$

   $$I(X^{t+1}, X^{0:t-1} | X^t)$$

   Esto es equivalente a que en la red bayesiana dinámica solo haya enlaces entre series temporales consecutivas (además de enlaces estáticos).

   Este concepto se puede extender con la suposición de Markov de orden α, con $\alpha \in \mathbb{N}^+$, que consiste en que un instante de tiempo t está condicionado por los α anteriores. Es análogo a

   $$I(X^{t+1+\alpha}, X^{0:t-1} | X^{t:t+\alpha})$$

2. Proceso estacionario en el tiempo. Un proceso dinámico, estocástico y de Markov de orden α, se dice que es estacionario si, para todo $t, t' \in \mathbb{N}^+$ se tiene

   $$p(X^{t+\alpha} \mid X^{t:t+\alpha-1}) = p(X^{t'+\alpha} \mid X^{t':t'+\alpha-1}) \tag{3.6}$$

   es decir, que todas las redes bayesianas entre distintas series temporales tienen la misma estructura, por lo que si esta condición se verifica, es posible hacer representaciones basadas en plantilla. En este trabajo $\alpha = 1$, por lo que

   $$p(X^{t+1} \mid X^t) = p(X^{t'+1} \mid X^{t'}) \tag{3.7}$$

La red bayesiana dinámica más sencilla es aquella para la cual se suponen ciertas estas dos hipótesis, y se representa con el par $B =< B_0, B_{\rightarrow} >$ donde B_0 es la red bayesiana inicial, que describe las relaciones sobre X^0 y define $p(X^0)$, mientras que $B_{\rightarrow}$ es la red bayesiana que describe las relaciones entre las variables en los tiempos sucesivos, y se conoce como red de transición. La red $B_{\rightarrow}$ define la probabilidad condicionada $p(X^{t+1} \mid X^t)$ y contiene tanto enlaces estacionarios, entre variables en un

mismo tiempo t, como enlaces dinámicos, entre variables en un tiempo t y variables en el tiempo anterior $t-1$, los cuales representan causalidad. Dado que se verifica la suposición de Markov, por 2.9 se tiene que

$$p(X^0, \ldots, X^T) = p(X^0) \prod_{t=1}^{T} p(X^t \mid X^{t-1}) \tag{3.8}$$

Por otra parte, de acuerdo con Murphy (2002)

$$p(X^t \mid X^{t-1}) = \prod_{i=1}^{n} p(X_i^t \mid \pi_i^t) \tag{3.9}$$

donde π_i^t son los padres de la variable X_i^t, los cuales pueden estar en la misma serie temporal t o en la anterior, $t-1$. Entonces, se obtiene que la función de probabilidad que define la red bayesiana $< B_0, B_{\rightarrow} >$, viene dada por

$$p(X^0, \ldots, X^T) = p(X^0) \prod_{t=1}^{T} \prod_{i=1}^{n} p(X_i^t \mid \pi_i^t) \tag{3.10}$$

3.3.1. Aprendizaje de redes bayesianas dinámicas

Los métodos de aprendizaje para redes bayesianas dinámicas son muy similares a los de redes bayesianas, con la diferencia de que, en el caso dinámico, hay que conocer la estructura y los parámetros de dos redes, B_0 y $B_{\rightarrow}$. En este trabajo emplearemos el algoritmo hill-climbing con los criterios de información de Akaike y bayesiano para el aprendizaje estructural, y el estimador de máxima verosimilitud para el aprendizaje paramétrico.

Aprendizaje paramétrico

Al igual que en las redes bayesianas, se va a emplear el estimador de máxima verosimilitud, el cual se basa en obtener los parámetros que maximizan la verosimilitud de los datos (2.17), que en este caso son los conjuntos $\hat{\theta}^0$ y $\hat{\theta}^{\rightarrow}$, los cuales se corresponden con los parámetros de las redes B_0 y $B_{\rightarrow}$ respectivamente. En las redes dinámicas, la función de verosimilitud a maximizar descompone como sigue (Friedman et al. (2013)):

$$L(\theta|G, D) = p(D|G, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{r_i} \prod_{k=1}^{s_i} \left(\theta_{ijk}^0\right)^{N_{ijk}^0} \cdot \prod_{i=1}^{n} \prod_{j'=1}^{r_i} \prod_{k'=1}^{s_i} \left(\theta_{ij'k'}^{\rightarrow}\right)^{N_{ij'k'}^{\rightarrow}} \tag{3.11}$$

donde la primera parte de productos se corresponde con los parámetros de B_0, y la segunda con los de $B_{\rightarrow}$. La notación empleada es muy similar, o la misma, que para el caso de redes estáticas, concretamente, n es el número de variables, r_i es el número

de posibles valores que puede tomar la variable X_i y s_i es el número de posibles realizaciones de los padres de X_i. Por otra parte, N^0_{ijk} es el número de ocurrencias en los datos en las que la variable X_i en el tiempo $t = 0$ toma su valor j-ésimo y sus padres toman los valores de su k-ésima realización, $N^{\rightarrow}_{ij'k'}$ es el número de veces que en los datos, la variable X_i, para cualquier tiempo $t \in \{1, \dots T\}$, toma su valor j-ésimo y sus padres toman su k-ésima realización. Por último,

$$\theta^0_{ijk} = p(X^0_i = x_{i_j} | \pi^0_i = w_k) \text{ y } \theta^{\rightarrow}_{ij'k'} = p(X^t_i = x_{i_{j'}} | \pi^t_i = w_{k'})$$

con $\theta^0 = \{\theta_{ijk} : i \in \{1 \dots n\}, j \in \{1, \dots r_i\}, k \in \{1 \dots s_i\}\}$ y $\theta^{\rightarrow} = \{\theta_{ij'k'} : i \in \{1 \dots n\}, j \in \{1, \dots r_i\}, k \in \{1 \dots s_i\}\}$. Al igual que se hace para las redes bayesianas, se tiene que el logaritmo de la verosimilitud es

$$\ell(\theta|G, D) = p(D|G, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{k=1}^{s_i} N^0_{ijk} \log \theta^0_{ijk} + \sum_{i=1}^{n} \sum_{j'=1}^{r_i} \sum_{k'=1}^{s_i} N^{\rightarrow}_{ij'k'} \log \theta^{\rightarrow}_{ij'k'} \quad (3.12)$$

En el caso de distribuciones multinomiales, los parámetros que maximizan la verosimilitud vienen dados por las siguientes expresiones

$$\hat{\theta^0_{ijk}} = \frac{N^0_{ijk}}{\sum_{k=1}^{s_i} N^0_{ijk}} \quad (3.13)$$

$$\hat{\theta^{\rightarrow}_{ij'k'}} = \frac{N^{\rightarrow}_{ij'k'}}{\sum_{k'=1}^{s_i} N^{\rightarrow}_{ij'k'}} \quad (3.14)$$

Aprendizaje estructural

Los algoritmos de aprendizaje estructural basados en puntuación y búsqueda que se emplean para redes bayesianas, se aplican de la misma manera para redes bayesianas dinámicas. No obstante, las medidas de calidad se han de adaptar. Al igual que en el caso estático, los criterios BIC y AIC se corresponden con el logaritmo de la verosimilitud de los datos junto a una penalización, proporcional a la complejidad del modelo, sin embargo, ahora se ha de tener en cuenta la calidad de ambas redes, la inicial y la de transición. Sea $dim(B)$ la dimensión del modelo, con $dim(B) = dim(B_0) + dim(B_{\rightarrow})$ y, N_0 y $N_{\rightarrow}$ el número de instancias en los conjuntos de datos utilizados para entrenar B_0 y $B_{\rightarrow}$ respectivamente (D_0 y $D_{\rightarrow}$). Entonces, los criterios BIC y AIC para redes dinámicas son los que siguen.

1. **BIC**.

$$BIC(B, D) = BIC(B_0, D_0) + BIC(B_{\rightarrow}, D_{\rightarrow}) \quad (3.15)$$

donde, aplicando (2.31) independientemente a cada red, se obtiene

$$BIC(B_0, D_0) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{k=1}^{s_i} N^0_{ijk} \log \theta^0_{ijk} - \frac{1}{2} dim(B_0) \log N_0 \quad (3.16)$$

$$BIC(B_{\rightarrow}, D_{\rightarrow}) = \sum_{i=1}^{n} \sum_{j'=1}^{r_i} \sum_{k'=1}^{s_i} N_{ij'k'}^{\rightarrow} \log \theta_{ij'k'}^{\rightarrow} - \frac{1}{2} dim(B_{\rightarrow}) \log N_{\rightarrow} \tag{3.17}$$

2. **AIC**.

$$AIC(B, D) = AIC(B_0, D_0) + AIC(B_{\rightarrow}, D_{\rightarrow}) \tag{3.18}$$

y, procediendo igual que en el método anterior, se tiene

$$AIC(B_0, D_0) = \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{k=1}^{s_i} N_{ijk}^{0} \log \theta_{ijk}^{0} - dim(B_0) \tag{3.19}$$

$$AIC(B_{\rightarrow}, D_{\rightarrow}) = \sum_{i=1}^{n} \sum_{j'=1}^{r_i} \sum_{k'=1}^{s_i} N_{ij'k'}^{\rightarrow} \log \theta_{ij'k'}^{\rightarrow} - \frac{1}{2} dim(B_{\rightarrow}) \log N_{\rightarrow} \tag{3.20}$$

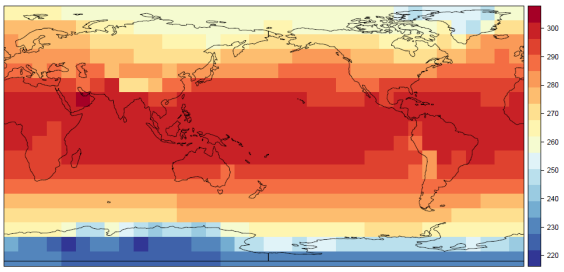
Capítulo 4

Experimentos y resultados

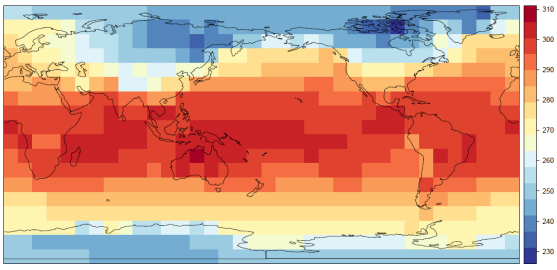
4.1. Datos climáticos de temperatura

Las observaciones de estaciones meteorológicas proporcionan registros históricos que describen la evolución de distintos parámetros climáticos (como la temperatura o la precipitación). Sin embargo, estos datos están disponibles sólo para localidades puntuales donde se emplazan estaciones meteorológicas y no permiten caracterizar las condiciones climáticas de forma regular (por ejemplo, sobre una cuadrícula global). Por lo tanto, para poder analizar de manera homogénea el clima a nivel global, se utilizan datos creados por un reanálisis. Estos productos son simulaciones llevadas a cabo por modelos globales del clima, que asimilan las observaciones disponibles para un día dado y proporcionan una estimación del estado de las variables sobre la rejilla regular del modelo (con una resolución típica de decenas de kilómetros); estos datos se consideran pseudo-observaciones y constituyen una de las referencias más utilizadas para estudiar la evolución del clima. En este trabajo se utilizan datos producidos por el reanálisis ERA-Interim, realizado por el Centro Europeo de Predicción a Medio Plazo, que proporciona valores diarios (agregados mensualmente en este trabajo) para un período de 30 años (1981-2010) representativo del clima actual; en este trabajo se utiliza la temperatura media sobre una red regular de 10°, correspondiente a una resolución espacial de 1000 km aproximadamente. Por lo tanto se obtiene una rejilla espacial con $p = 36 \times 18 = 648$ casillas con datos para una serie temporal mensual de longitud $n = 30 \times 12 = 360$.

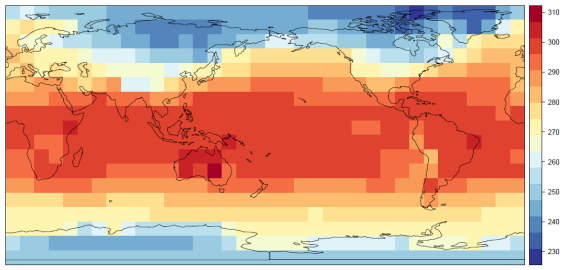
En la figura 4.1, se pueden observar los datos proporcionados por el reanálisis, con la media de las temperaturas mensuales en 4.1a y las temperaturas para dos meses concretos, enero de 1998 4.1b, durante el cual tiene lugar el fenómeno de El Niño, y enero de 1999 4.1c, en el que no. Por ello, se puede apreciar que la temperatura en los trópicos en enero de 1998 fue más elevada que en enero de 1999, un año después.



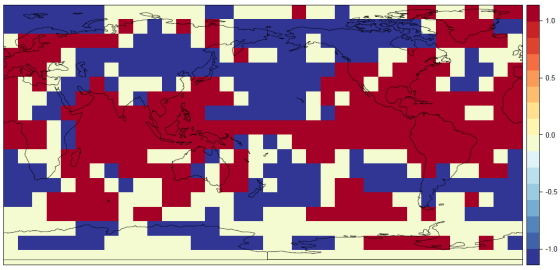
(a) Media de las temperaturas mensuales (1981-2010)



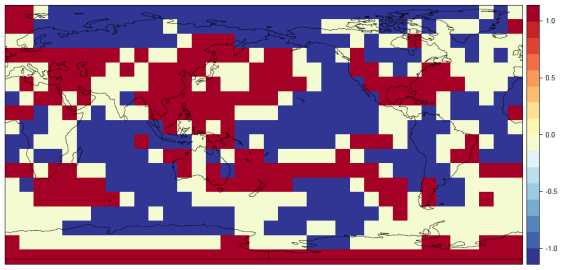
(b) Temperaturas globales de enero de 1998 dadas por el reanálisis ERA-Interim



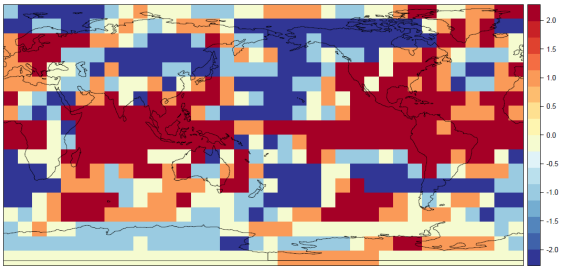
(c) Temperaturas globales de enero de 1999 dadas por el reanálisis ERA-Interim



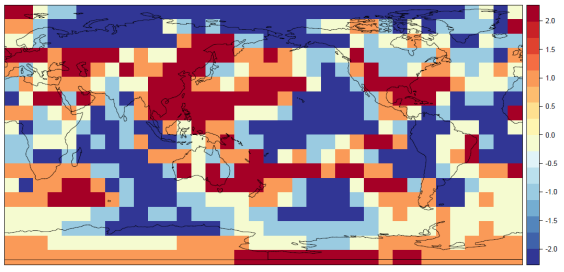
(d) Anomalías globales en enero 1998 para la discretización en tres estados



(e) Anomalías globales en enero 1999 para la discretización en tres estados



(f) Anomalías globales en enero 1998 para la discretización en cinco estados



(g) Anomalías globales en enero 1999 para la discretización en cinco estados

Figura 4.1: Mapas con distintas representaciones de la temperatura media y de las anomalías de esta para dos meses particulares: enero de 1998 y 1999.

4.1.1. Preprocesamiento de los datos: Anomalías

Como los datos diarios de temperatura varían estacionalmente, es común eliminar este factor considerando anomalías. Dada una casilla i, sean $Ti = (T_{i,1}, \ldots, T_{i,n})$ sus temperaturas mensuales; definimos la anomalía de temperatura como $A_i = T_i - \bar{T}_i$, con $\bar{T}_i$ la media de temperatura para cada mes del año (en la casilla i). Es decir, dado $k \in \{1, \ldots, 360\}$, que se corresponde con el mes $m \in \{1, \ldots, 12\}$ de un año y_0, la anomalía de la casilla i para la serie temporal k es $A_{i,k} = T_{i,k} - \sum_{y=1}^{30}(T_{i,y\cdot m}/30)$.

4.1.2. Discretización de los datos

Debido a que se van a emplear redes bayesianas multinomiales y los datos son continuos, se realizan dos discretizaciones de estos mediante métodos no supervisados. Se han dividido en tres y cinco intervalos, calculando los terciles y quintiles para cada variable en cada casilla, y clasificando los datos después. Para el primer caso, el dominio de cada nodo es anomalía positiva (1), anomalía negativa (-1) o anomalía nula (0), mientras que para el segundo, el dominio de cada variable es anomalía positiva (2), anomalía positiva leve (1), anomalía negativa (-2), anomalía negativa leve (-1) o anomalía nula (0).

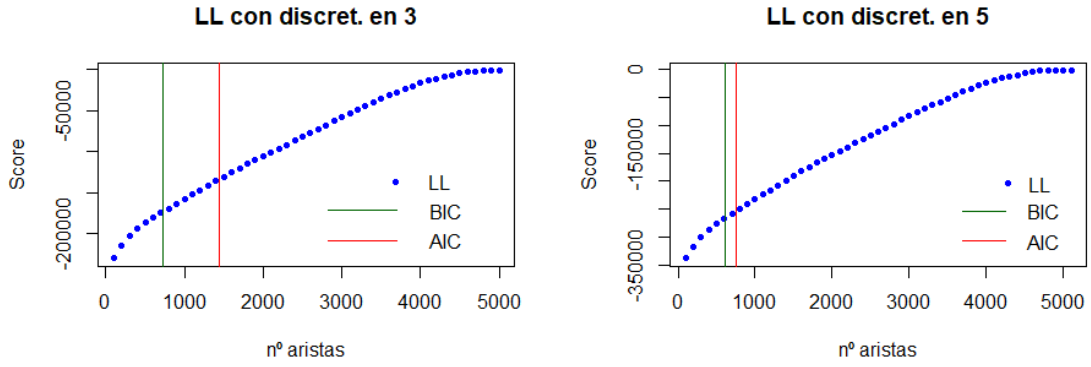
En la figura 4.1, se encuentran la representación de las anomalías para ambos conjuntos de datos discretos. Por una parte, las figuras 4.1d y 4.1e muestran las anomalías en cada casilla del mapa para la discretización en tres estados, también para enero de 1998 y 1999 respectivamente. En las figuras 4.1f y 4.1g se representan las anomalías de temperatura en esos mismos meses, pero para el conjunto de datos discretizado en cinco estados, por lo que hay un mayor nivel de detalle. En ambas discretizaciones se observa que en enero de 1998, mes en el que tuvo lugar el fenómeno El Niño, son más frecuentes las anomalías positivas en los trópicos, en contraposición con enero de 1999.

4.2. Aprendizaje de redes bayesianas

Se han entrenado distintas redes bayesianas mediante el paquete *bnlearn* de R, utilizando el algoritmo hill-climbing (HC), puesto que como se ha mencionado anteriormente, distintos estudios concluyen que en el caso del clima es el más adecuado. Para ello se han considerado distintas medidas de calidad de calidad de la red, como son el criterio de máxima verosimilitud de información (LL) (2.29), el criterio de información de Akaike (AIC) (2.30) o el criterio de información bayesiano (BIC) (2.31), aplicado a ambos conjuntos de datos discretos.

4.2.1. Niveles de discretización

En la figura 4.2 se muestra la evolución del criterio de máxima verosimilitud de información a medida que se añaden enlaces a las redes bayesianas entrenadas con dicha medida y el algoritmo hill-climbing, para la discretización en tres y cinco intervalos respectivamente. Las líneas verticales se corresponden con el tamaño de las redes que se obtienen aplicando, a cada conjunto de datos, los criterios AIC (rojo) y BIC (verde), sobre las cuales se profundiza más adelante en 4.2.3.



(a) Discretización en tres

(b) Discretización en cinco

Figura 4.2: Gráficas de evolución de la medida de calidad de una red bayesiana en función del número de aristas, usando el criterio LL.

La medida en términos absolutos difiere entre ambas discretizaciones, lo cual es esperable teniendo en cuenta que poseen un número distinto de posibles estados. También es de esperar la diferencia en el número de enlaces de las redes bayesianas utilizando los criterios AIC y BIC (las líneas verticales roja y verde respectivamente). Esto se debe a que para definir la función de probabilidad para la discretización en cinco estados son necesarios más parámetros (para definir la probabilidad condicional de un nodo con dos padres son necesarios 3^3 parámetros con 3 estados y 3^5 con 5), y estas medidas tienen una penalización proporcional a esa cantidad, lo que hace que paren antes. No obstante, estructuralmente tienen un comportamiento similar, con un crecimiento inicial muy rápido y finalizando alrededor de 5000 enlaces. Por lo tanto, en el resto del trabajo, todas las redes serán entrenadas utilizando la primera discretización, en tres intervalos.

Dado que LL es una medida de calidad que no penaliza la complejidad de la red, añade alrededor de 5000 aristas en ambos casos. No obstante, a pesar de que la puntuación de la red es creciente, parte de estos enlaces son redundantes, como se verá a continuación.

4.2.2. Validación cruzada

La validación cruzada es una técnica muy utilizada en el aprendizaje automático para comprobar la capacidad de generalización un modelo, si este permite extrapolar información correctamente a partir de datos nuevos, que no pertenecen al conjunto de entrenamiento. Un modelo está sobreajustado si explica muy bien los datos utilizados para entrenarlo pero no generaliza, es decir, no es capaz de realizar predicciones acertadas con nueva información.

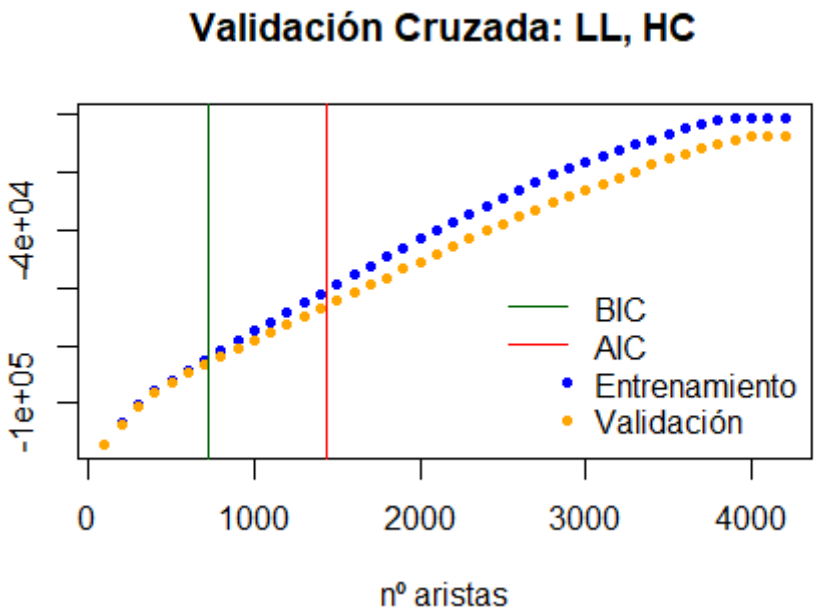


Figura 4.3: Medidas de máxima verosimilitud, $LL(B_t, D_t)$ y $LL(B_t, D_v)$, respecto al número de enlaces de la red B_T

El método se aplica entrenando el modelo, la red bayesiana, con la mitad de los datos, y evaluándolo utilizando la otra mitad. Los datos se separan de manera aleatoria en el conjunto de validación, D_V, y en el de entrenamiento, D_T, para después entrenar una red B_T utilizando D_T, el algoritmo hill-climbing y el criterio LL. Para evaluar como se ajusta el modelo a los datos se utiliza LL, calculando $LL(B_T, D_T)$ y $LL(B_T, D_V)$ (2.29), como se muestra en la figura 4.3. Se puede observar que $LL(B_T, D_T)$ es consistente con la puntuación de la red utilizando el conjunto de datos completo, que se muestra en la figura 4.2a.

Por otra parte, a pesar de que la verosimilitud es creciente también para el conjunto de validación, la separación que existe respecto a los datos de entrenamiento es debido a un sobreajuste en el modelo. Esta diferencia es pequeña debido a que los datos son muy homogéneos (están discretizados en solo tres estados) y tienen una

fuerte dependencia espacial, por lo que el modelo es muy robusto. El sobreajuste comienza alrededor de los 1000 enlaces, y aumenta en torno a los 1500. Después, se mantiene constante a medida que crece la red, puesto que estos enlaces nuevos son simplemente redundantes, pero no empeoran la generalización del modelo, siguen explicando bien el test.

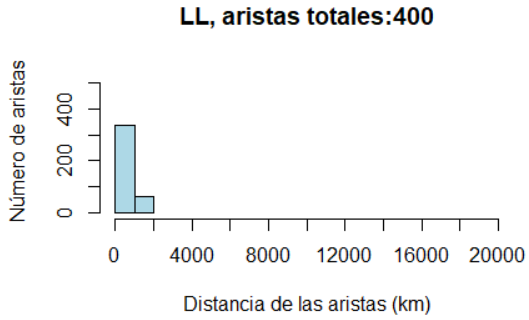
Para comprobar cómo crece la red anterior se realizan distintos histogramas para representar la distribución de distancias de sus enlaces. En la figura 4.4 se observa cómo el algoritmo comienza añadiendo enlaces cortos, los locales, para conseguir representar la consistencia espacial de los datos. Una vez ha añadido los enlaces cortos, comienza a ser más ventajoso incluir enlaces largos, que se corresponden con las teleconexiones (dinámica de la circulación que relaciona regiones distantes), los cuales comienzan alrededor de los 700 enlaces. El histograma 4.4d se corresponde con el momento en la figura 4.3 en el que comienza a aumentar la diferencia entre la verosimilitud del conjunto de entrenamiento y la del conjunto de validación. Se puede apreciar que, principalmente, se están añadiendo nuevos enlaces largos. Posteriormente, la red continúa añadiendo más enlaces de este tipo, al igual que otros de longitud media (entre dos y cuatro kilómetros), los cuales no empeoran el ajuste de los datos de validación, pero no son necesarios, ya que aumentan excesivamente la complejidad del modelo.

El objetivo de utilizar redes bayesianas es obtener un modelo que proporcione un balance entre su capacidad de generalización y su complejidad, que extraiga las características del sistema más relevantes, ya que en un sistema complejo como es el clima, tener demasiados parámetros conlleva que sea difícil de analizar. A pesar de que la calidad de la red sea creciente al aumentar el número de aristas, emplear el modelo obtenido con el criterio LL supone trabajar con 6213 parámetros. Por lo tanto, se puede concluir que la red entrenada con el criterio LL no es un modelo adecuado, lo cual queda reflejado en 4.5.

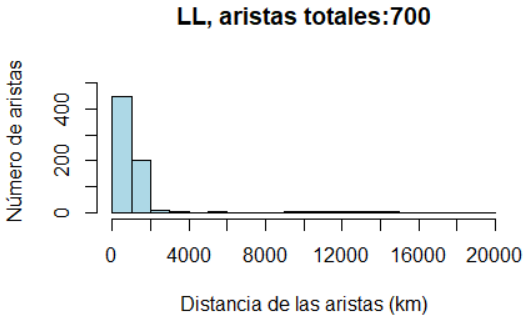
Es conveniente encontrar un modelo que sea informativo, en el que haya un balance entre la complejidad y el ajuste a los datos, por lo que se emplearán medidas de calidad que penalicen la estructura en función de los parámetros que necesite para definir la función de probabilidad asociada, es decir, BIC y AIC.

4.2.3. Modelos con AIC y BIC

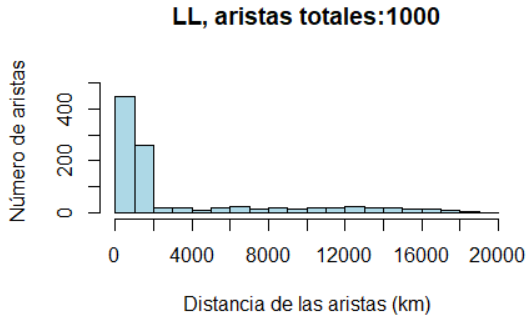
Como se puede observar en la figura 4.3, las redes bayesianas entrenadas con los criterios BIC y AIC, dejan de añadir enlaces antes de que se produzca un sobreajuste. A continuación, en la figura 4.6 se muestran dichas redes junto a histogramas



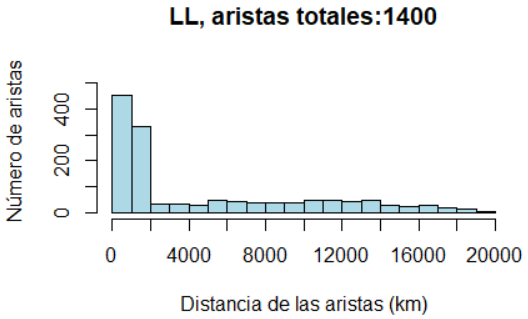
(a) Histograma de la red con 400 enlaces



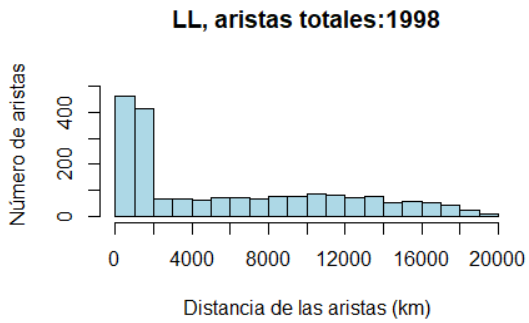
(b) Histograma de la red con 700 enlaces



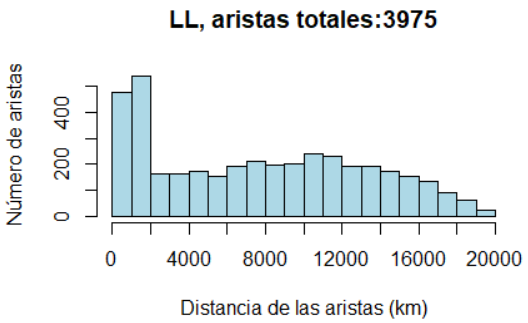
(c) Histograma de la red con 1000 enlaces



(d) Histograma de la red con 1400 enlaces



(e) Histograma de la red con 2000 enlaces



(f) Histograma de la red con 4000 enlaces

Figura 4.4: Histogramas de la distribución de los enlaces en una red bayesiana en función de su longitud, para distinto número total de aristas.

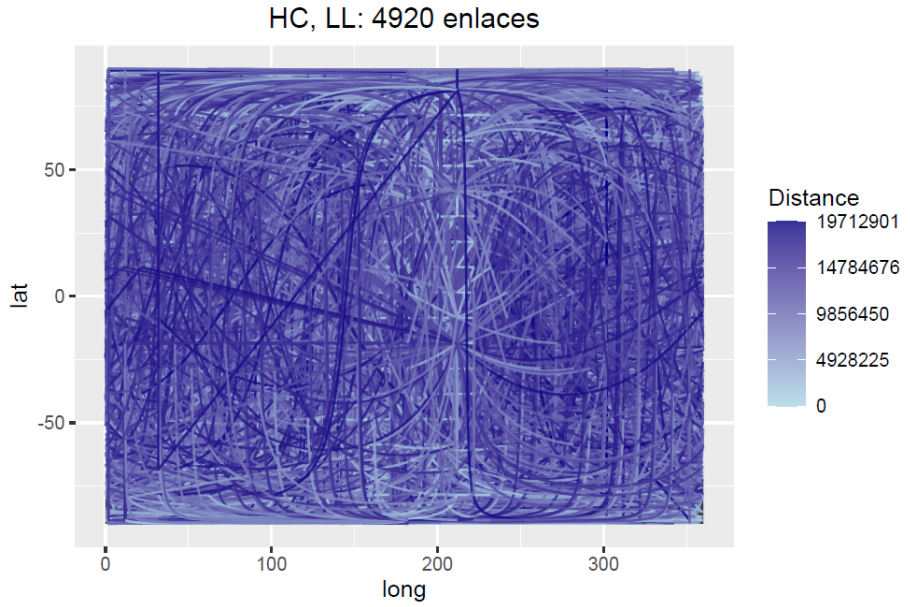
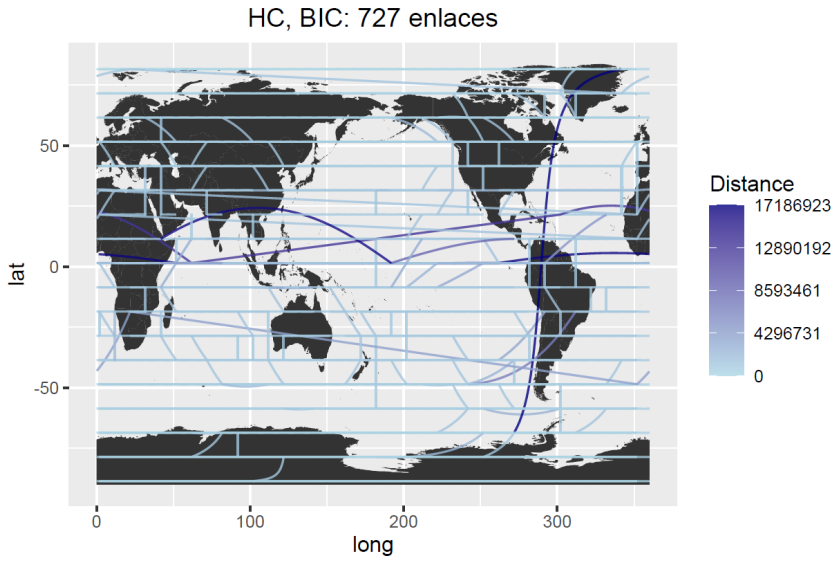


Figura 4.5: Red bayesiana entrenada con LL

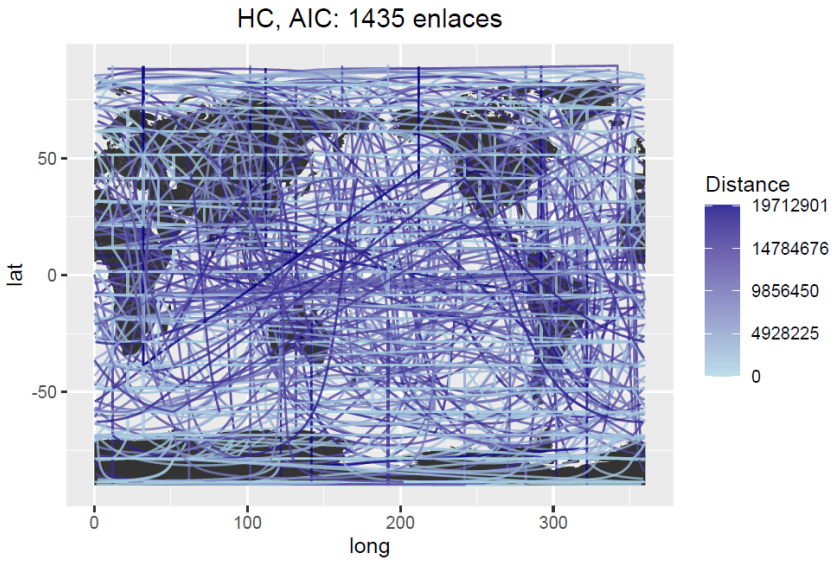
de sus enlaces en función de su longitud. Las redes están representadas sobre un mapamundi, en las que los enlaces representan relaciones de dependencia entre las temperaturas en los correspondientes puntos, y cuyo color varía en función de la distancia entre estos. Cuando se habla de enlaces cortos, se hace referencia a aquellos que conectan puntos a una distancia igual o menor a 2000km, siendo los largos el resto. En ambas redes, se encuentran más enlaces cortos horizontales que verticales, debido a que los puntos en un mismo paralelo tienen la misma latitud y por ello el sol incide con el mismo ángulo en estos, por lo que sus tiempos de insolación y la intensidad de la radiación son los mismos.

En la figura 4.6 está reflejado que BIC permite menos enlaces en la red, casi la mitad que AIC, debido a que penaliza más el número de parámetros (la penalización de AIC es la $dim(B)$, y la de BIC es $\frac{1}{2}dim(B)\log N$). Además, emplear BIC conlleva que el aprendizaje de la red finalice cuando se están empezando a añadir los primeros enlaces largos, mientras que AIC continúa añadiendo tanto enlaces cortos, como largos. No obstante, dado que BIC sí que añade enlaces largos en la zona de los trópicos, a pesar de que sigue pudiendo añadir cortos (como continúa haciendo AIC), queda reflejada la importancia de las teleconexiones para explicar el clima.

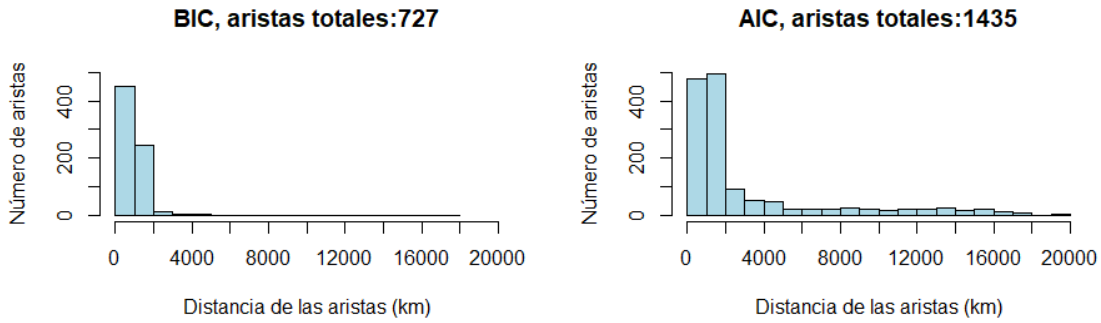
Por otra parte, la distribución de los enlaces de las redes en función de su distancia es similar a la que sigue la red entrenada con LL (figura 4.4), pero no idéntica, puesto que el orden en el que se añaden los enlaces no es el mismo con los distintos criterios. Dado que AIC y BIC penalizan el número de parámetros necesarios, es previsible



(a) Red bayesiana con BIC



(b) Red bayesiana con AIC



(c) Histograma de la red obtenida con el criterio BIC

(d) Histograma de la red obtenida con el criterio AIC

Figura 4.6: Modelos finales con los criterios BIC y AIC, junto a los histogramas de sus enlaces en función de su longitud.

que añadan antes enlaces que no supongan que un nodo tenga numerosos padres, mientras que LL simplemente añade los enlaces que más información proporcionen. Debido a que el score BIC restringe demasiado la complejidad de la red, y por ello añade menos enlaces, se utilizará el score AIC para las redes dinámicas, obteniendo así un mejor equilibrio entre la explicabilidad del modelo y su complejidad.

4.3. Análisis de causalidad

4.3.1. Causalidad de Granger

Aplicar la definición de Granger (3.1) para comprobar si las relaciones presentes en un conjunto de datos son causales es similar a aplicar tests de independencia en el aprendizaje basado en restricciones para redes bayesianas (ver Sec. 2.5.2). Sin embargo, en vez estudiar la independencia condicional de dos variables, condicionadas a un conjunto reducido de otras (típicamente limitado a tres o cuatro), se tendría que realizar el test condicionando a todas las variables del sistema, que en este caso son 648, para todas las series temporales anteriores a la actual. Este método no es viable estadísticamente, ya que al estar condicionado a tantas variables, con 360 instancias en los datos, la incertidumbre es muy alta, por lo que disminuye considerablemente el poder de detección.

De acuerdo con Runge et al. (2019), los métodos más utilizados para detectar relaciones causales entre series temporales en sistemas dinámicos complejos (como es el caso del clima) son modelos de autoregresión lineal implementados en el marco de la causalidad de Granger. Sin embargo, como ya se ha comentado, la alta dimensión de los datos conlleva un poder de detección muy bajo. Por otra parte, en Wismüller et al. (2021) se introduce un método de Granger no lineal y multivariado para series temporales a gran escala, pero aún aplicado a datos de poca dimensión, con un máximo de 34 variables, por lo que para este problema, con 648 variables, no es una opción adecuada.

Además, en Zou y Feng (2009), se realiza una comparación entre la capacidad de las redes bayesianas dinámicas y de la causalidad de Granger (mediante un modelo multivariado) para detectar relaciones causales en series temporales. El estudio concluye que, para problemas en los que el conjunto de datos es grande, el método de Granger proporciona mejores resultados, pero para aquellos en los que el conjunto de datos es menor, son mas adecuadas las redes bayesianas dinámicas.

Dado que para el problema que se trata en este trabajo el número de variables es

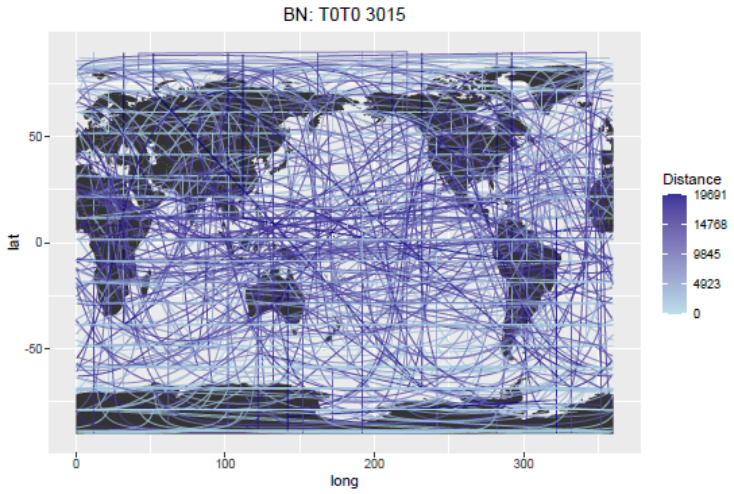
elevado, 648, y el conjunto de datos es pequeño, 360, se van a aplicar redes bayesianas dinámicas para detectar relaciones causales, puesto que es previsible que obtengan mejores resultados que la causalidad de Granger.

4.3.2. Redes bayesianas dinámicas

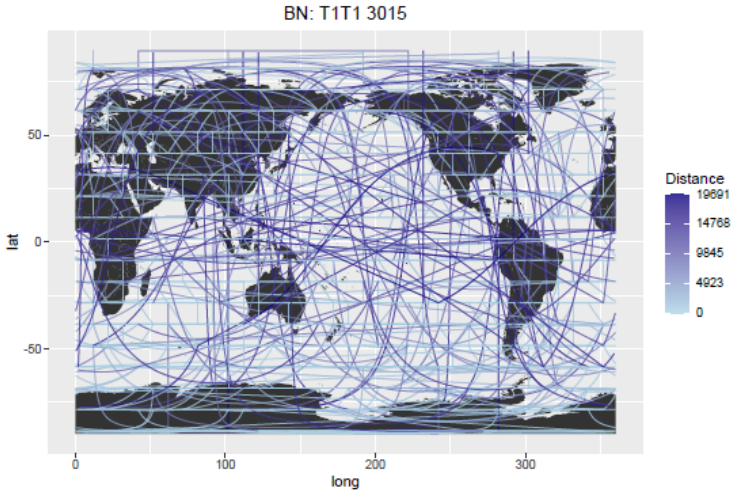
En base a los resultados obtenidos en la sección 4.2.3 sobre distintas configuracions de los métodos de aprendizaje de redes bayesianas, para el aprendizaje de redes bayesianas dinámicas se aplica el algoritmo hill-climbing y la medida de calidad basada en la medida de información de Akaike (AIC), 3.18. Se han entrenado diferentes redes con desfases temporales (DT) distintos (entre uno y seis meses). Por ejemplo, las figuras 4.7 y 4.8 muestran los resultados para desfases de uno y cinco meses, respectivamente.

En cada figura se muestran tres mapas cuyos títulos indican qué series temporales conectan los enlaces de la red representada y el número total de aristas de la red bayesiana dinámica completa. El mapa superior se corresponde con la red inicial, B_0, puesto que representa las relaciones intratemporales del tiempo T0. Los dos restantes, juntos, componen la red de transición $B_{\rightarrow}$. No obstante, se muestra separada en dos figuras para facilitar su comprensión. En la del medio están representados los enlaces intratemporales, estáticos, de la red de transición. La de abajo, en cambio, representa los enlaces intertemporales, dinámicos, que conectan nodos que se encuentran en series temporales consecutivas (puesto que se asume la suposición de Markov). Estos enlaces son aquellos que representan relaciones causalidad, y por ello están representados con flechas, para poder distinguir causa y efecto en la representación, que se corresponden con el pasado y el futuro respectivamente. En el estudio realizado por Zou y Feng (2009), se consideran como relaciones de causa-efecto aquellas existentes entre series temporales distintas que se verificaban para un 95 % de la población. Sin embargo, en este trabajo no se realizará dicha comprobación.

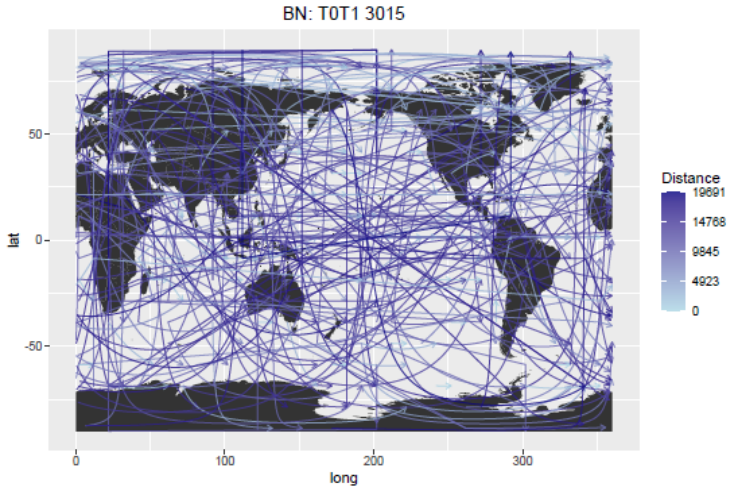
Además, junto a la representación de las redes, también se presentan los histogramas de la longitud de sus enlaces, estando separados de la misma manera que la estructura gráfica. Por una parte se muestra la distribución de los enlaces de B_0 y después la de la red de transición, otra vez, separada en aristas dinámicas y estáticas.



(a) Red inicial B_0

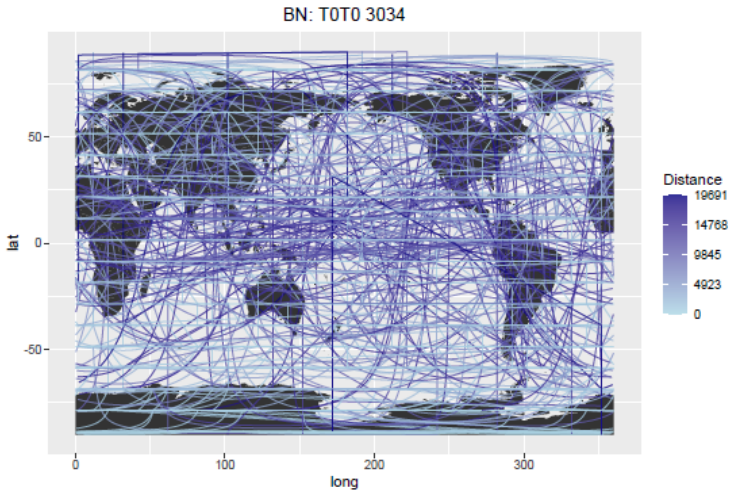


(b) Enlaces estáticos de la red de transición.

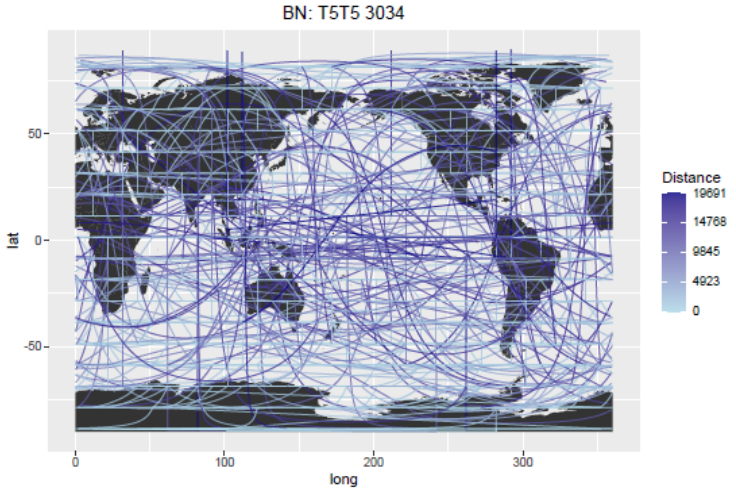


(c) Enlaces dinámicos de la red de transición.

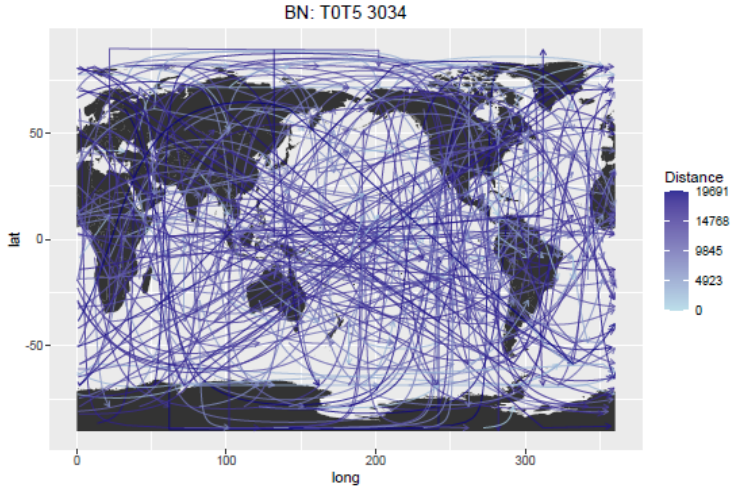
Figura 4.7: Red bayesiana Dinámica con AIC, con $TS = 1$



(a) Red inicial B_0



(b) Enlaces estáticos de la red de transición.



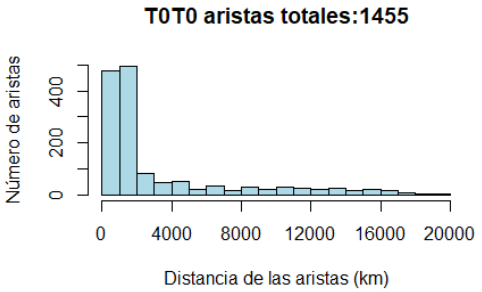
(c) Enlaces dinámicos de la red de transición.

Figura 4.8: Red bayesiana Dinámica con AIC, con $TS = 5$

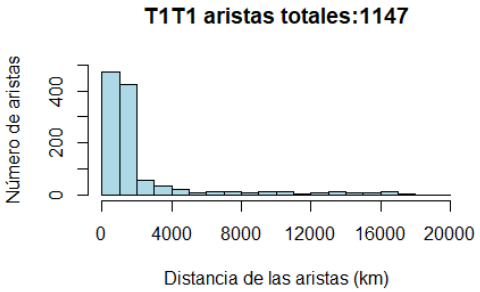
En la figura 4.7 se encuentra la estructura gráfica de la red bayesiana dinámica para la cual el desfase temporal es de un mes ($DT = 1$), es decir, la diferencia temporal entre los nodos de la red inicial y los de la red de transición es de un mes. Los histogramas de sus enlaces se muestran en la figura 4.9. Respecto a la distribución y número de los enlaces de la red inicial, se puede observar como el histograma 4.9a es prácticamente idéntico al histograma 4.6d de la red bayesiana entrenada con AIC. Es posible que las pequeñas diferencias que hay entre ambas se deban a que el conjunto de entrenamiento de B_0 difiere respecto al de la red bayesiana no dinámica. Por otra parte, también se muestra cómo, de los 413 enlaces intertemporales que tiene la red, aproximadamente 175 son locales, de 1000km o menos. Esto es coherente con que la diferencia entre las series temporales es de solo un mes y así, la persistencia del tiempo, la tendencia de las condiciones meteorológicas a mantenerse en tiempos sucesivos, y la fuerte dependencia espacial local de los datos siguen teniendo importancia. El resto de enlaces, de mayor longitud, representan las teleconexiones.

También se ha entrenado una red bayesiana dinámica con un desfase temporal de cinco meses ($DT = 5$). La estructura de la red bayesiana dinámica obtenida se presenta en la figura 4.8, y los histogramas de sus enlaces en función de la distancia en 4.10. Comparando esta red con la que se obtiene empleando $TS = 1$, se observa que una de las principales diferencias está en la distribución de los enlaces dinámicos. En el caso $TS = 1$, de los 413 enlaces dinámicos que tiene la red, alrededor de 175 son de locales. Sin embargo, si $TS = 5$, la diferencia temporal es de cinco meses, y en el histograma 4.10c se puede apreciar cómo apenas existen enlaces dinámicos locales, primando los que cubren distancias más largas. Esto se debe a que las dependencias locales debido a la persistencia ya no se mantienen, al ser cinco veces mayor la diferencia temporal, en cambio, las teleconexiones sí que continúan siendo importantes.

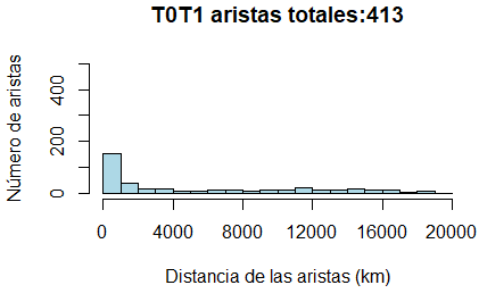
A continuación, la figura 4.11 muestra el distinto número de enlaces para cada red dinámica, dependiendo del desfase temporal empleado para entrenarla. Los enlaces se muestran agrupados según pertenecen a la red inicial B_0, a la parte estática de la red de transición o a su parte dinámica. En esta está reflejado cómo los enlaces dinámicos de $B_{\rightarrow}$ disminuyen a partir de un desfase de dos meses, para después estabilizarse, perdiendo los enlaces locales, puesto que deja de haber persistencia, y conservando únicamente las teleconexiones. Esto tiene un efecto en la parte estática de la red de transición, ya que aumentan los intratemporales si disminuyen los intertemporales. Respecto a la red inicial, la cantidad de enlaces varía ligeramente, debido a las diferencias en los conjuntos de datos mediante los cuales se realiza el entrenamiento.



(a) Histograma de las aristas de B_0

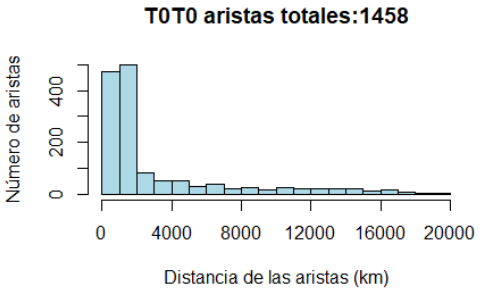


(b) Hist. de las aristas estáticas de $B_{\rightarrow}$

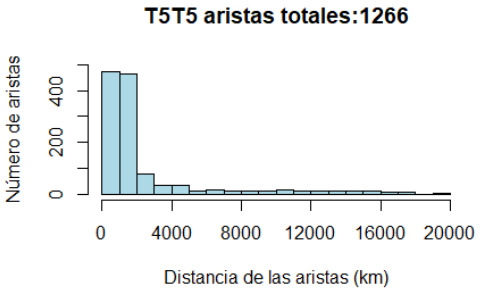


(c) Hist. de las aristas dinámicas de $B_{\rightarrow}$

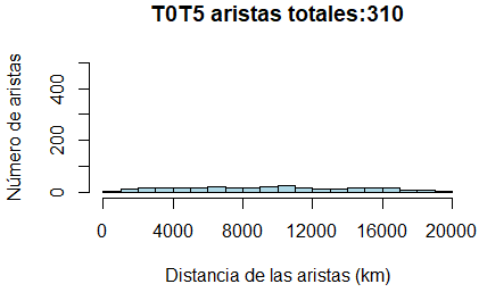
Figura 4.9: Histogramas de la longitud de los enlaces de la red representada en 4.7



(a) Histograma de las aristas de B_0



(b) Hist. de las aristas estáticas de $B_{\rightarrow}$



(c) Hist. de las aristas dinámicas de $B_{\rightarrow}$

Figura 4.10: Histogramas de la longitud de los enlaces de la red representada en 4.8

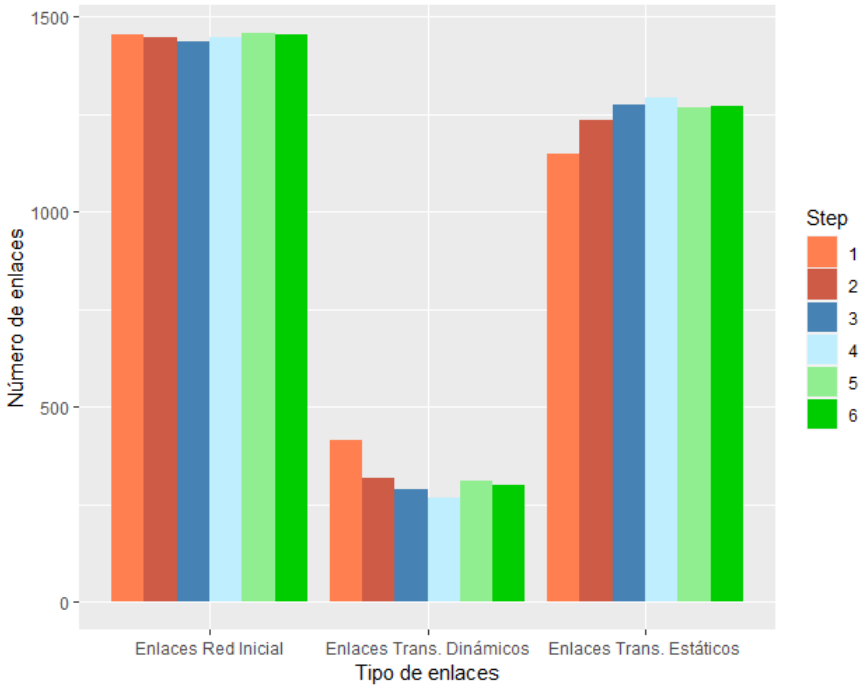
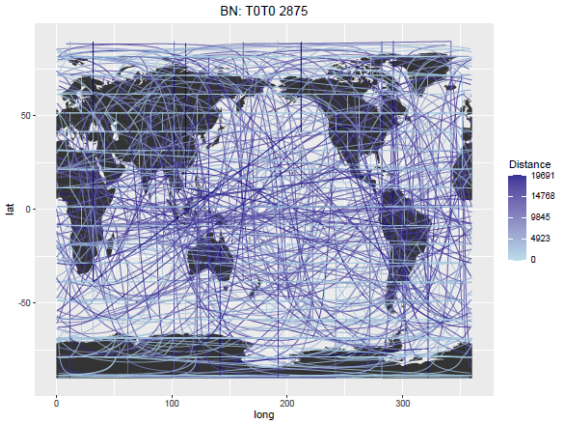


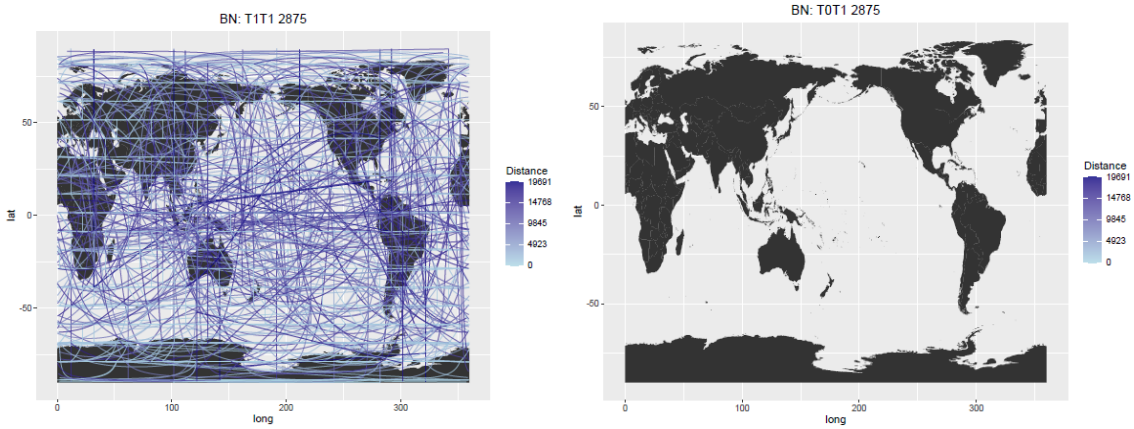
Figura 4.11: Gráfica de barras del número de enlaces de distintas redes dinámicas, según el tipo.

Adicionalmente, se ha probado otro método para realizar el aprendizaje de las redes bayesianas dinámicas, cuyos resultados se muestran en 4.12. Este consiste en realizar primero el aprendizaje de la red bayesiana inicial, B_0, y después asumir que los enlaces estáticos de esta son los mismos que aquellos en la red de transición, $B_{\rightarrow}$. Entonces, las relaciones entre los nodos en una misma serie temporal no varían respecto al instante inicial. Por último, se añaden los enlaces dinámicos de la red de transición. Utilizando esta técnica se obtienen menos enlaces dinámicos, puesto que, al imponer la red estática de B_0 en $B_{\rightarrow}$, se consigue que las relaciones de los datos estén suficientemente explicadas, perdiendo importancia y capacidad de explicación los enlaces dinámicos. Concretamente, su aplicación con el criterio AIC produce tan solo un único enlace dinámico, que además es de un nodo consigo mismo y por lo tanto, no sale reflejado en la figura 4.12, pero sí en el histograma correspondiente 4.13b (aunque es difícil de apreciar, se trata de un pequeño punto sobre el cero del eje horizontal).

Respecto a la causalidad en las redes bayesianas dinámicas en la aplicación en clima, los enlaces intertemporales para desfases temporales de dos meses se corresponden con teleconexiones y estas son, físicamente, relaciones causales. Para confirmar teóricamente si esa causalidad se da, es necesario algún criterio o definición de causalidad. Como ya se ha mencionado, en Zou y Feng (2009) se consideran como

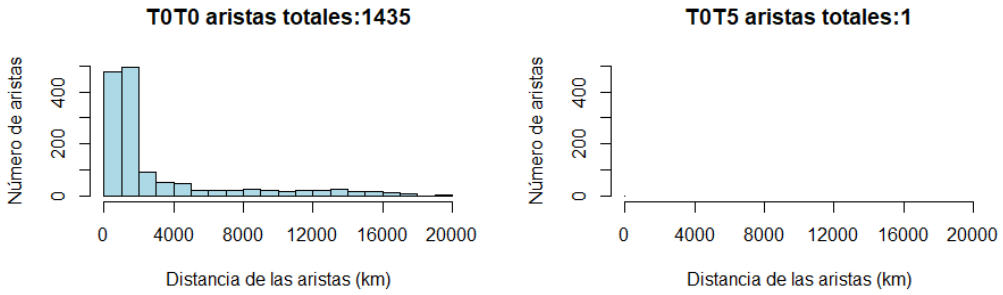


(a) Red inicial B_0.



(b) Enlaces estáticos de la red de transición. (c) Enlaces dinámicos de la red de transición.

Figura 4.12: Red bayesiana dinámica con AIC, $TS = 1$ e imponiendo que los enlaces estáticos de B_0 sean los de $B_{\rightarrow}$



(a) Histograma de las aristas de B_0 (b) Hist. de las aristas dinámicas de $B_{\rightarrow}$

Figura 4.13: Histogramas de la longitud de los enlaces de la red representada en 4.12

causales aquellos enlaces intertemporales que se verifican para un 95 % de los datos, pero existen otros métodos (ver Eichler y Didelez (2012)).

Capítulo 5

Conclusiones

En la actualidad, uno de los métodos más extendidos para detectar relaciones causales en clima es la causalidad de Granger. Sin embargo, para conjuntos de datos de alta dimensión, como es el caso del problema estudiado a lo largo de este trabajo, se ha visto que tanto aplicar métodos basados en Granger, como utilizar la definición original para realizar tests de independencia no es apropiado. La gran cantidad de variables, unido a que el conjunto de datos disponible es pequeño, causa que estas técnicas tengan un muy reducido poder de detección.

Como alternativa a la causalidad de Granger, se ha estudiado una opción más práctica, las redes bayesianas dinámicas, una extensión de las redes bayesianas en la que se introduce la dimensión del tiempo. En estas, las relaciones causales quedan determinadas mediante relaciones de dependencia entre series temporales distintas, en las cuales el tiempo establece la causa y el efecto (pasado y futuro respectivamente).

Tras analizar y explorar la aplicación de redes probabilísticas, en concreto redes bayesianas, para la modelización de datos climáticos (en particular datos globales de temperatura para un período de 30 años), se concluye que el método que mejor equilibrio proporciona entre la complejidad del modelo y su explicabilidad es aplicar el algoritmo hill-climbing, basado en puntuación y búsqueda, y el criterio de información de Akaike, una medida de calidad basada en teoría de la información. Por lo tanto, esta metodología es la que se ha empleado en las redes bayesianas dinámicas.

Utilizar redes bayesianas dinámicas para detectar relaciones causales ha producido los resultados esperados, detectando teleconexiones que modulan el comportamiento del clima de regiones distantes para distintos desfases temporales y omitiendo los enlaces locales debidos a la persistencia, a partir de un desfase de dos meses. Es-

tas relaciones coinciden con el conocimiento físico que se tiene del sistema y se deben a procesos causales. No obstante, para analizar más en detalle dicha causalidad, son necesarios criterios adicionales.

Bibliografía

Castillo, E., Gutiérrez, J. M., y Hadi, A. S. (1996). *Sistemas expertos y modelos de redes probabilísticas* (1st ed.). Academia de Ingeniería.

Cooper, G. F., y Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*, 309–347.

Eichler, M., y Didelez, V. (2012, 06). Causal reasoning in graphical time series models. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*.

Friedman, N., Murphy, K., y Russell, S. (2013). *Learning the structure of dynamic probabilistic networks.*

Graafland, C. E. (2022). *Probabilistic network modeling in complex systems* (Tesis Doctoral no publicada). Escuela de Doctorado de la Universidad de Cantabria.

Granger, C. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, *2*. doi: 10.1016/0165-1889(80)90069-x

Gutiérrez, J. M., Cano, R., Cofiño, A. S., y Sordo, C. M. (2004). *Redes probabilísticas y neuronales en las ciencias atmosféricas.* Universidad de Cantabria.

Heckerman, D., Geiger, D., y Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*, 197-–243.

Hitchcock, C. (2021). Probabilistic causation. En E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic/`.

Hitchcock, C., y Rédei, M. (2021). Reichenbach's Common Cause Principle. En E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2021/entries/physics-Rpcc/`.

Kitson, N., Constantinou, A., Zhigao, G., Liu, Y., y Chobtham, K. (2023). A survey of bayesian network structure learning. *Artificial Intelligence Review*, 1-94. doi: 10.1007/s10462-022-10351-w

Kleinberg, S. (2013). *Causality, probability, and time*. Cambridge University Press. doi: 10.1017/CBO9781139207799

Koller, D., y Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Koski, T. J., y Noble, J. (2012, Aug). A review of bayesian networks and structure learning. *Mathematica Applicanda*, *40*(1), 51–103. Descargado de `http://wydawnictwa.ptm.org.pl/index.php/matematyka-stosowana/article/view/278` doi: 10.14708/ma.v40i1.278

Murphy, K. (2002, 01). Dynamic bayesian networks: Representation, inference and learning. *Ph.D*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., y Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996. Descargado de `https://www.science.org/doi/abs/10.1126/sciadv.aau4996` doi: 10.1126/sciadv.aau4996

Santos Burguete, C., Simarro Grande, J. P., y Fuertes Marrón, D. (2018). Física del caos. En C. Santos Burguete (Ed.), *Física del caos en la predicción meteorológica* (p. 49–65). Agencia Estatal de Meteorología. Descargado de `http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/Fisica_del_caos_en_la_predicc_meteo/05_Fisica_del_caos.pdf` doi: 10.31978/014-18-009-X.05

Shojaie, A., y Fox, E. B. (2022, Mar). Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, *9*(1), 289–319. doi: 10.1146/annurev-statistics-040120-010930

Wismüller, A., Dsouza, A. M., Vosoughi, M. A., y Abidin, A. (2021, Apr). Large-scale nonlinear granger causality for inferring directed dependence from short multivariate time-series data. *Scientific Reports*, *11*(1), 7817. Descargado de `https://www.nature.com/articles/s41598-021-87316-6` doi: 10.1038/s41598-021-87316-6

Zou, C., y Feng, J. (2009, Dec). Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC Bioinformatics*, *10*(1), 122. Descargado de https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-122 doi: 10.1186/1471-2105-10-122