

**DESVELANDO PATRONES GLOBALES DE  
SINCRONICIDAD DE INCENDIOS MEDIANTE  
REDES COMPLEJAS**

**UNVEILING GLOBAL WILDFIRE SINCHRONICITY PATTERNS USING  
COMPLEX NETWORKS**

**MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS**

Presentado por

**SERGIO GRACIA BOROBIA**

Dirigido por

**JOAQUÍN BEDIA JIMÉNEZ**

Y

**CATHARINA ELISABETH GRAAFLAND**



**JULIO 2022**

**FACULTAD DE CIENCIAS**



---

## Agradecimientos

Me gustaría aprovechar para agradecer a todas las personas que me han acompañado durante mi etapa universitaria, que concluye con este trabajo. En primer lugar a mis padres y a mi hermano por su constante apoyo; también a mis compañeros del Máster con quienes he compartido incontables horas desvelando los misterios que encierra el mundo de los datos. Por último, quiero agradecer especialmente a mis dos directores, Joaquín y Lisette, quienes con su incansable apoyo, cariño y dedicación han conseguido despertar en mi un interés especial por la investigación.

Este trabajo va dedicado a todos vosotros.



---

# Índice general

<b>Índice de Figuras</b>	<b>vi</b>
<b>Resumen</b>	<b>xii</b>
<b>Abstract</b>	<b>xii</b>
<b>1. Introducción</b>	<b>1</b>
1.1. El fuego como fenómeno global y complejo . . . . .	1
1.2. Estudio de la sincronicidad de los incendios . . . . .	2
<b>2. Obtención y procesamiento de datos</b>	<b>5</b>
2.1. Base de datos global de área quemada . . . . .	5
2.2. Acceso y descarga de los datos . . . . .	6
2.3. Post-proceso: preparación y visualización de datos . . . . .	8
<b>3. Métodos</b>	<b>11</b>
3.1. Redes de correlación . . . . .	11
3.1.1. Correlación en datos de área quemada . . . . .	11
3.1.2. Construcción del grafo . . . . .	12
3.1.3. Descriptores de las conexiones . . . . .	13
3.1.4. Medidas globales de conectividad en la red de correlación . . . . .	14
3.1.5. Medidas de centralidad de la red de correlación . . . . .	15
3.1.6. Comunidades . . . . .	18
3.2. Redes bayesianas . . . . .	19
3.2.1. Codificación de las (in)dependencias del grafo . . . . .	19
3.2.2. Codificación de las (in)dependencias en la función de probabilidad .	20

---

3.2.3. Aprendizaje de redes bayesianas . . . . .	21
3.2.4. Capacidad de explicar los datos: Log-likelihood . . . . .	23
3.2.5. Inferencia en redes bayesianas gaussianas . . . . .	24
<b>4. Resultados</b>	<b>25</b>
4.1. Redes de correlación . . . . .	25
4.1.1. Elección del umbral de correlación . . . . .	25
4.1.2. Medidas de centralidad . . . . .	30
4.1.3. Búsqueda de comunidades . . . . .	31
4.2. Redes Bayesianas . . . . .	35
4.2.1. Construcción de la red bayesiana . . . . .	35
4.2.2. Inferencia bayesiana . . . . .	38
<b>5. Conclusiones generales y trabajo futuro</b>	<b>41</b>
5.1. Conclusiones . . . . .	41
5.2. Trabajo futuro . . . . .	43
5.3. Reproducibilidad de los resultados . . . . .	43
<b>Bibliografía</b>	<b>45</b>
<b>Anexo A: Material suplementario</b>	<b>51</b>
Descriptores de las conexiones: valores según umbral . . . . .	51
Representación espacial de diferentes CNs . . . . .	55
Medidas de centralidad para diferentes CNs . . . . .	58

## Índice de Figuras



4.8.	Mapas de las comunidades obtenidas para $\tau_c = 0.6$ (ordenadas por tamaño) en varios niveles de corte del dendrograma. Los mapas de comunidades con un número en la parte superior izquierda se corresponden con los niveles de corte indicados en la Figura 4.7. Para evitar ruido en el mapa se filtran las comunidades con un único nodo, dándoles el color beige. . . . .	35
4.9.	Dendrograma resultado de aplicar el algoritmo de detección de comunidades basado en el betweenness de los enlaces de la CN para el umbral $\tau_c = 0.7$ . La línea horizontal indica el nivel de corte para el que se obtiene la comunidad destacada con el rectángulo. Los dos mapas de comunidades resultantes por encima y por debajo de este nivel de corte quedan reflejados en la Figura 4.10 (mapas izquierdo y derecho respectivamente). . . . .	36
4.10.	Mapas espaciales de las comunidades obtenidas para $\tau_c = 0.7$ (ordenadas por tamaño) en varios niveles de corte del dendrograma (ver Fig. 4.9). Para evitar ruido en el mapa se filtran las comunidades con un único nodo, dándoles el color beige. . . . .	36
4.11.	Estudio de la log-likelihood de un conjunto de redes bayesianas según su tamaño. . . . .	37
4.12.	Diferencia (en unidades de desviación estándar) entre la probabilidad condicional (tras la propagación de la evidencia) y marginal (estado inicial) $P(X_i \geq 1   X_e = 2) - P(X_i \geq 1)$ (en rojo) y $P(X_i \leq 1   X_e = 2) - P(X_i \leq 1)$ (en azul). El nodo donde se da evidencia está marcado en verde. . . . .	39
4.13.	Diferencias entre las probabilidades (en unidades de desviación estándar) condicional (propagación de la evidencia) y marginal (estado inicial) $P(X_i \geq 1   X_e = 2) - P(X_i \geq 1)$ dada la evidencia en Indonesia (izquierda) y Sudamérica (derecha). Los nodos donde se da evidencia están marcados en verde. . . . .	40
A1.	Red de correlación pesada para un umbral $\tau_c = 0.5$ . . . . .	55
A2.	Red de correlación pesada para un umbral $\tau_c = 0.7$ . . . . .	56
A3.	Red de correlación pesada para un umbral $\tau_c = 0.8$ . . . . .	57
A4.	Medidas de centralidad para una CN con umbral $\tau_c = 0.5$ . . . . .	58

- A5. Medidas de centralidad para una CN con umbral  $\tau_c = 0.7$  . . . . . 59

---

## Resumen

En vista del actual cambio global, una mejor comprensión del fenómeno de los incendios puede resultar clave para anticipar futuros impactos y minimizar, en la medida de lo posible, las consecuencias negativas sobre los sistemas naturales y la economía humana, orientando adecuadamente las políticas de gestión, preparación y mitigación a diferentes niveles del proceso de toma de decisiones.

En este trabajo se presenta un análisis a escala global de la sincronicidad anual del fenómeno de los incendios, a través de la construcción de dos tipos de redes complejas que reflejan las estructuras subyacentes en los datos desde enfoques distintos: redes de correlación y redes bayesianas. Estudiando sus propiedades a través de medidas de centralidad y detección de comunidades para la red compleja y de inferencia para la red bayesiana se logra una mejor comprensión de la interrelación entre la actividad de los incendios en diferentes regiones del planeta, destacando aquellos patrones de teleconexión más relevantes que pueden dar lugar a análisis posteriores más detallados sobre la causalidad de dichas relaciones.

**Palabras clave:** redes de correlación, redes bayesianas, minería de datos, área quemada, patrones de teleconexión, centralidad, inferencia.

## *Abstract*

In light of the ongoing global change, a better understanding of global wildfire activity is the key to anticipate future impacts and minimize, as much as possible, their negative consequences on natural ecosystems and human economy, in order to adequately guide management practices, preparedness and mitigation policies at the different stages of the decision-making process.

This master's thesis presents a global-scale analysis of the inter-annual synchronicity of wildfires using two types of probabilistic networks, able to capture the underlying structures in the data from two different approaches: correlation-based and bayesian networks. By studying their properties through centrality measures and community detection for the complex network, and inference from the Bayesian network, we seek to gain a better understanding of the interrelationships between fire activity in different regions of the planet, highlighting the most important teleconnection patterns, helping to develop and test plausible hypothesis about the underlying mechanisms supporting these relationships.

**Keywords:** correlation networks, bayesian networks, data mining, burned area, teleconnection patterns, betweeness, inference.

## Introducción

### *1.1. El fuego como fenómeno global y complejo*

El fuego es un proceso integral del sistema global desde que aparecen las primeras plantas terrestres en el registro fósil, y desde entonces juega un rol importante en la distribución de los ecosistemas, induciendo alteraciones en el ciclo del carbono y en el propio sistema climático (Bowman et al., 2009). La actividad del fuego está controlada por una serie de factores biofísicos interrelacionados, incluyendo factores climáticos (Bedia et al., 2015; Abatzoglou et al., 2018), los tipos de combustible y la conectividad del mismo (Pausas y Ribeiro, 2013), así como los usos del suelo y la influencia (directa e indirecta) de la actividad humana (Chuvieco y Justice, 2010; Bowman et al., 2011), los cuales afectan de uno u otro modo a la probabilidad de que un fuego se inicie, así como a su capacidad de propagación y virulencia (Pausas y Keeley, 2021).

Como resultado de esta compleja interacción de procesos, que operan a diferentes escalas espaciales y temporales, la distribución de la actividad global de los incendios no es ni mucho menos homogénea (Chuvieco et al., 2008), y mientras que en algunas zonas del planeta los fuegos son, año tras año, invariablemente frecuentes y extensos, en otras regiones es un fenómeno infrecuente o simplemente inviable (ver a modo de ejemplo la Fig. 2.2).

Los mecanismos relacionados con la actividad de los incendios han despertado el interés de la comunidad científica en las últimas décadas, debido a la importancia global del fenómeno tanto desde el punto de vista de los ciclos naturales como de su impacto directo en las actividades humanas (Bowman et al., 2017), la calidad del aire y la atmósfera,

así como su enorme potencial para alterar los paisajes y ocasionar pérdidas económicas (Pausas y Keeley, 2021). La mayor parte de los enfoques del problema se basan en estudios de carácter correlativo entre una serie de variables predictoras y la actividad de incendios (el predictando en este caso puede ser el área quemada, el número de incendios u otras variables descriptoras del fenómeno), siendo la fuente primaria de información en este caso las bases de datos satelitales, que es la única capaz de proveer datos a una escala global de manera uniforme. Pese a la indudable relevancia y utilidad de este enfoque, existen limitaciones fundamentales que limitan nuestra capacidad de comprensión global del fenómeno por esta vía. La principal se debe a la limitada extensión temporal de las series satelitales, restringidas a la parte más reciente del desarrollo aeroespacial (ver p. ej. Lentile et al., 2006, para una visión general), lo que deriva en una escasa robustez de los análisis. Además, este tipo de análisis suelen ser de carácter univariado, lo que da lugar a una simplificación del problema al considerar uno sólo de los elementos como predictando y el resto como predictores, cuando la realidad es mucho más compleja (Archibald et al., 2013). Por otro lado, los modelos mecanísticos de simulación son capaces de representar parte de la complejidad del sistema de manera dinámica (ver p. ej. Hantson et al., 2020), aunque su complejidad impide su aplicación general y no quedan exentos de incertidumbres relacionadas con sesgos, parametrizaciones y otras fuentes de error inherentes a este tipo de modelos.

## 1.2. *Estudio de la sincronicidad de los incendios*

De lo anteriormente expuesto puede contemplarse el fenómeno del fuego como un sistema complejo en el que se encuentran implicados múltiples factores operando de forma simultánea a diversas escalas espacio-temporales. Desde un punto de vista espacial, los sistemas complejos a menudo presentan correlaciones de largo alcance, de modo que las variables observadas muestran dependencia estadística a muy largas distancias. Estas teleconexiones pueden resultar muy relevantes a la hora de comprender la dinámica de sistemas complejos, a modo de “canales” de información a través del sistema (Graafland et al., 2020). En este estudio, hablaremos de *sincronicidad*, como una forma particular de teleconexión en la que se analizará la asociación de la variable área quemada considerando

el mismo instante de tiempo (en este caso año a año), sin introducir de momento –para no extender excesivamente el estudio–, retardos (o *lags*) temporales en el análisis, que pudieran revelar otros patrones de teleconexión asíncronos (ver, p. ej. Valle, 2021).

Para el estudio de la sincronicidad de la actividad de los incendios a escala global, en este trabajo se emplearán dos enfoques basados en redes complejas, alternativos aunque relacionadas como veremos: las redes de correlación (CNs, del inglés *correlation networks*) y las redes bayesianas (BNs, *bayesian networks*). El enfoque más habitual para obtener patrones de teleconexión en sistemas complejos se basa en el desarrollo de CNs (Agarwal et al., 2019). La principal diferencia entre CN y BN es que, mientras que las primeras se construyen exclusivamente considerando asociaciones por pares (a través de correlaciones), las segundas utilizan métodos de aprendizaje más sofisticados para modelar también dependencias condicionales, es decir, dependencias entre dos nodos de la red, dado el estado de un tercer nodo (a su vez dependiente de otros), lo que lleva el análisis de la sincronicidad a un nivel más avanzado al considerar no sólo la información que fluye directamente entre dos nodos, sino también las relaciones de dependencia condicionada por el estado de otros nodos (Ebert-Uphoff y Deng, 2012; Graafland et al., 2020). Así, la principal ventaja de las CNs sobre las BNs proviene de su sencillez de construcción e interpretación directa, si bien su principal inconveniente reside en la relativa arbitrariedad en la elección del umbral de correlación que determina la existencia de un enlace, lo que tiene una repercusión en las relaciones establecidas dentro de la red y la interpretación de los resultados (Tsonis y Roebber, 2004). Por otro lado, las redes bayesianas son capaces de modelar de forma eficaz relaciones de dependencia condicional, si bien también son sensibles a la elección del algoritmo de aprendizaje de la estructura del grafo (Scutari et al., 2019), y a la robustez del ajuste de la probabilidad conjunta, particularmente cuando las series disponibles son cortas, como en este caso. En ambos casos, la información proporcionada por las redes complejas, si bien establece un nexo de tipo probabilístico y no causal, puede ser una fuente muy útil de información que permita plantear hipótesis sobre los mecanismos causa-efecto subyacentes (Cano et al., 2004; Ebert-Uphoff y Deng, 2012).

Por lo tanto, en este TFM se abordará el desarrollo metodológico de las CNs a lo largo de la Sección 3.1, y el de las BNs en la Sección 3.2. Posteriormente, se presentarán los resultados en el Capítulo 4 junto con una discusión de los mismos y se elaborarán una

serie de conclusiones finales en el Capítulo 5, que permitirá establecer aspectos comunes a ambos métodos así como el posible valor añadido de los modelos bayesianos sobre las redes de correlación. Por último, se concluirá con un análisis de las implicaciones de este trabajo en el estudio de la sincronicidad global de los incendios y su potencialidad como herramienta de estudio en este campo particular.

## Obtención y procesamiento de datos

### 2.1. Base de datos global de área quemada

Los principales datos utilizados para el desarrollo de este trabajo provienen de la Base de Datos satelitales *Fire burned area from 2001 to present derived from satellite observations* (FBA database, DOI: 10.24381/cds.f333cf85)<sup>1</sup>, accesible de manera abierta a través de la infraestructura de acceso a Datos “Copernicus Climate Data Store” (CDS<sup>2</sup>, ver Sec. 2.2).

La base de datos FBA abarca la totalidad del globo mediante una malla regular con una resolución espacial de  $0.25^\circ$  (unos 25 Km.), proporcionando para cada uno de los puntos terrestres (celdas) de la malla, una serie mensual continua de diferentes variables relacionadas con la actividad de los incendios, entre las que se encuentra el área quemada total estimada. Esta base de datos se extiende a lo largo del período de vida de los sensores MODIS (más antiguo) y OLCI (más reciente), abarcando desde Enero de 2001 hasta Octubre de 2020, sin que se hayan producido nuevas actualizaciones hasta la fecha de descarga de los datos para este TFM (Mayo 2022). Lamentablemente, al haberse realizado el estudio sobre series anuales, ha tenido que descartarse el año 2020 por encontrarse incompleto.

La variable “área quemada” (BA, *burned area*) se deriva a partir de un algoritmo especializado que procesa los datos recogidos por los sensores de resolución media Terra MODIS y Sentinel-3 OLCI, considerando cambios en la reflectancia, en combinación con

---

<sup>1</sup><https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-fire-burned-area?tab=overview>

<sup>2</sup><https://cds.climate.copernicus.eu/cdsapp#!/home>

información térmica también recogida por el sensor MODIS (Lizundia-Loiola et al., 2018). Los productos de área quemada también incluyen información relacionada con la cobertura terrestre que se ha quemado (tipos de vegetación y/o usos del suelo<sup>3</sup>), obtenida a partir de la base de datos de cobertura terrestre del *Copernicus Climate Change Service* (C3S), lo que garantiza la coherencia entre todos los conjuntos de datos. En particular, para este trabajo, además del área quemada por píxel, se ha considerado la fracción de área potencialmente quemable (BAF, *burnable area fraction*), con el fin de enmascarar en los análisis aquellas zonas del planeta que por su tipo cobertura superficial no permiten el desarrollo de incendios (glaciares, masas de aguas continentales grandes, desiertos, arenales, manglares, roquedos etc.), como se detalla en la Sección 2.3.

Durante julio de 2020 se identificó un error en algunos archivos en la versión v5.1cds de la base de datos FBA, afectando los archivos de área quemada y variables asociadas de enero 2018 y octubre, noviembre y diciembre 2019. Estos errores fueron solucionados posteriormente y una nueva versión del dataset (v5.1.1cds) fue creada para todo el período completo, que es la que se ha utilizado en este trabajo.

Una descripción más detallada del producto FBA y sus características, así como acceso a la descarga de datos, se proporciona en el enlace: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-fire-burned-area?tab=overview>.

## 2.2. Acceso y descarga de los datos

El CDS es una infraestructura de acceso a datos geocientíficos que forma parte del Programa de la Unión Europea para la Observación Terrestre (Copernicus), desarrollado a iniciativa de la Comisión Europea (<https://www.copernicus.eu/en>) e implementado por el ECMWF (Centro Europeo para la predicción climática de corto y medio alcance).

La arquitectura C3S está orientada a proporcionar acceso gratuito y sin restricciones a datos de calidad controlada y disponibles en un entorno operativo, como clave para la transformación hacia una sociedad más implicada en la acción climática. La base del programa Copernicus son los datos satelitales, como los utilizados en este trabajo, que representan la columna vertebral de todos los servicios. Teniendo en cuenta el carácter *big*

<sup>3</sup><https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=overview>

*data* de las bases de datos disponibles (por ejemplo, la última versión del reanálisis del ECMWF ronda los 7 PB), C3S definió el núcleo del Servicio en torno al CDS, un repositorio distribuído geográficamente mediante el que se puede no sólo acceder y descargar datos brutos, sino tambien manejar y postprocesar a través de un entorno de computación en la nube (Buontempo et al., 2020).

Para el desarrollo de este trabajo, primero se han descargado los datos brutos de área quemada, y después se han post-procesado localmente para adecuarlos a los análisis realizados. Así, la petición de datos al CDS se ha realizado de manera manual a través de la interfaz de usuario del CDS (requiere registro previo por parte del usuario), que inicia un *job* de descarga en la nube (Fig. 2.1). Estas peticiones se han realizado de manera anual (12 meses cada petición), considerando una extensión geográfica global, para no rebasar el límite de tamaño permitido en cada petición. Una vez preparados los datos en el entorno cloud del usuario, se ha procedido a su descarga de manera automatizada a través del comando `wget` en entorno Linux, apuntando a cada una de las URIs única de los ficheros generados, en formato NetCDF (<https://www.unidata.ucar.edu/software/netcdf/>).

(a)		(b)				
Product	Submission date	End date	Duration	Size	Status	
Fire burned area from 2001 to present derived from satellite observations	2022-07-10 08:59:13		0:01:16		In progress	
<b>Open request form</b> Request ID: 64959277-6c15-4c5f-93fb-787ce3797f21						
Origin:	C3S (Copernicus Climate Change Service)					
Sensor:	OLCI (Ocean and Land Colour Instrument)					
Variable:	Grid variables					
Version:	1.1					
Year:	2019					
Month:	January, February, March, April, May, June, July, August, September, October					
Nominal day:	01					
Format:	Zip file (.zip)					

```
import cdstoolbox as ct

@ct.application(title='Download data')
@ct.output.download()
def download_application():
    data = ct.catalogue.retrieve(
        'satellite-fire-burned-area',
        {
            'origin': 'c3s',
            'sensor': 'olci',
            'variable': 'grid_variables',
            'version': '1_1',
            'year': '2019',
            'month': [
                '01', '02', '03',
                '04', '05', '06',
                '07', '08', '09',
                '10',
            ],
            'nominal_day': '01',
        }
    )
    return data
```

**Figura 2.1:** Captura de pantalla de ejemplo de una petición de datos al CDS, correspondientes al área quemada y variables adicionales para todo el globo del año 2019 (sensor OLCI del C3S). (a) Código de petición autogenerado por la aplicación para realizar de forma automática la petición a través de la CDS Toolbox (entorno python). (b) Cuadro de control de peticiones del usuario, que muestra las características del job y el estado del mismo.

### 2.3. Post-proceso: preparación y visualización de datos

Una vez descargada la colección de ficheros NetCDF se ha procedido a su lectura empleando para ello las herramientas del entorno *climate4R*<sup>4</sup> (Iturbide et al., 2019). Se trata de un entorno basado en el entorno abierto R (R Core Team, 2019) compuesto por un conjunto de paquetes y librerías interconectadas para el acceso transparente, el post-proceso y la visualización de datos climáticos. En particular, la lectura se ha llevado a cabo mediante las herramientas del paquete *loadR* (ver p. ej. Cofiño et al., 2018), que proporciona una interfaz amigable para el acceso a NetCDF desde R a través de la API NetCDF-Java<sup>5</sup>, facilitando de este modo la lectura de subconjuntos lógicos a través de coordenadas geográficas y temporales en lugar de índices de posición.

Dados los requerimientos computacionales para la construcción y análisis de los modelos basados en grafos a desarrollar (redes de correlación y redes bayesianas), se ha optado por trabajar con una resolución espacial más grosera que la nativa. Para ello, se aplicó una técnica de interpolación de carácter conservativo para degradar la resolución inicial de la malla regular de 0.25 a 5 grados de resolución, que se ha considerado un buen compromiso entre resolución espacial y coste computacional. El algoritmo utilizado se encuentra implementado en la función *upscaleGrid* del paquete *transformer*<sup>6</sup> (Iturbide et al., 2019).

Todos los análisis se han llevado a cabo considerando las series anuales de anomalías estandarizadas, para lo cual se han calculado las series anuales de área quemada total a partir de los datos originales mensuales y se ha procedido al cálculo de las anomalías con la función *scaleGrid* del paquete *transformer*, considerando los argumentos apropiados (ver, p.ej. Bedia et al., 2020, para diferentes casos de aplicación). Así, dada la serie temporal de área quemada  $\overrightarrow{BA}_i = (BA_{i,2001}, \dots, BA_{i,2019})$  de una celda  $i$ , la anomalía estandarizada de área quemada se define como  $\overrightarrow{Xs_i} = \overrightarrow{(BA_i - \mu_i^{BA})}/\sigma_i^{BA}$  donde  $\mu_i^{BA}$  es la media temporal de serie anual de área quemada de la celda  $i$  y  $\sigma_i^{BA}$  es su desviación típica. Se muestra un ejemplo de los datos obtenidos en la Figura 2.2.

Una vez los datos se han escalado de la manera deseada se disponen en una matriz bidimensional  $M_{t \times n}$  donde  $t = 19$  es la dimensión temporal en años y  $n = 36 \times 72 = 2592$

---

<sup>4</sup><https://github.com/SantanderMetGroup/climate4R>

<sup>5</sup><https://www.unidata.ucar.edu/software/netcdf-java/>

<sup>6</sup><https://doi.org/10.5281/zenodo.598233>

celdas es la dimensión espacial; cada fila corresponde a un año concreto y cada columna a una celda de la malla geoespacial.

El hecho de trabajar con datos de área quemada conlleva una consecuencia importante: la matriz de datos  $M_{t \times n}$  será dispersa, es decir, tendrá muchas componentes nulas. Esto es así porque hay una gran cantidad de zonas del planeta donde no pueden ocurrir incendios, tal y como se detalla en la sección 2.1. No es necesario incluir estas zonas en el cálculo de las CNs y BNs, pues no tendrían ninguna conexión con el resto de la red y únicamente actuarían como una fuente de ruido, y también se aumentaría el coste computacional innecesariamente. Se decidió también definir un criterio de exclusión en base a los datos de BAF (fracción de área combustible) de manera que la celda  $i$  no se considera si su valor es menor del 10 %. Así:

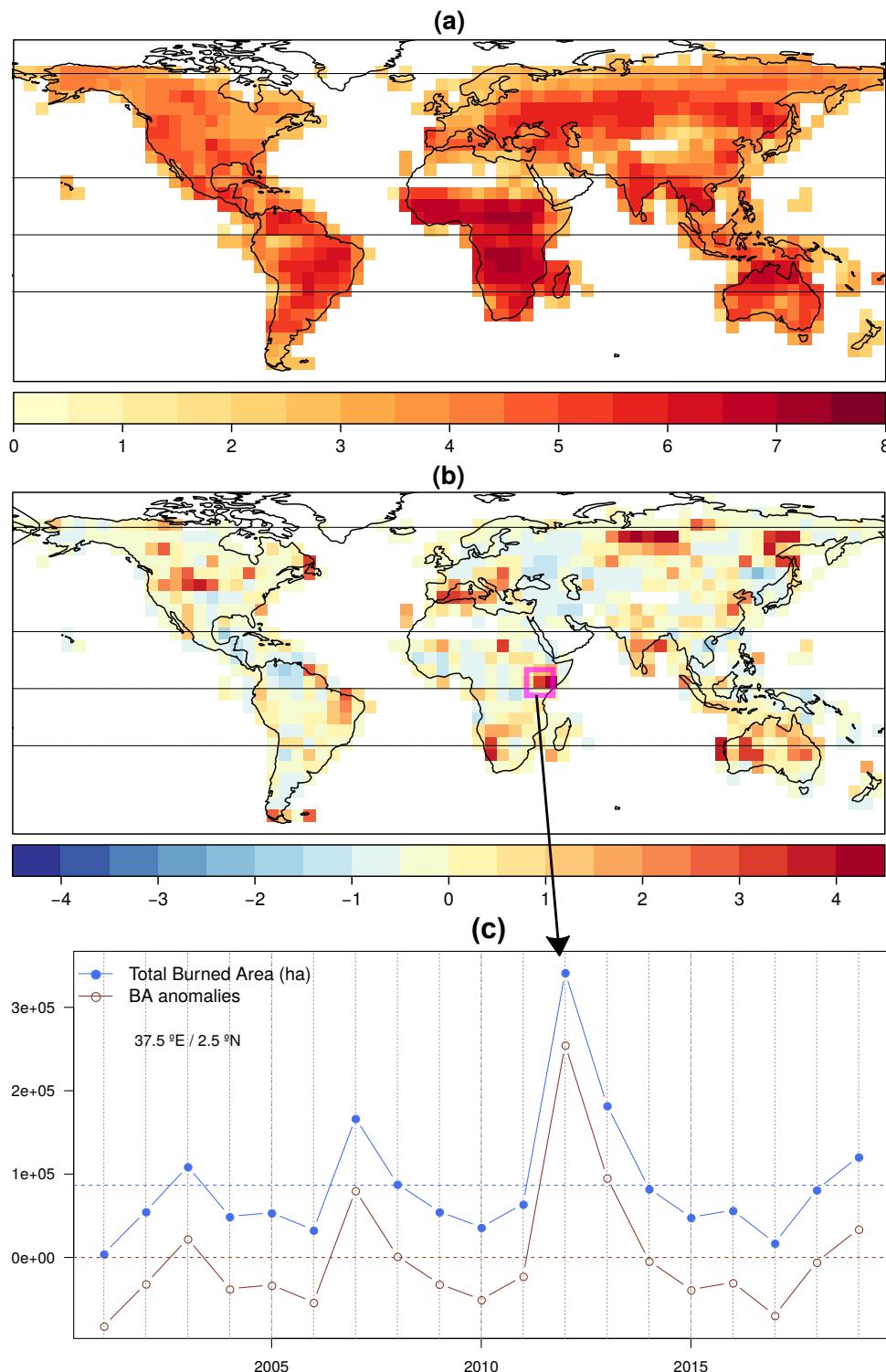
$$i \notin M_{t \times n} \quad \text{cuando} \quad \text{BAF}_i < 0.1 \quad (2.1)$$

Además, también existe la posibilidad de que en zonas del planeta con suficiente fracción de área combustible no haya ocurrido ningún incendio en el periodo de tiempo disponible, es decir, que su área quemada sea nula a lo largo de los casi 20 años de registro. Por tanto, se define un segundo criterio de exclusión, complementario a (2.1):

$$i \notin M_{t \times n} \quad \text{cuando} \quad \text{BA}_i = 0 \quad (2.2)$$

Tras aplicar los criterios de exclusión se obtiene una submatriz  $M_{t \times m}$ , con  $t = 19$  años y  $m = 645$  celdas, a partir de la cual se realizarán los análisis presentados en los siguientes capítulos.

Por último, es necesario indicar que la visualización de datos y generación de figuras se ha realizado principalmente con las funciones del paquete de climate4R *visualizeR* (Frías et al., 2018), salvo los mapas de enlaces de las CNs para los que se ha recurrido a utilidades del paquete *ggplot2* (Wickham, 2016). Además, la construcción y análisis de grafos se ha llevado a cabo contando con la librería *igraph*, a través de su interfaz en R (Csardi y Nepusz, 2006).



**Figura 2.2:** (a) Área quemada anual media (en hectáreas, log10-transformada) para el periodo 2001-2019, de acuerdo con la base de datos utilizada. Los datos han sido interpolados de manera conservativa desde la malla original ( $0.25^{\circ}$ ) a una malla más gruesa empleada para los análisis ( $5^{\circ}$ ). Los píxeles en los que el área quemada media es nula han sido enmascarados (en blanco) y excluidos de los análisis. (b) Mapa de anomalías estandarizadas del año 2012, calculadas sobre la serie interanual de área quemada 2001-2019. (c) Ejemplo de serie temporal interanual para uno de los píxeles del conjunto de datos (original y anomalía).

---

# 3

---

## Métodos

### 3.1. Redes de correlación

#### 3.1.1. Correlación en datos de área quemada

Partiendo de los datos explicados en el capítulo 2 y tras excluir aquellas celdas que no aportan información se procede a realizar la construcción del grafo. Dado que los datos de cada columna de la matriz  $M_{t \times m}$  pueden interpretarse como una serie temporal de área quemada en una celda  $\overrightarrow{BA_i}$ , es posible calcular la correlación entre las series temporales de diferentes celdas de la malla geoespacial. Para ello, se utiliza la correlación de Spearman, definida como

$$\rho_s = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \quad (3.1)$$

donde  $R(X)$ ,  $R(Y)$  denotan los rankings de las variables aleatorias  $X$  e  $Y$  respectivamente, y la covarianza se define como

$$\text{cov}(R(X), R(Y)) = \mathbb{E} [(R(X) - \mu_{R(X)})(R(Y) - \mu_{R(Y)})] \quad (3.2)$$

Así,  $\rho_s$  será una matriz simétrica de dimensiones  $(m \times m)$  en la que cada componente  $(i, j)$  es la correlación existente entre las series temporales de área quemada de las celdas  $i$ -ésima y  $j$ -ésima. Será esta matriz la que, tras ser sometida al tratamiento adecuado, permitirá la creación de los grafos de las CNs que se proponen estudiar en este trabajo.

### 3.1.2. Construcción del grafo

La construcción de la red de correlación es equivalente a la construcción de un grafo no dirigido compuesto de nodos y enlaces, el cuál se caracteriza a través de su matriz de adyacencia  $A$ , una matriz simétrica en la que se indica si los nodos del grafo se encuentran conectados entre sí, o no. Para el caso sometido a estudio los nodos se corresponden con las celdas de la malla geoespacial, de manera que la matriz de adyacencia indicará si dos puntos del planeta están conectados o no. Notar que además la diagonal de la matriz de adyacencia es nula por construcción. Esto se debe a que las celdas no pueden estar conectadas consigo mismas, ya que en este contexto los autoenlaces carecen de sentido.

$$\text{diag}(A) = \vec{0} \quad (3.3)$$

Así mismo, el grafo no dirigido puede ser pesado o no pesado en función de si a cada enlace se le asigna un valor (peso) en función de la importancia de las conexiones entre los nodos, medida a través de la correlación  $w_c$  o no, y quedan definidos a través de dos matrices de adyacencia distintas,  $(A_{ij})^w$  y  $(A_{ij})^u$  respectivamente.

Dado que la matriz de correlación obtenida en (3.1) es una matriz simétrica, es posible construir de manera inmediata la matriz de adyacencia, con dimensiones  $(m \times m)$ . Sin embargo, esto plantea un problema: por muy pequeña que sea la correlación entre dos nodos se va a crear un enlace entre ellos, por lo que prácticamente todos los nodos estarán interconectados (excepto aquellos pares cuya correlación sea exactamente nula), haciendo imposible estudiar las propiedades de la CN. La solución es sencilla y consiste en definir un umbral de correlación  $\tau_c$  de manera que aquellos nodos que tengan una correlación menor al umbral establecido se consideran inconexos:

$$(A_{ij})^w = \begin{cases} 0 & \text{si } \rho_{ij} \leq \tau_c \\ w_c & \text{si } \rho_{ij} > \tau_c \end{cases} \quad (3.3) \quad (A_{ij})^u = \begin{cases} 0 & \text{si } \rho_{ij} \leq \tau_c \\ 1 & \text{si } \rho_{ij} > \tau_c \end{cases} \quad (3.4)$$

Así, las CNs pesada y no pesada quedan completamente determinadas por las matrices de adyacencia (3.3) y (3.4) respectivamente.

### 3.1.3. Descriptores de las conexiones

Las conexiones en una red de correlación (en este caso con  $m = 645$  nodos) contienen una información básica que puede estudiarse a través de medidas simples que denominamos descriptores de las conexiones, definidas en esta sección. Notar que, individualmente, estos descriptores no son capaces de explicar la complejidad de la red, aunque en su conjunto proporcionan información indicativa de complejidad.

#### Número de enlaces

Podemos calcular el número total de enlaces  $n$  en la CN como:

$$n = \frac{1}{2} \sum_{i,j}^m A_{ij} \quad (3.5)$$

Además, podemos diferenciar estos enlaces según si la correlación entre los dos nodos que conecta es positiva o negativa, de manera que:

$$n^+ = \frac{1}{2} \sum_{i,j}^m A_{ij}^+ \quad (3.6) \qquad n^- = \frac{1}{2} \sum_{i,j}^m A_{ij}^- \quad (3.7)$$

donde  $A^+$  y  $A^-$  son las matrices de adyacencia de los subgrafos que consideran únicamente los enlaces con correlaciones positivas y negativas respectivamente. Notar además que se cumplirá  $n = n^+ + n^-$ .

#### Distancia geográfica media de los enlaces

Dado que los datos utilizados conforman una malla geoespacial, cada nodo de la CN queda explícitamente geolocalizado. Por este motivo, es posible estudiar la distancia geográfica que separa dos nodos como una propiedad del enlace que los une, asignándoles un peso alternativo  $(w_d)_{ij} = \text{dist}(i, j)$ , donde  $\text{dist}(i, j)$  es la distancia sobre el arco terrestre<sup>1</sup> y se define la distancia geográfica media de los enlaces de la CN como:

$$\mu_d = \frac{1}{m} \sum_{i,j}^m (w_d)_{ij} \quad (3.8)$$

---

<sup>1</sup>  $\text{dist}(i, j)$  se ha calculado con la función `spDists` del paquete de R `sp` (Bivand et al., 2013), sobre el elipsoide WGS84, sobre el que se define la malla de los datos.

donde  $m$  es el número de nodos de la CN, y coincide con las dimensiones de la matriz de adyacencia  $A$ . Procediendo de manera análoga al caso anterior, se definen la distancia geográfica media de los enlaces con correlación positiva y negativa respectivamente como:

$$\mu_d^+ = \frac{1}{m} \sum_{i,j}^m (w_d^+)^{ij} \quad (3.9)$$

$$\mu_d^- = \frac{1}{m} \sum_{i,j}^m (w_d^-)^{ij} \quad (3.10)$$

#### *Máxima y mínima distancia geográfica de los enlaces*

Al igual que para la distancia media se definen la máxima y mínima distancia geométrica de los enlaces de la red como:

$$(w_d)_{max} = \max\{w_d\} \quad (3.11)$$

$$(w_d)_{min} = \min\{w_d\} \quad (3.12)$$

donde  $\{w_d\}$  es el conjunto de todos los pesos alternativos basados en distancia geométrica para cada uno de los enlaces de la CN. Además, se definen la máxima y mínima distancia geométrica de los enlaces con correlación positiva y negativa respectivamente como:

$$\begin{aligned} (w_d)_{max}^+ &= \max\{w_d^+\} & (w_d)_{min}^+ &= \min\{w_d^+\} \\ (w_d)_{max}^- &= \max\{w_d^-\} & (w_d)_{min}^- &= \min\{w_d^-\} \end{aligned} \quad (3.13)$$

#### *3.1.4. Medidas globales de conectividad en la red de correlación*

Una de las características más importantes de la CN es que su configuración pone de manifiesto las estructuras subyacentes presentes en los datos, por lo que es necesario definir magnitudes que aporten información sobre la estructura global de la CN y el grado de conectividad presente en la misma. Para este trabajo, nos centraremos en dos: el coeficiente de clustering global y el diámetro de la red.

#### *Coeficiente de clustering global*

El coeficiente de clustering global  $C$  es una medida de la proporción en la que los nodos se juntan entre ellos formando agrupaciones a lo largo de la CN y se define como el

cociente entre el número de triángulos presentes en la red y el número total de triángulos que se formarían si todos los nodos estuviesen interconectados:

$$C = \frac{\sum_{i,j,k} A_{ij}A_{jk}A_{ik}}{\sum_i k_i(k_i - 1)} \quad (3.14)$$

donde  $k_i$  indica el grado (o número de enlaces) del nodo  $i$ -ésimo. Cuanto mayor sea el coeficiente de clustering, mayor es la tendencia de los nodos a conectarse con sus vecinos cercanos.

### *Diámetro*

El diámetro aporta información sobre cuál es el grado de desconexión de la red y se define como la longitud de la *geodésica* más larga del grafo, donde la geodésica  $g_{ij}$  es el camino más corto entre un par de nodos  $(i, j)$ . La expresión del diámetro se puede escribir como:

$$D = \max (\{g_{ij}\}_{i,j=1}^m) \quad (3.15)$$

donde  $\{g_{ij}\}_{i,j=1}^m$  es el conjunto de todas las geodésicas del grafo. Cuanto mayor sea el diámetro, mayor es el camino que es necesario recorrer para viajar entre dos nodos, y por lo tanto la CN estará menos conectada.

### *3.1.5. Medidas de centralidad de la red de correlación*

Las medidas globales de la CN no bastan para explicar su estructura, ya que son propiedades generales. Para poder estudiar cómo se comportan localmente los nodos de la red, identificando propiedades topológicas interesantes es necesario definir medidas adicionales “de centralidad”, que informan sobre la importancia de cada nodo en función de la información en la que se base dicha medida (para una revisión más profunda, consultar p. ej. Albert y Barabási, 2002; Boccaletti et al., 2006; Newman, 2010; Dijkstra et al., 2019). Se pone a disposición del lector una tabla resumiendo el tipo de información que aporta cada medida en el anexo A.

### *Grado*

El grado  $k_i$  del  $i$ -ésimo nodo se define como el número de enlaces que conectan dicho nodo con el resto de la CN. La expresión para su cálculo en una CN no pesada es

$$k_i = \sum_j^m A_{ij} \quad (3.16)$$

de manera que nodos con alta conectividad se consideran más importantes que aquellos con baja conectividad.

### *Betweenness*

La *betweenness*  $B_i$  del  $i$ -ésimo nodo se define como la proporción de geodésicas que lo atraviesan (definidas en el apartado 3.1.4). Es decir, de todas las geodésicas que existen entre un par de nodos  $(j, k)$ , se tienen en cuenta aquellas que atraviesan el nodo  $i$ . La expresión será (Newman, 2010):

$$B_i = \sum_{j,k \neq i}^m \frac{g_{jk}^i}{g_{jk}} \quad (3.17)$$

donde se escoge  $g_{jk}^i/g_{jk} = 0$  si  $g_{jk} = 0$ , es decir, cuando los nodos  $j$  y  $k$  no están conectados entre sí. Cuanto mayor sea la betweenness de un nodo, mayor será la cantidad de información que está pasando por el mismo; un ejemplo de nodo con alto betweenness sería aquel que conecta dos agrupaciones de nodos más entrelazados entre ellos que con el resto de la red, de manera que para transmitir la información de un agrupación a otra siempre es necesario atravesar dicho nodo.

### *Strength*

El concepto de *strength* es el mismo que el de grado solo que para una CN pesada. Es decir, el strength  $S_i$  de un nodo  $i$  es la suma de los pesos de cada uno de los links que conectan dicho nodo con el resto de la red. Por lo tanto, la expresión (3.16) se adapta a la matriz de adyacencia pesada en base a la correlación definida en (3.3), obteniendo la siguiente expresión:

$$(S_i)^c = \sum_j^m (A_{ij})^w = \sum_j^m (w_c)_{ij} \quad (3.18)$$

Paralelamente, y con los pesos alternativos basados en distancia geométrica  $w_d$  definidos en 3.1.3 se puede definir una medida de strength basado en distancia geométrica:

$$(S_i)^d = \sum_j^m (w_d)_{ij} \quad (3.19)$$

Cuanto más alto sea el strength de un nodo, mayor es su número de enlaces y/o mayor es la correlación (o distancia geométrica) de dichos enlaces.

#### *Area weighted connectivity*

El *area weighted connectivity* de un nodo  $i$  se entiende como la fracción de superficie terrestre a la que está conectado dicho nodo. Consiste en una corrección del grado  $k_i$  donde se tiene en cuenta que la superficie de una celda depende del coseno de su latitud  $\lambda_i$ , de manera que una celda que se encuentre en el ecuador ocupará mayor superficie que una que se encuentre cercana al polo, ya que ambas celdas tienen la misma resolución espacial ( $5^\circ$ ). La expresión para calcular esta magnitud es:

$$AWC_i = \frac{\sum_j^m A_{ij} \cos(\lambda_j)}{\sum_j^m \cos(\lambda_j)} \quad (3.20)$$

#### *Distancia geográfica media de conexión*

La distancia geográfica media  $MD_i$  de las conexiones de un nodo  $i$  se obtiene como el cociente del strength basado en distancia del nodo (3.19) y su grado (3.16), obteniendo la siguiente expresión:

$$MD_i = \frac{(S_i)^d}{k_i} = \frac{1}{k_i} \sum_j^{k_i} (w_d)_{ij} \quad (3.21)$$

Notar que esta expresión es similar a (3.8) pero teniendo en cuenta únicamente las conexiones pertenecientes al nodo  $i$ . Por último, indicar que se aporta un breve resumen de las diferentes medidas de centralidad utilizadas en este estudio en la Tabla A5 del Anexo.

### 3.1.6. Comunidades

En el contexto de las redes de correlación, las comunidades (o clústers) se definen como agrupaciones de nodos que se encuentran altamente conectados entre sí en comparación con el resto de la red. Existe una gran variedad de algoritmos para detectar comunidades, pero el más intuitivo es el basado en el betweenness de los links. En la expresión (3.17) se ha definido el betweenness de un nodo  $i$  como la proporción de geodésicas que pasan por  $i$ ; la definición para un link  $l$  es análoga ya que será la proporción de geodésicas atravesan dicho link, y viene dada por la expresión:

$$B_l = \sum_{j,k \neq l}^m \frac{g_{jk}^l}{g_{jk}} \quad (3.22)$$

donde nuevamente se escoge  $g_{jk}^l/g_{jk} = 0$  cuando  $g_{jk} = 0$ .

El algoritmo basado en el betweenness de los links que se ha empleado funciona de la siguiente manera: al inicio todos los nodos pertenecen a una misma comunidad y se calcula el valor de betweenness para cada enlace mediante (3.22). Se elimina el enlace con mayor betweenness y se recalcularon los valores de betweenness para los enlaces restantes. Repitiendo este proceso la comunidad inicial se va dividiendo en comunidades más pequeñas aislándose unas de otras, aunque si se permite iterar hasta el final cada nodo pertenecerá a una comunidad individual. Merece la pena mencionar que el hecho tener que recalcular los valores de betweenness para todos los enlaces restantes en cada iteración hace que este algoritmo sea de tipo “voraz”, que supone un aumento considerable de los requisitos computacionales conforme aumenta el número de enlaces de la red, esto es, para umbrales de correlación  $\tau_c$  progresivamente más bajos.

Las particiones llevadas a cabo por el algoritmo quedan reflejadas en un dendrograma, un gráfico jerárquico en forma de árbol donde las hojas (subdivisiones finales) representan a los nodos individuales, y el estudio de este dendrograma permite identificar fácilmente aquellas comunidades que resulten interesantes.

### 3.2. Redes bayesianas

Las redes probabilísticas (Gutiérrez et al., 2004; Graafland, 2022) consisten en un grafo  $\mathcal{G}$  y una función  $P$  (factorizada) de densidad de probabilidad conjunta (DPC) asociada, dada por un conjunto de parámetros  $\Theta$ . El grafo codifica las dependencias marginales y condicionales de orden superior<sup>2</sup> y por pares que son relevantes para el problema en cuestión, utilizando un criterio de separación particular, que depende del tipo de gráfico (dirigido o no dirigido). La separación del grafo se basa en la existencia o no de caminos específicos que unen las variables en el grafo, y determina una estructura local/agrupamiento de las variables. Estos “clusters” definen una factorización de la función DPC que preserva las dependencias codificadas. En este trabajo nos centramos en el caso continuo y, en particular, en las redes probabilísticas gaussianas dirigidas. Sea  $\mathbf{X}$  una variable gaussiana con  $d$  variables. La función DPC de  $\mathbf{X}$  viene dada entonces por:

$$P(\mathbf{X}) = (2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\{-1/2(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\}. \quad (3.23)$$

Aquí,  $\boldsymbol{\mu}$  es el vector de medias con  $d$  componentes y  $\boldsymbol{\Sigma}$  la matriz de covarianza de dimensiones  $d \times d$ .

#### 3.2.1. Codificación de las (in)dependencias del grafo

Las redes probabilísticas dirigidas, también denominadas **Redes Bayesianas (BN)**, utilizan Grafos Dirigidos Acíclicos (DAG, *Directed Acyclic Graph*) para representar las (in)dependencias entre las variables (utilizando el criterio de  $d$ -separación) y factorizan  $P(\mathbf{X})$  sobre un conjunto de coeficientes de regresión lineal y coeficientes de variación local ( $\boldsymbol{\beta}, \boldsymbol{\nu}$ ). A continuación se explica el criterio de  $d$ -separación que decodifica las independencias en un DAG (ver Castillo et al., 1997, para más detalles): Dos nodos  $X$  y  $Y \in \mathbf{X}$  (o subconjuntos de nodos) son condicionalmente dependientes dado un conjunto  $\mathcal{S}$  (denotado por  $D(X, Y | \mathcal{S})$ ) si y sólo si existe un camino D entre  $X$  y  $Y$  que satisface las dos condiciones siguientes:

---

<sup>2</sup>La mayoría de relaciones intensas que se reflejan en un grafo son de carácter local, de manera que orden superior se refiere a aquellas dependencias de larga distancia y no tan intensas.

1. Para cada “collider” <sup>3</sup>  $C$  en  $D$ , o bien  $C$  o un descendiente suyo está en  $\mathcal{S}$ .
2. Ningún nodo no-collider en  $D$  está en  $\mathcal{S}$ .

Obsérvese que, bajo el criterio de  $d$ -separación, la dependencia marginal entre dos nodos puede ser codificada por cualquier camino  $D$  sin V-estructuras.

### 3.2.2. Codificación de las (in)dependencias en la función de probabilidad

Los DAG con el criterio de  $d$ -separación determinan una factorización de la función JPD  $P(\mathbf{X})$  que codifica las dependencias representadas en el grafo con el mínimo número de parámetros (Castillo et al., 1997). La factorización está determinada por factores locales, uno por cada nodo del grafo, como

$$P(X_1, \dots, X_d) = \prod_{i=1}^d P_i(X_i | \Pi_i), \quad (3.24)$$

donde  $\Pi_i$  es el conjunto de padres del nodo  $X_i$ .

La presencia de un arco  $X_j \rightarrow X_i$  implica la presencia del factor  $P_i(X_i | \dots X_j \dots)$  en  $P(\mathbf{X})$ , y por tanto una dependencia condicional de  $X_i$  y  $X_j$ . Además, la ausencia de un arco entre  $X_i$  y  $X_j$  en el gráfico implica la ausencia de los factores  $P_i(X_i | \dots X_j \dots)$  o  $P_j(X_j | \dots X_i \dots)$  en  $P(\mathbf{X})$  y, por tanto, la existencia de un conjunto de variables  $\mathcal{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$  que hace que  $X_i$  y  $X_j$  sean condicionalmente independientes en el modelo probabilístico (véase Koller y Friedman, 2009; Castillo et al., 1997, para más información).

En el caso de BN Gaussianas,  $P(\mathbf{X})$  es la distribución normal multivariante (3.23) y los factores en (3.24) son variables aleatorias normales monovariables ligadas por dependencias lineales a sus padres (Geiger y Heckerman, 1994):

$$X_i | \Pi_i \sim N(\mu_i + \sum_{j | X_j \in \Pi_i} \beta_{ij}(X_j - \mu_j), \frac{1}{\nu_i}) \quad (3.25)$$

en lo que es esencialmente un modelo de regresión lineal de  $X_i$  contra el conjunto de padres  $\Pi_i$ , con coeficientes de regresión  $\beta_i = \{\beta_{i,j} : X_j \in \Pi_i\}$ ;  $\mu_i$  es la media incondicional de  $X_i$ ;  $\nu_i$  es la varianza condicional de  $X_i$  dado el conjunto  $\Pi_i$ .

---

<sup>3</sup>Un nodo  $C$  en un camino dirigido  $D$  se llama “collider” cuando la parte de  $D$  que pasa por encima de  $C$  tiene la forma de una “V-estructura”, es decir,  $\rightarrow C \leftarrow$

### 3.2.3. Aprendizaje de redes bayesianas

El aprendizaje de las redes bayesianas consta de dos partes esenciales: El **aprendizaje estructural** consiste en encontrar el grafo  $\mathcal{G}$  que codifica la estructura de dependencia subyacente a los datos; el **aprendizaje paramétrico** consiste en estimar los parámetros  $\Theta$  dado el grafo  $\mathcal{G}$  (es decir, dado la función P factorizada).

El aprendizaje de los parámetros es la parte fácil del proceso; si se dispone de la estructura del grafo y suponemos que los parámetros de las diferentes distribuciones locales son independientes, la factorización en (3.24) implica

$$P(\Theta | \mathcal{G}, \mathcal{D}) = \prod_{i=1}^d P(\Theta_i | \Pi_i, \mathcal{D}). \quad (3.26)$$

El conjunto de parámetros  $\Theta = (\boldsymbol{\beta}, \boldsymbol{\nu})$  puede aprenderse por separado y de forma eficiente para cada nodo según (3.26); en este caso los parámetros  $\beta_{ij}$  y  $\nu_i$  se obtienen mediante regresión lineal de  $X_i$  sobre su conjunto de padres  $\Pi_i$  utilizando como función de coste la medida de “verosimilitud” (*likelihood*). Para realizar el aprendizaje paramétrico se ha utilizado en este estudio la función `bn.fit` del paquete R bnlearn (Scutari, 2010).

Sin embargo, el aprendizaje estructural es un problema difícil desde el punto de vista computacional y se disponen de varios algoritmos para obtener la estructura de dependencia (es decir, el grafo  $\mathcal{G}$ ) a partir de los datos. Existen tres posibles enfoques de aprendizaje (Koller y Friedman, 2009; Scutari et al., 2019; Verma y Pearl, 1991): **basado en restricciones**, **basado en la puntuación** (*score*), e **híbrido**. Los algoritmos basados en restricciones utilizan pruebas de independencia condicional para aprender la estructura del grafo; los algoritmos basados en la puntuación utilizan las medidas de bondad de ajuste como funciones objetivo para maximizar; y los algoritmos híbridos combinan ambos.

En Scutari et al. (2019) se ha comprobado que los algoritmos basados en la puntuación son la opción más viable para el aprendizaje de estructuras de redes bayesianas de conjuntos de datos complejos, ya que son únicos en su capacidad de aprender redes en las que las dependencias de orden superior quedan profundamente representadas. Tras asignar una puntuación a cada red potencial, un algoritmo de búsqueda intenta maximizarla construyendo iterativamente una red. Algunos ejemplos son heurísticos como “greedy

search”, “simulated annealing” (Bouckaert, 1995), y los algoritmos genéticos (Larrañaga et al., 1997); una revisión exhaustiva de estos y otros enfoques se ofrece en Castillo et al. (1997).

Guiados por los resultados en (Graafland, 2022) en este trabajo se ha utilizado el algoritmo de búsqueda “Hill Climbing”, de tipo voraz, y que consta de una fase de inicialización (paso 1) seguida de una búsqueda “hill climbing” (paso 2). En cada iteración, “hill climbing” intenta eliminar e invertir cada arco del DAG actual  $\mathcal{G}_{max}$ , e intenta añadir cada posible arco que no esté ya presente en  $\mathcal{G}_{max}$  y que no introduzca ningún ciclo. Se trata de movimientos locales que sólo afectan a una o dos distribuciones locales en la BN, lo que reduce en gran medida la complejidad computacional de la búsqueda al evitar la necesidad de volver a puntuar todos los nodos en cada iteración. El  $\mathcal{G}$  resultante con la puntuación más alta  $S_{\mathcal{G}}$  se compara con  $\mathcal{G}_{max}$ ; si tiene una puntuación mejor ( $S_{\mathcal{G}} > S_{max}$ ) entonces  $\mathcal{G}$  se convierte en el nuevo  $\mathcal{G}_{max}$ . Si por el contrario  $S_{\mathcal{G}} < S_{max}$ , se ha alcanzado un óptimo y la búsqueda concluye.

No hay garantía de que  $\mathcal{G}$  sea un óptimo global; por lo tanto, el algoritmo puede realizar más pasos para reducir las posibilidades de que  $\mathcal{G}$  sea en realidad un óptimo local. Una opción es reiniciar la búsqueda en el paso 2 desde un punto de partida diferente, cambiando  $r$  arcos en  $\mathcal{G}$ , el óptimo actual. Otra opción es mantener una lista ‘tabú’ de los DAG visitados anteriormente y continuar buscando un DAG mejor que aún no se ha considerado. También es posible realizar ambos pasos simultáneamente y obtener una búsqueda tabú con reinicios aleatorios.

El criterio estadístico a utilizar en el aprendizaje estructural, la función de la puntuación de red, depende principalmente de la distribución de  $\mathbf{X}$ . En nuestro caso se utiliza el Criterio de Información Bayesiano (BIC):

$$\text{BIC}(\mathcal{G}; \mathcal{D}) = \sum_{i=1}^d \left[ \log P(X_i | \Pi_{X_i}) - \frac{|\Theta_{X_i}|}{2} \log n \right]. \quad (3.27)$$

BIC es una opción común tanto para BN discretas como para BN gaussianas, porque proporciona una aproximación simple a  $\log P(\mathcal{G} | \mathcal{D})$  que no depende de ningún hiperparámetro.

### 3.2.4. Capacidad de explicar los datos: Log-likelihood

Utilizamos la “log-likelihood” como medida probabilística de la capacidad de generalización de nuestras redes bayesianas. La log-likelihood de una BN se define como  $\log(P(\mathcal{D}|BN))$ , donde  $\mathcal{D}$  es el conjunto de datos. En esta expresión,  $P(\mathcal{D}|BN)$  es la densidad de probabilidad de los datos bajo el modelo BN. La log-likelihood puede interpretarse como la probabilidad de un conjunto de datos  $\mathcal{D}$  cuando  $P$  es modelado por una determinada BN (Para las funciones de densidad discretas,  $P(\mathcal{D}|BN)$  es igual a la probabilidad de los datos dado el modelo.) La probabilidad se simplifica casi siempre tomando el logaritmo natural ya que trabajando con probabilidad continua los valores suelen ser pequeños y la diferenciación de la función de likelihood (con el fin de optimizarla) suele ser difícil. Además, no afecta a la interpretabilidad de la comparación entre modelos, gracias al carácter monótono creciente del logaritmo.

La log-likelihood es capaz de comparar BNs que codifican el mismo tipo de función de densidad  $P$ , pero con diferentes parámetros. Esta medida debe interpretarse de forma comparativa: el valor de la log-likelihood del modelo  $BN_1$  no es muy significativo en términos absolutos. Sin embargo, si la log-likelihood del modelo  $BN_1$  es mayor que la del modelo  $BN_2$ , se puede concluir que la  $BN_1$  explica mejor los datos que la  $BN_2$ . En este trabajo se calculan BNs de diferentes tamaños que codifican todas una distribución gaussiana multivariable sobre un espacio de variables de dimensión constante, lo que hace que la log-likelihood sea una medida comparativa adecuada (Koller y Friedman, 2009). A continuación se detalla el cálculo de la log-likelihood  $\log P(\mathcal{D}|BN)$  para un conjunto de datos  $\mathcal{D}$  formado por  $t$  realizaciones de datos independientes  $\mathcal{D}_k$ ,  $k \in \{1, \dots, t\}$ , del vector aleatorio  $m$ -dimensional  $\mathbf{X}$  con  $\mathcal{D}_k = \{d_1^k \dots d_m^k\}$  y  $d_i^k$  la realización  $k$ -ésima de la variable  $X_i \in \mathbf{X}$ . En el caso de una BN gaussiana (BNG), a partir de (3.24), tenemos:

$$\begin{aligned}
 \log P(\mathcal{D}|BNG) &= \sum_{k=1}^t \log P(\mathcal{D}_k | BNG) \\
 &= \sum_{k=1}^t \log \prod_{i=1}^m P_i(X_i = d_i^k | \Pi_{X_i} = d_{\Pi_{X_i}}^k) \\
 &= \sum_{k=1}^t \sum_{i=1}^m \log P_i(X_i = d_i^k | \Pi_{X_i} = d_{\Pi_{X_i}}^k),
 \end{aligned} \tag{3.28}$$

donde  $d_{P_i X_i}^k$  es un subconjunto de  $\mathcal{D}_k$  que contiene el dato  $k$ -ésimo del conjunto de padres  $\Pi_{X_i}$  de  $X_i$ . De (3.25) sabemos que las densidades condicionales univariadas en la suma en (3.28) son normales univariadas, y las ejecutamos con el paquete básico de R **stats**.

### 3.2.5. Inferencia en redes bayesianas gaussianas

La estimación de probabilidades condicionales es uno de los principales problemas en el aprendizaje automático, ya que permite consultar el modelo para aplicaciones particulares, cuando se dispone de alguna evidencia. En el caso de una BN que modela una distribución gaussiana multivariable, existe una aproximación directa para este problema utilizando la expresión cerrada para condicionar las distribuciones gaussianas multivariable en (3.24) (ver Castillo et al., 1997, para más detalles). Utilizamos un mecanismo sencillo para estimar el impacto de una variable (o variables) probatoria(s)  $X_e$  (con valor conocido) sobre las demás variables (grid-boxes de área quemada) en la red. Por ejemplo, para simular eventos de incendio extremos en una región, podríamos introducir condiciones extremas en una casilla particular  $X_e$  situada en esa región (por ejemplo, un fuerte aumento de los eventos de incendio, digamos  $X_e = 2\sigma_{X_e}$ ). Entonces, la probabilidad condicional de las otras casillas  $P(X_i|X_e)$  proporciona una cuantificación del impacto de esta evidencia en el resto de nodos de la red. Esto permitirá, por ejemplo, estudiar las teleconexiones de  $X_e$  con otras regiones si se detecta un cambio significativo de estado en un nodo geográficamente distante de la red.

---

# 4

---

## Resultados

En este capítulo se detalla el proceso de construcción de las redes complejas y se recogen todos los resultados obtenidos utilizando la metodología descrita en el capítulo 3, tanto para redes de correlación como para redes bayesianas.

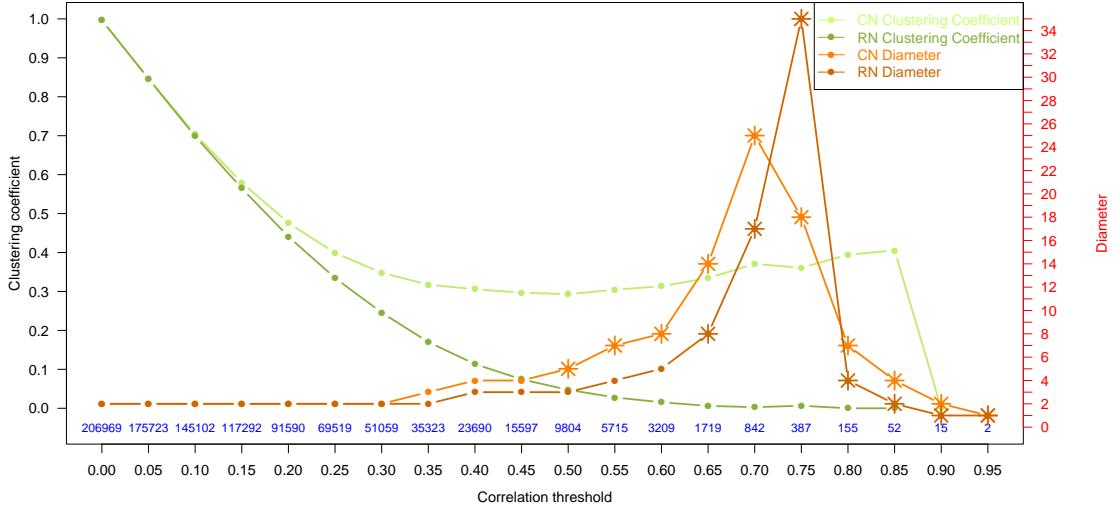
### 4.1. Redes de correlación

#### 4.1.1. Elección del umbral de correlación

Como se explica en la sección 3.1.2, para construir la red de correlación es necesario definir un umbral de correlación  $\tau_c$ . Esta elección es crítica para el estudio de la CN, ya que un umbral muy bajo resultará en una red muy conectada y con mucho ruido, mientras que un umbral demasiado alto apenas permitirá la creación de conexiones, simplificando excesivamente la red. Por este motivo es necesario realizar un estudio de las propiedades globales de la red, definidas en (3.14) y (3.15), para diferentes umbrales de correlación, lo que permitirá determinar cuál es la CN óptima para el estudio. Además, los resultados obtenidos pueden compararse con las medidas globales para una red aleatoria (RN, de *random network*, también denominada grafo de Erdős-Rényi<sup>1</sup>) para justificar la complejidad de la red construida. La RN se genera imponiendo como única restricción que tenga el mismo número de enlaces que la CN correspondiente. Procediendo de esta manera, se obtiene la gráfica 4.1, cuyos resultados se desarrollan a continuación.

---

<sup>1</sup>En este tipo de grafos se define un enlace entre dos nodos con una probabilidad  $2l/(m(m - 1))$  donde  $m$  es el número de nodos y  $l$  es el número total de enlaces

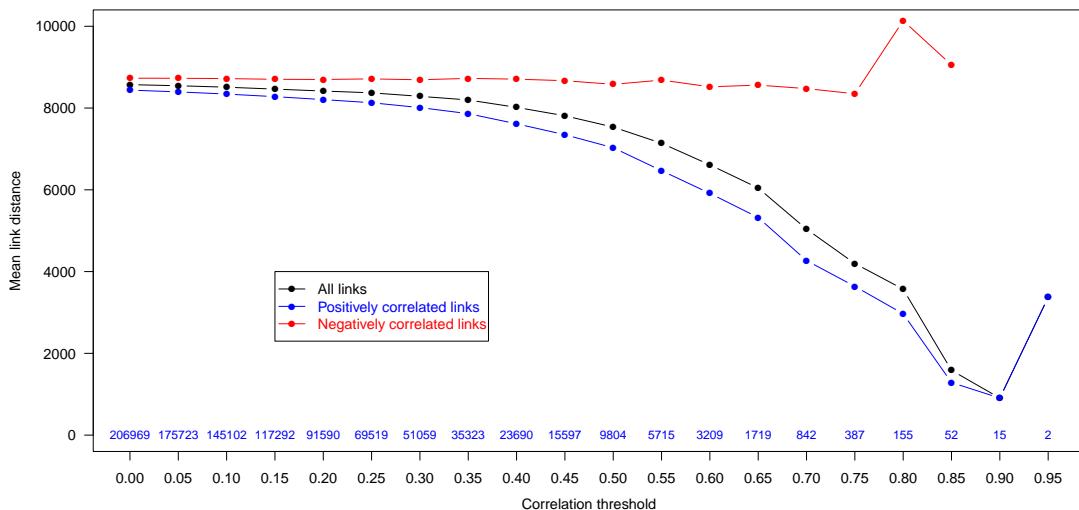


**Figura 4.1:** Coeficiente de clustering global (verdes) y diámetro (naranjas) para la red de correlación (CN) y la red aleatoria (RN), considerando un umbral de correlación  $\tau_c$  variable. Los puntos estrellados en el diámetro indican que la red no está completamente conectada. Las cifras en azul sobre el eje X indican, para cada umbral  $\tau_c$ , el número de enlaces de las redes generadas.

Para  $\tau_c = 0$  se observa que el coeficiente de clustering global es igual a la unidad (todos los nodos están interconectados entre sí), decreciendo a partir de este punto. Para la red aleatoria este decrecimiento es monótono y se alcanzan rápidamente valores muy bajos de  $C$ , mientras que la CN comienza con un descenso rápido pero en torno a  $\tau_c = 0.35$  se estabiliza (de hecho experimenta un ligero crecimiento) hasta finalmente hacerse nulo por la falta de enlaces en la red. Por otro lado, para el diámetro de la CN se observa un valor pequeño para umbrales bajos, indicando que hay una gran conectividad en la red ya que se puede viajar de un nodo a cualquier otro en un único paso; conforme aumenta el umbral (y disminuye el número de enlaces) el diámetro aumenta de manera sostenida hasta  $\tau_c = 0.6$  indicando que aumenta la distancia media para viajar entre dos nodos, y experimenta un crecimiento más acentuado hasta  $\tau_c = 0.7$  donde el diámetro alcanza su máximo. A partir de este punto la red comienza a desconectarse en mayor medida (aparecen grupos de nodos interconectados aislados de otros grupos) y el escaso número de enlaces provoca un descenso en el diámetro. Por otro lado, la red aleatoria experimenta un comportamiento similar, pero tarda más en iniciar el crecimiento, lo que indica que es más difícil de desconectar, aunque crece más rápidamente la CN llegando a

un máximo más grande. En la Figura 4.1, los asteriscos indican que la red es inconexa, es decir, existen nodos aislados sin conexiones con el resto de la red y se observa que este fenómeno se produce antes para la CN que para la red aleatoria. Este comportamiento diferencial entre la red aleatoria y la generada a partir de los datos de área quemada, pone de manifiesto la existencia de una estructura no aleatoria subyacente en los datos, que justifica el empleo de las técnicas basadas en grafos para su análisis.

Para obtener un conocimiento más detallado del tipo de relaciones existentes en la CN generada, se estudia a continuación la distancia geográfica media de todos los enlaces, y además se separan los positivos de los negativos, obteniéndose la gráfica de la Fig. 4.2.



**Figura 4.2:** Distancia geográfica media de la totalidad de los enlaces, diferenciando además los positivos (azul) y los negativos (rojo), para diferentes umbrales de correlación  $\tau_c$ . En la zona inferior de la gráfica, en azul, el número total de enlaces de la CN.

En esta gráfica se puede apreciar que en promedio, los enlaces negativos son más distantes que los positivos, y mantienen una tendencia estable debido a que en este caso, la longitud de la conexión se distribuye de manera similar para enlaces con alta y baja correlación. El decrecimiento de la distancia que experimentan los enlaces positivos al aumentar el umbral se produce porque las correlaciones positivas más fuertes son de carácter local (existencia de autocorrelación espacial), de manera que para umbrales altos la mayoría de los enlaces serán muy cercanos, aunque sigan existiendo algunos de larga distancia. Además,

como hay una mayor proporción de enlaces positivos, la media total se ve influenciada por estos, manifestando un comportamiento similar.

En base a lo observado en la figura 4.1, se decide estudiar a fondo la red construida con un umbral de  $\tau_c = 0.6$  por dos motivos: el primero es que su coeficiente de clustering se encuentra en una zona estable, por lo que se tratará de una CN interesante; el segundo, a partir de este punto el diámetro crece mucho más rápido, de manera que para  $\tau_c = 0.6$  la red no es ni demasiado conexa ni demasiado inconexa, permitiendo la visualización de patrones relevantes ya que la cantidad de información no satura pero tampoco es insuficiente. Representando esta red en un mapa sobre coordenadas geográficas<sup>2</sup> se obtiene la figura 4.3; las tablas 4.1 y 4.2 complementan esta información aportando los valores de los descriptores de los enlaces y las medidas de conectividad respectivamente.

Enlaces totales	Enlaces Positivos	Enlaces Negativos
3209	2351	858
Distancia media total	Distancia media positiva	Distancia media negativa
6610.33	5913.46	8519.82
Máxima distancia total	Máxima distancia positiva	Máxima distancia negativa
20030.35	20030.35	19582.97
Mínima distancia total	Mínima distancia positiva	Mínima distancia negativa
213.55	213.55	555.66

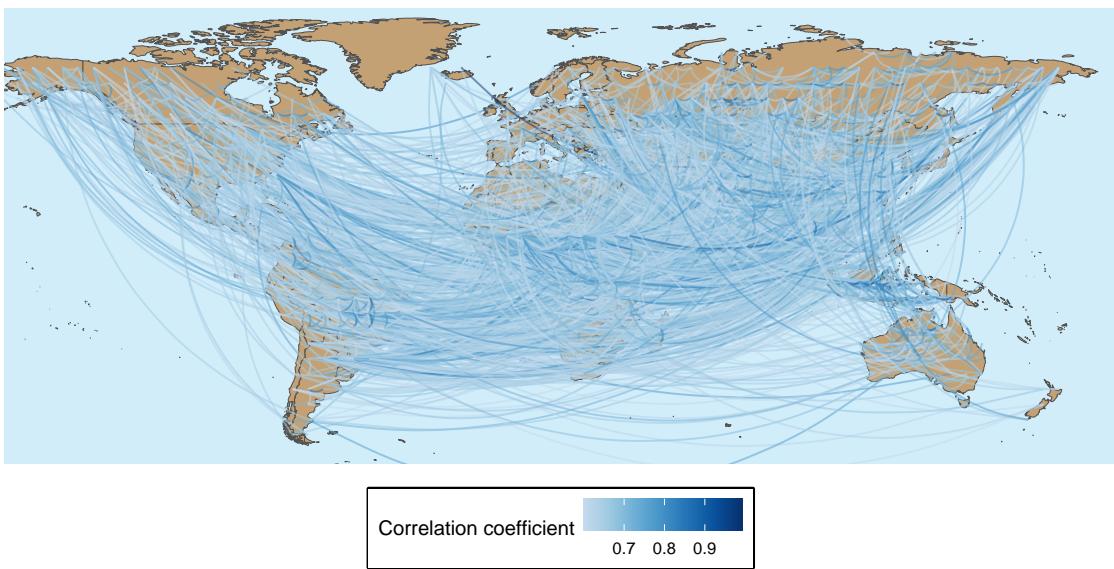
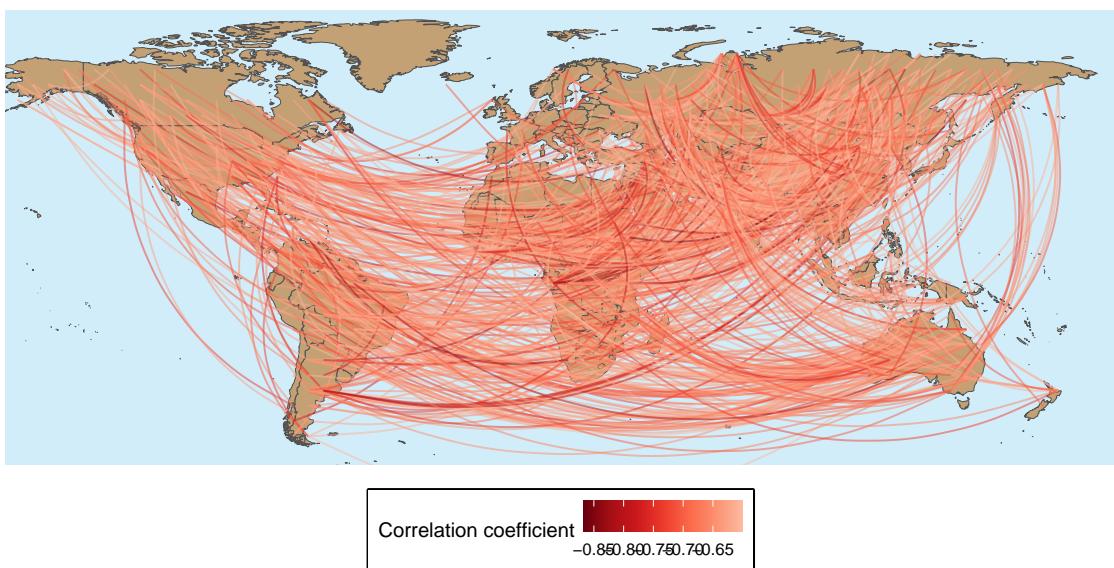
**Tabla 4.1:** Descriptores de las conexiones para  $\tau_c = 0.6$

Coeficiente de custering	Diametro
0.31	8.00

**Tabla 4.2:** Medidas de conectividad para  $\tau_c = 0.6$

En la figura 4.3 se aprecia un mayor número de enlaces positivos en comparación con los negativos, ya que en la gráfica superior existe una densidad de enlaces mucho mayor. Además, se aprecian conexiones locales con alta correlación positiva en Norteamérica, Sudamérica, Siberia, Indonesia y Australia. Esto no se observa en la gráfica inferior, poniendo de manifiesto que las conexiones locales con correlación negativa son menos frecuentes y que el fenómeno estudiado presenta cierta autocorrelación espacial, como parece esperable dada la homogeneidad y continuidad relativas tanto de los combustibles (tipos de vegeta-

<sup>2</sup>Las mismas representaciones para CNs construidas con otros umbrales de correlación pueden consultarse en el anexo A.

Positive Spatial Network for  $\tau_c = 0.6$ Negative Spatial Network for  $\tau_c = 0.6$ 

**Figura 4.3:** Representación espacial de la CN pesada según  $w_c$ , considerando el umbral de correlación  $\tau_c = 0.6$ . En azul/rojo, se representan los enlaces con correlación positiva/negativa.

ción), como de las situaciones meteorológicas conducentes a la ocurrencia de incendios.

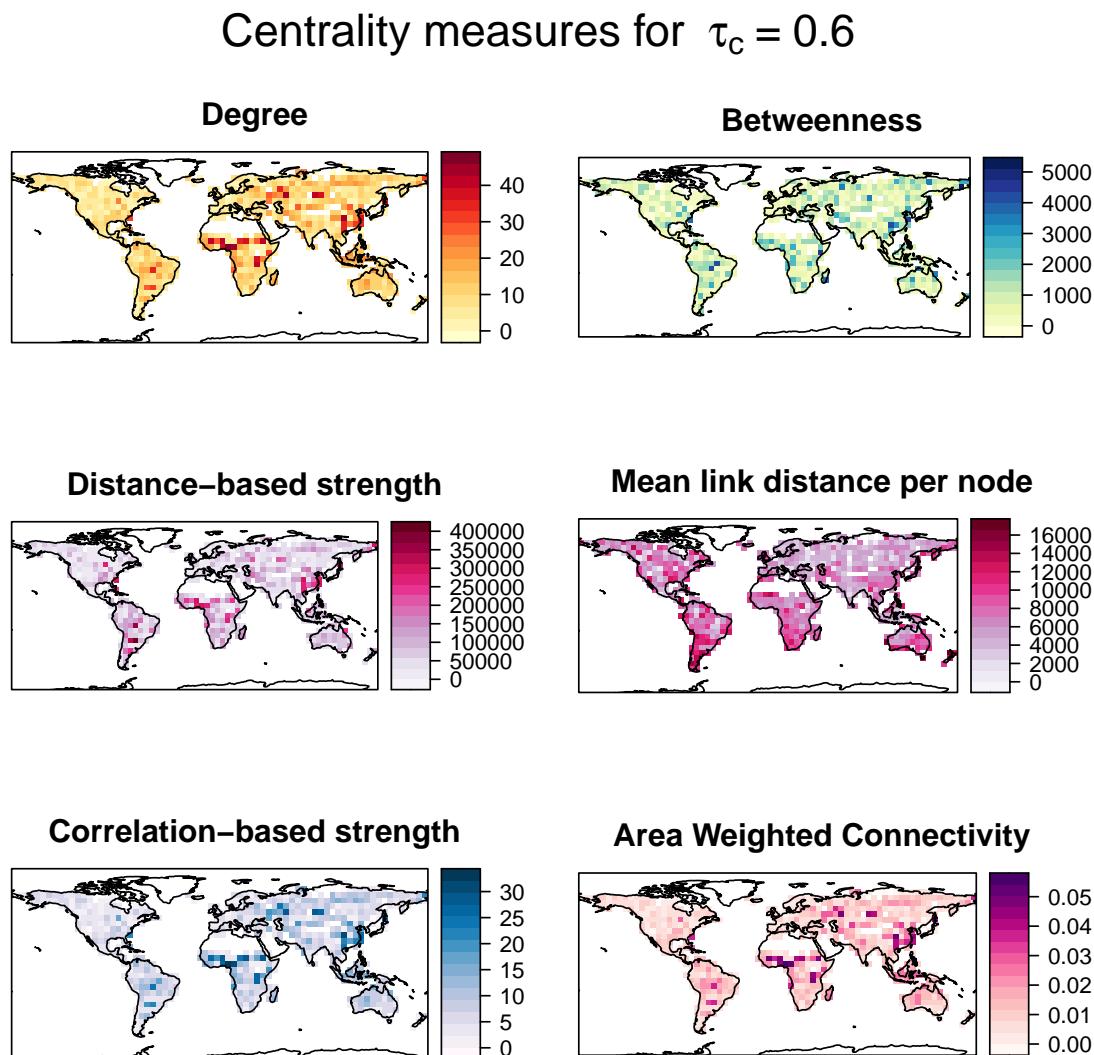
#### 4.1.2. *Medidas de centralidad*

Una vez escogida la CN se estudian las medidas de centralidad para esta red<sup>3</sup>, definidas en la sección 3.1.5 y representadas en la figura 4.4

Comenzando por el grado se observa que las zonas de mayor conectividad son el África Subsahariana, China oriental, parte de los Urales, el estrecho de Bering y algunos puntos de Sudamérica y Norteamérica. Por lo tanto los nodos localizados en estas zonas son los que más conexiones tienen con otros nodos, aunque no es posible discernir si son de corta o larga distancia, o si son de alta correlación, por lo que será necesario estudiar otras medidas de centralidad. Con respecto al betweenness destacan nodos en la cuenca del Amazonas, África oriental y Madagascar y (nuevamente) el este de China y el estrecho de Bering; su importancia radica en que, aunque algunos no sean los más conectados (Madagascar y Bering son un claro ejemplo pues tienen bajo grado), son puntos por los que circula una gran cantidad de información y son posibles nexos de unión entre diferentes comunidades.

Por otra parte, en la gráfica para el strength basado en distancia se observa que los nodos de Sudamérica muestran valores más altos que aquellos en África o China, que los superan en grado; esto significa que aunque tengan menos conexiones, éstas son de una distancia considerablemente más larga ya que las conexiones locales contribuyen menos que las de larga distancia. Además, esto también queda reflejado en la distancia media por nodo, donde se registran valores más altos en América en comparación con Eurasia; sin embargo, es necesario resaltar que para zonas aisladas geográficamente es mucho más fácil que las conexiones sean de larga distancia, pues hay menos posibilidades de formar conexiones locales simplemente por falta de espacio. Este hecho se aprecia especialmente bien en islas como Nueva Zelanda. El mapa de strength basado en correlación es muy similar al de grado en este caso, lo cual puede significar dos cosas: (i) para los nodos con muchas conexiones éstas tienen una correlación elevada, o bien (ii) la distribución de correlaciones en los enlaces es relativamente uniforme y esta medida no aporta información. Finalmente, el mapa de AWC es bastante similar al de grado, lo que indica que los enlaces

<sup>3</sup>Las medidas de centralidad para CNs construidas con otros umbrales de correlación pueden consultarse en el anexo A.



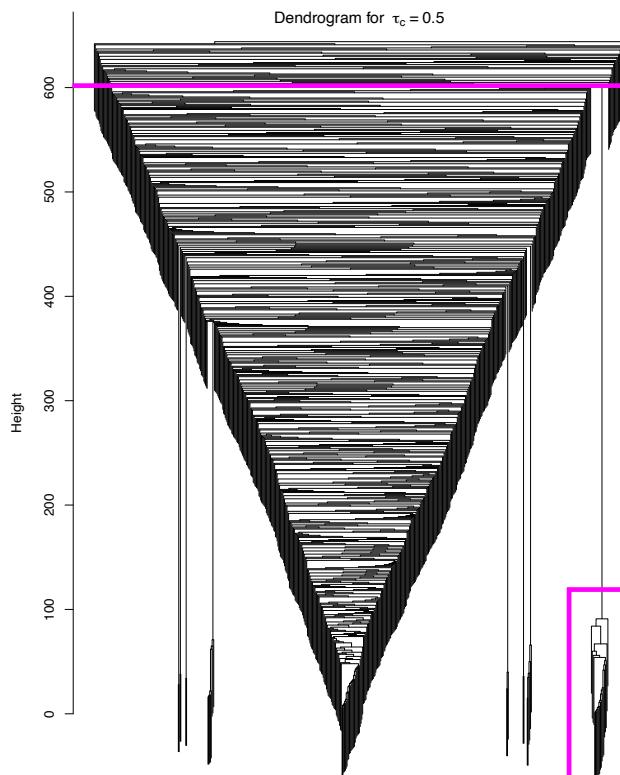
**Figura 4.4:** Medidas de centralidad para la CN, que aportan información sobre la importancia de cada uno de los nodos. De izquierda a derecha, y de arriba a abajo: Grado, Betweenness, Strength basado en distancia, Distancia geográfica media por nodo, Strength basado en correlación y Conectividad pesada por el área (AWC).

tienden a conectarse sobre la franja intertropical (sobre todo en África).

#### 4.1.3. Búsqueda de comunidades

Tras estudiar las propiedades de la CN es momento de intentar localizar comunidades; el propósito de este estudio es explorar los resultados en busca de nodos distantes que presenten una sincronicidad fuerte. Para ello, se aplica el algoritmo de detección de comu-

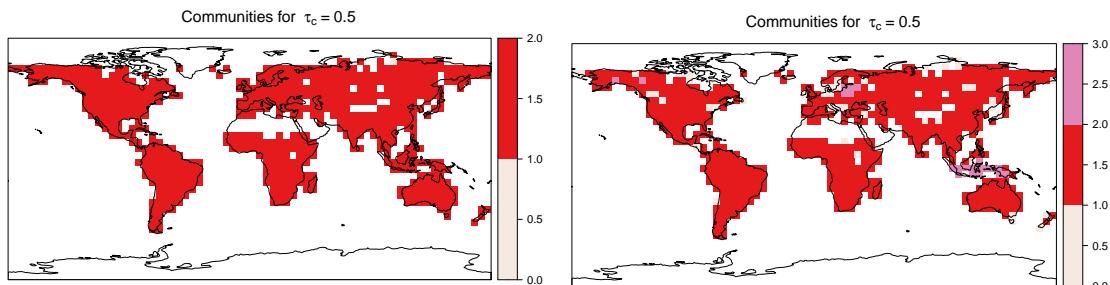
nidades definido en la sección 3.1.6 para las redes construidas con umbrales de correlación  $\tau_c > 0.4$ . Tras explorar los dendrogramas se decide estudiar las redes con umbrales de correlación  $\tau_c$  de 0.5, 0.6 y 0.7; para umbrales inferiores la red es tan densa que el algoritmo únicamente separa nodos individuales de la comunidad inicial (no siendo efectivo por lo tanto para localizar comunidades propiamente dichas), mientras que para umbrales superiores la red tiene tan pocos enlaces (Fig. 3.14) que la mayoría de nodos son inconexos y forman comunidades individuales. Los dendrogramas obtenidos para los umbrales escogidos se presentan en las gráficas 4.5, 4.7 y 4.9 respectivamente.



**Figura 4.5:** Dendrograma resultado de aplicar el algoritmo de detección de comunidades basado en el betweenness de los enlaces de la CN para el umbral  $\tau_c = 0.5$ . La línea horizontal indica el nivel de corte para el que se obtiene la comunidad destacada con el rectángulo. Los dos mapas de comunidades resultantes por encima y por debajo de este nivel de corte quedan reflejados en la Figura 4.6 (mapas izquierdo y derecho respectivamente).

Para  $\tau_c = 0.5$  se observa el comportamiento descrito para umbrales bajos: la red está muy conectada y el algoritmo va separando uno a uno los nodos de la comunidad inicial en cada paso. Sin embargo, ya se distingue una comunidad muy diferente que se separa en un punto muy inicial del proceso divisivo. En el mapa de comunidades (Fig 4.6) se observa

que esta comunidad (en verde) se localiza en Indonesia y Europa oriental, con un par de nodos en Norteamérica.

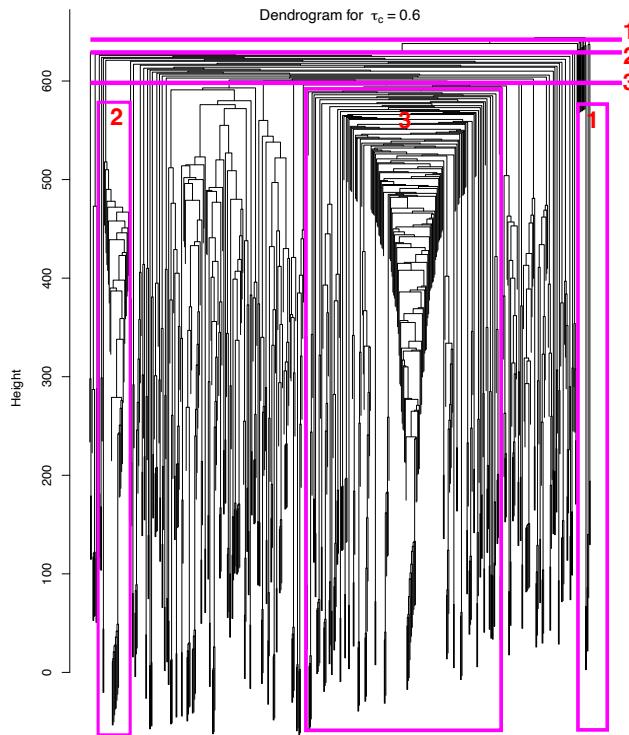


**Figura 4.6:** Mapas espaciales de las comunidades obtenidas para  $\tau_c = 0.5$  (ordenadas por tamaño) en varios niveles de corte del dendrograma (ver Fig. 4.5). Para evitar ruido en el mapa se filtran las comunidades con un único nodo, dándoles el color beige.

En el dendrograma para  $\tau_c = 0.6$  se puede observar que se separan un mayor número de comunidades notables, así como una zona central (comunidad inicial) de la que se subdividen los grupos en comunidades pequeñas o individuales. Para estudiar las comunidades de mayor tamaño se procede como en el caso anterior, definiendo diferentes niveles de corte y representando los mapas espaciales obtenidos, representados en la Fig. 4.8

Para la situación inicial (1) el algoritmo obtiene una comunidad central y varias comunidades pequeñas y dispersas que no aportan demasiada información. En el primer nivel de corte (2) aparece una comunidad en la zona de Italia (rosa) con algunos nodos dispersos por el planeta. El siguiente nivel de corte (3) muestra la comunidad de Indonesia (rosa), obtenida también para el caso anterior, lo que indica que las dos zonas pertenecientes a esta comunidad se encuentran muy fuertemente conectadas. En sucesivos cortes (4 y 5) aparecen una comunidad dispersa por Eurasia y África (dorado) y una comunidad localizada en México respectivamente. Finalmente, para el último nivel de corte estudiado (6) la comunidad inicial (roja) se fragmenta, dividiendo en mitades las zonas de Norteamérica, Sudamérica y Australia.

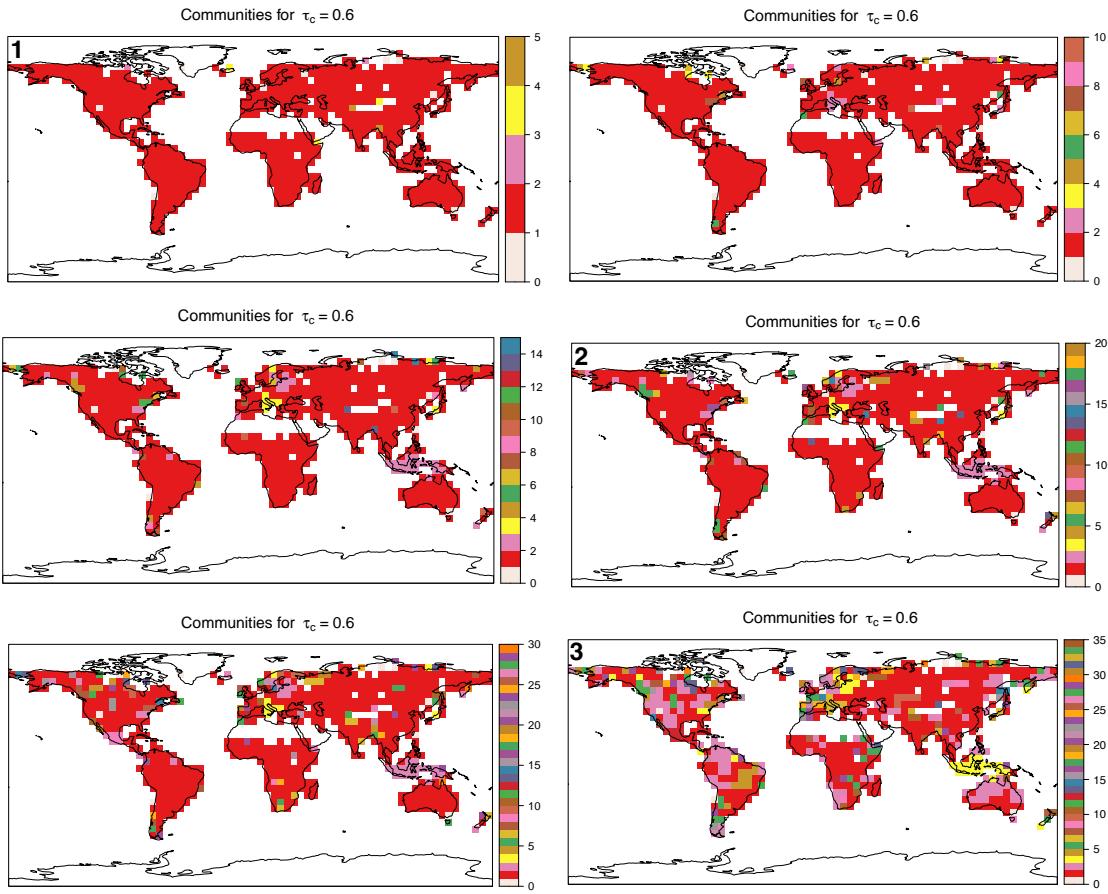
Por último, para el umbral  $\tau_c = 0.7$  se obtiene un dendrograma con dos comportamientos claramente distintos: por una parte, la zona central corresponde a la comunidad inicial, que se dividirá en comunidades cada vez más pequeñas al igual que en los casos anteriores; por la otra, en las zonas de la izquierda y la derecha se distinguen subdivisiones en mitades hasta llegar a comunidades individuales, siendo la manera que tiene el algo-



**Figura 4.7:** Dendrograma resultado de aplicar el algoritmo de detección de comunidades basado en el betweenness de los enlaces de la CN para el umbral  $\tau_c = 0.6$ . Las líneas horizontales representan niveles de corte que dan lugar a los mapas de comunidades representados en la Figura 4.8. Las comunidades más relevantes para cada uno de los niveles de corte se destacan mediante los rectángulos numerados.

ritmo de representar lo que ocurre con los nodos aislados, ya que al no disponer de más enlaces para eliminar, les asigna comunidades individuales desde un primer momento. Los mapas para dos niveles de corte distintos del dendrograma de la Fig. 4.9 se recogen en la figura 4.10.

En la situación inicial (1) se observa una gran comunidad inicial, pero el mapa ya está dividido en comunidades dispersas, destacando la de Centroamérica (verde) que aparece también en la Fig. 4.8. Realizando el corte (2) vuelve a aparecer la comunidad localizada en Indonesia y Europa, obtenida en los dos casos anteriores, indicando nuevamente un grado de conexión entre estas zonas y cuyo estudio será interesante desde el punto de vista de la red bayesiana dado que será posible observar cómo se propaga la información una vez dada la evidencia en uno de los nodos pertenecientes a dicha comunidad.

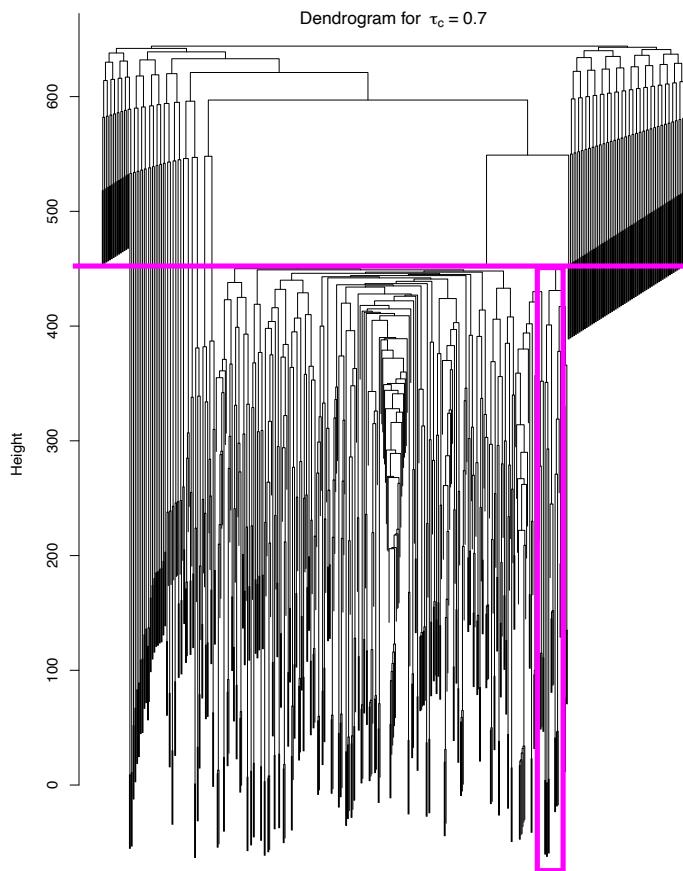


**Figura 4.8:** Mapas de las comunidades obtenidas para  $\tau_c = 0.6$  (ordenadas por tamaño) en varios niveles de corte del dendrograma. Los mapas de comunidades con un número en la parte superior izquierda se corresponden con los niveles de corte indicados en la Figura 4.7. Para evitar ruido en el mapa se filtran las comunidades con un único nodo, dándoles el color beige.

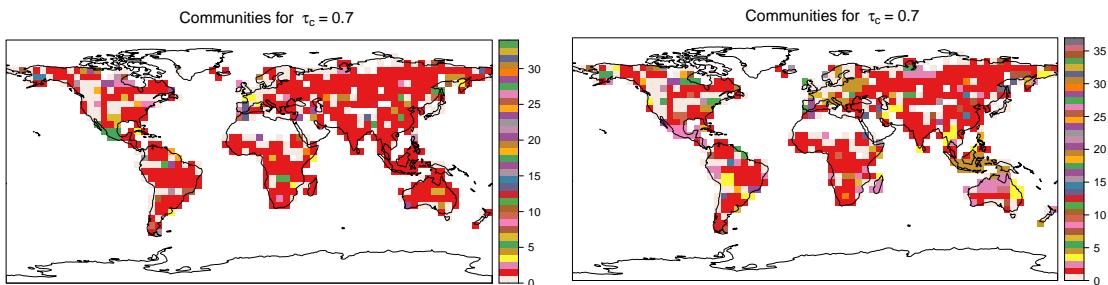
## 4.2. Redes Bayesianas

### 4.2.1. Construcción de la red bayesiana

Como se ha detallado en la sección 3.2 la construcción de la red bayesiana se compone de dos partes: el aprendizaje estructural y el aprendizaje paramétrico. Mediante el primero se obtiene la estructura del grafo que, de manera similar a la CN, cuenta con tantos nodos como celdas haya presentes en la malla geoespacial ( $m = 645$  nodos), ya que se utiliza el mismo conjunto de datos para su construcción ( $\mathcal{D} = M_{t \times m}$ ). Por tanto, el tamaño de la red bayesiana dependerá únicamente del número de enlaces entre los  $m$  nodos considerados. Naturalmente, cuanto mayor sea el tamaño de la red bayesiana, mejor explicará las



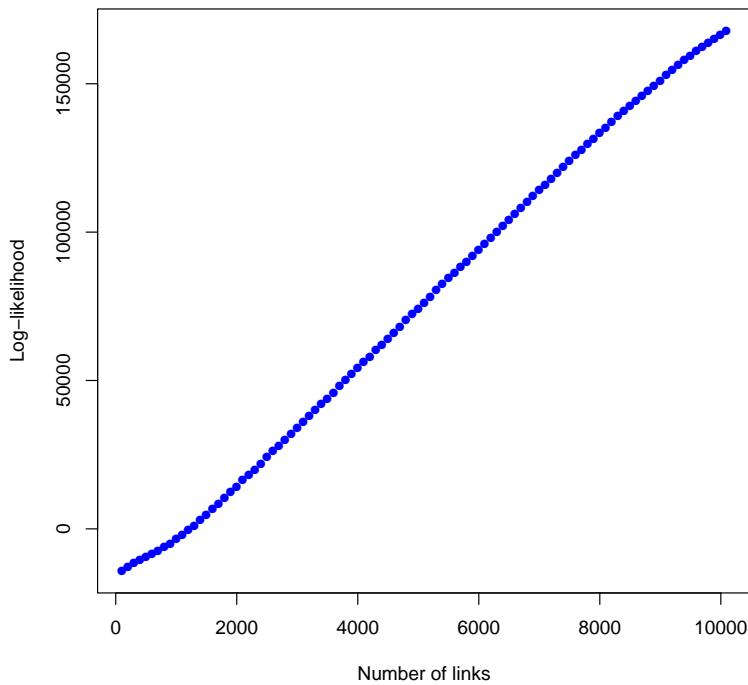
**Figura 4.9:** Dendrograma resultado de aplicar el algoritmo de detección de comunidades basado en el betweenness de los enlaces de la CN para el umbral  $\tau_c = 0.7$ . La línea horizontal indica el nivel de corte para el que se obtiene la comunidad destacada con el rectángulo. Los dos mapas de comunidades resultantes por encima y por debajo de este nivel de corte quedan reflejados en la Figura 4.10 (mapas izquierdo y derecho respectivamente).



**Figura 4.10:** Mapas espaciales de las comunidades obtenidas para  $\tau_c = 0.7$  (ordenadas por tamaño) en varios niveles de corte del dendrograma (ver Fig. 4.9). Para evitar ruido en el mapa se filtran las comunidades con un único nodo, dándoles el color beige.

estructuras subyacentes de los datos; sin embargo, también aumentará considerablemente el coste computacional, tanto para la propia construcción de la red como para los estudios

de inferencia posteriores. Por este motivo y con el objetivo de encontrar una red bayesiana en la que se alcance un compromiso entre tamaño y coste computacional, se decide realizar el aprendizaje estructural de manera secuencial, permitiendo en cada iteración del entrenamiento la adición de 100 arcos a la red anterior hasta un tamaño máximo de 10000 arcos, obteniendo un conjunto de 100 grafos. El aprendizaje se realiza con el algoritmo de “Hill-climbing” explicado en la sección 3.2.3, utilizando el BIC (3.27) como función de puntuación de la red. A continuación se realiza el aprendizaje paramétrico para cada uno de los grafos, obteniendo la distribución de probabilidad en cada uno de los nodos según (3.26), y por tanto un total de 100 redes bayesianas de diferentes tamaños.



**Figura 4.11:** Estudio de la log-likelihood de un conjunto de redes bayesianas según su tamaño.

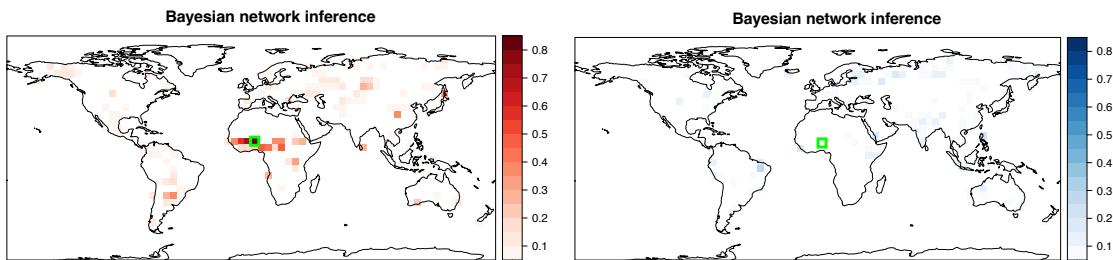
Por la definición de log-likelihood el hecho de añadir más arcos a la red permite que esta explique los datos de mejor manera y que la distribución obtenida sea creciente, hecho que se observa en la figura 4.11, donde cada adición de 100 arcos supone una mejora en la capacidad explicativa de la red. No se observa una estabilización en el crecimiento, “saturando” el aumento de información que se obtiene al aumentar el tamaño, lo que implica

que la red de mayor tamaño con 10000 arcos sería la que mejor explica los datos, pero sin llegar a una situación de sobreajuste. Nuestra hipótesis es que este fenómeno de aumento lineal es debido al bajo número de datos temporales considerados (cada serie dispone únicamente de 19 años). Paralelamente, en un estudio realizado para datos mensuales (aumentando el número de datos temporales a  $19 \text{ años} \times 12 \text{ meses} = 228$ ) sí que se observa una estabilización en la curva de log-likelihood indicando que, en caso de disponer de más datos, la red llega a sobreajustar a partir de un cierto nivel de complejidad. Este fenómeno también se observa en (Graafland, 2022) donde también se emplearon series temporales más extensas de temperatura.

Aunque la red de mayor tamaño todavía no sobreajusta los datos, no es viable escogerla para el estudio de inferencia que se desea realizar a continuación, ya que supondría un coste computacional inasumible. Por este motivo, se decide escoger una red bayesiana con 2000 nodos, lo que acelerará en gran medida los tiempos de computación necesarios para realizar la inferencia.

#### 4.2.2. *Inferencia bayesiana*

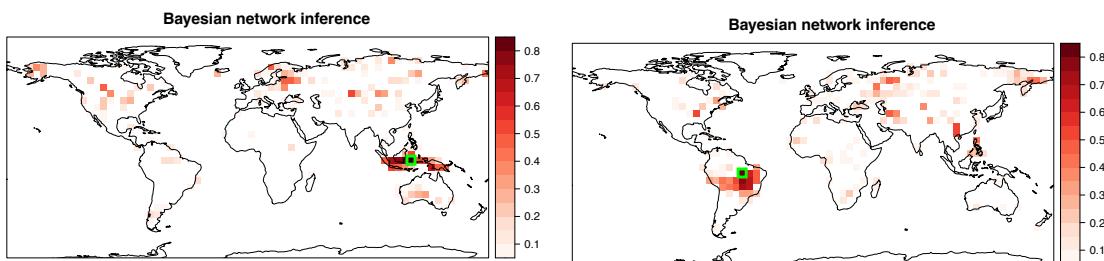
Para este estudio, explicado en la sección 3.2.5, se escogen varios nodos que resultan interesantes por tener alta centralidad o por su pertenencia a determinadas comunidades para utilizarlos como nodos de evidencia, de acuerdo con los resultados descritos en la Sec. 4.1.3. Estos nodos se localizan en: (i) África, ya que es una zona de gran conectividad y los nodos tienen un grado alto; (ii) Sudamérica, escogiendo un nodo con alto betweenness, del que se espera que se propague la evidencia dado que circula mucha información por él; y (iii) Indonesia, ya que pertenece a una comunidad que se ha detectado de forma consistente en CNs de varios umbrales. Así, la evidencia para estos nodos será que se desvíen una magnitud  $2\sigma$  respecto de su media (simulando un año extremo en cuanto a área quemada total), y se estudia cómo se propaga esa evidencia a todos los puntos del planeta estimando la probabilidad de que cada nodo se desvíe una cantidad  $\sigma$  tanto positiva como negativamente; en resumen, se estudia la probabilidad de que dado un año extremo en un punto, éste facilite o perjudique la probabilidad de incendios en otros puntos del planeta. Por ejemplo, dando evidencia en el nodo escogido en África se obtienen los resultados recogidos en la figura 4.12.



**Figura 4.12:** Diferencia (en unidades de desviación estándar) entre la probabilidad condicional (tras la propagación de la evidencia) y marginal (estado inicial)  $P(X_i \geq 1 | X_e = 2) - P(X_i \geq 1)$  (en rojo) y  $P(X_i \leq 1 | X_e = 2) - P(X_i \leq 1)$  (en azul). El nodo donde se da evidencia está marcado en verde.

Se observa que la evidencia se propaga con mucha intensidad en el África subsahariana indicando una gran correlación local y favoreciendo años con más incendios en esa zona, pero también se propaga en gran parte de Eurasia, Sudamérica y algunas zonas de Norteamérica, aunque con una intensidad significativamente menor. Esto se debe a que esta zona de África tiene una conectividad muy elevada (como ya se veía en la centralidad de grado para la CN en la figura 4.4) de manera que la información repercuta en muchas zonas diferentes del planeta, ya sea en mayor o en menor medida. Por otro lado, se observa que, si bien hay zonas en las que dada la evidencia se dificulta la aparición de incendios, la probabilidad de este suceso es muy pequeña (siendo el valor más alto de 0.25); el resultado es congruente con el obtenido por las CNs, en las que se observa que el número de enlaces con correlación negativa es mucho menor que aquellos con correlación positiva, de manera que es menos probable que un incendio en una zona penalice la aparición de otro incendio en una zona diferente. Además, al igual que en el caso de la CN, no se observan correlaciones negativas locales: esto tiene sentido, ya que si una zona está ardiendo lo más probable es que una zona contigua arda también (proceso de autocorrelación espacial anteriormente descrito en la Sec. 4.1.1), y no al contrario. Los resultados de propagación negativa aportan de nuevo escasa información, por lo que se decide omitirlos en los siguientes ejemplos; para estos, los resultados obtenidos se recogen en la figura 4.13.

Dada la evidencia de año extremo en el nodo señalado, localizado en Indonesia, se observa que la información se propaga localmente en todo el archipiélago, pero también a grandes distancias en la zona de Europa oriental. La relación entre estas dos zonas es un resultado que ya se ha observado en la búsqueda de comunidades de la CN en la



**Figura 4.13:** Diferencias entre las probabilidades (en unidades de desviación estándar) condicional (propagación de la evidencia) y marginal (estado inicial)  $P(X_i \geq 1 | X_e = 2) - P(X_i \geq 1)$  dada la evidencia en Indonesia (izquierda) y Sudamérica (derecha). Los nodos donde se da evidencia están marcados en verde.

sección 4.1.3, donde se obtiene una comunidad que agrupa estos dos conjuntos de nodos, y además aparece de manera consistente para diferentes redes. Por tanto, este nuevo resultado refuerza todavía más la idea de una relación muy estrecha entre Indonesia y Europa del Este. Por otro lado, en comparación con la propagación dada la evidencia en África, se observa que influye en menos zonas, pero lo hace con mayor intensidad, sobre todo en el centro de Norteamérica y en el centro de Asia.

Por último, dada la evidencia en un nodo de Sudamérica (localizado concretamente en la cuenca Amazónica), se observa que la información se propaga muy intensamente de manera local, pero que también existen repercusiones significativas a larga distancia, resaltando nodos particulares de Norteamérica y China. El hecho de que la información se transmita con mayor intensidad a largas distancias puede deberse a que sea un nodo con alta centralidad en términos de betweenness.

---

# 5

---

## Conclusiones generales y trabajo futuro

A la vista de los resultados presentados, en este capítulo se realiza, a modo de síntesis, un resumen de las principales conclusiones extraídas del estudio, así como una propuesta de las líneas de trabajo futuro que sugiere el trabajo realizado hasta el momento.

### 5.1. *Conclusiones*

1. El análisis de robustez de la red de correlación frente a la red aleatoria permite concluir que la base de datos de fuegos contiene una estructura espacial subyacente susceptible de ser analizada en profundidad mediante las técnicas empleadas en este estudio.
2. La aplicación de la red de correlación ha permitido tanto la identificación de los patrones de teleconexión más evidentes e inmediatos, como la definición de comunidades a través del algoritmo de agrupamiento. Las ventajas de esto han sido por un lado la posibilidad de formular hipótesis de partida sobre potenciales teleconexiones, posteriormente confirmadas con la red bayesiana, así como la identificación de los clústeres más fuertemente conectados, que proporcionan una indicación sobre potenciales nodos de la red bayesiana sobre los que realizar la propagación para la búsqueda de sincronicidades.
3. La concordancia en algunos de los resultados obtenidos entre la red de correlación y la red bayesiana revela la consistencia de ambas aproximaciones en la detección de patrones espaciales subyacentes en los datos. Se refuerza la idea de que, pese

al escaso número de datos para realizar el ajuste de la BN ( $n = 19$  años), se han obtenido resultados razonables que permiten extraer conocimiento útil de las redes desarrolladas, las cuales tienen un carácter probabilístico y no causal. Esto permite desarrollar confianza en los resultados arrojados por la red bayesiana.

4. En relación con lo anterior, las redes bayesianas parecen una opción preferible en el estudio de la sincronicidad, al ser capaces de eliminar las redundancias de las redes de correlación, evitando el problema de la selección del umbral de correlación. Además, la red bayesiana permite tareas de predicción, actuando como “sistema experto” de testeo de hipótesis mediante la presentación de evidencias y la propagación, lo que la dota de capacidad de generalización.
5. La principal desventaja encontrada con la red bayesiana no proviene de su potencial para la extracción de información relevante, sino del elevado coste computacional derivado del proceso de propagación de la evidencia e inferencia a lo largo de la red creada, debido a su complejidad estructural. Dada la imposibilidad de influir en el tamaño muestral para obtener un ajuste más robusto –y eventualmente una red más parsimoniosa–, una posibilidad de mejora de este aspecto pasaría por el análisis de patrones a una escala espacial aún más grosera (menos nodos), si bien esto podría redundar en la dilución de información relevante.
6. Los resultados revelan una sincronicidad significativa en la actividad anual de los incendios entre regiones distantes del planeta, tales como (i) África ecuatorial y Sudamérica, (ii) Indonesia, el Norte de Europa y Norteamérica/Alaska, o (iii) la Cuenca Amazónica y Filipinas, entre otros hallazgos ilustrados en el Capítulo 4. Cabe hipotetizar que dichas sincronicidades pueden venir en gran medida marcadas por teleconexiones climáticas de larga escala tales como la PDO (*Pacific Decadal Oscillation*) o ENSO (*El Niño Southern Oscillation*, ver p. ej. Chen et al., 2017), en las que la temperatura del océano juega un papel decisivo al influir sobre las principales vías físicas y biológicas que regulan el peligro de incendios (*fire weather*) y las propiedades de los combustibles (productividad primaria y contenido de humedad). Independientemente de los mecanismos causales de estos patrones, el resultado es en sí mismo valioso a la hora de abordar experimentos de predicción estacional o

modelización de incendios, y puede sentar la base para estudiar la influencia de los patrones de teleconexión atmosférica de manera más enfocada a regiones particulares (ver p. ej. Duffy et al., 2005; Harley et al., 2014).

### *5.2. Trabajo futuro*

En los resultados obtenidos para la construcción de la red bayesiana se observa que el bajo número de datos en las series temporales ( $t = 19$  años) tiene un impacto negativo en la elección de la red óptima. Por este motivo sería conveniente realizar un estudio utilizando como datos las series temporales mensuales aumentando; notar que dicho estudio debería extenderse a las redes de correlación en busca de unos resultados coherentes entre ambas redes y facilitando la interpretabilidad de los mismos.

Por otro lado, en este trabajo se han revelado patrones de sincronicidad robustos que, como se comenta en las conclusiones, pueden estar regulados por teleconexiones climáticas. Un próximo enfoque posible puede consistir en la construcción de un modelo híbrido que considere tanto la variable área quemada como otras variables auxiliares (por ejemplo, temperatura, humedad relativa, presión atmosférica etc.) de las que existen datos en toda la superficie del planeta, por lo que permitirían complementar la información incluyendo de forma explícita descriptores climáticos relacionados con dichos patrones de teleconexión. Es necesario tener en cuenta que este estudio únicamente es posible llevarlo a cabo a través de las redes bayesianas, gracias a su capacidad de combinar la información de variables distintas, propiedad que no presentan las redes de correlación.

Finalmente, la influencia de los patrones de teleconexión puede manifestarse de manera diferida en el tiempo en distintas regiones del planeta (ver p. ej. Chen et al., 2017), por lo que la metodología desarrollada en este TFM se puede extender fácilmente incluyendo retardos temporales en el análisis, en busca de patrones asíncronos de teleconexión y ventanas de oportunidad en el campo de la predicción estacional.

### *5.3. Reproducibilidad de los resultados*

Se ha realizado un esfuerzo adicional para garantizar la reproducibilidad de todos los resultados presentados en este trabajo, adoptando en lo posible los principios FAIR para

la gestión de datos científicos (Wilkinson et al., 2016). Como resultado, el código necesario para reproducir los resultados está disponible en un repositorio abierto de GitHub, <https://github.com/sergrabo/MastersThesis>, que sirven de ayuda en la total reproducibilidad de los resultados y su escrutinio, y constituyen un extenso material complementario a este estudio. Por otra parte, los datos pueden obtenerse de la manera detallada en el capítulo 2.

---

## Bibliografía

- Abatzoglou, J. T., A. P. Williams, L. Boschetti, M. Zubkova, y C. A. Kolden, 2018: Global patterns of interannual climate–fire relationships. *Global Change Biology*, **24**, 5164–5175, doi:10.1111/gcb.14405.
- Agarwal, A., L. Caesar, N. Marwan, R. Maheswaran, B. Merz, y J. Kurths, 2019: Network-based identification and characterization of teleconnections on different scales. *Scientific Reports*, **9**, 8808, doi:10.1038/s41598-019-45423-5.
- Albert, R. y A.-L. Barabási, 2002: Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**, 47–97.
- Archibald, S., C. E. R. Lehmann, J. L. Gomez-Dans, y R. A. Bradstock, 2013: Defining pyromes and global syndromes of fire regimes. *Proceedings of the National Academy of Sciences*, **110**, 6442–6447, doi:10.1073/pnas.1211466110.
- Bedia, J., J. Baño-Medina, M. N. Legasa, M. Iturbide, R. Manzanas, S. Herrera, A. Casanueva, D. San-Martín, A. S. Cofiño, y J. M. Gutiérrez, 2020: Statistical downscaling with the downscaleR package (v3.1.0): contribution to the VALUE intercomparison experiment. *Geoscientific Model Development*, **13**, 1711–1735, doi:10.5194/gmd-13-1711-2020.
- Bedia, J., S. Herrera, J. Gutierrez, A. Benali, S. Brands, B. Mota, y J. Moreno, 2015: Global patterns in the sensitivity of burned area to fire-weather: implications for climate change. *Agricultural and Forest Meteorology*, **214–215**, 369–379, doi:10.1016/j.agrformet.2015.09.002.
- Bivand, R. S., E. Pebesma, y V. Gomez-Rubio, 2013: *Applied spatial data analysis with R, Second edition*. Springer, NY.

- Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, y D. Hwang, 2006: Complex networks: Structure and dynamics. *Physics Reports*, **424**, 175–308.
- Bouckaert, R. R., 1995: *Bayesian Belief Networks: from Construction to Inference*. Ph.D. thesis, Utrecht University, The Netherlands.
- Bowman, D. M. J. S., J. Balch, P. Artaxo, y et al., 2011: The human dimension of fire regimes on Earth: The human dimension of fire regimes on Earth. *Journal of Biogeography*, **38**, 2223–2236, doi:10.1111/j.1365-2699.2011.02595.x.
- Bowman, D. M. J. S., J. K. Balch, P. Artaxo, y et al., 2009: Fire in the Earth System. *Science*, **324**, 481–484, doi:10.1126/science.1163886.
- Bowman, D. M. J. S., G. J. Williamson, J. T. Abatzoglou, C. A. Kolden, M. A. Cochrane, y A. M. S. Smith, 2017: Human exposure and sensitivity to globally extreme wildfire events. *Nature Ecology & Evolution*, **1**, 0058, doi:10.1038/s41559-016-0058.
- Buontempo, C., J.-N. Thépaut, y C. Bergeron, 2020: Copernicus Climate Change Service. *IOP Conference Series: Earth and Environmental Science*, **509**, 012005, doi:10.1088/1755-1315/509/1/012005.
- Cano, R., C. Sordo, y J. M. Gutiérrez, 2004: Applications of Bayesian networks in meteorology. *Advances in Bayesian networks*, Springer, 309–328.
- Castillo, E., J. M. Gutiérrez, y A. S. Hadi, 1997: *Expert Systems and Probabilistic Network Models*. Springer Publishing Company, Incorporated, New York, 1st edition.
- Chen, Y., D. C. Morton, N. Andela, G. R. van der Werf, L. Giglio, y J. T. Randerson, 2017: A pan-tropical cascade of fire driven by El Niño/Southern Oscillation. *Nature Climate Change*, **7**, 906–911, doi:10.1038/s41558-017-0014-8.
- Chuvieco, E., L. Giglio, y C. Justice, 2008: Global characterization of fire activity: toward defining fire regimes from Earth observation data. *Global Change Biology*, **14**, 1488–1502, doi:10.1111/j.1365-2486.2008.01585.x.

- Chuvieco, E. y C. Justice, 2010: Relations Between Human Factors and Global Fire Activity. *Advances in Earth Observation of Global Change*, E. Chuvieco, J. Li, y X. Yang, eds., Springer Netherlands, Dordrecht, 187–199.
- Cofiño, A., J. Bedia, M. Iturbide, M. Vega, S. Herrera, J. Fernández, M. Frías, R. Manzanas, y J. Gutiérrez, 2018: The ECOMS User Data Gateway: Towards seasonal forecast data provision and research reproducibility in the era of Climate Services. *Climate Services*, **9**, 33–43, doi:10.1016/j.cliser.2017.07.001.
- Csardi, G. y T. Nepusz, 2006: The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.
- Dijkstra, H. A., E. Hernández-García, C. Masoller, y M. Barreiro, 2019: *Networks in Climate*. Cambridge University Press, Cambridge.
- Duffy, P. A., J. E. Walsh, J. M. Graham, D. H. Mann, y T. S. Rupp, 2005: Impacts of large-scale atmospheric–ocean variability on alaskan fire season severity. *Ecological Applications*, **15**, 1317–1330, doi:<https://doi.org/10.1890/04-0739>.
- Ebert-Uphoff, I. y Y. Deng, 2012: Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate*, **25**, 5648–5665, doi:10.1175/JCLI-D-11-00387.1.
- Frías, M. D., M. Iturbide, R. Manzanas, J. Bedia, J. Fernández, S. Herrera, A. S. Cofiño, y J. M. Gutiérrez, 2018: An R package to visualize and communicate uncertainty in seasonal climate prediction. *Environmental Modelling & Software*, **99**, 101–110, doi:10.1016/j.envsoft.2017.09.008.
- Geiger, D. y D. Heckerman, 1994: Learning Gaussian Networks. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, 235–243.
- Graafland, C. E., 2022: *Probabilistic Networks In Complex Systems*. Ph.D. thesis, University of Cantabria, Spain.
- Graafland, C. E., J. M. Gutiérrez, J. M. López, D. Pazó, y M. A. Rodríguez, 2020: The probabilistic backbone of data-driven complex networks: an example in climate. *Scientific Reports*, **10**, 11484, doi:10.1038/s41598-020-67970-y.

- Gutiérrez, J., R. Cano, A. Cofiño, y C. Sordo, 2004: *Redes probabilísticas y neuronales en las ciencias atmosféricas*. Centro de Publicaciones, Ministerio de Medio Ambiente, Madrid, Spain, in Spanish.
- Hantson, S., D. I. Kelley, A. Arneth, y et al., 2020: Quantitative assessment of fire and vegetation properties in simulations with fire-enabled vegetation models from the Fire Model Intercomparison Project. *Geoscientific Model Development*, **13**, 3299–3318, doi:10.5194/gmd-13-3299-2020.
- Harley, G. L., H. D. Grissino-Mayer, S. P. Horn, y C. Bergh, 2014: Fire synchrony and the influence of pacific climate variability on wildfires in the florida keys, united states. *Annals of the Association of American Geographers*, **104**, 1–19.
- Iturbide, M., J. Bedia, S. Herrera, J. Baño-Medina, J. Fernández, M. Frías, R. Manzanas, D. San-Martín, E. Cimadevilla, A. Cofiño, y J. Gutiérrez, 2019: The R-based climate4R open framework for reproducible climate data access and post-processing. *Environmental Modelling & Software*, **111**, 42–54, doi:10.1016/j.envsoft.2018.09.009.
- Koller, D. y N. Friedman, 2009: *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge, Massachusetts.
- Larrañaga, P., B. Sierra, M. J. Gallego, M. J. Michelena, y J. M. Picaza, 1997: Learning Bayesian Networks by Genetic Algorithms: A Case Study in the Prediction of Survival in Malignant Skin Melanoma. *Proceedings of the 6th Conference on Artificial Intelligence in Medicine in Europe (AIME'97)*, Springer, 261–272.
- Lentile, L. B., Z. A. Holden, A. M. S. Smith, y et al., 2006: Remote sensing techniques to assess active fire characteristics and post-fire effects. *International Journal of Wildland Fire*, **15**, 319, doi:10.1071/WF05097.
- Lizundia-Loiola, J., M. Pettinari, E. Chuvieco, T. Storm, y J. Gomez-Dans, 2018: CCI ECV Fire Disturbance: D2.1.3 Algorithm Theoretical Basis Document – MODIS. Technical report, European Space Agency (ESA), version 2.0.
- Newman, M. E. J., 2010: *Networks: an introduction*. Oxford University Press, New York.

- Pausas, J. G. y J. E. Keeley, 2021: Wildfires and global change. *Frontiers in Ecology and the Environment*, **19**, 387–395, doi:10.1002/fee.2359.
- Pausas, J. G. y E. Ribeiro, 2013: The global fire-productivity relationship. *Global Ecology and Biogeography*, **22**, 728–736, doi:10.1111/geb.12043.
- R Core Team, 2019: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- URL <https://www.R-project.org/>
- Scutari, M., 2010: Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, **35**, 1–22.
- Scutari, M., C. E. Graafland, y J. M. Gutiérrez, 2019: Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, **115**, 235–253, doi:10.1016/j.ijar.2019.10.003.
- Tsonis, A. A. y P. J. Roebber, 2004: The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, **333**, 497–504, doi:<https://doi.org/10.1016/j.physa.2003.10.045>.
- Valle, M., 2021: *Automated wildfire season detection at a global scale: Application for the development of a predictive system of fire activity*. Master's thesis, Facultad de Ciencias, Universidad de Cantabria.
- URL [https://repositorio.unican.es/xmlui/bitstream/handle/10902/25076/2021\\_TFM\\_MarcosValle.pdf?sequence=1](https://repositorio.unican.es/xmlui/bitstream/handle/10902/25076/2021_TFM_MarcosValle.pdf?sequence=1)
- Verma, T. y J. Pearl, 1991: Equivalence and synthesis of causal models. *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, Elsevier Science Inc., New York, UAI '90, 255–270.
- Wickham, H., 2016: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, y et al., 2016: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018, doi:10.1038/sdata.2016.18.



---

## Anexo A: Material suplementario

*Descriptores de las conexiones: valores según umbral*

Umbral $\tau_c$	Enlaces totales	Enlaces Positivos	Enlaces Negativos
0.00	206969	112103	94866
0.05	175723	96318	79405
0.10	145102	80687	64415
0.15	117292	66284	51008
0.20	91590	52633	38957
0.25	69519	40628	28891
0.30	51059	30401	20658
0.35	35323	21596	13727
0.40	23690	14971	8719
0.45	15597	10133	5464
0.50	9804	6608	3196
0.55	5715	3986	1729
0.60	3209	2351	858
0.65	1719	1336	383
0.70	842	688	154
0.75	387	341	46
0.80	155	142	13
0.85	52	50	2
0.90	15	15	0
0.95	2	2	0

**Tabla A1:** Número de enlaces para los diferentes umbrales de correlación considerados

Umbral $\tau_c$	Distancia media total	Distancia media positiva	Distancia media negativa
0.00	8572.47	8438.92	8730.29
0.05	8545.51	8395.15	8727.89
0.10	8509.82	8341.80	8720.29
0.15	8463.01	8278.51	8702.77
0.20	8416.03	8206.75	8698.77
0.25	8372.04	8129.56	8713.03
0.30	8286.44	8011.38	8691.22
0.35	8198.03	7862.81	8725.42
0.40	8016.46	7614.57	8706.52
0.45	7810.45	7347.56	8668.88
0.50	7534.54	7028.88	8580.04
0.55	7138.81	6468.58	8683.96
0.60	6610.33	5913.46	8519.82
0.65	6043.79	5322.38	8560.28
0.70	5038.59	4269.18	8475.97
0.75	4197.48	3637.42	8349.24
0.80	3569.20	2967.96	10136.65
0.85	1584.92	1286.19	9053.39
0.90	906.48	906.48	
0.95	3371.17	3371.17	

**Tabla A2:** Distancia media de los enlaces para los diferentes umbrales de correlación considerados

Umbral $\tau_c$	Máxima distancia total	Máxima distancia positiva	Máxima distancia negativa
0.00	20037.44	20037.44	20037.44
0.05	20037.44	20037.44	20037.44
0.10	20037.44	20037.44	20037.44
0.15	20037.44	20037.44	20037.44
0.20	20037.44	20037.44	20037.44
0.25	20037.44	20037.44	20037.44
0.30	20037.44	20037.44	20037.44
0.35	20037.44	20037.44	20037.44
0.40	20037.44	20037.44	20025.06
0.45	20037.44	20037.44	20025.06
0.50	20037.44	20037.44	19582.97
0.55	20030.35	20030.35	19582.97
0.60	20030.35	20030.35	19582.97
0.65	19582.97	19536.31	19582.97
0.70	19582.97	18933.22	19582.97
0.75	19582.97	18933.22	19582.97
0.80	19294.84	18933.22	19294.84
0.85	10662.44	10090.18	10662.44
0.90	6484.71	6484.71	
0.95	6484.71	6484.71	

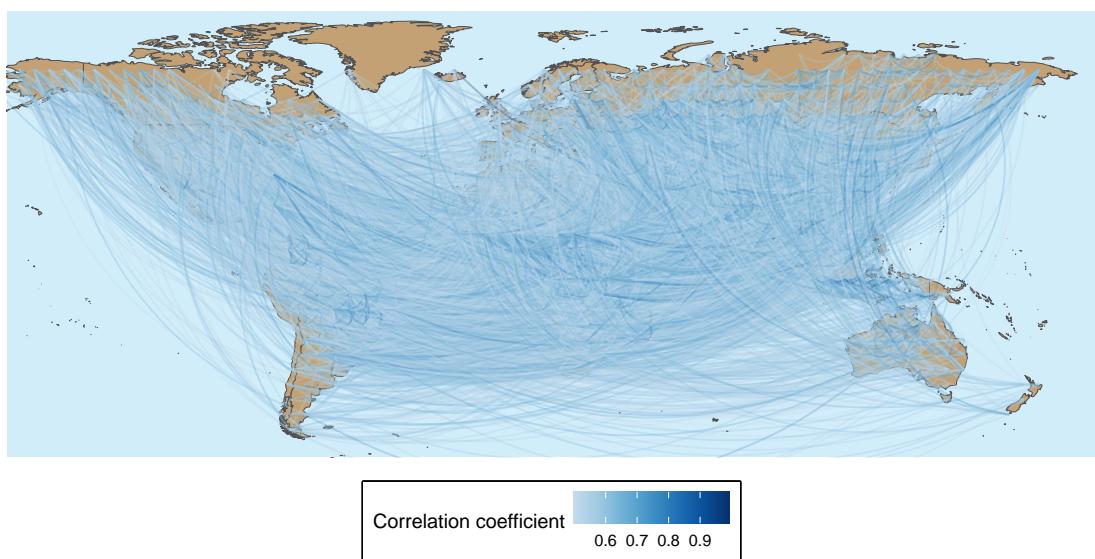
**Tabla A3:** Distancia máxima de los enlaces para los diferentes umbrales de correlación considerados

Umbral $\tau_c$	Mínima distancia total	Mínima distancia positiva	Mínima distancia negativa
0.00	167.83	167.83	167.83
0.05	167.83	167.83	167.83
0.10	167.83	167.83	167.83
0.15	167.83	167.83	213.55
0.20	167.83	167.83	426.76
0.25	167.83	167.83	469.84
0.30	167.83	167.83	469.84
0.35	167.83	167.83	469.84
0.40	167.83	167.83	553.04
0.45	167.83	167.83	554.71
0.50	167.83	167.83	554.71
0.55	167.83	167.83	554.71
0.60	213.55	213.55	555.66
0.65	213.55	213.55	1194.11
0.70	213.55	213.55	1194.11
0.75	213.55	213.55	1424.26
0.80	213.55	213.55	5873.37
0.85	213.55	213.55	7444.34
0.90	257.62	257.62	
0.95	257.62	257.62	

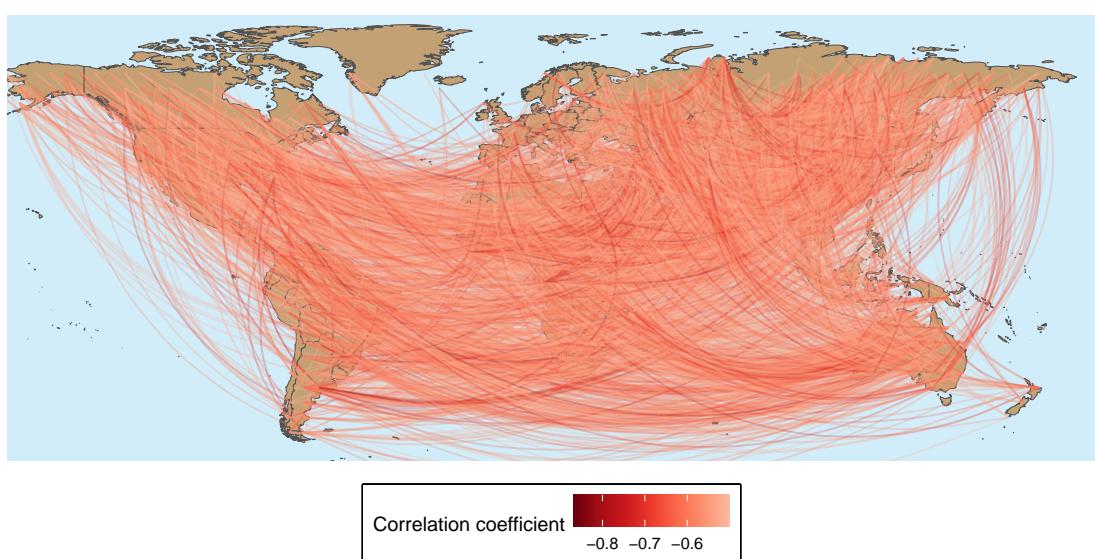
**Tabla A4:** Distancia mínima de los enlaces para los diferentes umbrales de correlación considerados

*Representación espacial de diferentes CNs*

Positive Spatial Network for  $\tau_c = 0.5$

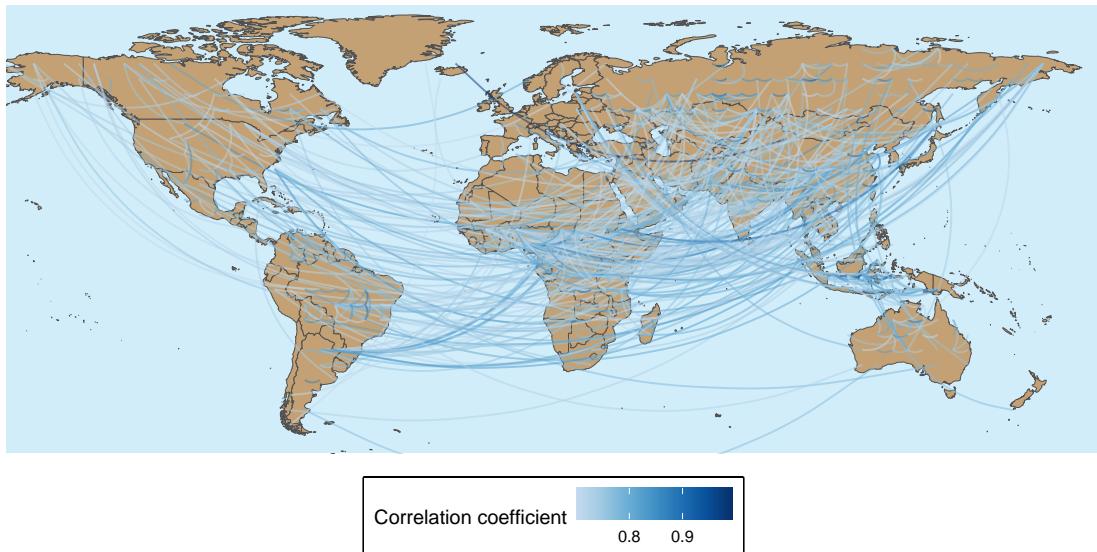


Negative Spatial Network for  $\tau_c = 0.5$

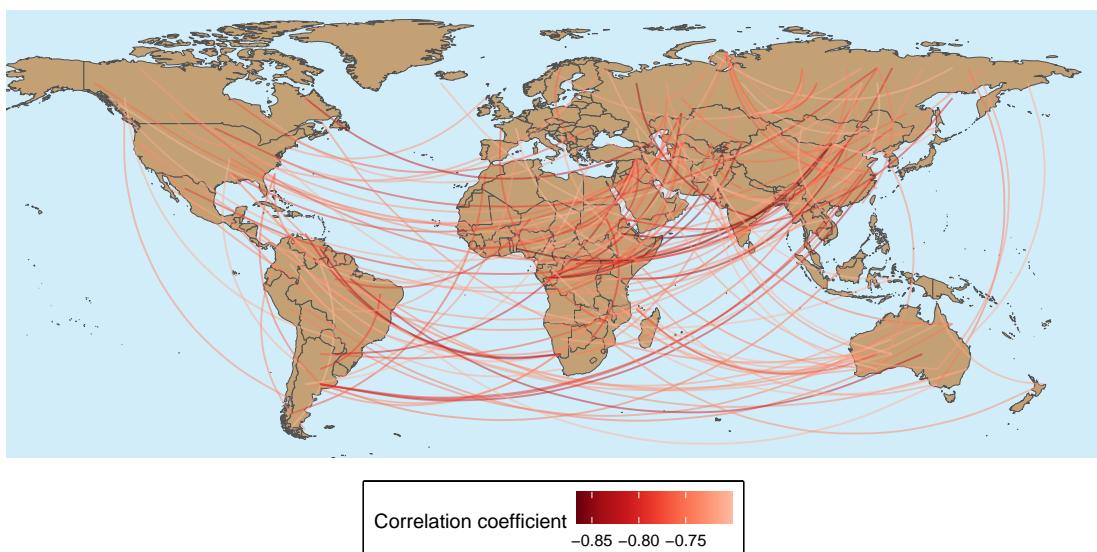


**Figura A1:** Red de correlación pesada para un umbral  $\tau_c = 0.5$

Positive Spatial Network for  $\tau_c = 0.7$

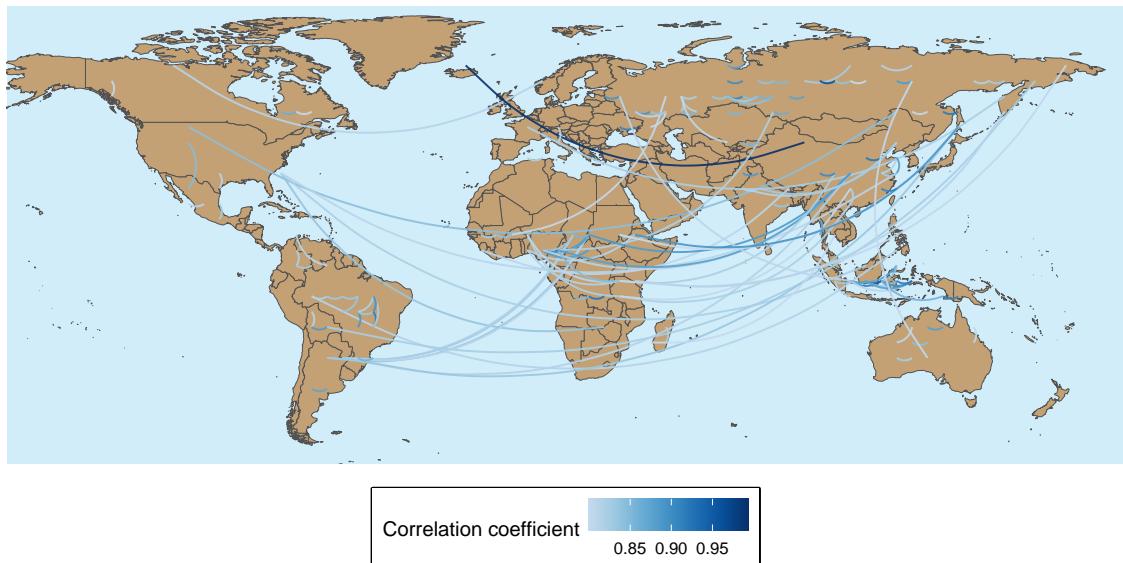


Negative Spatial Network for  $\tau_c = 0.7$

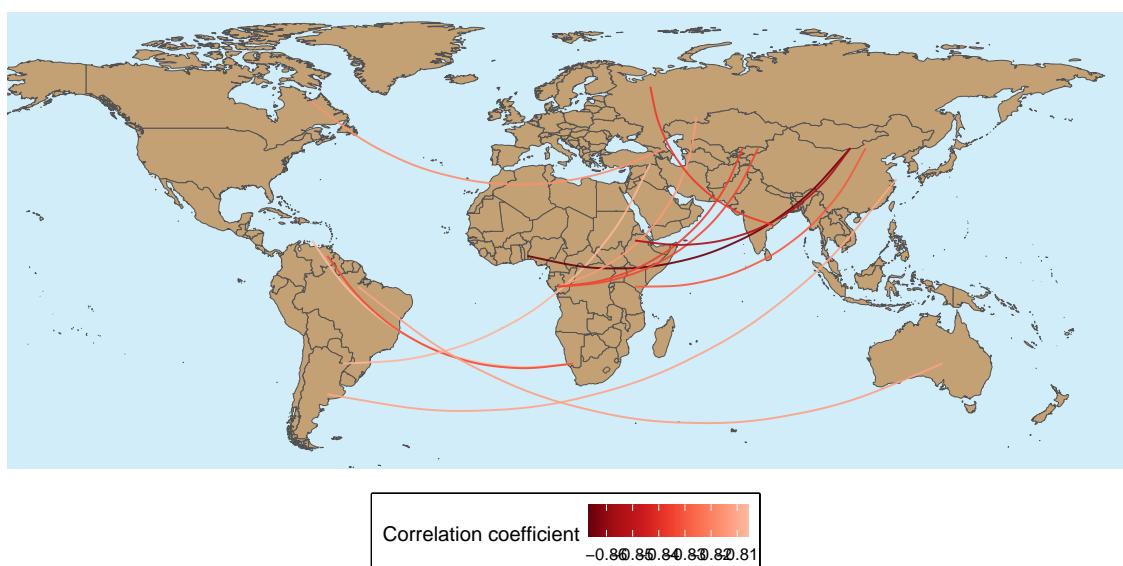


**Figura A2:** Red de correlación pesada para un umbral  $\tau_c = 0.7$

Positive Spatial Network for  $\tau_c = 0.8$



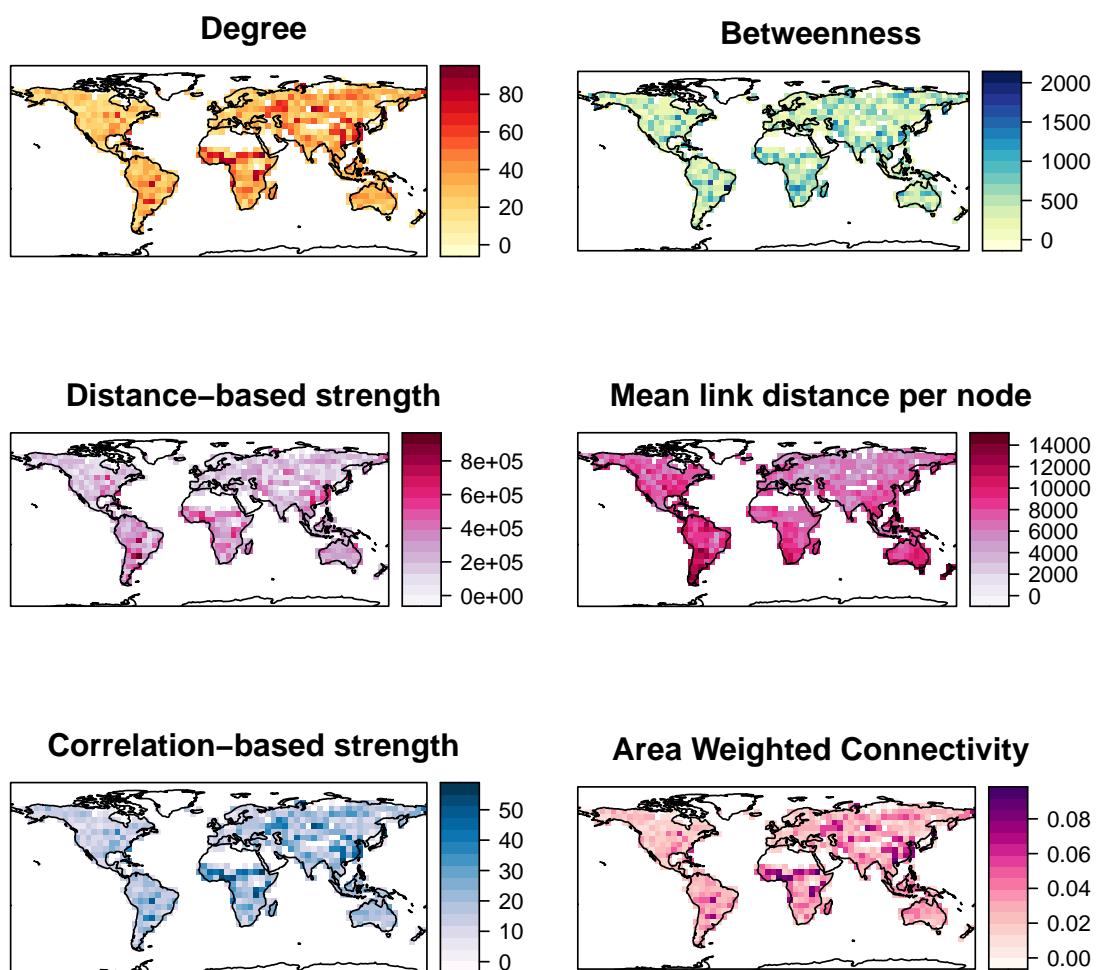
Negative Spatial Network for  $\tau_c = 0.8$



**Figura A3:** Red de correlación pesada para un umbral  $\tau_c = 0.8$

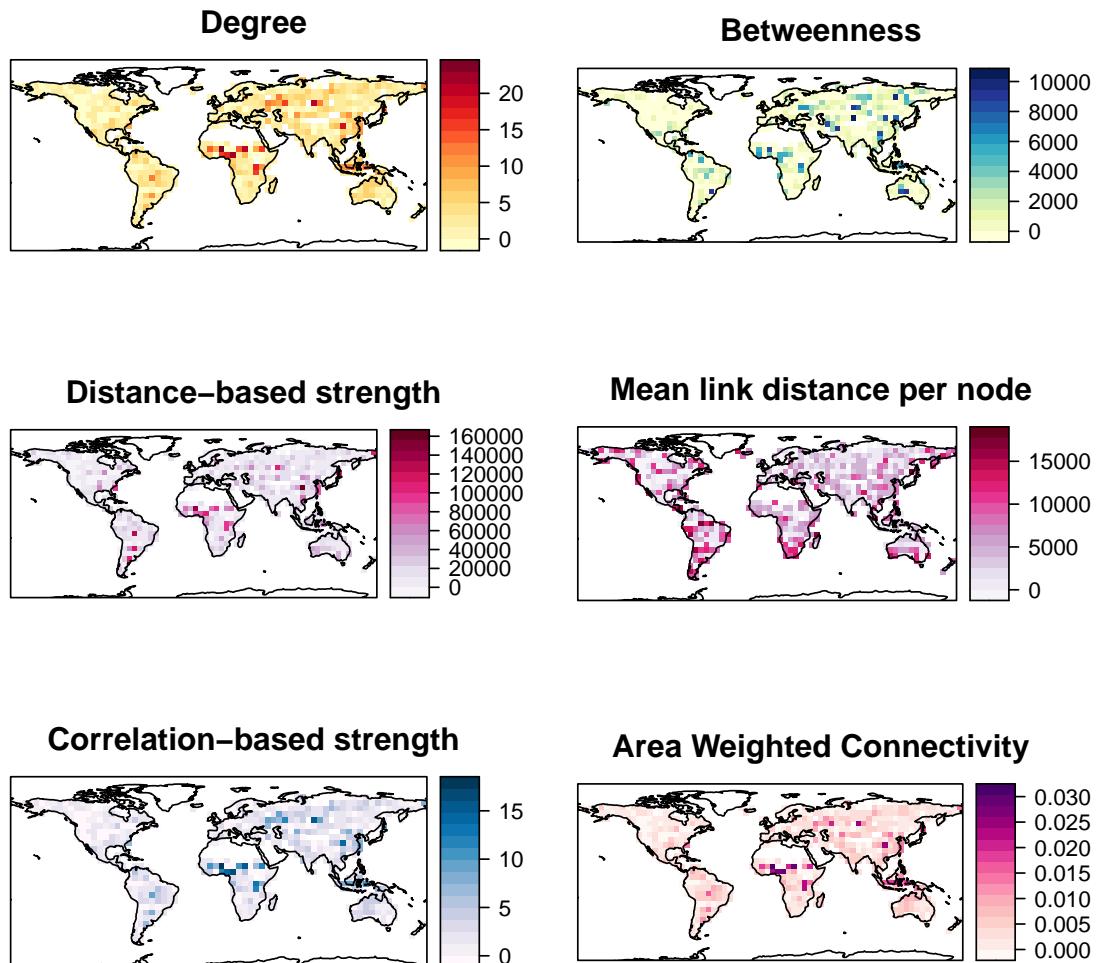
*Medidas de centralidad para diferentes CNs*

Centrality measures for  $\tau_c = 0.5$



**Figura A4:** *Medidas de centralidad para una CN con umbral  $\tau_c = 0.5$*

## Centrality measures for $\tau_c = 0.7$



**Figura A5:** *Medidas de centralidad para una CN con umbral  $\tau_c = 0.7$*

Medida de centralidad	Descripción
Grado	Número de enlaces del nodo
Betweenness	Número de geodésicas que atraviesan el nodo
Strength basado en correlación	Suma de los pesos $w_c$ de los enlaces del nodo
Strength basado en distancia	Suma de los pesos $w_d$ de los enlaces del nodo
AWC	Proporción de área a la que se conecta el nodo
Distancia media	Media de las distancias geográficas de los enlaces del nodo

**Tabla A5:** *Resumen del tipo de información que aporta cada medida de centralidad*