COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

Modelación Estadística de la Génesis de los Ciclones Tropicales en el Atlántico Norte

Julio César Buendía Espinoza

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO,TEXCOCO, EDO. DE MÉXICO
2016

La presente tesis titulada: **Modelación Estadística de la Génesis de los Ciclones Tropicales en el Atlántico Norte**, realizada por el alumno: **Julio César Buendía Espinoza**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTOR EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO Pérez Rdz.

Dr. Paulino Pérez Rodríguez

ASESOR

Dr. Michel Rosengaus Moshinsky

ASESOR

Dr. Sergio Pérez Elizalde

ASESOR

Dr. Malaquías Peña Méndez

ASESOR

Dr. Adolfo A. Exebio García

Modelación Estadística de la Génesis de los Ciclones Tropicales en el Atlántico Norte

Julio César Buendía Espinoza

Colegio de Postgraduados, 2016

Resumen

Los modelos de mezclas son una herramienta flexible para el agrupamiento no supervisado que ha encontrado popularidad en una amplia gama de áreas de investigación. Cuando se desarrollan en el marco bayesiano, los modelos de mezclas proporcionan un medio natural para la captura y la propagación de la incertidumbre en los diferentes aspectos de una solución de agrupación, indiscutiblemente resulta un análisis más rico de la población bajo estudio.

Este trabajo de investigación tiene como primer objetivo dar a conocer el uso de métodos estadísticos clásicos y bayesianos no paramétricos mediante los modelos de mezclas para agrupar datos. En particular, se examinan tres variantes comunes en el modelo de mezclas, es decir, *i) "Mezclas Finitas Gaussianas" (MMG)*, *ii) "Mezclas de Procesos Dirichlet" (DPMM)* y *iii) "Mezclas de Modelos de Regresión Lineal" (MMR)*. Más allá del desarrollo y la aplicación de estos modelos, esta tesis también se centra en proponer un estadístico de prueba para comparar los grupos. Para hacer frente a estos objetivos, se consideran tres estudios de caso con los datos de génesis de los ciclones tropicales en el Atlántico Norte. El primero es utilizar MMG para determinar tanto el número de grupos como sus centroides en la cuenca oceánica del Atlántico Norte. El segundo caso es similar al anterior pero con la aplicación del DPMM. Y finalmente, el tercer caso es la aplicación del MMR para determinar cómo la temperatura de la superficie del mar influye en la ubicación de los centroides.

Los resultados muestran que las diferencias en las estimaciones de densidad y las representaciones predictivas de los métodos analizados conducen a un comportamiento diferente del modelo en los mismos datos. Sin embargo; los DPMM son más convenientes de utilizar, ya que incorpora un procedimiento para determinar el número de componentes en la mezcla. De acuerdo a los DPMM existen tres grupos de génesis dentro de la cuenca oceánica del Atlántico Norte; en contraste, de acuerdo al MMG y al MMR, solo existen dos regiones de génesis. Respecto a los centroides de los grupos, los resultados muestran que éstos han experimentado cambios en su localización tal y como lo mencionan Mori *et al.* (2013) y se están desplazando hacia las zonas de mayor temperatura del océano.

Palabras clave: Mezclas, algoritmo de Gibbs en Bloques, pruebas de hipótesis, bootstrap, algoritmo EM, ciclones tropicales.

Statistical Modeling of the Genesis of Tropical Cyclones in the North Atlantic.

Julio César Buendía Espinoza

Colegio de Postgraduados, 2016

Abstract

Mixture models are a flexible tool for unsupervised clustering that have found popularity in a vast array of research areas. Furthermore, when developed in the Bayesian framework, mixture models provide a natural means for capturing and propagating uncertainty in different aspects of a clustering solution, arguably resulting in richer analyses of the population under study.

This research work aims mainly to publicize the use of statistical methods of clustering and nonparametric Bayesian using mixture models for grouping data. In particular, three common variants are discussed in the model mixtures, i.e. *i) "Finite Gaussian Mixtures" (MMF)*, *ii) "Dirichlet Process Mixtures" (DPMM)* and *iii) "Mixtures of Linear Regression Models" (MMR)*. Beyond the development and application of these models, this thesis also focuses on proposing a statistical test to compare the groups. To address these objectives, they considered three case studies with data genesis of tropical cyclones in the North Atlantic. To address these objectives, three case studies with data from genesis of tropical cyclones in the North Atlantic are considered. The first is to use *MMF* to determine both the number of groups and their centroids in the North Atlantic oceanic basin. The second case is similar to the above but this time with the application of a *DPMM*. And finally, the third case is the application of a *MMR* to determine how the temperature of the sea surface influences the location of the centroid.

The results show that differences in density estimates and the predictive representations of the analyzed methods lead to different behavior model on the same data. However; the *DPMM* are more convenient to use, because automatically determines the number of components in the mixture. According to the *DPMM*, there are three groups of genesis within the North Atlantic oceanic basin; in contrast, according to the *MMG* and *MMR*, there are only two regions of genesis. With respect to the centroids of the groups, the results show that these have experienced changes in their location as mentioned Mori *et al.* (2013) and are moving towards areas of higher ocean temperatures.

Key words: Mixtures, Gibbs algorithm in blocks, hypothesis tests, bootstrap, EM algorithm, tropical cyclones.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado durante la realización de mis estudios de postgrado.

Al Colegio de Postgraduados por haberme brindado la oportunidad de seguir mi formación académica.

A los integrantes de mi Consejo Particular:

Dr. Paulino Pérez Rodríguez por la dirección de este trabajo, pero principalmente por su apertura al diálogo, evitando imponer y manipular y generando una conversación.

Dr. Sergio Pérez Elizalde por sus observaciones y comentarios a esta investigación.

Dr. Michel Rosengaus Moshinsky por sus observaciones y comentarios en la realización de este trabajo.

Dr. Malaquías Peña Méndez por sus comentarios, observaciones y revisión detallada en el desarrollo de este trabajo de investigación.

Dr. Adolfo Exebio García por sus observaciones y comentarios a esta investigación.

Dr. Juan Manuel González Camacho por sus observaciones y comentarios en el desarrollo de este trabajo.

A mis profesores, compañeros de clases y todos aquellos que de alguna u otra manera fueron parte de esta aventura llamada doctorado.

DEDICATORIA

A mis **padres:** Manuel Buendia De La Rosa y Bertha Espinoza Montiel por todo el cariño y apoyo brindado para cumplir otro reto más en mi vida.

Gracias Padre y Madre, los amo.

A mis **hermanos:** Lidia, Patricia, Héctor, Angélica, Marco Antonio y Edgar Hugo por lo que representan para mí y por ser parte importante de una familia unida.

A mi **esposa:** Elisa del Carmen y mis **hijas:** Romina y Regina con mucho amor porque ellas han dado razón a mi vida, las amo.

A toda mi **familia** que es lo mejor y más valioso que Dios me ha dado.

"No permitas que nadie
te diga que eres incapaz
de hacer algo.

Si tienes un sueño debes
conservarlo.

Si quieres algo ve tras ello y
punto.

¿Sabes?, la gente que no logra
conseguir sus sueños suele
decirles a los demás que
tampoco cumplirán lo suyos".

Will Smith.
En Busca de la Felicidad.

Índice

Índice de tablas

Índice de figuras

Índice de figuras

Capítulo 1

Introducción

El debate sobre el cambio climático con frecuencia combina temas de la ciencia y la política. Debido a sus impactos significativos y vehementes, el análisis de eventos extremos es un lugar frecuente de dicha fusión. Linda Mearns, del Centro Nacional para la Investigación Atmosférica (NCAR), acertadamente caracteriza este contexto: "Hay una presión sobre los climatólogos para decir algunas cosas acerca de los extremos, ya que son tan importantes, pero eso puede ser muy peligroso sí realmente no se sabe la respuesta" (Henson, 2005). Este trabajo se centra en un tipo particular de evento extremo -los ciclones tropicales que se forman en las aguas del Atlántico Norte- en el contexto del calentamiento global.

Las coacciones de los vínculos entre el calentamiento global y los efectos de los huracanes son prematuras por dos razones. En primer lugar, no se ha establecido ninguna conexión entre las emisiones de gases de efecto invernadero y el comportamiento observado de los huracanes (Houghton *et al.*, 2001, Walsh, 2004). Emanuel (2005) sugiere tal conexión, pero no en modo definitivo. En el futuro, esta conexión puede ser establecida [por ejemplo, en el caso de las observaciones de Emanuel (2005) o las proyecciones Knutson y Tuleya (2004)] o hecha en el contexto de otras métricas de duración e intensidad de los ciclones tropicales que continúan para ser cercanamente examinados. En segundo lugar, los estudios de Henderson-Sellers *et al.* (1998) citados también por Pielke Jr. *et al.* (2005) señalan que existe un consenso científico de que los cambios futuros en la intensidad de los huracanes probablemente serán pequeños en el contexto de la variabilidad observada, mientras que el problema científico de la ciclogénesis se han centrado por una parte en el desarrollo de modelos de predicción sobre el número ciclones tropicales por temporada, debido al aumento de la frecuencia y la intensidad de los huracanes del Atlántico en los últimos años (Webster *et al.*, 2005), y por otra parte en el desarrollo de modelos estadísticos que simulan sitios de génesis alrededor de los sitios muestreados de la génesis histórica (Hall y Jewson, 2007, McDonnell y Holbrook, 2004, Rumpf *et al.*, 2007, Tippett *et al.*, 2011, Werner y Holbrook, 2011, Yonekura y Hall, 2011).

1. Introducción

Recientemente, Mori *et al.* (2013) proyectaron el impacto del calentamiento global sobre los centroides de la ciclogénesis de las diferentes cuencas oceánicas para el siglo XXI. Encotraron que los centroides se desplazarán hacia el centro de las cuencas y que los cambios futuros en las condiciones dinámicas y termodinámicas en los océanos influirán en la frecuencia de la ciclogénesis. Además que los ciclones que se desarrollan en la parte central del océano durarán más tiempo debido a que las temperaturas de la superficie del mar son más cálidas que las de las orillas (Chan, 2007, Yokoi y Takayabu, 2009), y que los cambios en su intensidad estarán más relacionados con el desplazamiento de los centroides que con el cambio de la temperatura del mar.

Bajo este contexto, en este trabajo de investigación se pretende aplicar el uso métodos estadístico de agrupación y métodos estadísticos bayesianos no paramétricos mediante modelos de mezclas, por una parte para determinar los cambios temporales y espaciales de los centroides de las regiones de génesis de los ciclones tropicales en la cuenca oceánica del Atlántico Norte, durante los últimos 60 años. Y por otra, para determinar si la temperatura de la superficie del mar influye en la ubicación de los mismos.

Los modelos de mezclas son modelos probabilísticos para representar la presencia de sub poblaciones dentro de una misma población. La partición de datos en sub grupos homogéneos es una tarea trascendental, debido a su alta dimensión y heterogeneidad significativa en las respuestas observadas. Además, los modelos de mezclas son utilizados para crear inferencias estadísticas, aproximaciones y predicciones acerca de las propiedades de las sub poblaciones a partir de las observaciones de la población estudiada sin necesidad de información que identifique a la sub población.

La estructura del trabajo es como sigue. En el Capítulo 2 se presentan los objetivos del trabajo. En el Capítulo 3 se describen los métodos estadístico de agrupación y bayesianos no paramétricos mediante los modelos de mezclas, incluyendo las razones de su uso, especificaciones de los modelos y sus métodos de inferencia. Específicamente, en esta tesis se tomaron en cuenta los *"Modelos de Mezclas finitas Gaussianas" (MMG)*, los *"Modelos de Mezclas de Procesos Dirichlet" (DPMM)* y los *"Modelos de Mezclas Gaussianas de Modelos de Regresión Lineal" (MMR)*. En el Capítulo 4 se describen los datos utilizados en el presente trabajo. En los Capítulos 5 y 6, se modelan la mismos puntos de génesis de los ciclones mediante diferentes modelos, MMG y DPMM, con el propósito de determinar la ubicación de los centroides en la cuenca del Atlántico Norte. El número de componentes de la mezcla en el primer se determina mediante un método heurístico y en en el segundo automáticamente. Los datos de los puntos de génesis fueron obtenidos de las "mejores trayectorias" o IBTrACS (por siglas en inglés: International Best Track Archive for Climate Stewardship, se puede consultar en: `https://www.ncdc.noaa.gov/ibtracs/index.php?name=ibtracs-data-access`). En el Capítulo 7 considera la aplicación de un modelo de MMR para simular el efecto de la temperatura de la superficie del mar sobre la ubicación de los centroides de los ciclones Tropicales. Para finalizar, el Capítulo 8 resume las principales aportaciones de esta tesis y trabajos futuros por realizar.

Capítulo 2

Objetivos

2.1. Objetivo General

Examinar y mostrar el uso de métodos estadísticos de agrupación y bayesianos no paramétricos mediante modelos de mezclas para agrupar datos de génesis de ciclones tropicales en el Atlántico Norte.

2.2. Objetivos Particulares

- Comparar las estimaciones de las funciones de densidad y las representaciones predictivas entre los métodos estadísticos de agrupación y los métodos estadístico bayesianos no paramétricos en los datos de génesis de ciclones tropicales del Atlántico Norte.
- Modelar el efecto de la temperatura de la superficie del mar en la génesis de los ciclones tropicales mediante métodos estadísticos de agrupación.
- Implementar un estadístico de prueba mediante la técnica re-muestro paramétrico para comparar en tiempo y espacio las funciones de densidad de probabilidades de las regiones ciclogenéticas determinadas en ambos métodos.

Capítulo 3

Modelos Estadísticos

En este capítulo, se presentan tres técnicas estadística que permiten analizar la ciclogénesis y están basadas en: Modelos de Mezclas *Gaussianas*, Modelos de Mezclas de Procesos *Dirichlet* y Mezclas *Gaussianas* para Modelos de Regresión Lineal.

3.1. Modelos de Mezclas Gaussianas

Esta investigación se centra en los modelos de mezclas paramétricas que modelan la densidad de probabilidad asociada a un conjunto de datos como una superposición lineal de funciones denominadas núcleos.

3.1.1. Modelo de Mezclas Gaussianas Finitas (MMG)

Una variable aleatoria $\boldsymbol{Y}$ d-dimensional sigue una distribución de mezclas finitas cuando su función de densidad de probabilidad $p(\boldsymbol{Y}|\boldsymbol{\phi})$ se expresa como una suma ponderada de funciones de densidad de probabilidad denominadas núcleos. Cuando esas funciones núcleo son Gaussianas se llama "*Mezcla Gaussiana*". Luego entonces la distribución de una variable aleatoria $\mathbf{Y}$ cuya función de densidades es:

$$p\left(\boldsymbol{Y}|\boldsymbol{\phi}\right) = \sum_{j=1}^{K} \pi_j \cdot p\left(\boldsymbol{y}|\boldsymbol{\theta}_j\right), \text{ con } j = 1, \ldots, K \text{ y } \sum_{j=1}^{K} \pi_j = 1, \tag{3.1}$$

se denomina distribución de mezclas finitas de K componentes, donde $\boldsymbol{\pi}_j$ son las probabilidades de pertenencia a cada uno de los núcleos, y $\boldsymbol{\theta}_j$ es el conjunto de paráme-

tros que describe a cada núcleo. En el caso de mezclas Gaussianas $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ está conformado por la media y la matriz de varianzas y covarianzas que caracterizan a la distribución normal. Redner y Walker (1984) muestran una revisión detallada de las técnicas de máxima verosimiltud para estimar estos parámetros en modelos de mezclas. En el caso de mezclas finitas Gaussianas, la función de verosimiltud y log-verosimilitud está dada por:

$$\begin{aligned} L\left(\boldsymbol{\phi}|\boldsymbol{Y}\right) &= \prod_{i=1}^{n}\left[\sum_{j=1}^{K}\pi_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right], \\ l\left(\boldsymbol{\phi}|\boldsymbol{Y}\right) &= \log L\left(\boldsymbol{\phi}|\boldsymbol{Y}\right) = \log\left\{\prod_{i=1}^{n}\sum_{j=1}^{K}\pi_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right\}, \\ &= \sum_{i=1}^{n}\log\left\{\sum_{j=1}^{K}\pi_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right\}, \end{aligned} \tag{3.2}$$

donde $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ son realizaciones de la variable $\boldsymbol{Y}$ bajo el supuesto de que son independientes e idénticamente distribuidas. El estimador de máxima verosimiltud se obtiene al maximizar $l(\boldsymbol{\phi}|\boldsymbol{Y})$, es decir:

$$\hat{\boldsymbol{\phi}} = \arg\max_{\phi}\left\{l(\boldsymbol{\phi}|\boldsymbol{Y})\right\}.$$

Usualmente no es posible obtener el máximo de forma analítica, por tanto es necesario utilizar algún método numérico como el algoritmo Esperanza-Maximización (EM) (Dempster *et al.*, 1977, McLachlan y Peel, 2000, McLachlan y Basford, 1988, Redner y Walker, 1984, Titterington *et al.*, 1985). Sin embargo, existen otros algoritmos de optimización que permiten calcular los estimadores por máxima verosimilitud tal como el de Newton-Raphson que está basado en derivadas y el de Nelder-Mead que es de búsqueda directa (Nelder y Mead, 1965).

3.1.2. Maximización vía el algoritmo EM

El algoritmo EM (Esperanza-Maximización) es uno de los algoritmos más usados para obtener estimaciones de máxima verosimiltud de los parámetros de las mezclas (Dempster *et al.*, 1977, McLachlan y Peel, 2000, McLachlan y Basford, 1988). EM es un procedimiento iterativo que permite encontrar los máximos verosímiles a problemas en los que es posible utilizar *"aumentación de datos"* para simplificar el problema original. En el caso de modelos de mezclas Gaussianas, el principio básico detrás es

introducir y hacer inferencias sobre un conjunto de variables indicadoras no observadas dado un conjunto de observaciones, $\boldsymbol{Z}|\boldsymbol{Y}$ (Redner y Walker, 1984). En este contexto, $\boldsymbol{Z}_i$ representa a una variable indicadora binaria K-dimensional cuyo elemento j-ésimo, Z_{ij}, indica la pertenencia de la observación $\boldsymbol{y}_i$ al j-ésimo componente de la mezcla ($i = 1, \ldots, n$, $j = 1, \ldots, K$). Es decir, $z_{ij} \in \{0, 1\}$ y:

$$Z_{ij} = \begin{cases} 1 & \text{si } \boldsymbol{y}_i \text{ es generado del } j\text{-ésimo componente.} \\ 0 & \text{de otro modo.} \end{cases}$$

La representación matricial de los datos incompletos y completos es:

$$\boldsymbol{y}^T = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} \text{ y } \begin{pmatrix} \boldsymbol{y}^T & \boldsymbol{z}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{z}_1 \\ \boldsymbol{y}_2 & \boldsymbol{z}_2 \\ \vdots & \vdots \\ \boldsymbol{y}_n & \boldsymbol{z}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_1 & z_{11} & \cdots & z_{1K} \\ \boldsymbol{y}_2 & z_{21} & \cdots & z_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{y}_n & z_{n1} & \cdots & z_{nK} \end{pmatrix}.$$

Dada la naturaleza categórica de las variables Z_{ij} al indicar la pertenencia de los puntos muestrales a una componente u otra de la mezcla, puede asumirse que $\boldsymbol{Z}_i$ sigue una distribución *Multinomial* de solo una realización sobre K categorías con probabilidades $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, es decir, la función de probabilidad de $\boldsymbol{Z}_i$ será:

$$\begin{aligned} &\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim Multinomial(1, \boldsymbol{\pi}), \\ &P(\boldsymbol{Z}_i = \boldsymbol{z}_i) = \begin{pmatrix} 1 \\ z_{i1}, \ldots, z_{iK} \end{pmatrix} \pi_1^{z_{i1}} \ldots \pi_K^{z_{iK}} = \prod_{j=1}^{K} \pi_j^{z_{ij}}, \end{aligned} \tag{3.3}$$

donde: $\sum_{j=1}^{K} z_{ij} = 1$ y $\sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} = n$.

De este modo, la verosimilitud de los datos completos se expresa de la siguiente manera:

$$p(\boldsymbol{Y}, \boldsymbol{Z}) = p(\boldsymbol{Y}|\boldsymbol{Z}) \cdot p(\boldsymbol{Z}).$$

Note que las variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, con $\boldsymbol{Z}_i = \{Z_{i1}, \ldots, Z_{iK}\}$, están relacionadas con la observación muestral $\boldsymbol{Y}_i$ condicionalmente, siendo ésta la única información de la que se dispone de la distribución de $\boldsymbol{Z}_i$:

$$p\left(\boldsymbol{Y}_i|Z_{ij}=1\right) \sim p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right).$$

Luego entonces, se tiene que:

$$\begin{aligned} p\left(\boldsymbol{Y},\boldsymbol{Z}\right) &= \left\{\prod_{j=1}^{K}\left[p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)^{z_{ij}}\right]\right\}\cdot\left\{\prod_{j=1}^{K}\pi_j^{z_{ij}}\right\}, \\ &= \prod_{j=1}^{K}\left[\pi_j\cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right]^{z_{ij}}. \end{aligned} \tag{3.4}$$

La función de verosimilitud conjunta para todos los valores observados $\boldsymbol{y}$ y para el vector $\boldsymbol{z}$ de todos los valores no observados z_{ij} será, por tanto:

$$L\left(\boldsymbol{\phi}|\boldsymbol{Y},\boldsymbol{Z}\right) = \prod_{i=1}^{n}\prod_{j=1}^{K}\left[\pi_j\cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right]^{z_{ij}}. \tag{3.5}$$

En consecuencia la log-verosimiltud es igual a:

$$\begin{aligned} l\left(\boldsymbol{\phi}|\boldsymbol{Y},\boldsymbol{Z}\right) = \log L\left(\boldsymbol{\phi}|\boldsymbol{Y},\boldsymbol{Z}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log\left[\pi_j\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right], \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\left[\log\pi_j + \log p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right], \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log\pi_j + \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right). \end{aligned} \tag{3.6}$$

Después de haber definido las variables Z_{ij}, se introduce el concepto de agrupamiento sobre los datos observados. Uno de los propósitos de los modelos de mezclas es el de proporcionar una partición de los datos en K grupos, siendo K un número previamente establecido. La j-ésima proporción de la mezcla ($\pi_j,\ j=1,\ldots,K$) puede interpretarse como la probabilidad a priori de que una observación muestral pertenezca a la población K, luego entonces:

$$P\left(Z_{ij}=1\right)=\pi_j \quad \text{para}\, j=1,\ldots,K.$$

Bajo los datos completos, el procedimiento de agrupamiento tiene como objetivo asociar cada una de las variables $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$ con los datos observados $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$. Una vez que el modelo de mezclas ha sido ajustado y su parámetro estimado, se proporciona un agrupamiento probabilístico de las observaciones en términos de las probabilidades a posteriores (Teorema de Bayes) de pertenencia a uno u otro grupo:

$$\hat{\tau}_{ij}=P\left\{Z_{ij}=1|\boldsymbol{Y}_i=\boldsymbol{y}_i\right\}=\frac{\hat{\pi}_j\cdot p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\phi}}_j\right)}{\sum_{l=1}^{K}\hat{\pi}_l\cdot p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\phi}}_l\right)} \quad i=1,\ldots,n \text{ y } l=1,\ldots,K.$$

Por tanto, $\hat{\tau}_{i1},\ldots,\hat{\tau}_{iK}$ representan las probabilidades (a posteriores) de que la observación $\boldsymbol{y}_i$ pertenezca al K-ésimo componente de la mezcla. Finalmente, la asignación de una observación a uno u otro grupo se decide mediante la mayor de estas probabilidades:

$$\hat{z}_{ij}=\begin{cases}1 & \text{si } j=\arg\underset{l}{\text{máx}}\left\{\hat{\tau}_{il}\right\} \quad i=1,\ldots,n \text{ y } j=1,\ldots,K,\\ 0 & \text{de otro modo.}\end{cases}$$

Luego entonces, el algoritmo EM se deriva de la siguiente manera. Sea Q una función auxiliar, la esperanza condicional de los datos completos $(\mathbf{Y},\mathbf{Z})$, dados los datos observados $\mathbf{Y}$ y una parametrización $\phi^{(t-1)}$, se tiene que:

$$\begin{aligned}Q\left(\boldsymbol{\phi},\boldsymbol{\phi}^{(t-1)}\right)&=E\left[l\left(\boldsymbol{\phi}|\boldsymbol{Y},\boldsymbol{Z}\right)|\boldsymbol{Y},\boldsymbol{\phi}^{(t-1)}\right],\\&=E\left[\sum_{i=1}^{n}\sum_{j=1}^{K}z_{ij}\log\left[\pi_j\cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right]|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\phi}^{(t-1)}\right], \qquad (3.7)\\&=\sum_{i=1}^{n}\sum_{j=1}^{K}E\left[z_{ij}|\boldsymbol{Y}_i=\boldsymbol{y}_i,\boldsymbol{\phi}^{(t-1)}\right]\left[\log\pi_j+\log p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right].\end{aligned}$$

En el *paso-E* se estima z_{ij} de la ecuación 3.7:

$$
\begin{aligned}
E\left[Z_{ij}|\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{\phi}^{(t-1)}\right] &= P\left(Z_{ij} = z_{ij} = 1|\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{\phi}^{(t-1)}\right), \\
&= \frac{p\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|Z_{ij} = z_{ij} = 1\right) \cdot p\left(Z_{ij} = z_{ij} = 1\right)}{\sum_{l=1}^{K} p\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|Z_{il} = z_{ij} = 1\right) \cdot p\left(Z_{il} = z_{ij} = 1\right)}|_{\boldsymbol{\phi}^{(t-1)}}, \\
&= \frac{\hat{\pi}_k \cdot p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_j\right)}{\sum_{j=1}^{k} \hat{\pi}_j \cdot p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_j\right)}|_{\boldsymbol{\phi}^{(t-1)}} := \hat{\tau}_{ij}^{(t-1)}.
\end{aligned} \tag{3.8}
$$

Por lo tanto, la ecuación 3.7 se puede reescribir de la manera siguiente:

$$
\begin{aligned}
Q\left(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t-1)}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{K} \hat{\tau}_{ij}^{(t-1)} \left[\log \pi_j + \log p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right], \\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} \hat{\tau}_{ij}^{(t-1)} \left[\log \pi_j\right] + \sum_{i=1}^{n}\sum_{j=1}^{K} \hat{\tau}_{ij}^{(t-1)} \left[\log p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right].
\end{aligned} \tag{3.9}
$$

En el *paso-M* se maximiza de la función Q con respecto a $\boldsymbol{\phi}$. Ahora, note que π_j aparece solo en el primer término de la ecuación 3.9 y $\boldsymbol{\theta}_j$ en el segundo, en consecuencia, se puede maximizar de manera independiente. Note que el algoritmo EM genera una secuencia de estimaciones de conjunto de parámetros $\left\{\hat{\boldsymbol{\phi}}^{(t-1)},\ t = 1, 2, \ldots\right\}$ alternando los pasos E (Esperanza) y M (Maximización) hasta lograr la convergencia (para más detalles sobre convergencia ver: McLachlan y Peel, 2000). Finalmente, para mayor información sobre el algoritmo EM en mezclas Gaussianas revisar Redner y Walker (1984).

3.1.2.1. Inicialización en el algoritmo EM

El éxito del algoritmo EM depende de los valores iniciales del conjunto de parámetros. La mayoría de las propuestas incluyen las siguientes estrategias: a) Emplear varias inicializaciones aleatorias y seleccionar la que concluya con un mayor valor de verosimilitud o bien b) Realizar un agrupamiento previo con algún algoritmo existente (McLachlan y Krishnan, 1997, McLachlan y Peel, 2000).

3.1.2.2. Determinación del número de componentes

McLachlan y Peel (2000) proporcionan una interpretación detallada de los diferentes enfoques disponibles para determinar el número óptimo de componentes. La mayoría de estos enfoques se dividen generalmente en dos categorías: a) Modelos basados en el principio de la parsimonia y b) Modelos basados en procedimientos de prueba, ambos sustentados en la función de log-verosimilitud. Sin embargo; en este estudio número de componentes (K) se determinó mediante un método heurístico, conocido como partición alrededor de los medoides (PAM por sus siglas en inglés, Partitioning Around Medoids).

Kaufman y Rousseeuw (2005) menciona que el algoritmo de la PAM se basa en la formación de K particiones u objetos representativos, llamados *"medoides"*, de n observaciones de un conjunto de datos. Se eligen aleatoriamente K medoides de un conjunto de datos. Cada observación se asigna al grupo correspondiente al medoide más cercano en distancia, estimada mediante la distancia *"Euclidiana"*. Es decir, la observación i se coloca en el grupo v_i cuando el medoide m_{v_i} está más cerca que cualquier otro medoide m_w:

$$d\left(i, m_{v_i}\right) \leq d\left(i, m_w\right) \ \forall \ w = 1, \ldots, K.$$

Los K objetos representativos deben minimizar la suma de las diferencias, de todos los objetos a su medoide más cercano:

$$\text{Función objetivo} = \sum_{i=1}^{n} d\left(i, m_{v_i}\right).$$

El algoritmo consiste de dos pasos:

1. Seleccionar de forma consecutiva K objetos situados en el centro que se utilizarán como medoides iniciales.
2. Si la función objetivo se puede reducir intercambiando un objeto seleccionado con un objeto no seleccionado, entonces el intercambio se lleva a cabo. Esto se continúa hasta que la función objetivo ya no cambia.

Una partición de los datos, tal como la agrupación encontrada por el algoritmo de la PAM, se puede visualizar a través de el diagrama de la silueta.

Para cada observación i, se calcula el valor de la silueta $s(i)$ y luego se representa en la gráfica como una barra de longitudes $s(i)'s$. Para definir $s(i)$ se procede de la siguiente manera:

1. Considere cualquier objeto i del conjunto de datos, y sea A el grupo al cual está asignado, entonces:

   $$a(i) = \text{es el promedio de disimilitud de } i \text{ a todos los demás objetos de } A.$$

2. Considere ahora cualquier grupo C diferente de A y defina:

   $$d(i, C) = \text{es el promedio de disimilitud de } i \text{ a todos los objetos de } C.$$

3. Calcule $d(i, C)$ para todos los grupos de $C \neq A$, y luego seleccione el más pequeño de éstos:

   $$b(i) = \min_{C \neq A} d\left(i, C\right).$$

El grupo B que alcanza el mínimo, es decir $d(i, B) = b(i)$, se llama el *"vecino"* del objeto i. Este es el segundo mejor grupo para el objeto i.

El valor del ancho de i-ésima silueta ($s(i)$) ahora se puede definir como sigue:

$$s\left(i\right) = \frac{b\left(i\right) - a\left(i\right)}{\max\left\{a\left(i\right), b\left(i\right)\right\}}.$$

Puede verse fácilmente que $s(i)$ se encuentra entre -1 y 1. El valor de $s(i)$ puede ser interpretado como indica la Tabla 3.1 (Kaufman y Rousseeuw, 2005).

Tabla 3.1: Clasificación del ancho de la silueta

$s(i) = 1$	La disimilaridad "dentro" $a(i)$ es mucho más pequeña que la más pequeña disimilitud "entre". En otras palabras, el objeto i ha sido asignado a un grupo apropiado. El segundo mejor grupo B no es tan cerca como el grupo actual A.
$s(i) = 0$	$a(i)$ y $b(i)$ son aproximadamente iguales. Por lo tanto, no está claro si debería asignarse a A o B. Se puede considerar como un caso "intermedio".
$s(i) = -1$	El objeto i está mal clasificado. Cuando s es cercano a -1, el objeto no está bien clasificado. Su disimilitud con otros objetos en su grupo es mucho mayor que su disimilitud con objetos en el grupo más cercano.

Por lo tanto, el valor de la silueta resume cuán apropiado está cada objeto en su grupo.

La silueta de un grupo es un gráfico de $s(i)'s$ clasificadas en orden decreciente de todos los objetos i. La gráfica es una línea horizontal, cuya longitud es proporcional a $s(i)$. Las siluetas muestran cuales observaciones se encuentran dentro del grupo y cuáles están simplemente en algún lugar entre los grupos.

Una silueta amplia indica valores de $s(i)$ grandes y, por tanto, un grupo pronunciado. La altura de un grupo es simplemente igual al número de observaciones en el grupo.

La gráfica de siluetas entera muestra las siluetas de todos los grupos contiguos, por lo que se puede comparar la calidad de los grupos. El ancho promedio de silueta de la gráfica de la silueta es el promedio de las $s(i)'s$ sobre todos los objetos en el conjunto de datos.

La gráfica de siluetas es muy útil para decidir el número de grupos. Se puede correr el algoritmo de la PAM varias veces, cada vez para diferentes valores de K, y luego comparar las gráficas resultantes de la silueta. El ancho promedio de silueta se puede utilizar para seleccionar el "número óptimo de grupos", eligiendo la k que produzca el mayor ancho de silueta.

$$SC = \max_{K} \left\{ \bar{s}\left(K\right) \right\},$$

donde el máximo se toma sobre todos los K, para el que la silueta se puede construir, lo que significa $k = 2, \ldots, n-1$. Este coeficiente de Silueta (SC) es una medida adimensional de la extensión de la estructura de agrupamiento que ha sido descubierto por el algoritmo de agrupamiento. De acuerdo con Kaufman y Rousseeuw (2005), el SC se puede interpretar como se muestra en la Tabla 3.2.

Tabla 3.2: Rango del coeficiente de silueta.

Rango del coeficiente de silueta	Interpretación
0.71-1.00	Estructura fuerte
0.51-0.70	Estructura razonable
0.26-0.50	Estructura débil y podría ser artificial
≤ 0.25	Ninguna estructura sustancial

3.2. De la Estadística Bayesiana Paramétrica a la No-Paramétrica

3.2.1. Modelo Beta-Bernoulli

Sea Y una variable aleatoria que toma dos valores en el dominio $\mathcal{Y} = \{1, 2\}$ con probabilidad p_1 y $p_2 = (1 - p_1)$, respectivamente. Esto es:

$$Y|p_1, p_2 \sim Bernoulli\,(p_1)\,,$$

con $p_j > 0$ tal que $j = 1, 2$ y $\sum_{j=1}^{2} p_j = 1$. Ahora, defina una medida de probabilidad sobre $\boldsymbol{\mathcal{Y}}$ para cada valor de $\boldsymbol{p} = (p_1, p_2)$. Dado que los parámetros p_1 y p_2 son desconocidos entonces puede utilizarse una distribución a priori sobre p_1 $(0 \leq p_1 \leq 1)$ que puede ser la distribución *Beta*:

$$p\,(p_1) = \frac{\Gamma\,(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p_1^{\alpha_1 - 1}\,(1 - p_1)^{\alpha_2 - 1} \text{ con } \alpha_1, \alpha_2 > 0,$$

donde $\Gamma\,(z)$ es la función gamma que por definición es $\Gamma(z) = \int_0^\infty t^{z-1} \exp^{-t} dt$, para enteros positivos las cantidades integran a $\Gamma(n) = (n-1)!$.

Note que α_j está asociada con cada elemento del domino de $\boldsymbol{\mathcal{Y}}$, tal que: $\alpha\,(\boldsymbol{\mathcal{Y}}_j) = \alpha_j$ con $j = 1, 2$. Esto significa que α funciona como una medida sobre $\boldsymbol{\mathcal{Y}}$, donde $\alpha\,(\boldsymbol{\mathcal{Y}}) = \sum_{\boldsymbol{\mathcal{Y}}_j \in \boldsymbol{\mathcal{Y}}} \alpha_j$. Luego entonces, el valor esperado de p_j es igual a:

$$E\,(p_j) = \frac{\alpha_j}{\alpha\,(\boldsymbol{\mathcal{Y}})}.$$

Entonces, si $Y_1, \ldots, Y_n|p_1, p_2 \overset{iid}{\sim} Bernoulli(p_1, 1 - p_1)$ y $p_1 \sim Beta(\alpha_1, \alpha_2)$, la distribución a posteriori de p_1 está dada también por la distribución *Beta* y es igual a:

$$\begin{aligned} p\,(p_1|Y_1, \ldots, Y_n) &\propto L\,(p_1|Y_1, \ldots, Y_n) \cdot p\,(p_1)\,, \\ &\propto p_1^{\alpha_1 + \sum_{i=1}^n \delta_{Y_i}(1) - 1}\,(1 - p_1)^{\alpha_2 + \sum_{i=1}^n \delta_{Y_i}(2) - 1}\,, \\ &\propto Beta\left(\alpha_1 + \sum_{i=1}^{n} \delta_{Y_i}\,(1)\,, \alpha_2 + \sum_{i=1}^{n} \delta_{Y_i}\,(2)\right), \end{aligned}$$

donde $\delta_{Y_i}\,(y)$ es la función indicadora que es:

$$\delta_{Y_i}\,(y) = \begin{cases} 1 & Y_i = y, \\ 0 & \text{de otro modo.} \end{cases}$$

Por su parte, la distribución condicional predictiva a posteriori está definida por:

$$p\left(Y_{n+1}|Y_1,\ldots,Y_n\right) = \frac{p\left(Y_1,\ldots,Y_n,Y_{n+1}|\alpha_1,\alpha_2\right)}{p\left(Y_1,\ldots,Y_n|\alpha_1,\alpha_2\right)},$$
$$= \frac{\Gamma\left(\alpha\left(Y\right)+n\right)\Gamma\left(\alpha_1+\sum_{i=1}^{n+1}\delta_{Y_i}\left(1\right)\right)\Gamma\left(\alpha_2+\sum_{i=1}^{n+1}\delta_{Y_i}\left(2\right)\right)}{\Gamma\left(\alpha\left(Y\right)+n+1\right)\Gamma\left(\alpha_1+\sum_{i=1}^{n+1}\delta_{Y_i}\left(1\right)\right)\Gamma\left(\alpha_2+\sum_{i=1}^{n+1}\delta_{Y_i}\left(2\right)\right)},$$

si $Y_{n+1}=1$, entonces:

$$p\left(Y_{n+1}|Y_1,\ldots,Y_n\right) = \frac{\alpha_1+\sum_{i=1}^{n+1}\delta_{Y_i}\left(1\right)}{\alpha_1+\alpha_2+n},$$

y si $Y_{n+1}=2$ entonces:

$$p\left(Y_{n+1}|Y_1,\ldots,Y_n\right) = \frac{\alpha_2+\sum_{i=1}^{n+1}\delta_{Y_i}\left(2\right)}{\alpha_1+\alpha_2+n}.$$

3.2.2. Modelo Dirichlet-Multinomial

La distribución *Dirichlet* es la versión multivariada de la distribución *Beta*, así como la *Multinomial* lo es de la *Binomial*. Continuando con el esquema del modelo *Beta-Bernoulli*, considere $Y_1,\ldots,Y_n$ variables aleatorias que toman K valores en el dominio $\boldsymbol{\mathcal{Y}}=\{1,\ldots,K\}$ con probabilidades $p_1,\ldots,p_k$. Entonces:

$$\left(Y_1,\ldots,Y_n\right)|p_1,\ldots,p_K \sim Multinomial\left(p_1,\ldots,p_K\right),$$

con $p_j>0$ tal que $j=1,\ldots,K$ y la $\sum_{j=1}^{K}p_j=1$. Esto significa que el vector $(p_1,\ldots,p_K)$ pertenece a un simplex Δ^{K-1}.

Análogamente como en el modelo *Beta-Bernoulli*, $p_1,\ldots,p_K$ son desconocidos; por consiguiente, se les puede asignar una distribución a priori, que en este caso sería la distribución *Dirichlet*:

$$p\left(p_1,\ldots,p_K\right) = \frac{\Gamma\left(\sum_{j=1}^{K}\left(\alpha_j\right)\right)}{\prod_{j=1}^{K}\left(\Gamma\left(\alpha_j\right)\right)}\prod_{j=1}^{K}p_j^{\alpha_j-1}. \tag{3.10}$$

Cuando hay sólo dos parámetros, la distribución también se llama una distribución

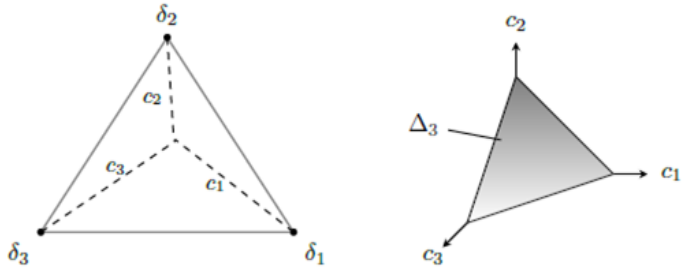


Figura 3.1: El simplex Δ^3. Cada punto en el conjunto puede interpretarse como una medida de probabilidad en tres eventos disjuntos. Para cualquier K finito, el simplex Δ^{K-1} puede considerarse como un subconjunto del espacio euclidiano $\mathbb{R}^K$ (Orbanz, 2014).

Beta. Por conveniencia, los exponentes de una distribución son definidos igual a $\alpha_j - 1$, para que la media y la varianza de la densidad tengan la forma simple siguiente (Sudderth, 2006):

$$E(p_j) = \frac{\alpha_j}{\sum_{j=1}^{K} \alpha_j} \qquad \text{y} \qquad Var(p_j) = \frac{E(p_j)(1 - E(p_j))}{\sum_{j=1}^{K} \alpha_j + 1}. \tag{3.11}$$

El parámetro $\sum_{j=1}^{K} \alpha_j$ se denomina típicamente como *el parámetro de concentración* y controla qué tan concentrada la distribución está alrededor de su valor esperado (Huang, 2005). La Figura 3.2 muestra las gráficas de la distribución *Dirichlet* sobre el simplex Δ^3 para diferentes configuraciones del parámetro de concentración $\sum_{j=1}^{K} \alpha_j$.

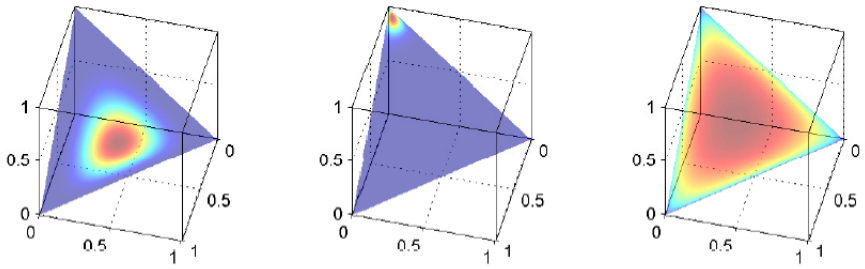


Figura 3.2: Distribución *Dirichlet* para diferentes configuraciones del parámetro de concentración $\sum_{j=1}^{K} \alpha_j$. El gráfico de la derecha tiene una $\sum_{j=1}^{K} \alpha_j$ pequeña que se traduce en una distribución difusa, mientras que el gráfico central tiene una $\sum_{j=1}^{K} \alpha_j$ grande, por lo que se concentra alrededor de la media. Figura tomada de Huang (2005), página 3.

Del mismo modo en que se obtuvieron algunas propiedades de la distribución *Beta-Bernoulli*, se obtienen de la distribución *Dirichlet-Multinomial*.

La distribución a posteriori de $p_j|Y_1,\ldots,Y_n$ de la distribución *Multinomial* es también *Dirichlet*.

$$p\left(p_1,\ldots,p_K|Y_1,\ldots,Y_n\right) \propto \prod_{j=1}^{K} p_j^{\alpha_j+\sum_{i=1}^{n}\delta_{Y_i}(j)-1}.$$

Es decir:

$$p_1,\ldots,p_K|Y_1,\ldots,Y_n \sim Dirichlet\big(\alpha_1+\sum_{i=1}^{n}\delta_{Y_i}\left(1\right),\ldots,\alpha_K+\sum_{i=1}^{n}\delta_{Y_i}\left(K\right)\big).$$

Respecto a la distribución condicional de una nueva observación $Y_{n+1}|Y_1,\ldots,Y_n$, ésta es:

$$p\left(Y_{n+1}|Y_1,\ldots,Y_n\right) = \frac{\Gamma\left(\alpha\left(\boldsymbol{\mathcal{Y}}\right)+n\right)}{\Gamma\left(\alpha\left(\boldsymbol{\mathcal{Y}}\right)+n+1\right)}\prod_{j=1}^{K}\frac{\Gamma\left(\alpha_j+\sum_{i=1}^{n+1}\delta_{Y_i}\left(j\right)\right)}{\Gamma\left(\alpha_j+\sum_{i=1}^{n}\delta_{Y_i}\left(j\right)\right)}.$$

También se puede expresar:

$$p\left(Y_{n+1}=j|Y_1,\ldots,Y_n\right) = \frac{\alpha_j+\sum_{i=1}^{n}\delta_{Y_i}\left(j\right)}{\alpha\left(\boldsymbol{\mathcal{Y}}\right)+n}. \tag{3.12}$$

Cabe mencionar que la distribución a posteriori predictiva también se puede describir de acuerdo a un esquema de **"Urna de Polya"**.

Otra propiedad importante de la distribución *Dirichlet-Multinomial* es que cumple con el Teorema de Consistencia de Kolmogorov que a la letra dice:

Teorema 3.1 *Sea $B_1,\ldots,B_m$ cualquier partición del dominio $\boldsymbol{\mathcal{Y}}$, entonces:*

$$\left(P\left(Y\in B_1\right),\ldots,P\left(Y\in B_m\right)\right) \sim Dirichlet\left(\alpha\left(B_1\right),\ldots,\alpha\left(B_m\right)\right),$$

donde $\alpha\left(B_i\right)=\sum_{\boldsymbol{\mathcal{Y}}_j\in B_i}\alpha_j$.

Finalmente, es importante preguntarse ¿Qué ocurre si $\boldsymbol{\mathcal{Y}}=\mathbb{R}, \mathbb{R}^K$?. En este caso donde $\boldsymbol{\mathcal{Y}}=\{1,2,\ldots\}=\mathbb{N}$ se requiere el uso de ***"Procesos Estocásticos"***, ya que $\boldsymbol{p}=(p_1,p_2,\ldots)^T$ con $p_j>0$ tal que $j=1,2,\ldots$ y la $\sum_{j=1}^{\infty}p_j=1$. Luego entonces, es necesario especificar una distribución a priori para el proceso p_j.

3.2.3. Procesos Dirichlet (*DP*)

De acuerdo con la sección anterior, en ambos modelos lo que se hizo fue asignar una medida de probabilidad a un espacio finito de probabilidad. En esta sección, la idea es extender esta forma de entender dichas distribuciones al caso donde $\boldsymbol{\mathcal{Y}}$ sea un conjunto infinito dimensional, particularmente un espacio separable. Es en la realización de esta idea donde la definición de los procesos *Dirichlet* toma importancia como un caso particular de una definición más general como son medidas de probabilidad aleatorias. El *DP* como una distribución a priori se propuso por primera vez por Ferguson (1973) y sus propiedades han sido ampliamente estudiadas desde entonces por: Blackwell y MacQueen (1973), Antoniak (1974), Sethuraman (1994), Walker *et al.* (1999), entre muchos otros.

3.2.3.1. Definición y propiedades del *DP*

Sea $\boldsymbol{\mathcal{Y}}$ un conjunto con un número infinito de elementos (por ejemplo $\mathbb{R}$, el intervalo $[0,1]$, $\mathbb{R}^K$) y sea $\boldsymbol{\mathcal{A}}(\boldsymbol{\mathcal{Y}}) = \boldsymbol{\mathcal{A}}$ un σ-álgebra de subconjuntos $\boldsymbol{\mathcal{Y}}$. Ferguson (1973) define un proceso *Dirichlet* como sigue:

Definición 3.1 *Sea α una medida finita no nula (no negativa y finitamente aditiva) en $(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{A}})$. Se dice que P es un proceso Dirichlet en $(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{A}})$ con parámetro α si para cada $j = 1, 2, \ldots$ y partición medible $(B_1, \ldots, B_K)$ de conjuntos de Borel de $\boldsymbol{\mathcal{Y}}$ y $K \in \mathbb{N}$, la distribución del vector aleatorio $(P(B_1), \ldots, P(B_K))$ es Dirichlet, Dirichlet $(\alpha(B_1), \ldots, \alpha(B_k))$.*

Con base en lo anterior y con la aclaración del cambio de nomenclatura $\boldsymbol{\mathcal{Y}} = \boldsymbol{\Theta}$ y $\boldsymbol{\mathcal{A}} = \boldsymbol{\Sigma}$ debido a que actualmente es la que la literatura reporta (Fox, 2009, Rodríguez, 2007, Sudderth, 2006, Teh, 2010), se re escribe la definición 3.1 como sigue:

"Sea H una distribución de probabilidades sobre un espacio medible $(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$ y α un escalar positivo. Una medida de probabilidad G sobre $\boldsymbol{\Theta}$ se llama proceso Dirichlet si cada partición finita medible $A_1, \ldots, A_K$ en $\boldsymbol{\Theta}$ sigue una distribución Dirichlet K-dimensional:

$$p(G(A_1), \ldots, G(A_K) | \alpha, H) \sim Dirichlet(\alpha H(A_1), \ldots, \alpha H(A_K)), \tag{3.13}$$

para alguna medida base H y parámetro de concentración α, hay un proceso estocástico único satisfaciendo estas condiciones, el cual se denota por $DP(\alpha, H)$, ver Figura 3.3".

Luego entonces, combinando las ecuaciones de la media y la varianza de la distribución *Dirichlet* vistas en la sección 3.2.2 y la definición 3.1, se obtiene el valor esperado y

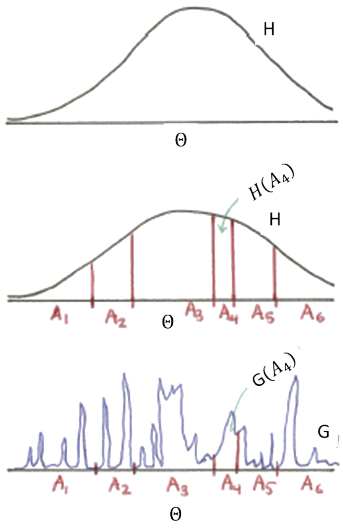


Figura 3.3: Ejemplo de una medida base H sobre un espacio $\boldsymbol{\Theta}$, con $K = 6$ particiones. El peso que asigna una medida aleatoria, $G \sim DP(\alpha, H)$, a dichas particiones sigue una distribución *Dirichlet* (véase la ecuación 3.13). Fuente: Tomada del Tutorial de Ghahramani (2005).

la varianza del DP para cualquier región $A \in \boldsymbol{\Theta}$:

$$\begin{aligned} E\left(G\left(A\right)\right) &= \frac{\alpha H\left(A_j\right)}{\alpha H\left(A_1\right) + \ldots + \alpha H\left(A_K\right)} = H\left(A\right), \\ Var\left(G\left(A\right)\right) &= \frac{\alpha H\left(A_j\right)\left(\alpha - \alpha H\left(A_j\right)\right)}{\alpha^2\left(\alpha + 1\right)} = \frac{H\left(A\right)\left(1 - H\left(A\right)\right)}{\alpha + 1}, \end{aligned} \tag{3.14}$$

donde la distribución base H es el valor medio del DP y el parámetro de concentración α se considera como una varianza inversa que determina la desviación promedio de las muestras de la medida base y cuando es grande el proceso es altamente concentrado alrededor de H.

De lo anterior, se puede resumir de una manera más sencilla y más intuitiva la definición de DP:

"Sea $\boldsymbol{\Theta}$ un espacio particionado en cualquier forma finita $(A_1, \ldots, A_K)$ y sea G una distribución de probabilidad específica que asigna probabilidades a dichas particiones. El DP es una distribución sobre todas las distribuciones de probabilidad posibles sobre el espacio $\boldsymbol{\Theta}$ y se parametriza con la función base H y el parámetro de concentración α. Luego entonces, se puede decir que $G \sim DP(\alpha, H)$, si la distribución conjunta de las probabilidades que G asigna a las particiones de $\boldsymbol{\Theta}$ sigue la distribución Dirichlet".

Ahora, sea $y = (y_1, \ldots, y_n)$ las observaciones hechas a partir de G, la cual es una distribución de probabilidad sobre $\boldsymbol{\Theta}$, entonces se pueden extraer muestras independientes e idénticamente distribuidas $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim G$, y además sea G una muestra de un $DP(\alpha, H)$, entonces para cualquier partición medible finita $\{A_1, \ldots, A_K\}$ en $\boldsymbol{\Theta}$, la distribución a posteriori de G dada la secuencia de observaciones es dada por:

$$
\begin{aligned}
& p\left(G(A_1), \ldots, G(A_K) \mid y_1, \ldots, y_n\right) \\
& \quad \propto \quad p\left(y_1, \ldots, y_n | G(A_1), \ldots, G(A_K)\right) \cdot p\left(G(A_1), \ldots, G(A_K)\right), \\
& \quad \propto \quad \left[\prod_{j=1}^{K} G(A_j)^{\sum_{i=1}^{n} \delta_{y_i}(A_j)}\right] \left[\prod_{j=1}^{K} G(A_j)^{\alpha H(A_j)-1}\right], \qquad (3.15) \\
& \quad = \quad DP\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{y_i}}{\alpha + n}\right), \\
& \quad = \quad DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \sum_{i=1}^{n} \frac{\delta_{y_i}}{n}\right),
\end{aligned}
$$

donde $\sum_{i=1}^{n} \delta_{y_i}(A_j)$ indica cuántas observaciones caen dentro de la partición A_j, y además defina a δ_{y_i} por $\delta_{y_i} = 1$ si $y_i \in A_j$ y 0 de otro modo. Este resultado se debe al hecho de que cada partición finita del *DP* sigue una distribución *Dirichlet* y por la conjugación *Dirichlet-Multinomial* descrita en la sección 3.2.2, sólo se tiene que actualizar los parámetros de la distribución con el número de observaciones en cada partición correspondiente (Ferguson, 1973).

De lo anterior se puede ver que la medida base a posteriori es una media ponderada entre la medida base a priori H y la distribución empírica $\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$ y el peso asociado con la medida base a priori es proporcional a α, mientras que la distribución empírica tiene un peso proporcional al número de observaciones n. Por lo tanto, α puede interpretarse como la fuerza asociada con la a priori.

Hasta ahora se han estudiado algunas características del *DP*, incluyendo la de conjugación, pero ninguna de éstas proporcionan directamente un mecanismo para la toma de muestras o la predicción de futuras observaciones. En la siguiente sección, se analizará una de las diferentes representaciones importantes del proceso *Dirichlet* que existen, y la cual se le conoce como *"stick-breaking"*.

3.2.3.2. Representación del *DP* mediante el Stick-Breaking

Los primeros indicios de la existencia del *DP* tal y como ya se mencionó fueron argumentados por Ferguson (1973), quien utilizó el Teorema 3.1, y proporcionó su defini-

ción y describió su proceso a posteriori. Posteriormente, Blackwell y MacQueen (1973) siguieron con la investigación de Ferguson y mediante el Teorema de Finetti de Intercambiabilidad demostraron la existencia de tal medida de probabilidad e introdujeron el esquema de ***"Urna de Polya"***. Más tarde, Aldous (1985) proporcionó otra forma de construir el *DP*, nombrada ***"Proceso del Restaurante Chino"***. Finalmente, Sethuraman (1994) proporciona otra manera de construir un *DP*, llamada ***"Stick-Breaking"***. Las diversas representaciones del *DP* son matemáticamente equivalentes, pero su formulación es diferente porque examinan el problema desde diferentes puntos de vista. En este trabajo de investigación únicamente se abordará la construcción *stick-breaking*.

Luego entonces, utilizando la ecuación 3.15 en conjunto con la ecuación 3.14, se puede observar que para cualquier $A \in \boldsymbol{\Theta}$:

$$E\left(G\left(A\right) \mid y_1, \ldots, y_n, H, \alpha\right) = \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \sum_{i=1}^{n} \frac{\delta_{y_i}}{n}, \tag{3.16}$$

Ahora, sí se toma el límite cuando el número de observaciones tiende a infinito se obtiene que (Fox, 2009, Sudderth, 2006):

$$\lim_{n \to \infty} E\left(G\left(A\right) \mid y_1, \ldots, y_n, H, \alpha\right) = \sum_{j=1}^{\infty} \pi_j \delta_{\boldsymbol{\theta}_j^*} \tag{3.17}$$

donde el vector $\boldsymbol{\theta^*} = \left(\boldsymbol{\theta}_1^*, \ldots\right)^T$ contiene los distintos valores entre el conjunto de observaciones, y el $\boldsymbol{\pi} = \left(\pi_1, \ldots\right)^T$ representa su correspondiente frecuencia empírica. Suponiendo que la distribución a posteriori se concentra en torno a su media, la ecuación 3.17 sugiere que las medidas *Dirichlet* son discretas con probabilidad uno.

Demostración. *"Sea* $\boldsymbol{\theta} \sim H$ *y la partición* $\{\boldsymbol{\theta}, \boldsymbol{\Theta} \backslash \boldsymbol{\theta}\}$ *de* $\boldsymbol{\Theta}$*. El proceso a posteriori* $G|\boldsymbol{\theta} \sim DP(\alpha + 1, \frac{\alpha H + \delta_{\boldsymbol{\theta}}}{\alpha + 1})$ *implica que:*

$$\begin{aligned} \left(G\left(\boldsymbol{\theta}\right), G\left(\boldsymbol{\Theta} \backslash \boldsymbol{\theta}\right)\right) | \boldsymbol{\theta} &\sim Dirichlet\left(\alpha H + \delta_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}\right), \alpha H\left(\boldsymbol{\Theta} \backslash \boldsymbol{\theta}\right) + \delta_{\boldsymbol{\theta}}\left(\boldsymbol{\Theta} \backslash \boldsymbol{\theta}\right)\right), \\ &= Dirichlet\left(1, \alpha\right), \\ \iff \qquad G\left(\boldsymbol{\theta}\right) | \boldsymbol{\theta} &\sim Beta\left(1, \alpha\right). \end{aligned}$$

Por lo tanto, G *tiene una masa puntual ubicada en* $\boldsymbol{\theta}$*, de tal manera que:*

$$G = \beta \delta_{\boldsymbol{\theta}} + \left(1 - \beta\right) G', \qquad \beta \sim Beta\left(1, \alpha\right),$$

donde G' *es la medida de probabilidad (normalizado de nuevo) sin la masa puntual" (Archambeau, 2008).*

Ahora, la medida de probabilidad G' tiene la forma siguiente:

"Sea $\{A_1, \ldots, A_K\}$ *una partición* $\boldsymbol{\Theta}\backslash\boldsymbol{\theta}$. *Dado* $G(\boldsymbol{\Theta}\backslash\boldsymbol{\theta}) = \sum_{j=1}^{K} G(A_k)$, *la propiedad de aglomeración Dirichlet conduce a:*

$$(G(\boldsymbol{\theta}), G(A_1), \ldots, G(A_K)) | \boldsymbol{\theta} \sim Dirichlet(\alpha H(A_1), \ldots, \alpha H(A_K)).$$

Por lo tanto, note que:

$$\begin{aligned} G(\boldsymbol{\theta}) | \boldsymbol{\theta} &= \beta, \\ G(A_1) | \boldsymbol{\theta} &= (1-\beta) G'(A_1), \\ &\vdots \\ G(A_K) | \boldsymbol{\theta} &= (1-\beta) G'(A_K). \end{aligned}$$

De la propiedad decimativa de la distribución Dirichlet, se tiene que:

$$(G'(A_1), \ldots, G'(A_K)) \sim Dirichlet(\alpha H(A_1), \ldots, \alpha H(A_K)).$$

En otras palabras: $G' \sim DP(\alpha, H)$*".*

Luego entonces, la medida aleatoria G tiene la forma siguiente:

$$\begin{aligned} &G \sim DP(\alpha, H), \\ &G = \beta_1 \delta_{\boldsymbol{\theta}_1^*} + (1-\beta_1) G_1 && G_1 \sim DP(\alpha, H), \\ &G = \beta_1 \delta_{\boldsymbol{\theta}_1^*} + (1-\beta_1)\left(\beta_2 \delta_{\boldsymbol{\theta}_2^*} + (1-\beta_2) G_2\right) && G_2 \sim DP(\alpha, H), \\ &\vdots \\ &G = \sum_{j=1}^{\infty} \pi_j \delta_{\boldsymbol{\theta}_j^*}, \end{aligned}$$

donde:

$$\pi_j = \beta_j \prod_{l=1}^{K-1} (1-\beta_l), \qquad \beta_j \sim Beta(1, \alpha), \qquad \boldsymbol{\theta}_j^* \sim H.$$

El siguiente teorema verifica esta hipótesis, y proporciona una construcción explícita para el conjunto infinito de pesos de mezclas.

Teorema 3.2 *Sea* $\boldsymbol{\pi} = \{\pi_j\}_{j=1}^{\infty}$ *una secuencia infinita de pesos de mezclas derivados*

del siguiente proceso "stick-breaking", con parámetro $\alpha > 0$*:*

$$\beta_j \sim Beta(1, \alpha), \qquad j = 1, \ldots,$$

$$\pi_j = \beta_j \prod_{l=1}^{K-1} (1 - \beta_l) = \beta_l \left(1 - \sum_{l=1}^{K-1} \beta_l\right). \tag{3.18}$$

Dada una medida base H *en* $\boldsymbol{\Theta}$*, supóngase que se extrae un número infinito de muestras* $\boldsymbol{\theta}_j^* \overset{iid}{\sim} H$ *y se forma la siguiente medida de probabilidad discreta:*

$$G(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} \pi_j \delta_{\boldsymbol{\theta}_j^*} \qquad y \qquad \boldsymbol{\theta}_j \sim H. \tag{3.19}$$

Por lo tanto, por una parte esta construcción devuelve un DP con parámetro de concentración α *y distribución base* H*, es decir:* $G \sim DP(\alpha, H)$*. Por otra, todas las muestras de un DP son casi seguramente discretas y tienen una representación como en la ecuación 6.4 y ver Figura 3.4.*

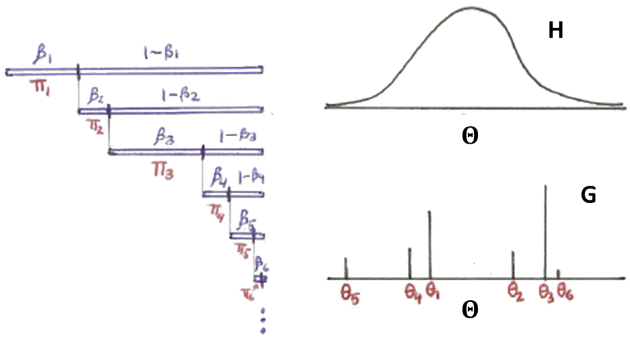


Figura 3.4: Construcción *stick-breaking* del *DP*. Sea un palillo de longitud uno, el cual se rompe en la posición β_1 con probabilidad π_1. Ahora, se rompe el resto del palillo en β_2, con probabilidad π_2. Se continúa este proceso hasta obtener una sucesión de π_i's. Este proceso generalmente se denota por $\pi \sim GEM$, donde GEM proviene de las primeras letras de Griffiths, Engen y McCloskey (Pitman, 2006). Fuente: Tomada del Tutorial de Ghahramani (2005).

3.2.4. Modelos de Mezclas de Procesos Dirichlet (*DPMM*)

Tal y como ya se mencionó en la sección 3.1, el principio básico detrás de los modelos de mezclas es introducir y hacer inferencias sobre un conjunto de variables indicadoras no observadas dado un conjunto de observaciones, $\mathbf{Z}|\mathbf{Y}$. Luego entonces, en las siguientes secciones, primero se considerará el caso donde se conoce a priori el número

de grupos del modelo de mezclas, K, y más tarde, se incorporará incertidumbre en el número de grupos aplicando el proceso *"stick-breaking"* descrito en la sección 3.2.3.2.

3.2.4.1. Modelos de mezclas finitas

De acuerdo con la ecuación 3.5, la verosimilitud conjunta de un modelo de mezclas finitas se expresa como sigue:

$$L\left(\boldsymbol{\phi}|\boldsymbol{Y},\boldsymbol{Z}\right)=\prod_{i=1}^{n}\prod_{j=1}^{K}\left[\pi_j\cdot p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right)\right]^{z_{ij}},$$

hay dos parámetros que requieren de una distribución a priori; las proporciones de las mezclas π_j y los parámetros de la función de distribución de probabilidad correspondiente a cada componente $\boldsymbol{\theta}_j$. Como se describió anteriormente, el vector de variables indicadoras $(z_{i1},\ldots,z_{iK})$ de cada observación se considera como una variable aleatoria con una distribución $Multinomial(\pi_1,\ldots,\pi_K)$. Por lo tanto, una elección natural para la a priori en π es la distribución $Dirichlet(\alpha_1,\ldots,\alpha_K)$, la cual es una distribución conjugada de la distribución $Multinomial$.

En general, no hay restricciones sobre la a priori para los parámetros de los componentes $\boldsymbol{\theta}_j$. En la mayoría de las aplicaciones de los modelos de mezclas, $p(\boldsymbol{y}|\boldsymbol{\theta})$ se elige de la familia exponencial. Por ejemplo, las observaciones euclidianas son comúnmente modeladas mediante mezclas *Gaussianas*, de modo que el parámetro $\boldsymbol{\theta}_j=(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ especifica la media $\boldsymbol{\mu}_j$ y covarianza $\boldsymbol{\Sigma}_j$ de cada grupo. Además, cuando se están obteniendo mezclas a partir de datos, comúnmente se utiliza una a priori conjugada independiente H, con hiper-parámetro α, en los parámetros de cada grupo:

$$\boldsymbol{\theta}_j\sim H\left(\alpha\right),\ \text{para } j=1,\ldots,K.$$

Respecto a la distribución a priori de los pesos de las mezclas $\boldsymbol{\pi}$, se les puede asignar una a priori *Dirichlet* simétrica con precisión α:

$$\boldsymbol{\pi}\sim Dirichlet\left(\frac{\alpha}{K},\ldots,\frac{\alpha}{K}\right). \tag{3.20}$$

Con los comentarios antes mencionados, a continuación se muestra la siguiente representación jerárquica para el modelo de mezclas finitas que ayudará a entender mejor la estructura y la interdependencia de los parámetros del modelo:

$$
\begin{aligned}
\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\Theta} &\sim p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right) && 1 \leq i \leq n, \\
\boldsymbol{z}_i &\sim Multinomial\left(\pi_1,\ldots,\pi_K\right), && \\
\boldsymbol{\pi} &\sim Dirichlet\left(\alpha_1,\ldots,\alpha_K\right), && \\
\boldsymbol{\theta}_j &\sim H && 1 \leq j \leq K,
\end{aligned}
\tag{3.21}
$$

donde $\boldsymbol{\theta}_j$ se refiere al vector de parámetros correspondiente a los componentes de las mezclas z_i, ver Figura 3.5.

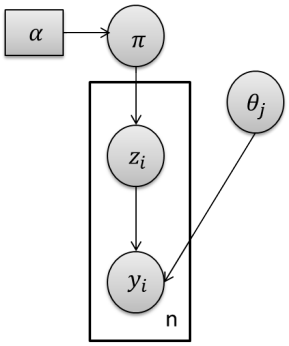


Figura 3.5: Representación de la variable indicadora, en la que $z_i \sim \pi$ es el grupo que genera $\mathbf{y}_i \sim F(\theta)$, donde $F(\theta)$ es alguna familia exponencial de densidades. Fuente: Sudderth (2006).

Ahora, suponga que hay K componentes en un modelo de mezclas finitas y que es posible asignar una distribución a priori en los pesos de las mezclas como la que se muestra en la ecuación 3.20, entonces el modelo de mezclas finitas se puede reescribir en términos de extracciones de una distribución discreta G^K y puede muestrear la matriz de parámetros desde esta distribución como sigue:

$$
\begin{aligned}
\boldsymbol{y}_i|\hat{\boldsymbol{\theta}} &\sim p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_i\right), \\
\hat{\boldsymbol{\theta}}_i &\sim G^K, \\
G^K\left(\boldsymbol{\theta}\right) &= \sum_{j=1}^{K} \pi_j \delta_{\boldsymbol{\theta}_j}, \\
\boldsymbol{\pi} &\sim Dirichlet\left(\frac{\alpha}{K},\ldots,\frac{\alpha}{K}\right), \\
\boldsymbol{\theta}_j &\sim H_0 \qquad 1 \leq j \leq K.
\end{aligned}
\tag{3.22}
$$

Estas dos representaciones son estadísticamente equivalentes, ver ejemplo que demuestra ésto en Ishwaran y Zarepour (2002) o Sudderth (2006).

3.2.4.2. Modelos de mezclas infinitas

Para generar modelos de mezclas con un número infinito de grupos se puede hacer uso de la naturaleza no paramétrica del *DP*. Al igual que con modelos de mezclas finitas, se supone que se tiene un modelo jerárquico en el que cada una de las observaciones $\boldsymbol{y}_i$ tiene una función de densidad p con vector de parámetros muestreado independientemente $\boldsymbol{\theta}_i$. Sin embargo, en el límite $K \longrightarrow \infty$, las etiquetas correspondientes pierden significado cuando el espacio de etiquetas posibles se convierte en continuo (Ranganathan, 2006). Esto significa que se pueden descartar las etiquetas, y mientras el límite infinito de una distribución *Dirichlet* es un *DP*, la parte semejante de dimensión infinita del modelo de mezclas explicado anteriormente sería como sigue:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{\Theta} &\sim p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_i\right), \\ \boldsymbol{\theta}_i|G &\sim G \qquad\qquad i = 1, \ldots, n, \\ G &\sim DP\left(\alpha, H\right), \end{aligned} \tag{3.23}$$

donde G se conoce como la distribución de mezclas de la que independientemente se extraen valores para los parámetros $\boldsymbol{\theta}_i$ para modelar una sola observación, $\boldsymbol{y}_i$. H es la distribución base que actúa como la distribución a priori para los parámetros del componente del modelo de mezclas del *DP*. α controla el nivel de variación de G sobre H. La incertidumbre en α se puede incorporar en el modelo 3.23.

Esto se llama modelo de mezclas de procesos *Dirichlet* (*DPMM*, Antoniak, 1974, Neal, 2000, Sudderth, 2006). Para ver que en realidad es el límite infinito del modelo de mezclas finitas, suponga que coloca una distribución a priori *Dirichlet* simétrica, ver ecuación 3.20, sobre las proporciones de la mezcla en la ecuación 3.21.

Tal y como ya se mencionó en la sub sección 3.2.3.2, existen tres representaciones constructivas del *DP* que permiten la estimación del *DPMM*; sin embargo, en este trabajo de investigación solo se abordará una de ellas, que es la del ***"stick-breaking"***.

Conexión del stick-breaking y los DPMM. Dicha conexión, definida en la ecuación 3.23, se hace a través de los pesos *"stick-breaking"* $\pi_1, \ldots$, los cuales corresponden a los pesos de los componentes de la mezcla. El *DPMM*, por tanto, puede interpretarse como un modelo de mezclas finitas cuando $K \longrightarrow \infty$, de tal manera que estos pesos actúan como una a priori de los indicadores de los componentes ocultos z_i, resultando el modelo generativo y gráfico acíclico dirigido (ver Figura 3.6):

$$
\begin{aligned}
\boldsymbol{Y}|\mathbf{Z},\boldsymbol{\Theta} &\sim p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right), \\
\boldsymbol{z}_i &\sim Multinomial\left(\pi_1,\ldots,\pi_K\right), \\
\boldsymbol{\pi} &\sim GEM\left(\alpha\right), \\
\boldsymbol{\theta}_j &\sim H.
\end{aligned}
\tag{3.24}
$$

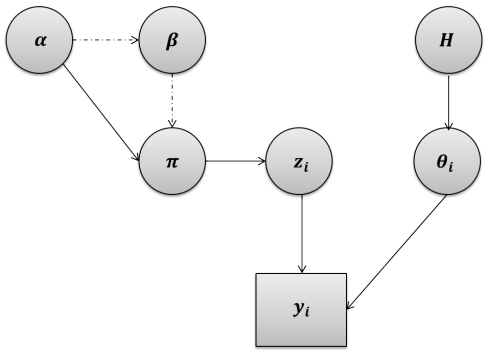


Figura 3.6: Gráfico acíclico dirigido por un modelo de mezcla *DP*, basado en la construcción *"stick-breaking"*. Las líneas discontinuas representan las dependencias implícitas en la construcción de los pesos *"stick-breaking"* (Sudderth, 2006).

Al revisar el comportamiento del *DP* para el *"stick-breaking"*, es de destacar la influencia de α en la determinación del número de grupos, ya que cuando $\alpha \longrightarrow 0$, la a priori *DP* asigna todas las observaciones a un solo grupo. En contraste, cuando $\alpha \longrightarrow \infty$, K también aumenta hasta el punto donde todas las observaciones se les asigna su propio grupo. Este comportamiento indica que α debe ser tratado como desconocido en un *DPMM* y estimado de los datos. Dentro del marco Bayesiano, la incertidumbre que rodea a α se explica fácilmente a través de una distribución a priori. Explícitamente, se puede poner una distribución $Gama(a, b)$ en α con parámetro de forma a y parámetro de escala b mayor que cero (Escobar y West, 1995).

Finalmente, la elección de la representación del *DP* conduce a aplicar diferentes algoritmos de inferencia. Existen muestreadores Monte Carlo que simulan eficazmente la a posteriori con *DPMM* y que se dividen en dos categorías: ***"Métodos Marginales"*** en los cuales el *DP* se integra analíticamente (Escobar, 1994, Escobar y West, 1995, MacEachern y Muller, 1998, Neal, 2000) y ***"Métodos Condicionales"*** utilizando la representación *"stick-braking"* en la que los parámetros de la a priori *DP* se imputan de forma explícita (Ishwaran y James, 2001, Papaspiliopoulos y Roberts, 2008, Walker, 2007).

En este trabajo de investigación se utilizó el **"Muestreador Gibbs en Bloques"** que es un método condicional y para entender su motivación es necesario recordar el comportamiento de los pesos *"stick-breaking"*, ver Figura 3.6. Independientemente del

valor asignado a α, se puede ver que el número de pesos de los componentes mayores que cero disminuye rápidamente. Esto sugiere que, con el fin de hacer factible el cálculo utilizando esta representación, puede ser introducido un nivel de truncamiento L, esencialmente colocando un límite superior sobre el número de grupos ocupados. El nivel de truncamiento debe ser mayor que el número esperado de componentes, o a lo más al tamaño de la muestra, tal que la distribución de la mezcla sea aproximada por:

$$G \approx \sum_{j=1}^{L} \pi_j \delta_{\boldsymbol{\theta}_j}.$$

La ventaja principal de este truncamiento es que, en cada iteración Gibbs, sólo hay que calcular K pesos *"stick-breaking"*, por lo tanto, reduce lo que fue una vez una suma infinita a un problema de dimensión finita. Esta simplificación fue primero introducida por Ishwaran y Zarepour (2000) y su implementación en Cadenas de Markov Monte Carlo (MCMC) se le conoce como muestreador Gibbs en Bloques (Ishwaran y James, 2001).

Para estimar un *DPMM* mediante el muestreador Gibbs en Bloques, debe muestrear de las condicionales completas:

$$\boldsymbol{z} \sim p\left(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{\theta}\right), \tag{3.25}$$

$$\boldsymbol{\theta} \sim p\left(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}, H\right), \tag{3.26}$$

$$\boldsymbol{\beta} \sim p\left(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z}, \alpha\right), \tag{3.27}$$

$$\alpha \sim p\left(\alpha|\boldsymbol{z}, a, b\right), \tag{3.28}$$

$$\pi_j \sim \beta_j \prod_{l=1}^{L-1} \left(1 - \beta_l\right) \quad j = 1, \ldots, L. \tag{3.29}$$

Análogamente como en el caso de modelos de mezclas finitas, la derivación de las condicionales completas para el muestreo Gibbs envuelve la verosimilitud de los datos completos. Para los indicadores de los componentes latentes, con la condicional completa dada en la ecuación 3.25, éstos son muestreados de manera idéntica a los modelos de mezclas finitas. Asimismo, los parámetros del componente $\boldsymbol{\theta}_j$ son actualizados como en un modelo de mezclas finitas:

$$p\left(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z},H\right) \propto H \prod_{\substack{i:\\ z_i=j}} p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right) \quad j=1,\ldots,L, \tag{3.30}$$

donde L es el nivel de truncación.

Sin embargo, a diferencia del modelo de mezclas finitas, la incertidumbre en los pesos de los componentes se cambia a β, las entradas en la construcción de los pesos *"stick-breaking"*, definido en el Teorema 3.2. Para obtener su condicional completa, se re-escribe la probabilidad de los datos completos en términos de β:

$$p\left(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\pi},\boldsymbol{\theta}\right) = \prod_{i=1}^{n} \beta_{z_i} \prod_{l=1}^{z_i-1} \left(1-\beta_l\right) p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_{z_i}\right). \tag{3.31}$$

Expandiendo el producto sobre i para un nivel de truncamiento de L dada:

$$p\left(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\beta},\boldsymbol{\theta}\right) = \beta_1^{n_1}\left(1-\beta_1\right)^{\sum_{l=2}^{L} n_l} \beta_2^{n_2}\left(1-\beta_2\right)^{\sum_{l=3}^{L} n_l} \ldots \beta_L \prod_{j=1}^{L} \prod_{\substack{i:\\ z_i=j}} p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_j\right), \tag{3.32}$$

donde n_j es definida como el número de observaciones asignadas al j-ésimo grupo. La incorporación de la distribución a priori asignada en cada β_j, se convierte en su condicional completa:

$$\begin{aligned} p\left(\beta_j|\boldsymbol{y},\boldsymbol{z},\alpha\right) &\propto \beta_j^{n_j}\left(1-\beta_j\right)^{\sum_{l=j+1}^{L} n_l}\left(1-\beta_j\right)^{\alpha-1}, \\ &\sim Beta\left(1+n_j, \alpha+\sum_{l=j+1}^{L} n_l\right). \end{aligned} \tag{3.33}$$

Por último, si α se trata como desconocido, ésta se actualiza por un distribución *Gama*, con la ayuda de variables auxiliares, es decir:

$$
\begin{aligned}
p\left(\alpha|\mathbf{z}, a, b\right) &\sim Gama\left(K^* + a - \rho, b - \log\left(\varphi\right)\right), \\
\varphi &\sim Beta\left(1 + \alpha, n\right), \\
\rho &\sim Bernoulli\left(\frac{n}{n + \alpha}\right),
\end{aligned} \tag{3.34}
$$

donde K^* es el número de componentes con pesos distintos de cero resultante de la iteración actual del Gibbs.

A continuación se muestra la aplicación del muestreador Gibbs en Bloques para el *DPMM* con un nivel de truncación K y D iteraciones:

1. Extraer los valores iniciales para α, $\boldsymbol{\pi}$, y $\boldsymbol{\theta}$; $\alpha \sim Gamma(a, b)$, $\pi^0 \sim GEM(\alpha)$, $\boldsymbol{\theta} \sim H$. Para $d = 1, \ldots, D$ iteraciones,
2. Calcular las probabilidades a posteriori de los miembros para $i = 1, \ldots, n$ y $j = 1, \ldots, K$,
3. Muestrear el indicador del componente latente para $i = 1, \ldots, n$, de:

$$
\begin{aligned}
z_i^{(d)} \sim Multinomial(&p(z_i^{(d)} = 1|\boldsymbol{y}_i, \pi^{(d-1)}, \boldsymbol{\theta}^{(d-1)}), \\
&\ldots, p(z_i^{(d)} = K|\boldsymbol{y}_i, \pi^{(d-1)}, \boldsymbol{\theta}^{(d-1)})),
\end{aligned}
$$

4. Simular los pesos del *"stick-breaking"* $\boldsymbol{\beta}^{(d)}$ de la ecuación 3.33,
5. Calcular las proporciones de los componentes $\boldsymbol{\pi}$ de la ecuación 3.29, dado $\boldsymbol{\beta}^{(d)}$,
6. Simular $\boldsymbol{\theta}_j^{(d)}$ de la ecuación 3.30 para $j = 1, \ldots, K$,
7. Actualizar el parámetro de concentración de la ecuación 3.34, y
8. Regresar al paso 2.

3.3. Modelos de Mezclas para Regresión (MMR)

El análisis de regresión es el proceso de modelar la relación de cierta variable aleatoria $\boldsymbol{Y}$ contra algún vector de covariables $\boldsymbol{X} \in \mathbb{R}^q$ via funciones de densidad para $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ de la forma $p\left(\boldsymbol{y}|\boldsymbol{x}\right)$, y es más comúnmente conducido via modelos paramétricos de la forma $p\left(\boldsymbol{y}|\boldsymbol{x}\right) = p\left(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\phi}\right)$, donde $\boldsymbol{\phi}$ es el vector de parámetros. El más popular de tales modelos es el *"Modelo de Regresión Lineal Gaussiano"* de la forma:

$$p\left(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\beta}\right)=p\left(\boldsymbol{y};\boldsymbol{x}^{T}\boldsymbol{\beta},\sigma^{2}\right), \tag{3.35}$$

para $\boldsymbol{Y}$ univariada, donde $\boldsymbol{\beta}\in\mathbb{R}^{q}$ y $\sigma^{2}>0$. Sin embargo; hay también bastante literatura en modelos de regresión paramétrica y semi-paramétrica, ver por ejemplo Green y Silverman (1994), Gyorfi *et al.* (2002) y Ruppert *et al.* (2003).

Al igual que en el modelo de mezclas finitas general, si se supone que existe alguna variable latente aleatoria adicional $\boldsymbol{Z}$ con función de densidad de probabilidad $P(\boldsymbol{Z}=\boldsymbol{z}_{j})=\pi_{j}$, para cada j, tal que $\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}$, $\boldsymbol{Z}=\boldsymbol{z}_{j}$ tiene la densidad $p_{j}(\boldsymbol{y}|\boldsymbol{x})$, entonces se tiene la mezcla de modelos de regresión (MMR):

$$p\left(\boldsymbol{y}|\boldsymbol{x}\right)=\sum_{j=1}^{K}\pi_{j}p_{j}\left(\boldsymbol{y}|\boldsymbol{x}\right). \tag{3.36}$$

Ahora, la mezcla de modelos de regresión lineal Gaussiano (MMRLG) univariado más general es:

$$p\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\phi}\right)=\sum_{j=1}^{K}\pi_{j}p\left(\boldsymbol{y}_{i},\boldsymbol{\beta}_{j}^{T}\boldsymbol{x}_{i},\sigma_{j}^{2}\right), \tag{3.37}$$

donde $\boldsymbol{\phi}=\left(\pi_{1},\ldots,\pi_{K-1},\beta_{1}^{T},\ldots,\beta_{K}^{T},\sigma_{1}^{T},\ldots,\sigma_{K}^{T}\right)^{T}$ fue propuesto en Quandt (1972) y la estimación de los parámetros por máxima verosimiltud via el algoritmo EM fue dada en DeSarbo y Cron (1988).

Aquí, para mayor claridad, se definirá una mezcla lineal de modelos de regresión como algún modelo para el caso de $\boldsymbol{Y}\in\mathbb{R}^{d}$ y $E(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}_{j}=\boldsymbol{z}_{j})=\boldsymbol{B}_{j}^{T}\boldsymbol{x}_{i}$, para cada j, donde $\boldsymbol{B}_{j}\in\mathbb{R}^{q\times d}$. Luego entonces, de acuerdo con Jones y McLachlan (1989), la mezcla lineal de modelos de regresión multivariado Gaussiano es igual a:

$$p\left(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\phi}\right)=\sum_{j=1}^{K}\pi_{j}p\left(\boldsymbol{y}_{i};\boldsymbol{B}_{j}^{T}\boldsymbol{X}_{i},\boldsymbol{\Sigma}_{j}\right), \tag{3.38}$$

donde $\boldsymbol{\phi}=(\pi_{j},\boldsymbol{\theta}_{j})$, es decir $\pi_{j}=(\pi_{1},\ldots,\pi_{K-1})$ y $\boldsymbol{\theta}_{j}=(\mathrm{vec}^{T}(B_{1}),\ldots,\mathrm{vec}^{T}(B_{K})$, $\mathrm{vec}^{T}(\Sigma_{1}),\ldots,\mathrm{vec}^{T}(\Sigma_{K}))^{T}$, y $\boldsymbol{B}_{j}\in\mathbb{R}^{q\times d}$ es una matriz de coeficientes, para cada j. La m-ésima columna de $\boldsymbol{B}_{j}$ corresponde a la relación entre el componente del vector $\boldsymbol{y}_{m}$ y las covariables $\boldsymbol{x}$.

3.3.1. Maximización vía el algoritmo EM en MMR

La estimación de los parámetros de las mezclas se hace utilizando técnicas de máxima verosimilitud vía el algoritmo EM (Dang *et al.*, 2014, Dempster *et al.*, 1977, McLachlan y Krishnan, 1997). Sea $S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ una muestra de n observaciones independientes del modelo 3.38. Entonces, su función de verosimilitud es igual a:

$$L\left(\boldsymbol{\phi}|\boldsymbol{S}\right) = \prod_{i=1}^{n}\left[\sum_{j=1}^{K}\pi_j p\left(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_j\right)\right],$$

donde $\boldsymbol{S}$ son los datos incompletos de acuerdo al contexto del algoritmo EM. Luego entonces, los datos completos son $\boldsymbol{S}_c = \{(\boldsymbol{x}_1, \boldsymbol{y}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{z}_n)\}$, donde la variable latente $\boldsymbol{z}_i$ es el vector del componente etiquetado tal que $z_{ij} = 1$ si $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ provienen del j-ésimo componente y 0 en caso contrario.

Ahora, la verosimilitud correspondiente de los datos completos es:

$$L_c\left(\boldsymbol{\phi}|\boldsymbol{S}_c\right) = \prod_{i=1}^{n}\prod_{j=1}^{K}\left[\pi_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_j\right)\right]^{z_{ij}}. \tag{3.39}$$

La función de la log-verosimilitud de los datos completos es:

$$\begin{aligned} l_c\left(\boldsymbol{\phi}|\boldsymbol{S}_c\right) &= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\left[\log p\left(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_j\right) + \log \pi_j\right], \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log p\left(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_j\right). \end{aligned} \tag{3.40}$$

El *paso-E* consiste en calcular la esperanza de la log-verosimilitud de los datos completos:

$$Q\left(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t-1)}\right) = E\left\{l_c\left(\boldsymbol{\phi}|\boldsymbol{S}_c\right)\right\}$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{K}\hat{\tau}_{ij}^{(t-1)}\left[\log p\left(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_j\right) + \log\pi_j\right], \tag{3.41}$$

donde:

$$\hat{\tau}_{ij}^{(t-1)} = E\left\{\boldsymbol{Z}_{ij} = \boldsymbol{z}_{ij}|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{\phi}^{(t-1)}\right\},$$
$$= \frac{\hat{\pi}_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_j\right)}{\sum_{j=1}^{K}\hat{\pi}_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_j\right)}|_{\boldsymbol{\phi}^{(t-1)}} := \hat{\tau}_{ij}^{(t-1)}, \tag{3.42}$$

proporciona el valor actual de z_{ij} en la k-ésima iteración.

El *paso-M* consiste en maximizar la función $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t-1)})$ con respecto al conjunto de parámetros que conforman $\boldsymbol{\phi}$. Posteriormente, se repite el proceso hasta lograr convergencia.

3.3.2. Inferencias sobre el coeficiente de regresión

Los estimadores $\hat{\boldsymbol{\beta}}'s$ dependen de la muestra seleccionada, por lo tanto son variables aleatorias y presentarán una distribución de probabilidad. Estas distribuciones de probabilidad de los estimadores pueden utilizarse para construir intervalos de confianza o contrastes sobre los parámetros del modelo de regresión. Con el propósito de decidir si el efecto de la variable independiente es o no significativo para la variable dependiente. Esto es equivalente a contrastar si el coeficiente β_{ij} para $i = 1\ldots, q$ (q variables explicativas) y $j = 1, \ldots, K$ es o no significativamente distinto de cero. Luego entones:

I. La hipótesis a probar es sí el coeficiente es significativo, es decir:

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0,$$

II. El estadístico de prueba es:

$$t = \frac{\left|\hat{\beta}_j\right|}{\sqrt{\hat{V}(\hat{\beta}_j)}},$$

en la que $\hat{V}(\hat{\beta}_j)$ es la varianza asintótica del estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}_{ML}$ (ML por sus siglas en inglés, Maximum Likelihood) de acuerdo con las propiedades de normalidad asintótica de los estimadores de máxima verosimilitud (Lehmann y Casella, 1998).

Ahora, bajo ciertas condiciones generales:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}) \xrightarrow{D} N\left(0, \sqrt{I^{-1}(\boldsymbol{\beta})}\right),$$

donde:

$$I(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\left[-\frac{\partial^2 \log f(y_1, \ldots, y_n; \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right],$$

es la matriz de información de Fisher.

Luego entonces, la varianza asintótica del $\hat{\boldsymbol{\beta}}_{ML}$ es igual a:

$$V\left(\hat{\boldsymbol{\beta}}_{ML}\right) = \frac{1}{E_{\boldsymbol{\beta}}\left[-\frac{\partial^2 \log f(y_1,\ldots,y_n;\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right]} \approx \frac{1}{-\frac{\partial^2}{\partial(\boldsymbol{\beta})^2} \log(\boldsymbol{\beta}) \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ML}}},$$

III. La regla de de decisión es: **Rechazar** H_0 **al nivel** α **si** $t > t_{n-2,1-\frac{\alpha}{2}}$.

3.3.3. Selección del número de grupos

El algoritmo de agrupación llamado K-medias, propuesto por MacQueen (1967), consiste en definir grupos de manera que la varianza total dentro de los grupos se reduzca al mínimo (conocido como variación total dentro del grupo, ver Figura 3.7).

La ecuación a resolver se puede definir como sigue:

$$\text{mín}\left(\sum_{j=1}^{K}(W_j)\right),$$

donde C_j es el j-ésimo grupo y $W(C_j)$ es la varianza del grupo dentro del grupo C_j.

Hay muchas maneras de definir la varianza dentro del grupo ($W(C_j)$). El algoritmo de Hartigan y Wong (1979), utilizado por defecto en el software R (R Core Team, 2016), aplica la medida de la distancia Euclideana entre los puntos de los datos para

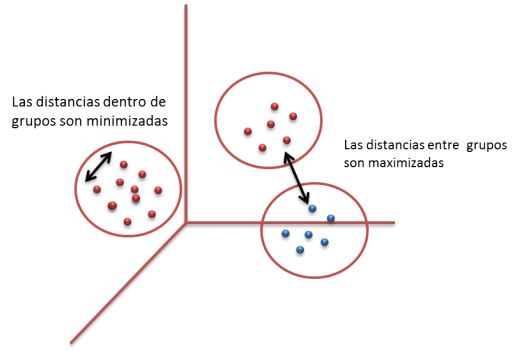


Figura 3.7: Gráfico del algoritmo K-medias. Fuente: Elaboración Propia.

determinar la similitud dentro y entre grupos. Cada observación se asigna a un grupo determinado de tal manera que la suma de cuadrados (SS) de la observación a su centros de los grupos asignados sea mínima, ver Figura 3.8.

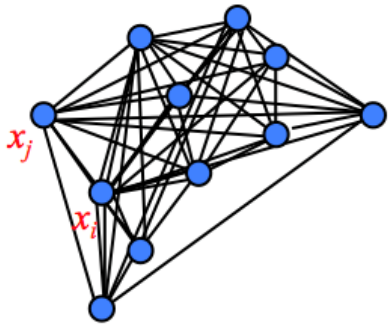


Figura 3.8: Gráfico de asignación de observaciones. Fuente: <http://cambioclimaticoglobal.com/>[Consulta: 3 febrero 2015].

Para resolver la ecuación presentada anteriormente, la variación dentro del grupo ($W(C_j)$) para un grupo determinado C_j, que contiene n_j puntos, pueden ser definida como sigue:

$$W\left(C_j\right) = \frac{1}{n_j} \sum_{\boldsymbol{y}_i \in C_j} \sum_{\boldsymbol{y}_j \in C_j} \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right)^2 = \sum_{\mathbf{y}_i \in C_j} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right)^2,$$

donde: $\boldsymbol{y}_i$ es un punto de los datos que pertenece al grupo C_j y $\boldsymbol{mu_j}$ es la media de los puntos asignados al grupo C_j.

Luego entonces; la variación dentro del grupo para un grupo C_j con n_j número de puntos se define como la suma de todos los pares de las distancias cuadradas Euclideanas entre las observaciones de C_j, dividida por n_j.

Finalmente, la suma de cuadrados total dentro del grupo (SSW, es decir, la variación

total dentro del grupo) se define como sigue:

$$SSW = \sum_{j=1}^{K} W(C_j) = \sum_{j=1}^{K} \sum_{\boldsymbol{y}_i \in C_j} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right)^2.$$

La suma de cuadrados total dentro de grupos mide la compacidad (es decir, la bondad) del grupo y se quiere que sea lo más pequeño posible <http://cambioclimaticoglobal.com/>[Consulta: 3 febrero 2015].

Capítulo 4

Datos

En este capítulo se describe la base de datos que se utiliza en la modelación de la ciclogénesis.

4.1. Descripción y Alcance Geográfico de la Base de Datos de Ciclones Tropicales

Los datos de ubicación de ocurrencia (*Longitud* y *Latitud*) de los ciclones tropicales sobre el Atlántico Norte fueron obtenidos de la base de datos de las "mejores trayectorias" o IBTrACS (ver Figura 4.1 y Figura A.2. del Apéndice A), que es un proyecto colaborativo y coordinado a escala mundial para recopilar los datos de varios de los Centros Meteorológicos Regionales Especializados de la Organización Mundial de Meteorología y otros organismos, y desde 1945 emitió registros globales (Knapp *et al.*, 2010).

4.1.1. Ciclones tropicales

En el panel a) de la Figura 4.2 se muestra el número de ciclones por año para el intervalo 1951-2013, mientras que en el panel b) se muestra la representación gráfica de la serie. En el eje horizontal se representa la escala del tiempo, y en el vertical, las frecuencias de los ciclones tropicales. En dicha gráfica se puede observar su tendencia, su variación, su estacionalidad y sus ciclos. Cabe mencionar que debido a que los datos son anuales se descarta la estacionalidad.

Asimismo, se puede observar que el número de ciclones tropicales descendió durante

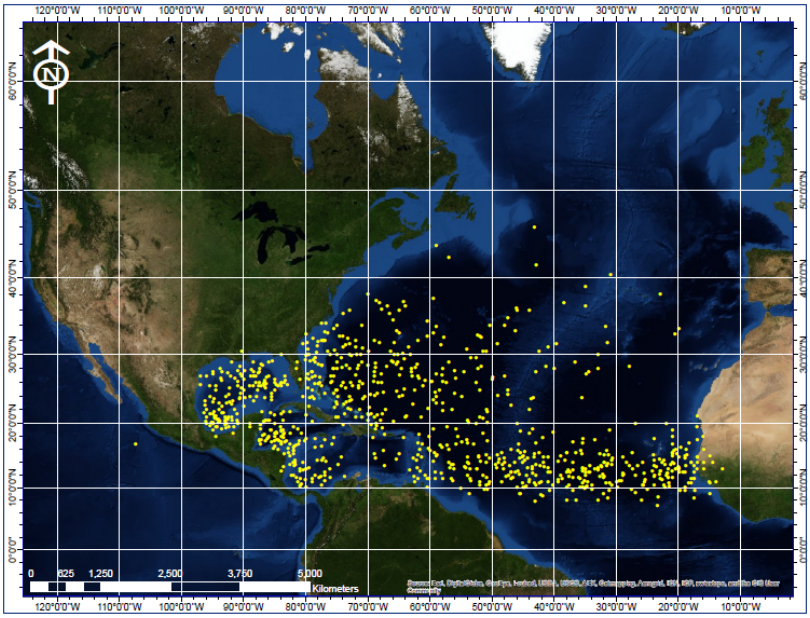
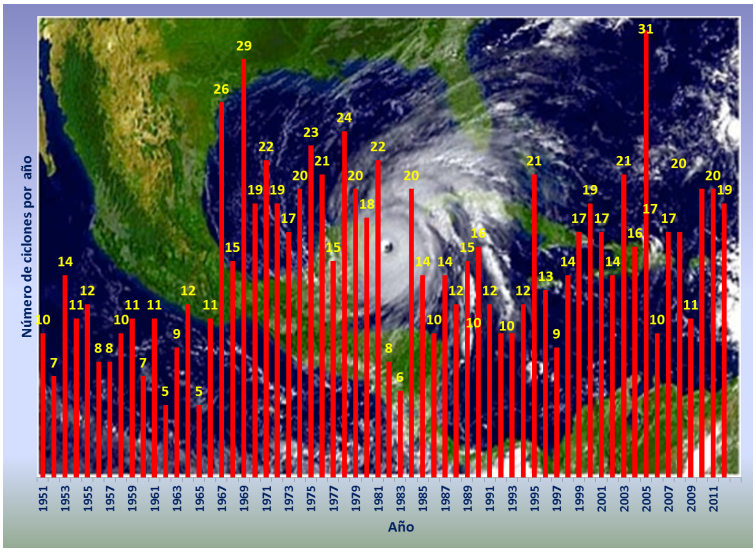


Figura 4.1: Distribución espacial de los puntos de ubicación de ocurrencia de los ciclones tropicales en la región del Atlántico Norte, desde el año de 1951 a 2013. Fuente: Elaboración Propia con base en los datos del IBTrACS.

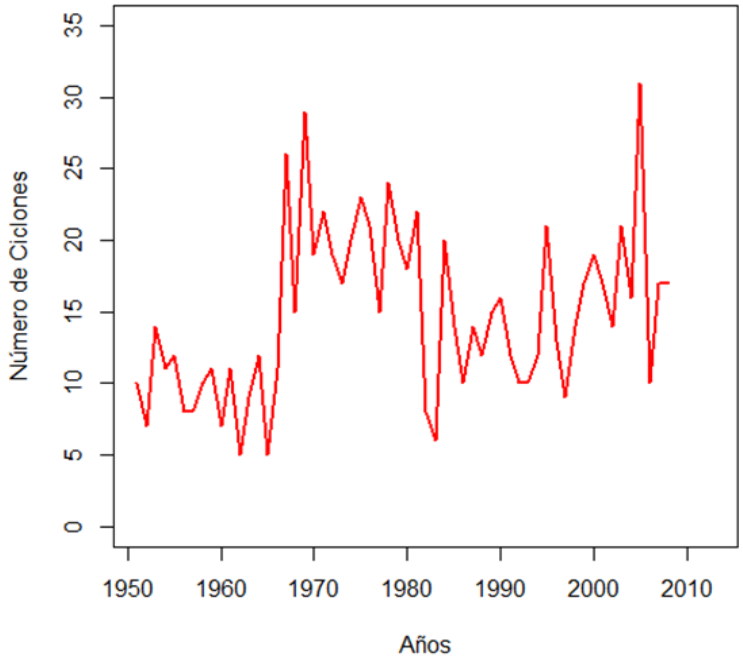
el intervalo 1951-1965 pero aumentó durante el intervalo 1966-1981, alcanzando un valor máximo de 29 ciclones en el año de 1969. Para el intervalo 1982-2013, el número de ciclones nuevamente descendió, alcanzando un valor máximo de 31 ciclones en el año 2005. Lo anterior indica que la serie observada presenta diferentes cambios estructurales debido a que presenta diferentes cambios de tendencia, es decir, de una tendencia decreciente pasa a una creciente, y así sucesivamente. Estos cambios afectan la identificación del modelo así como su pronóstico, generando el ajuste de un modelo para cada una de las partes.

Como primer paso se estimaron los puntos de cambio tanto en media como en varianza de la serie a través del algoritmo implementado por Killick y Eckley (2013) en el paquete R-ChangePoint. Ahora, dado que los datos son recuentos del número de ciclones en cada año desde 1951 hasta 2013, el modelo probabilístico ajustado a cada uno de los segmentos entre los diferentes puntos de cambio fue el Poisson con su propio parámetro de intensidad.

El número de puntos cambio encontrados en la serie de tiempo ciclones, {ciclones}, fueron tres y corresponden a los años 1966, 1981 y 1998. En consecuencia, el número de segmentos son cuatro (1951-1966, 1967-1981, 1982-1998 y 1999-2013) y sus valores promedio estimados de ciclones (λ) fueron: 9.43750, 20.66667, 12.70588 y 17.78571, respectivamente, ver Figura 4.3.



(a) Frecuencia de los ciclones tropicales por año en la región del Atlántico Norte para el intervalo 1951-2013. Fuente: Elaboración Propia con base en los datos del IBTrACS.



(b) Serie simple de los ciclones tropicales (1951-2008).

Figura 4.2: En el panel a) se muestra la gráfica de las frecuencias de los ciclones tropicales por año en la región del Atlántico Norte para el intervalo 1951-2013 y en el panel b) se muestra la gráfica de la serie simple de los ciclones tropicales. Fuente: Elaboración Propia con base en los datos del IBTrACS.

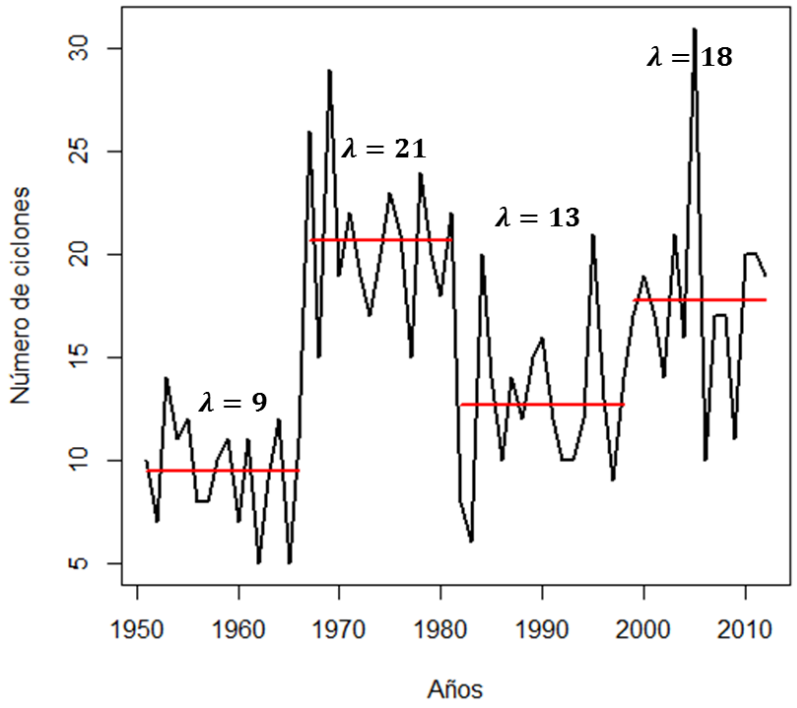


Figura 4.3: Serie de tiempo simple de los ciclones con sus puntos de cambio y sus valores promedio de ciclones por segmento. Fuente: Elaboración Propia.

4.1.2. Temperatura media mundial y temperatura media de las superficie del mar y del Atlántico Norte

De acuerdo con el panel a) de la Figura 4.4, las variaciones de la temperatura media mundial de la superficie de la tierra, que es el promedio de la temperatura del aire cerca de la superficie de la tierra y la temperatura de la superficie del mar, han aumentado desde el decenio de 1860 a 2000. Durante el siglo XX, este aumento fue de 0.6°C y 1998 y 2005 los años más calurosos desde que se tienen registros instrumentales. Además, la tasa promedio de calentamiento durante los últimos 50 años (0.13°C ± 0.03°C por decenio) es casi el doble de la tasa de los últimos 100 años (Houghton *et al.*, 2001).

Análogamente como en la temperatura media mundial de la superficie de la tierra, la tasa promedio de calentamiento de la temperatura media mundial de la superficie del mar es positiva. Siendo 1935 el año umbral de dicho aumento, ver inciso b) de la Figura 4.4. Por el contrario, la temperatura media de la superficie del mar en el Atlántico Norte presenta un comportamiento diferente a la de la temperatura media global de la superficie de la tierra y a la de la temperatura media global de la superficie del mar.

De acuerdo con el panel c) de la Figura 4.4, durante el intervalo 1980-1927, se presentaron TSM por abajo de cero, a diferencia del intervalo 1928-1968 que registró temperaturas por arriba de cero. Sin embargo; durante el intervalo 1970-1980 nuevamente la temperatura descendió por debajo del cero y la del intervalo 1981-2008 nuevamente registró temperaturas por encima de cero.

4.1.3. Definición de los intervalos de estudio

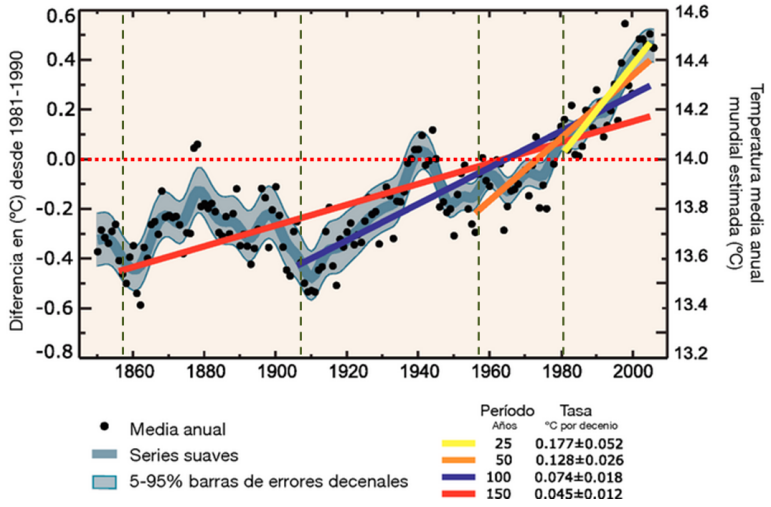
De acuerdo con la información de la Figura 4.3 y el panel c) de la Figura 4.4 de las secciones 4.1.1 y 4.1.2, respectivamente, se construyó la Figura 4.5, en la cual se puede observar la frecuencia de ciclones por año en cada segmento de los tres puntos de cambio versus la temperatura de la superficie del mar. En el panel a) de dicha figura se puede ver que cuando la TSM del aumenta el número de ciclones disminuye y viceversa cuando la TSM disminuye la frecuencia de ciclones aumenta.

Luego entonces, el primer intervalo se determinó con base en el calentamiento ocurrido durante el siglo XX en los intervalos de 1928-1975 y 1976-2013, ver panel b) de la Figura 4.5. Cabe mencionar que se tomó como valor de inicio el año 1950 porque a partir de este año fue usado rutinariamente el reconocimiento aéreo para supervisar los ciclones tropicales. Esto significa que la información sobre ciclones medidos antes de esta fecha es menos confiable (Vecchi y Knutson 2008 citados por Villarini *et al.*, 2011).

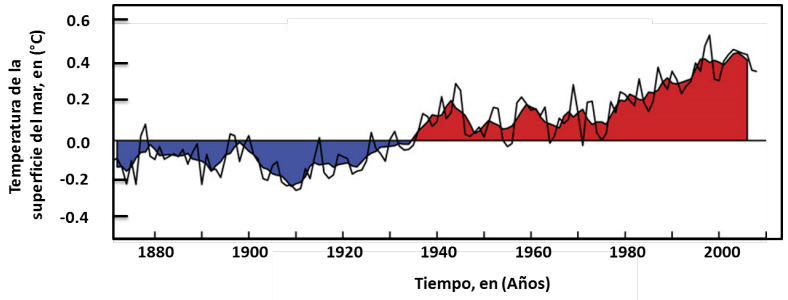
Respecto al segundo intervalo, éste se determinó de acuerdo con Webster *et al.* (2005), quienes observaron un aumento considerable en los ciclones tropicales en todas las cuencas oceánicas durante los últimos 30 años para las categorías más fuertes (4 y 5 de acuerdo con la escala de Saffir-Simpson) debido al incremento de la TSM. Cabe mencionar que dicho intervalo de tiempo lo dividieron en dos intervalos 1975-1989 vs 1990-2004.

4.2. Descripción y Alcance Geográfico de la Base de Datos de las Temperaturas de la Superficie Mar (TSM)

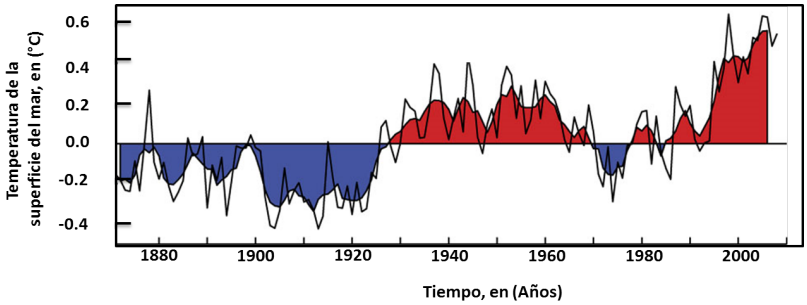
La TSM de cada uno de los puntos de génesis de los ciclones tropicales en el Atlántico Norte se generó sobreponiendo las coordenadas de *Longitud* y *Latitud* de cada uno de éstos en el mapa de TSM promedio correspondiente al día en que se originó el ciclón. Este proceso se hizo con el Software ArcGIS [software GIS]. Versión 10.0. Redlands, CA: Environmental Systems Research Institute, Inc., 2010. Los mapas de TSM se ge-



(a) Variaciones de la temperatura media mundial de la superficie de la tierra para el intervalo 1860-2000. Fuente: Trenberth *et al.* (2007).

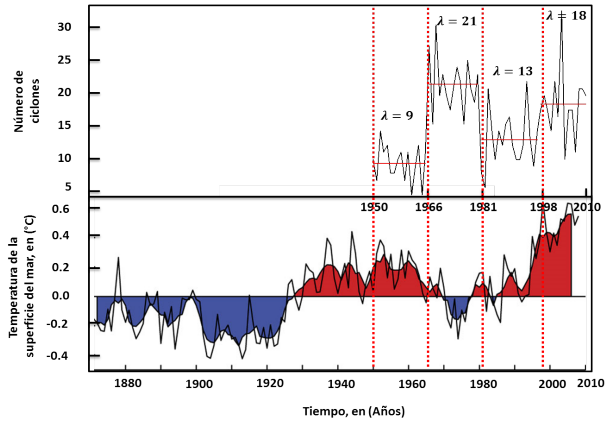


(b) Variaciones de la temperatura media mundial de la superficie del mar para el intervalo 1870-2008. Fuente: Knudsen *et al.* (2011).

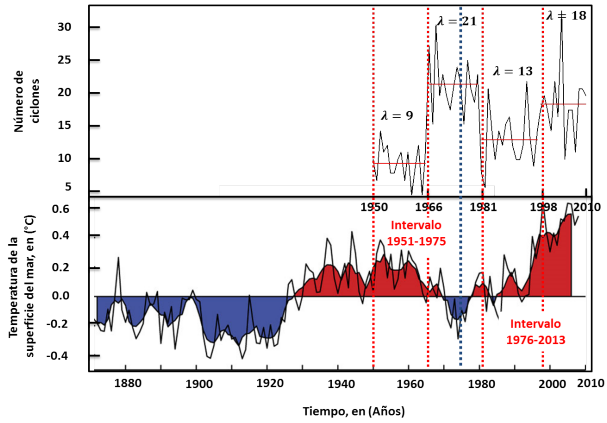


(c) Variaciones de la temperatura media de la superficie del mar en el Atlántico Norte para el intervalo 1870-2008. Fuente: Knudsen *et al.* (2011).

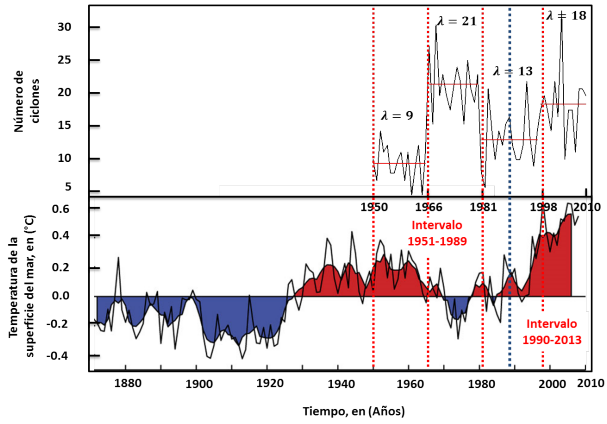
Figura 4.4: Variaciones de la temperatura media mundial y variaciones de la temperatura media de la superficie del mar global y del Atlántico Norte.



(a) Puntos de cambio en la serie ciclones tropicales vs Temperatura de la superficie del mar en el Atlántico Norte.



(b) Serie simple de los ciclones tropicales con puntos de cambio e intervalos de estudio 1951-1975 vs 1976-2013.



(c) Serie simple de los ciclones tropicales con puntos de cambio e intervalos de estudio 1951-1989 vs 1990-2013.

Figura 4.5: Gráfica de la serie simple de los ciclones tropicales con puntos de cambio e intervalos de estudio. Fuente: Elaboración Propia con gráfica tomada de Knudsen *et al.* (2011).

neraron de la base de datos de la Administración Oceánica y Atmosférica Nacional de los Estados Unidos (NOAA, por sus siglas en inglés: National Oceanic and Atmospheric Administration, puede consultarse en: http://www.ospo.noaa.gov/Products/ocean/sst/contour/), y las coordenadas de la ciclogénesis tal y como ya se mencionó se obtuvieron de la base de datos del IBTrACS.

En la Figura 4.6, se muestra las temperaturas de cada uno de los ciclones desde 1951 hasta 2013.

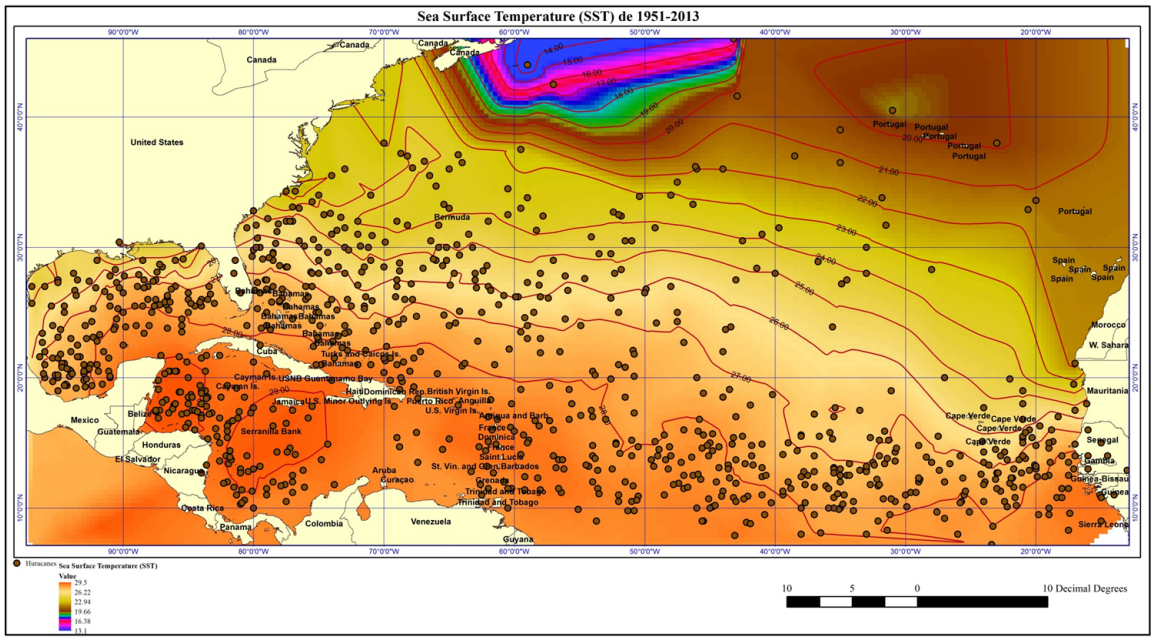


Figura 4.6: Distribución espacial la temperatura de la superficie del mar de los puntos de ubicación de ocurrencia de los ciclones tropicales para el intervalo 1951-2013. Fuente: Elaboración Propia con datos de la TSM, generados de la base de datos de la NOAA.

Capítulo 5

Modelo de Mezclas Gaussianas para determinar las Regiones de Ciclogénesis del Atlántico Norte e identificar sus cambios

5.1. Introducción

Existe un interés creciente en conocer el impacto del calentamiento global en la actividad de los ciclones tropicales, mismo que surge del hecho de que los ciclones tropicales están dentro de los fenómenos climatológicos más destructivos entre los desastres naturales. A medida que la temperatura global promedio de la superficie del planeta se incrementa, se espera que la intensidad, la frecuencia, las trayectorias, la ubicación de ocurrencia y la lysis (llegada a tierra de los ciclones) de estos fenómenos meteorológicos se alteren por el clima de hoy en día. Walsh (2004) menciona que aunque no hay en este momento cambios perceptibles en las características de los ciclones tropicales, que razonablemente podrán atribuirse al calentamiento global, predicciones de simulaciones de modelos de circulación general sugieren incrementos en su intensidad máxima entre 5 y 10 % para el 2050. Respecto a las regiones de formación de huracanes, éstas probablemente no cambien, y respecto a los cambios en el número de ciclones o trayectorias poco consenso ha habido al respecto. Cabe mencionar que existe incertidumbre en las predicciones climáticas por algunas deficiencias en los modelos; por lo tanto, si las predicciones de intensidades son correctas sus cambios se detectarán en el Atlántico después de 2050. Utilizando los registros de las mejores trayectorias del Centro de Advertencia de Tifones de la Marina Estadounidense (Joint Typhoon Warning Center, JTWC) y de la Agencia Atmosférica y Oceanográfica Nacional (National Oceanographic and Atmospheric Administration, NOAA), Emanuel

(2005), Webster *et al.* (2005), y más recientemente Elsner *et al.* (2008) demostraron que la intensidad histórica de las tormentas ha aumentado tanto en el Pacífico Nor-occidental (WNP) como en el Atlántico Norte. Webster *et al.* (2005) observaron un aumento considerable en los ciclones tropicales en todas las cuencas oceánicas durante los últimos 30 años para las categorías más fuertes (4 y 5 de acuerdo con la escala de Saffir-Simpson). En particular, en las regiones ciclogenéticas del Pacífico Nor-occidental y del Atlántico Norte, 25 y 20 %, respectivamente, de estas tormentas se presentaron en el intervalo de 1975-1989 mientras que de 1990-2004 se produjeron 41 y 25 %, respectivamente; esto significa que hubo aumentos de 16 y 5 %, respectivamente. Sin embargo, los resultados obtenidos para la cuenca oceánica del Pacífico Nor-occidental han sido cuestionados, ya que la aparente tendencia que se observa en la intensidad de los ciclones es parte de un gran oscilación interdecadal (Chan, 2006) o de posibles errores de medición en el conjunto de datos (Knaff y Zehr, 2007). Adicionalmente, Klotzbach (2006) con base en el análisis de los registros de las mejores trayectorias del intervalo de 1986-2006 encontró que la tendencia de la intensidad de los ciclones tropicales para la cuenca del Atlántico Norte no muestra evidencia de que haya cambiado, y la tendencia para el Pacífico Nor-occidental presenta una baja considerable. Kossin *et al.* (2007), también con base en el análisis de los registros de las mejores trayectorias encontraron que no hay un aumento en la intensidad de los huracanes en cualquier cuenca distinta a la del Atlántico Norte en las últimas dos décadas (1985-2005).

Recientemente, Mori *et al.* (2013) simularon mediante un modelo estadítico el impacto del calentamiento global sobre los centroides de la ciclogénesis y de la ciclolysis de las diferentes cuencas oceánica para finales del siglo XXI. Ellos encotraron que los centroides de la ciclogénesis se desplazarán hacia el centro de las cuencas oceánicas y que los cambios futuros en las condiciones dinámicas y termodinámicas en los océanos influirán en la frecuencia de la génesis de los ciclones tropicales. Además que los ciclones que se desarrollan en la parte central del océano durarán más tiempo debido a que las temperaturas de la superficie del mar son más cálidas que las de las orillas (Chan, 2007, Yokoi y Takayabu, 2009), y que los cambios en la intensidad de los ciclones tropicales estarán más relacionados con el desplazamiento de los centroides que con el cambio de la temperatura del océano.

Por lo antes mencionado, es de interés estudiar el número regiones de ocurrencia de los ciclones tropicales y si sus centroides han experimentado algún cambio dentro de la cuenca oceánica del Atlántico Norte.

Bajo este contexto, en este trabajo de investgación se aplicó un modelo de mezclas *Gaussianas* (Sección 5.2), por una parte para determinar el número de regiones ciclogenéticas y por otra para determinar los cambios temporales y espaciales de los centroides de dichas regiones en la cuenca oceánica del Atlántico Norte, para dos intervalos (1951-1975 versus 1976-2013 y 1951-1989 versus 1990-2013). Se utilizaron los puntos de inicio de los datos de las "mejores trayectorias" o IBTrACS (por siglas en inglés: International Best Track Archive for Climate Stewardship, se puede consultar en:

https://www.ncdc.noaa.gov/ibtracs/index.php?name=ibtracs-data-access). Se determinaron las funciones de densidad de probabilidades de las regiones de génesis a través del algoritmo Esperanza-Maximización (EM), las cuales se evaluaron y compararon para verificar los cambios espacio-temporales. En la Sección 5.3, se muestran los resultados y su relación e interpretación con el fenómeno natural en estudio. Finalmente, en la Sección 5.4 se concluye con una breve discusión.

5.2. Materiales y Métodos

5.2.1. Descripción de la Base de Datos de la Ciclogénesis Tropical

Los datos de ciclones tropicales en Atlántico Norte provienen de la "mejores trayectorias" o IBTrACS. y cubren el intervalo 1951-2013. Dicho intervalo fue dividido en dos intervalos: 1951-1975 vs 1976 vs 2013 y 1951-1989 vs 1990-2013. El conjunto de datos contienen mediciones de la ubicación del centro de los ciclones tropicales en *Latitud* y *Longitud*, para mayor detalle ver el Capítulo 4.

5.2.2. Modelos de Mezclas Gaussianas (*MMG*)

Bajo el modelo de mezclas finitas de distribuciones normales, ajustado en este estudio, cada punto $\mathbf{y}_i = (Latitud_i, Longitud_i)$ de ubicación de ocurrencia de los ciclones se considera que proviene de una súper población, la cual es una mezcla de un número finito K de poblaciones en algunas proporciones o pesos $\pi_1, \ldots, \pi_K$, respectivamente, donde: $\sum_{j=1}^{K} \pi_j = 1$ y $\pi_j > 0$.

De acuerdo con la ecuación 3.1, la función de densidad probabilidad (fdp) de una observación $\boldsymbol{y}$ (de dimensión-d) en la forma de mezcla finita Gaussiana es:

$$p\left(\boldsymbol{y}|\boldsymbol{\phi}\right) = \sum_{j=1}^{K} \pi_j \cdot \frac{1}{(2\pi)^{\frac{d}{2}} \left|\boldsymbol{\Sigma}_j\right|^{\frac{1}{2}}} \cdot \exp^{-\frac{1}{2}\left[\left(\boldsymbol{y}-\boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1} \left(\boldsymbol{y}-\boldsymbol{\mu}_j\right)\right]} \tag{5.1}$$

donde $p(\boldsymbol{y}|\boldsymbol{\phi})$ es la función de densidad de probabilidade correspondiente a K-ésimo componente, π_j es el peso de la mezcla y $\boldsymbol{\theta}$ denota los parámetros desconocidos de los elementos de los vectores de las medias $\boldsymbol{\mu}_j$, y elementos distintos de matrices de covarianzas $\boldsymbol{\Sigma}_j$ para $j = 1, \ldots, K$, que pertenece a algún espacio de parámetros $\boldsymbol{\Theta}$.

5.2.2.1. Estimación de los parámetros utilizando el algoritmo EM

Existen varios procedimientos para determinar los parámetros de un modelo de mezclas Gaussianas (*MMG*) de un conjunto de datos (McLachlan y Basford, 1988). Sin embargo, el método más popular y mejor establecido es el de máxima verosimilitud. En este trabajo, la estimación de los parámetros se hizo utilizando técnicas de máxima verosimilitud vía el algoritmo EM (Dempster *et al.*, 1977, Redner y Walker, 1984).

El algoritmo EM en mezclas Gaussianas es un proceso iterativo que consiste de dos etapas: Esperanza (E) y Maximización (M). En el caso de componentes Gaussianos, la densidad de la mezcla contiene los siguientes parámetros ($\boldsymbol{\phi}$): π_j, $\boldsymbol{\mu}_j$ y $\boldsymbol{\Sigma}_j$ donde $j = 1, \ldots, K$. Ahora, de acuerdo con la ecuación 3.6, la log-verosimilitud condicionada esperada para el conjunto de datos completos, conocida como función Q, es igual a:

$$Q\left(\boldsymbol{\phi}; \boldsymbol{y}\right) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \left\{ \log \left|\boldsymbol{\Sigma}_j\right| + \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right) \right\} + \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log\left(\pi_j\right) - \frac{dn}{2} \log\left(2\pi\right) \quad (5.2)$$

El *paso-E* consiste en la estimación de las probabilidades de pertenencia $\tau_{ij}^{(t-1)}$ de acuerdo con la ecuación 3.8 y 3.9.

$$\hat{\tau}_{ij}^{(t-1)} = \frac{\hat{\pi}_j^{(t-1)} \cdot p\left(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_j^{(t-1)}, \hat{\boldsymbol{\Sigma}}_j^{(t-1)}\right)}{\sum_{j=1}^{K} \hat{\pi}_j^{(t-1)} \cdot p\left(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_j^{(t-1)}, \hat{\boldsymbol{\Sigma}}_j^{(t-1)}\right)} \quad (5.3)$$

El *paso-M* se maximiza la ecuación 3.9 con respecto a ϕ:

$$\begin{aligned} \hat{\pi}_j^{(t)} &= \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{ij}^{(t)} \\ \hat{\boldsymbol{\mu}}_j^{(t)} &= \frac{\sum_{i=1}^{n} \hat{\tau}_{ij}^{(t-1)} \boldsymbol{y}_i}{\sum_{i=1}^{n} \hat{\tau}_{ij}^{(t-1)}} \\ \hat{\boldsymbol{\Sigma}}_j^{(t)} &= \frac{\sum_{i=1}^{n} \hat{\tau}_{ij}^{(t-1)} \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_j^{(t-1)}\right) \left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_j^{(t-1)}\right)^T}{\sum_{i=1}^{n} \hat{\tau}_{ij}^{(t-1)}} \end{aligned} \quad (5.4)$$

El algoritmo EM comienza con una estimación inicial del valor de los parámetros $\boldsymbol{\phi}$, llamada $\boldsymbol{\phi}^{(t-1)}$. Luego, mediante las ecuaciones 5.3 y 5.4 se estima el valor de los nuevos parámetros, llamado $\boldsymbol{\phi}^{(t)}$. El proceso se repite hasta que la diferencia entre dos evaluaciones sucesivas de la log-verosimilitud sea tan pequeña como se quiera, es decir:

$$|l\left(\boldsymbol{\phi}^{(t+1)}|\boldsymbol{y}\right) - l\left(\boldsymbol{\phi}^{(t)}|\boldsymbol{y}\right)| < \varepsilon. \tag{5.5}$$

Cabe mencionar que este resultado depende fuertemente de la selección de los parámetros iniciales (Seidel *et al.*, 2000).

5.2.2.2. Inicialización en el algoritmo EM

El algoritmo EM es un procedimiento iterativo de maximización que depende de los parámetros iniciales, ya que la función de verosimilitud puede tener máximos locales (McLachlan y Peel, 2000). Por lo tanto, una buena inicialización es crucial para encontrar los estimadores de máxima verosimilitud.

Se han sugerido diferentes procedimientos de inicialización en la literatura (Figueiredo y Jain, 2000, Maitra, 2009); sin embargo, ningún método supera uniformemente los demás. En este trabajo, se utilizó el procedimiento de Fraley y Raftery (2006), implementado en el paquete R-MCLUST, para encontrar los valores de los parámetros iniciales que permiten obtener el valor máximo en el marco de mezclas Gaussianas multivariadas, ya que es uno de los algoritmos más comunes y que mejor funcionan.

5.2.2.3. Identificación del número óptimo de componentes o grupos (K)

Hay una lista vasta de literatura dedicada al tema de la elección de K. McLachlan y Peel (2000) proporcionan una interpretación detallada de los diferentes enfoques disponibles para abordar este problema. La mayoría de los métodos destinados a la estimación de K se dividen generalmente en dos categorías: modelos basados en el principio de la parsimonia y modelos basados en procedimientos de prueba, ambos sustentados en la función de log-verosimilitud. Sin embargo; en este estudio K se determinó mediante un método heurístico, conocido como partición alrededor de los medoides (PAM por sus siglas en inglés, Partitioning Around Medoids).

El algoritmo de la PAM se basa en la formación de K particiones u objetos representativos (medoides) de n observaciones de un conjunto de datos. Se eligen aleatoriamente K medoides de un conjunto de datos. El medoide, el cual representa un grupo, se ubica en el centro del grupo. Los objetos restantes se agrupan con el medoide al que son

más similares basándose en la distancia entre el objeto y el medoide. La estrategia entonces es reemplazar uno de los medoides por los no medoides, siempre y cuando la calidad del agrupamiento mejore. Esta calidad es estimada usando una función de costo (distancia Euclidiana) que mide el promedio de disimilaridad o diferencia entre un objeto y el medoide de su grupo. Un medoide se define como la observación de un agrupamiento cuya diferencia promedio, con respecto a todas las observaciones en el grupo, es mínima (Kaufman y Rousseeuw, 2005).

El método PAM genera una gráfica, conocida como gráfica de "siluetas", para cada observación, la cual muestra una medida que indica la calidad de la clasificación. Valores cercanos a 1 indican que la observación está bien situada en su grupo, valores cercanos a 0 significa que la observación podría pertenecer a otro grupo y valores cercanos a -1 la observación es pobremente clasificada. En la gráfica hay una medida resumen denominada Ancho Promedio de Silueta o Coeficiente de Silueta que se interpreta de acuerdo con la Tabla 3.2.

5.2.2.4. Medida de la distancia entre *MMG*

De acuerdo con Sfikas *et al.* (2005) y Fukunaga (1993), la distancia de Bhattacharyya puede ser utilizada para comparar fdp de modelos de mezclas Gaussianas, y además tiene una expresión en forma cerrada. Esta distancia se utilizó para medir la distancia entre los grupos de los ciclones tropicales para los dos intervalos de estudio y se define como sigue:

$$d_B(f,g) = \frac{1}{8}\left(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g\right)^T \left(\frac{\boldsymbol{\Sigma}_f - \boldsymbol{\Sigma}_g}{2}\right)^{-1} \left(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g\right) + \frac{1}{2}\ln\left[\frac{\left|\frac{\boldsymbol{\Sigma}_f + \boldsymbol{\Sigma}_g}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_f|\,|\boldsymbol{\Sigma}_g|}}\right] \tag{5.6}$$

donde $\boldsymbol{\mu}_f$, $\boldsymbol{\Sigma}_f$ y $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ corresponden a los vectores de medias y las matrices de varianzas y covarianzas de los núcleos de las densidades Gaussianas, respectivamente.

5.2.2.5. Estadístico de prueba para la comparación de las fdp de los *MMG*

Se aplicó la técnica de re-muestreo paramétrica para obtener el percentil 95 de la distribución empírica del estadístico de prueba, que es la distancia de Bhattacharyya entre los componentes de los modelos de mezclas Gaussianas, para verificar si existen diferencias estadísticamente significativas entre ellos (Engel, 2010).

El método de re-muestreo fue propuesto por Efron (Efron, 1979, Efron y Tibshirani,

1993) para encontrar intervalos de confianza en situaciones donde es imposible obtener analíticamente la distribución muestral del estimador (Para mayor detalle ver DiCiccio y Tibshirani, 1987, Hall, 1988). Esta técnica se sustenta teóricamente en dos consideraciones: 1) La Función de Distribución verdadera $F(\mathbf{y})$ se estima mediante la Función de Distribución Empírica $\hat{F}(\mathbf{y})$ de acuerdo con el teorema Glivenko-Cantelli que muestra que $\hat{F}(\mathbf{y}) \xrightarrow{P} F(\mathbf{y})$ conforme $n \longrightarrow \infty$ (Bickel y Freedman, 1981), y 2) De acuerdo a la propiedad de consistencia, la $F(\hat{\theta})$ de una muestra dada puede aproximarse mediante la distribución muestral del re-muestreo $\hat{F}^*(\hat{\theta}^*)$ cuando el número de re-muestreos es grande y puede también aproximarse $\hat{F}(\mathbf{y})$ a $F(\mathbf{y})$. Bajo estos supuestos, Babu y Singh (1983) demostraron que $\hat{F}^*(\hat{\theta}^*) \approx F(\hat{\theta})$ cuando el número de re-muestreos es suficientemente grande.

Procedimiento re-muestreo. 1) Se genera una variable aleatoria $U \sim Uniforme(0,1)$, 2) Sí $U \in$ al intervalo $\left[\sum_{j=1}^{K} \pi_j, \sum_{j=1}^{K+1} \pi_{j+1}\right)$, donde π_j corresponde a la probabilidad del j-ésimo componente del modelo de mezclas, entonces, generar variables aleatorias a partir de la distribución del j-ésimo componente. Dicho componente tiene una distribución normal bivariada, obtenida mediante el algoritmo EM. 3) Se repiten los pasos 1) y 2) hasta que tenga la cantidad deseada de muestras de la mezcla de la distribución. 4) Se definen a las variables aleatorias $\mathbf{Y}_i \overset{iid}{\sim} N_2\left(\binom{\mu_1}{\mu_2}, \mathbf{\Sigma}\right)$, $i = 1, \ldots, n$ obtenidas en el punto anterior como muestra aleatoria; 5) Se obtiene un re-muestreo $\mathbf{Y}^* = \{\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*\}$; muestreando aleatoriamente n veces con reemplazo los datos originales $\mathbf{y}_1, \ldots, \mathbf{y}_n$, el tamaño de la muestra aleatoria es el mismo de la muestra de re-muestreo, y las $\mathbf{Y}_i^*$ tienen probabilidad n^{-1}, siendo igual en cada una de las $\mathbf{Y}_i$; 6) Se calcula el estadístico $\hat{d_B}(f,g)$ (Distancia de Bhattacharyya, ver sección 5.2.2.4) de este re-muestreo, produciendo $\hat{d_B}^*(f,g)$; 7) Se repite el paso 4 B veces, donde $B = 1000$ para esta investigación; y 6) Se construye la distribución de probabilidad de las B $\hat{d_B}^*(f,g)$, asignando probabilidad B^{-1} a cada $\hat{d_{B_i}}^*(f,g)$. Esta es la estimación de la distribución muestral de $\hat{d_B}(f,g)$, $\hat{F}^*(\hat{d_B}^*(f,g))$.

Prueba de hipótesis a partir del re-muestreo. La técnica de re-muestreo permite realizar la prueba de hipótesis de similitud de las fdp de los grupos de los modelos de mezclas Gaussianas entre los diferentes intervalos, es decir: $H_0 : d_B(f,g) = 0$. La regla de decisión es rechazar H_0 si $d_B(f,g)$ es grande. El procedimiento consiste en: 1) De los datos originales de la muestra se obtienen los estimadores (π, θ) de los componentes del modelo de mezclas normal bivariado mediante el algoritmo EM; 2) Mediante el procedimiento de re-muestreo descrito arriba se obtiene $d_B(f,g)$ bajo H_0 (es decir: $\hat{d_B}(f,g)_{H_0}$) realizando 1000 muestras de re-muestreo de la distribución normal bivariada bajo $H_0 : d_B(f,g) = 0$. Cada re-muestreo deberá ser del mismo tamaño que el de la muestra inicial. 4) Se calcula el estadístico $\hat{d_B}(f,g)$ para cada muestra de re-muestreo, y con ellos se construye la función de distribución empírica de $\hat{d_B}(f,g)$. 5) La prueba de hipótesis de la distancia de las fdp de los grupos es: $H_0 : \hat{d_B}(f,g)$, con $\alpha = 0.05$, esto equivale a obtener el percentil 95, y rechazar H_0 si $\hat{d_B}(f,g)$ es mayor que éste.

5.3. Resultados

5.3.1. Estimación de la función de densidad de probabilidad

De acuerdo con la Figura 5.1, el número de grupos por intervalo (1951-1975 versus 1976-2013 y 1951-1989 versus 1990-2013) es dos (es decir $K = 2$ componentes). Su Ancho Promedio de silueta fue 0.585 para ambos intervalos (ver Figura 5.2), lo que significa que tienen una estructura razonable (ver Tabla 3.2). Los valores de los parámetros iniciales para cada uno de los componentes de las mezclas ($K = 2$) se determinaron mediante el algoritmo M-CLUST. Los parámetros ϕ se estimaron iterativamente mediante el algoritmo EM. La función de densidad estimada para cada uno de los modelos de mezclas Gaussianas ajusta a la distribución espacial de los puntos de ubicación de ocurrencia de los ciclones tropicales en ambos intervalos (ver Figura 5.3, 5.4, y 5.5).

Adicionalmente, en la Figura 5.6 se muestra la ubicación de los centroides de cada uno de los grupos para ambos intervalos de estudio.

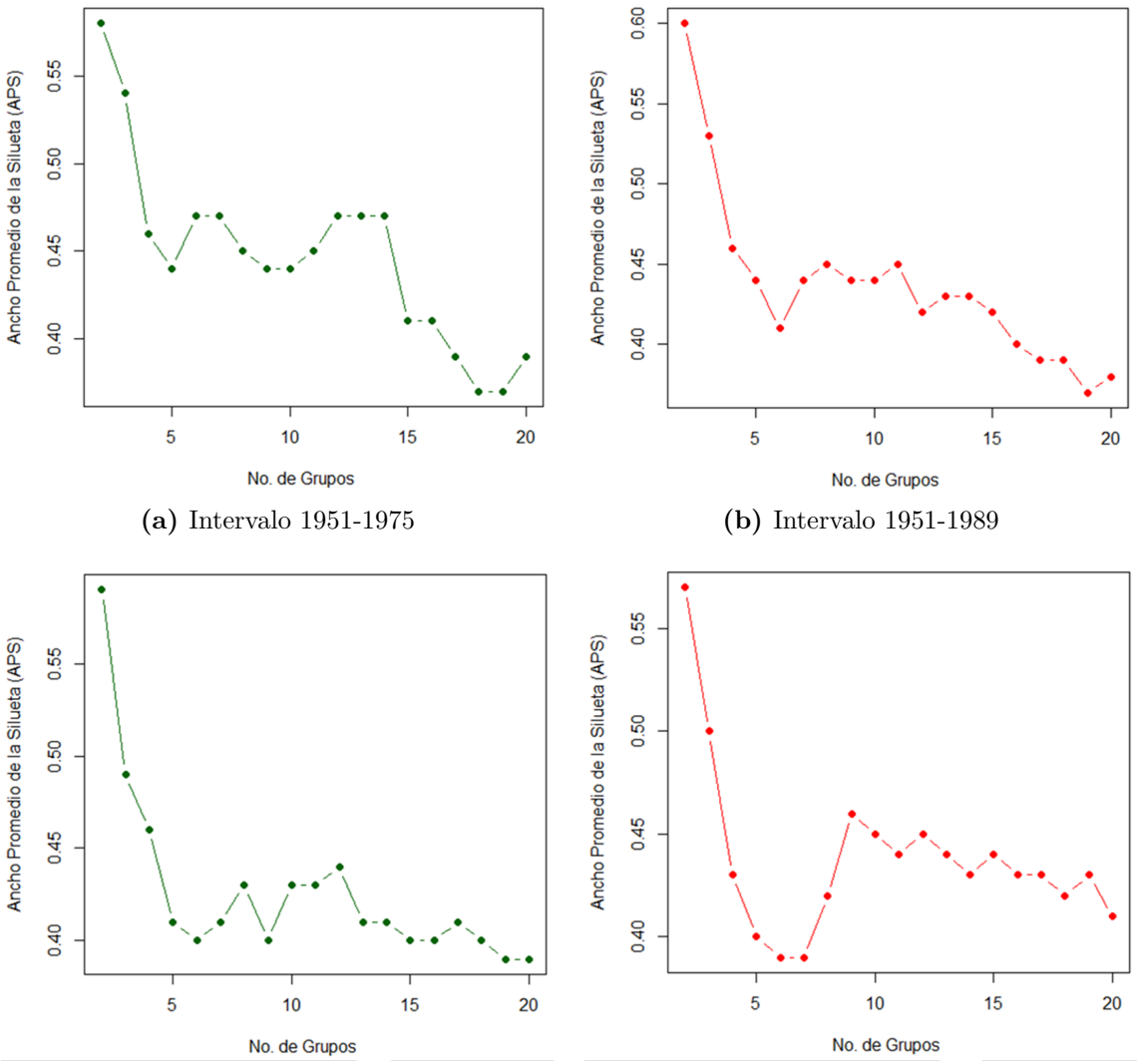
5.3.1.1. Prueba de re-muestreo paramétrica para comparar las fdp de los *MMG*

La obtención del percentil 95 de la distribución empírica del estadístico $\hat{d_B}(f, g)$, equivale a rechazar H_0 si $\hat{d_B}(f, g) > q_{0.95}$. El estadístico y el valor critico al 5 % de la distancia de Bhattacharyya de los grupos fue de $\hat{d_B}(f, g) = 0.081$, por lo que se rechaza la hipótesis nula de las distancias de las fdp de los grupos de los diferentes modelos de mezclas Gaussianas, indicando que estadísticamente existen diferencias significativas entre ellas.

5.4. Discusión y conclusiones

De acuerdo con los resultados obtenidos, en cada uno de los intervalos de estudio se encontraron dos regiones de génesis de ciclones tropicales en la cuenca oceánica del Atlántico Norte. Por lo tanto se concluye que solamente hay dos regiones que generan los ciclones tropicales desde 1951 hasta el 2013.

Los centroides de las dos regiones ciclogenéticas ubicadas en la cuenca oceánica del Atlántico Norte han experimentado cambios en su localización en el intervalo de estudio. Específicamente, el centroide de la región ciclogenética del Golfo de México se desplaza hacia el nor-este de la Costa-Este mientras que el de la región ubicada en



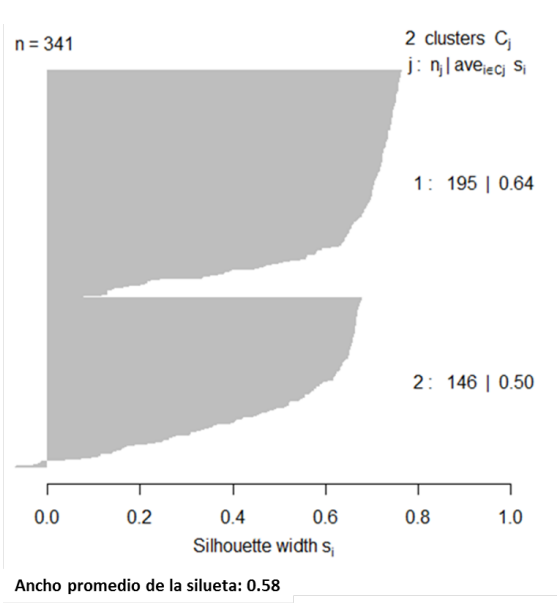
(a) Intervalo 1951-1975

(b) Intervalo 1951-1989

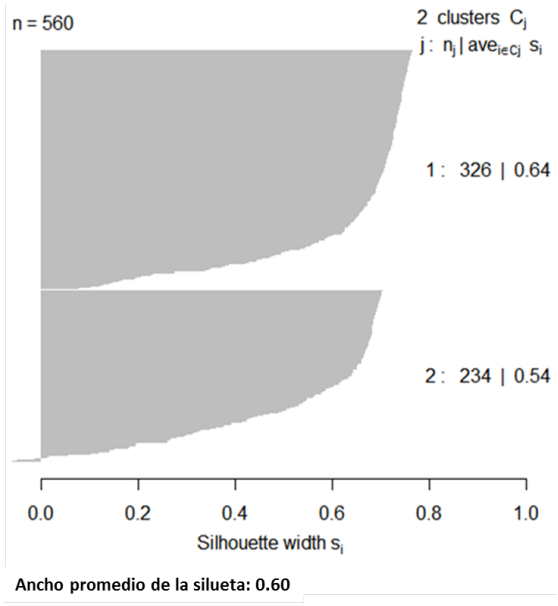
(c) Intervalo 1976-2013

(d) Intervalo 1990-2013

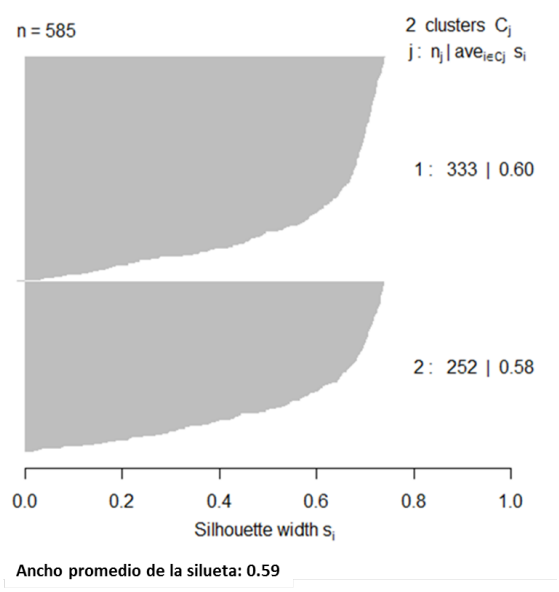
Figura 5.1: La gráfica número de grupos versus ancho promedio de silueta (APS) muestra que a medida que aumenta el número de grupos por intervalo disminuye el APS en las cuatro gráficas, pero disminuye la consistencia en su estructura de acuerdo con la Tabla 3.2.



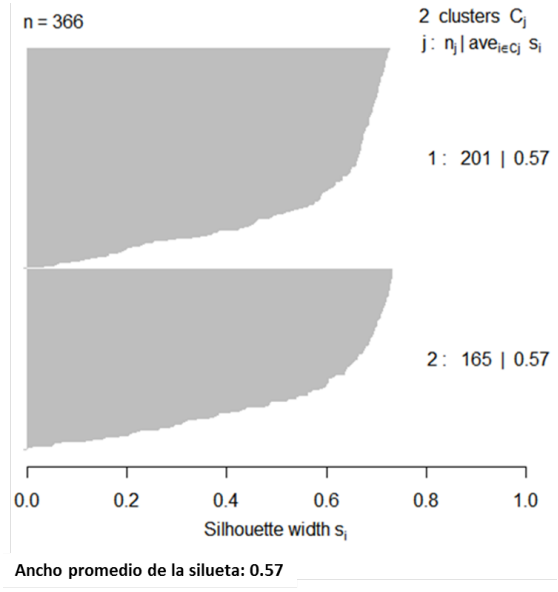
(a) Intervalo 1951-1975.



(b) Intervalo 1951-1989.

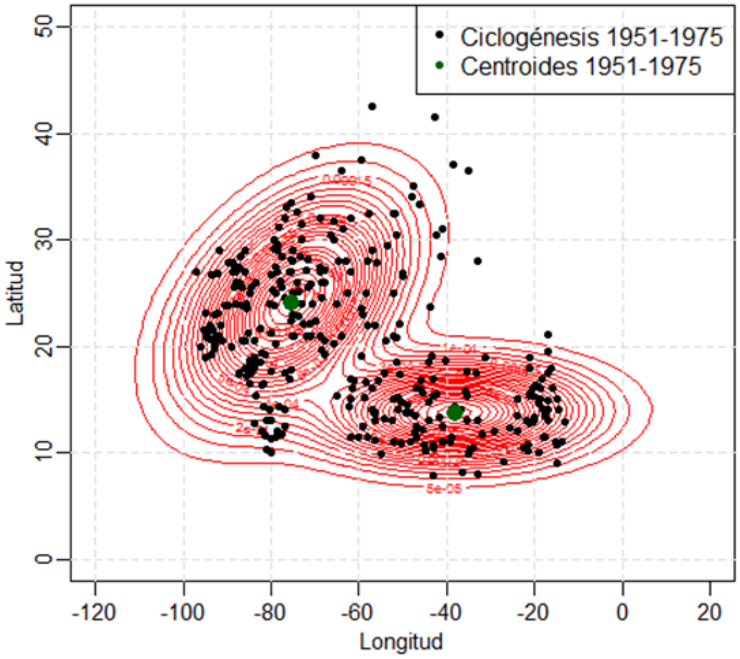


(c) Intervalo 1976-2013.

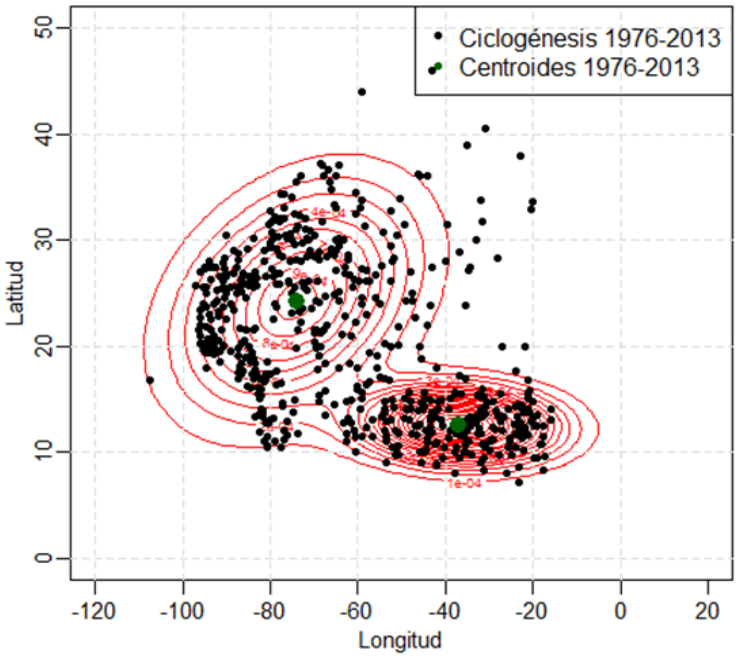


(d) Intervalo 1990-2013.

Figura 5.2: El ancho promedio de silueta, de manera general, indica que hay una buena estructura en los grupos elegidos, con la mayoría de las observaciones que parecen pertenecer a la agrupación en que están de acuerdo con la Tabla 3.2.

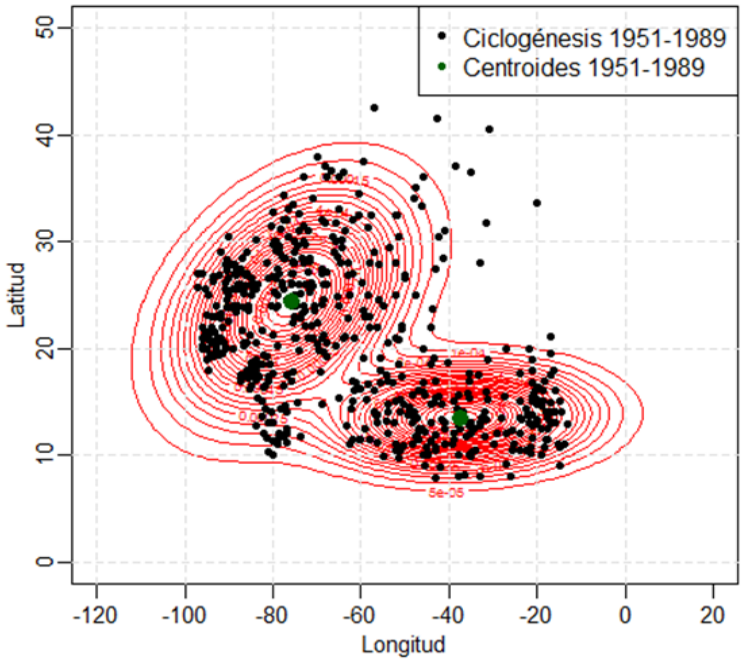


(a) Intervalo 1951-1975.

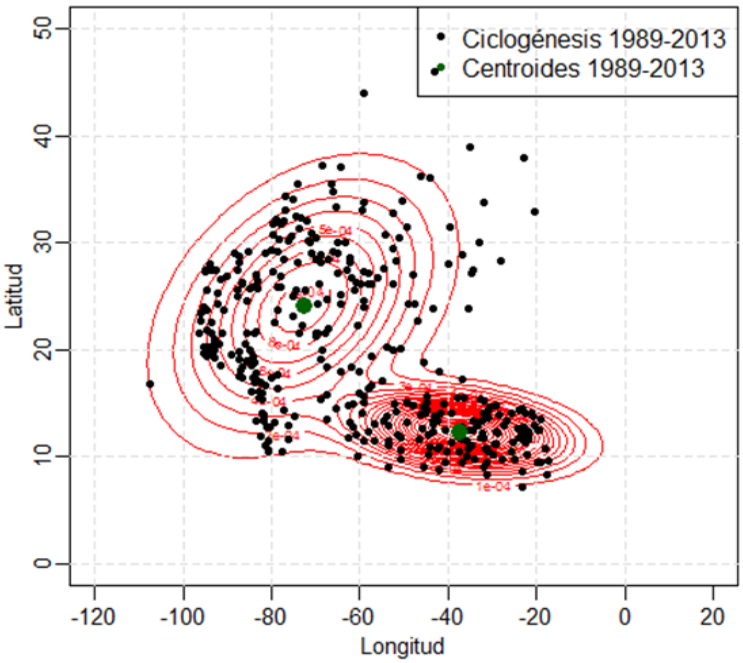


(b) Intervalo 1976-2013.

Figura 5.3: (a) Muestra la gráfica de contornos de la longitud y latitud para el intervalo 1951-1975 y, (b) Longitud y la latitud para el intervalo 1976-2013. Los puntos negros son los puntos de los datos. Los dos grupos en el intervalo 1951-1975 versus 1976-2013 indican que tienen la misma orientación.

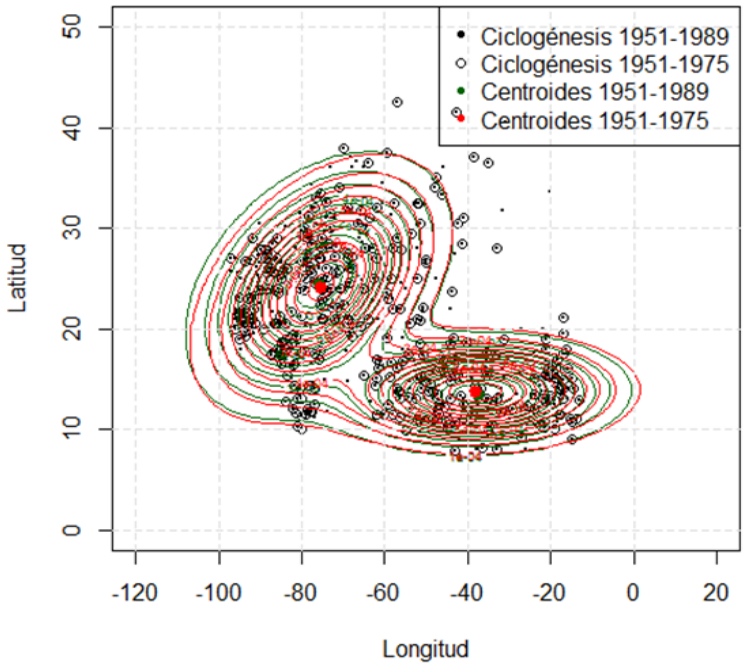


(a) Intervalo 1951-1989.

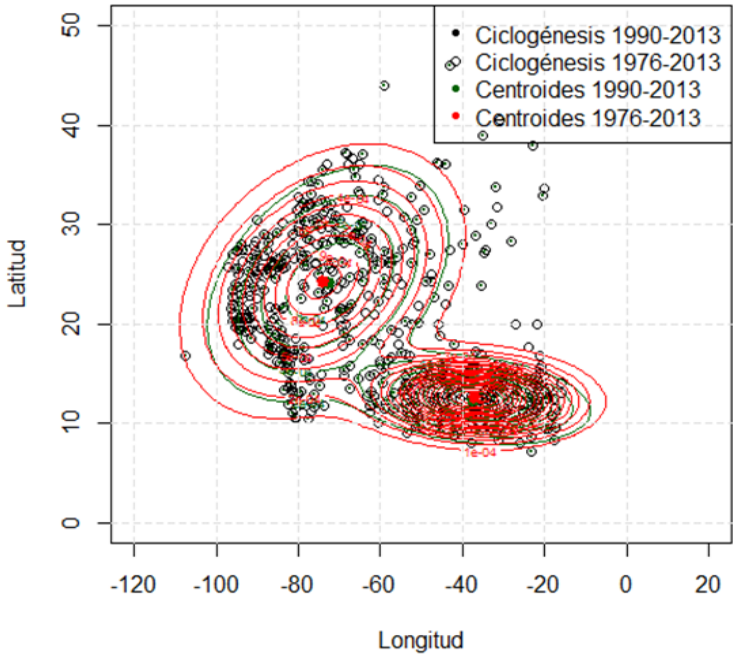


(b) Intervalo 1990-2013.

Figura 5.4: (a) Muestra la gráfica de contornos de la longitud y latitud para el intervalo 1951-1989 y, (b) Longitud y latitud para el intervalo 1990-2013. Los puntos negros son los puntos de los datos. Los dos grupos en el intervalo 1951-1989 versus 1990-2013 indican que tienen la misma orientación.

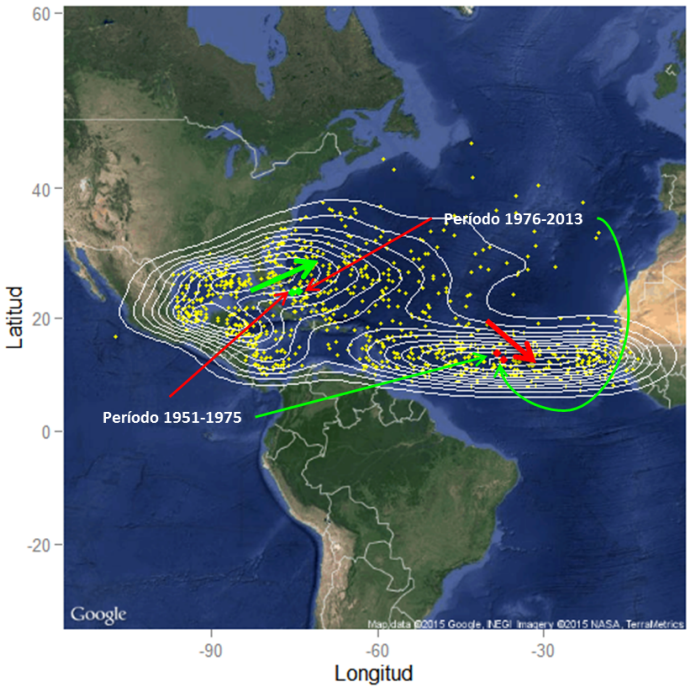


(a) Intervalo 1951-1975 vs 1951-1989.

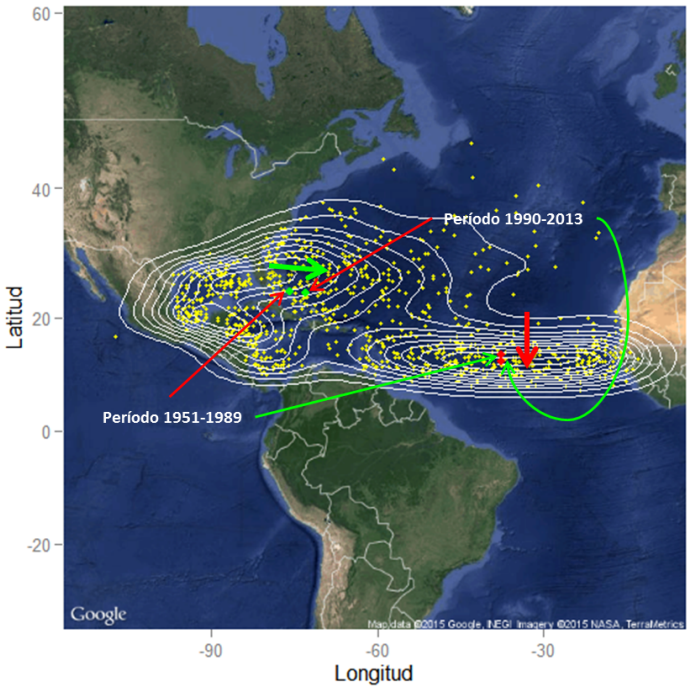


(b) Intervalo 1976-2013 vs 1990-2013.

Figura 5.5: (a) Muestra la gráfica de contornos de la longitud y la latitud para el intervalo 1951-1975 versus 1951-1989, y (b) Longitud y latitud para el intervalo 1976-2013 versus 1990-2013. Los puntos negros son los puntos de los datos. Los dos grupos en los intervalos de estudio indican que tienen la misma orientación.



(a) Intervalo 1951-1975 vs 1976-2013.



(b) Intervalo 1951-1989 vs 1990-2013.

Figura 5.6: Centroides de las regiones ciclogenéticas. Hay dos regiones de génesis de los ciclones tropicales y sus centroides aparentemente se han movido hacia el centro de la cuenca oceánica del Atlántico Norte en los dos intervalos de estudio.

el Atlántico-Tropical se desplaza hacia el sur de la misma (ver Figura 5.6) tal y como lo mencionan Mori *et al.* (2013) en sus proyecciones hechas para finales del siglo XXI respecto al impacto del calentamiento global sobre la génesis de los ciclones tropicales.

El hecho de que los centroides de las regiones ciclogenéticas se desplacen hacia el noroeste y sur de la cuenca oceánica la precipitación inducida por los ciclones tropicales sufrirá cambios tanto en distribución espacial como en eventos y cantidad, tal y como lo mencionan Kim *et al.* (2006) y Lau *et al.* (2008). Esto significa que la región de Centro América, zona donde se ubica México, disminuirá ligeramente su precipitación, lo cual coincide con lo que mencionan Houghton *et al.* (2001) en sus proyecciones de la precipitación para el siglo XXI debidas al cambio climático.

Finalmente, otra consecuencia importante debida también al desplazamiento de los centroides es que los ciclones tropicales durarán más tiempo ya que se van a generar más hacia el centro de la cuenca oceánica, lo cual coincide con lo que mencionan Chan (2007) y Yokoi y Takayabu (2009).

Capítulo 6

Modelo de Mezclas de Procesos Dirichlet para determinar las Regiones de Ciclogénesis del Atlántico Norte e identificar sus cambios

6.1. Introducción

Como una consecuencia del calentamiento global, varios modelos climáticos sugieren que la frecuencia e intensidad de los ciclones tropicales cambiarán a finales del siglo XXI. En general, evaluar cambios en la climatología de los ciclones requiere una comprensión de la física de los ciclones y de las proposiciones del modelo. Luego, las proyecciones del modelo en conjunto son particularmente útiles para predecir estos cambios, y varios estudios se han llevado a cabo utilizando diferentes modelos de circulación general.

Aunque en este momento no hay cambios perceptibles en las características de los ciclones tropicales, atribuidos al calentamiento global, predicciones de simulaciones de modelos de circulación general sugieren un aumento en su intensidad máxima (Walsh, 2004). En relación a las regiones de génesis de huracanes, éstas probablemente no cambien, y respecto a los cambios en el número de ciclones o trayectorias poco consenso ha habido al respecto. Emanuel (2005) discutió cómo el calentamiento global afecta a la intensidad de los ciclones tropicales futuros. Webster *et al.* (2005), y más recientemente Elsner *et al.* (2008) demostraron que la intensidad histórica de las tormentas ha aumentado tanto en el Pacífico Nor-occidental (WNP) como en el

Atlántico Norte. Al mismo tiempo, Webster *et al.* (2005) observaron un aumento considerable en los ciclones tropicales en todas las cuencas oceánicas para las categorías 4 y 5, pero específicamente en las regiones ciclogenéticas del Pacífico Nor-occidental y del Atlántico Norte, éste fue de 25 y 20 %, respectivamente, durante los últimos 30 años. Sin embargo, los resultados obtenidos para la cuenca oceánica del Pacífico Nor-occidental han sido cuestionados por Chan (2006) y por Knaff y Zehr (2007), quienes atribuyen esto en gran medida a la oscilación interdecadal y a posibles errores de medición en el conjunto de datos, respectivamente. Bajo este mismo contexto, Klotzbach (2006) y Kossin *et al.* (2007) encontraron que la tendencia de la intensidad de los ciclones tropicales para la cuenca del Atlántico Norte y del Pacífico Nor-occidental no muestran evidencia de que hayan cambiado durante los intervalos de 1986-2006 y 1985-2005, respectivamente.

En la actuliadad, Mori *et al.* (2013) simularon por medio de un modelo estadítico el impacto del calentamiento global sobre los centroides de la ciclogénesis y de la ciclolysis de las diferentes cuencas oceánica para finales del siglo XXI. Encotraron que los centroides de la génesis de los ciclones se desplazarán hacia el centro de las cuencas oceánicas y que los cambios futuros en las condiciones dinámicas y termodinámicas en los océanos influirán en la frecuencia de la génesis de los ciclones tropicales. Igualmente que los ciclones que se desarrollan en la parte central del océano durarán más tiempo debido a que las temperaturas de la superficie del mar son más cálidas que las de las orillas (Chan, 2007, Yokoi y Takayabu, 2009), y que los cambios en la intensidad de los ciclones tropicales estarán más relacionados con el desplazamiento de los centroides que con el cambio de la temperatura del océano.

Por consiguiente, es de interés estudiar si las regiones de génesis de los ciclones tropicales han experimentado algún cambio en espacio y en tiempo dentro de la cuenca oceánica del Atlántico Norte.

Bajo este contexto, en este trabajo de investgación se aplicó un modelo de mezclas de Procesos Dirichlet (sección 6.2), con el propósito por una parte para determinar el número de regiones ciclogenéticas y por otra para determinar los cambios temporales y espaciales de los centroides de dichas regiones de génesis en la cuenca oceánica del Atlántico Norte, para dos intervalos (1951-1975 versus 1976-2013 y 1951-1989 versus 1990-2013). Se utilizaron los datos de *Longitud* y *Latitud* de los ciclones tropicales de la base de datos de las "mejores trayectorias" o IBTrACS (por siglas en inglés: International Best Track Archive for Climate Stewardship, puede consultarse en: `https://www.ncdc.noaa.gov/ibtracs/index.php?name=ibtracs-data-access`). Se determinaron las funciones de densidad de probabilidades de las regiones de génesis por medio del muestreador gibbs, las cuales se evaluaron y compararon para verificar los cambios espacio-temporales. Los resultados estadísticos y su relación e interpretación con el fenómeno natural en estudio se presentan en la sección 6.3. Para terminar, se presenta una breve discusión y se presentan las conclusiones en la Sección 6.4.

6.2. Materiales y Métodos

6.2.1. Descripción de la Base de Datos de la Ciclogénesis Tropical

Así como en el Capítulo 5, los datos de ciclones tropicales se dividieron en dos intervalos: 1951-1975 vs 1976 vs 2013 y 1951-1989 vs 1990-2013 y el conjunto de datos contienen la ubicación del centro de los ciclones tropicales en *Latitud* y *Longitud*, para mayor detalle ver el Capítulo 4.

6.2.2. Modelos de mezclas de Procesos Dirichlet *DPMM*

Muchos algoritmos de agrupamiento populares requieren que el número de grupos sea conocido a priori o utilizan métodos heurísticos para elegir un número aproximado. Por el contrario, los *DPMM* proporcionan un marco Bayesiano no paramétrico para describir las distribuciones sobre modelos de mezclas con un número infinito de componentes.

6.2.2.1. Especificaciones del modelo

Los datos observados se denotan por la siguiente matriz:

$$\mathbf{Y} = \begin{pmatrix} \text{Latitud}_1 & \text{Longitud}_1 \\ \text{Latitud}_2 & \text{Longitud}_2 \\ \vdots & \vdots \\ \text{Latitud}_n & \text{Longitud}_n \end{pmatrix}$$

con filas que corresponden a las observaciones y con columnas que corresponden a coordenadas de la ubicación de los ciclones tropicales (*Latitud y Longitud*).

Para motivar el uso de un *DPMM*, se considera en primer lugar la verosimiltud para un modelo de mezclas finitas como el de la ecuación 3.2 donde las densidades de sus componentes depende de la forma de los datos observados en el estudio. En este estudio, los datos se consideran en forma de respuesta continua; por lo tanto, cada punto de ubicación se modela de una distribución Normal Bivariada.

$$p\left(\boldsymbol{Y}|\boldsymbol{\phi}\right) = \sum_{j=1}^{K} \pi_j N\left(\boldsymbol{y}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right), \tag{6.1}$$

donde $\boldsymbol{\mu}_j$ es el vector de medias, $\boldsymbol{\Sigma}_j$ es la matriz de varianza y covarianzas, π_j las proporciones de las mezcla (que deben ser positivas y sumar uno) y N es una Gaussiana (normalizada) con media y varianza especificada.

Este modelo es un enfoque alterno a la modelación de mezclas finitas Gaussianas que trata el número de componentes latentes como desconocido por medio de un *DP* a priori, originalmente definido por Ferguson (1973). La forma general del *DPMM*, tal y como ya se mencionó en la ecuación 3.23, es dada por:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{\Theta} &\sim p\left(\boldsymbol{y}_i|\boldsymbol{\theta}_i\right), \\ \boldsymbol{\theta}_i|G &\sim G, \\ G &\sim DP\left(\alpha, H\right), \end{aligned} \tag{6.2}$$

donde $\boldsymbol{\Theta}$, específica los parámetros que describen las y_i y se muestrea de una distribución de mezclas G, que a su vez se modela por un *DP* con distribución base H y parámetro de concentración $\alpha > 0$.

Cabe mencionar que una propiedad sustancial del *DP* dentro de un ambiente de modelos de mezclas es su propiedad de discretización (Ferguson, 1973), mediante la cual múltiples extracciones de un *DP*, denotado aquí por G, tienen una probabilidad no nula de tomar el mismo valor (Sethuraman, 1994). Como resultado, el comportamiento agrupación inducida de la *DP* elude antes de la necesidad de especificar K a priori, en comparación con el caso de dimensión finita. Como resultado de ello, el comportamiento de agrupación inducido del *DP* a priori evita la necesidad de especificar previamente K, en comparación con el caso de dimensión finita.

Ahora, tal y como se mencionó en la sub sección 3.2.3.2, existen tres representaciones constructivas del *DP* que permiten la estimación del *DPMM*; las cuales, a su vez, han dado lugar a una serie de diferentes algoritmos para la inferencia Bayesiana. En este trabajo sólo se abordará la representación *"stick-breaking"* desarrollada por Sethuraman (1994), y la cual consiste en sustituir la distribución de la mezcla G en la ecuación 6.2 con una suma infinita de masas puntuales en los parámetros de los componentes de acuerdo con el Teorema 3.2 dice que:

$$\begin{aligned} \beta_j &\sim Beta(1,\alpha), \qquad j = 1,\ldots, \\ \pi_j &= \beta_j \prod_{l=1}^{K-1} (1-\beta_l)\,, \end{aligned} \tag{6.3}$$

y sea H una medida base en $\boldsymbol{\Theta}$, entonces pueden extraerse un número infinito de muestras $\boldsymbol{\theta}_j^* \overset{iid}{\sim} H$, formándose la medida de probabilidad discreta siguiente:

$$G(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} \pi_j \delta_{\boldsymbol{\theta}_j^*} \qquad \text{y} \qquad \boldsymbol{\theta}_j \sim H. \tag{6.4}$$

Esta representación generalmente se escribe como $\pi \sim GEM(\alpha)$ (Pitman, 2002).

Ahora, relacionando esta recursividad de nuevo al modelo de la ecuación 6.1, estos pesos *"stick-breaking"* corresponden a los pesos de los componentes del modelo de mezclas. Bajo este contexto, se puede observar que la representación *"stick-breaking"* es como un modelo de mezcla finita cuando $K \longrightarrow \infty$. Además, su inferencia puede hacerse mediante el muestreador Gibbs en Bloques (Ishwaran y James, 2001, Ishwaran y Zarepour, 2000). Sin embargo; Ishwaran y Zarepour (2000) observaron que los valores muestreados para pesos sucesivos *"stick-breaking"* normalmente tienden a cero para $K << \infty$ y, a lo más, $K = n$. Por lo tanto, es razonable sugerir que un número infinito de pesos *"stick-breaking"* no tiene porque ser calculado. Por el contrario, dado el conocimiento del número de componentes ocupados que uno espera bajo el modelo, el proceso puede ser truncado en L-componentes, tal que L exceda el número de componentes únicos que uno espera dados los datos observados. Esto da lugar a una aproximación al DP que permite el cálculo más factible.

Al establecer $L >> E[K]$, el $DPMM$, debido en gran parte al parámetro de concentración y su papel en la construcción de los pesos de los componentes, automáticamente vacía componentes con parámetros que no son apoyados por los datos. Además, la naturaleza de esta construcción permite al número total de pesos de componentes no nulos variar entre iteraciones, a su vez, propagando la incertidumbre en el número verdadero de los componentes de la mezcla, K. Esto está en contraste con el modelo de mezclas finitas que, en ausencia de un parámetro de concentración, no expone este comportamiento dado $K = L$ y K se supone fija. La ventaja fundamental de imponer un nivel de truncamiento -con la única condición de que se supere el número esperado de componentes a priori- es la conveniencia computacional, ya que sólo hay que calcular los L pesos *"stick-breaking"*. Esto se traduce en un costo computacional comparable con la estimación de un modelo de mezclas finitas de L-componentes.

Luego entonces, de acuerdo con la ecuación 3.24, el *DPMM* adoptando la representación *"stick-breaking"* con distribución base $H = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L) = \prod_{j=1}^{L} N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ es igual a:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{z}_i = j, \boldsymbol{\Theta} &\sim \prod_{i=1}^{n} \sum_{j=1}^{L} \pi_j N(\boldsymbol{y}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \\ \boldsymbol{z}_i &\sim Multinomial(\boldsymbol{\pi}), \\ \boldsymbol{\pi} &\sim GEM(\alpha), \\ \boldsymbol{\theta}_j &\sim H \quad j = 1, \ldots, L. \end{aligned} \tag{6.5}$$

Note que la introducción de la variable latente $\boldsymbol{z}_i$ en este modelo resalta la discretización de las extracciones del *DP* a priori y por consiguiente el efecto del agrupamiento. Su inclusión también ofrece una nueva interpretación del *DP* a priori como una a priori sobre infinitos numerables, o en este caso, un máximo de L particiones no superpuestas con un peso distinto de cero (Ferguson, 1983).

6.2.2.2. Inferencia del modelo

La inferencia del modelo, para un nivel de truncamiento L, se realiza a través del muestreador Gibbs en Bloques que consiste en la iteración entre muestras de la condicional total de $\boldsymbol{Z}$, π, y $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y especificar al modelo el parámetro de concentración α. Para las estimaciones actuales de $\boldsymbol{\pi}$ y $\boldsymbol{\theta}$, la variable latente para la i-ésima observación se actualiza de la manera siguiente:

$$\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\pi}, \boldsymbol{\theta}, \alpha \sim Multinomial\left(\frac{\pi_1 N(\boldsymbol{y}_i|\boldsymbol{\theta}_1)}{\sum_{j=1}^{L} \pi_j N(\boldsymbol{y}_i|\boldsymbol{\theta}_j)}, \ldots, \frac{\pi_L N(\boldsymbol{y}_i|\theta_L)}{\sum_{j=1}^{L} \pi_j N(\boldsymbol{y}_i|\theta_j)}\right). \tag{6.6}$$

En consecuencia, para una iteración dada, cada $\boldsymbol{z}_i$ toma un valor sobre $1, \ldots, L$. Ahora, relacionando esto con el número real inferido de componentes, el número de valores únicos de $\{\boldsymbol{z}\}$ corresponde a la estimación del número de grupos ocupados (K) para la misma iteración. Dada esta asignación actualizada de las observaciones, las proporciones de los componentes se actualizan utilizando valores muestreados para $\beta_1, \ldots, \beta_L$, a través de la condicional completa:

$$\beta_j|\alpha, \boldsymbol{Z} \sim Beta\left(1+\sum_{i=1}^{n}\delta_{(\boldsymbol{z}_i=j)}, \alpha+\sum_{i=1}^{n}\delta_{(\boldsymbol{z}_i>j)}\right) \quad \text{con} \quad j=1,\ldots,L,$$

donde: $n_j = \sum_{i=1}^{n}\delta_{(\boldsymbol{z}_i=j)}$ y $n-\sum_{l=1}^{L-1}n_l = \sum_{i=1}^{n}\delta_{(\boldsymbol{z}_i>j)}$ o $\sum_{l=j+1}^{L}n_l = \sum_{i=1}^{n}\delta_{(\boldsymbol{z}_i>j)}$, entonces:

$$\beta_j|\alpha, \boldsymbol{Z} \sim Beta\left(1+n_j, \alpha+n-\sum_{l=1}^{L-1}n_l\right), \tag{6.7}$$

que luego se utilizan para actualizar cada π_j, utilizando la ecuación 6.3. Aquí, n_j representa el número de sujetos asignados al grupo j.

Ahora, para actualizar la media $\boldsymbol{\mu}_j$ se muestrea independientemente de la a posteriori normal:

$$\boldsymbol{\mu}_j|\text{resto} \sim N\left(\frac{\boldsymbol{B}_0^{-1}\cdot\boldsymbol{b}_0+n_j\cdot\boldsymbol{\Sigma}_j^{-1}\cdot\bar{y}_j}{\boldsymbol{B}_0^{-1}+n_j\cdot\boldsymbol{\Sigma}_j^{-1}}, \frac{1}{\boldsymbol{B}_0^{-1}+n_j\cdot\boldsymbol{\Sigma}_j^{-1}}\right) \quad j=1,\ldots,L,$$

donde: n_j es el número de observaciones que pertenecen al grupo j, y $\bar{y}_j = \frac{1}{n_j}\sum_{i:z_i=j}\boldsymbol{y}_i$ es la media de estas observaciones (Fruhwirth-Schnatter, 2006, Malsiner-Walli *et al.*, 2014, Muller *et al.*, 1996, Yau y Holmes, 2011). La a priori usual para μ_j es la a priori independiente $N(\boldsymbol{b}_0, \boldsymbol{B}_0)$, $j=1,\ldots,L$, donde $N(\cdot,\cdot)$ denota la distribución normal multivariada (Malsiner-Walli *et al.*, 2014). Es normal suponer que todos los componentes $\boldsymbol{\mu}_j$ son a priori independientes, dados los hiperparámetros dependientes de los datos $\boldsymbol{b}_0$ y $\boldsymbol{B}_0$. Posteriormente, se llamará a esta a priori "a priori estándar" y se elige la mediana para definir $\boldsymbol{b}_0 = \text{mediana}(\boldsymbol{y})$ y el rango R_r de los datos en cada dimensión r para definir $\boldsymbol{B}_0 = \boldsymbol{R}_0$ donde $\boldsymbol{R}_0 = Diag(R_1^2,\ldots,R_r^2)$.

Finalmente, para actualizar la varianza $\boldsymbol{\Sigma}$, se muestrea la a posteriori *Wishart (W)* conjugada condicional:

$$\boldsymbol{\Sigma}_j|\text{resto} \sim W\left(c_0+n_j, \frac{1}{c_0\cdot\boldsymbol{C}_0+\sum_{i:\boldsymbol{z}_i=j}(\boldsymbol{y}_i-\mu_j)\left(\boldsymbol{y}_i-\boldsymbol{\mu}_j\right)^T}\right) \quad j=1,\ldots,L,$$

donde la apriori jerárquica conjugada $\boldsymbol{\Sigma}_j^{-1} \sim W(c_0, \boldsymbol{C}_0)$, con $\boldsymbol{C}_0 \sim W(\boldsymbol{g}_0, \boldsymbol{G}_0)$ (Fruhwirth-Schnatter, 2006, Malsiner-Walli *et al.*, 2014, Muller *et al.*, 1996, Yau

y Holmes, 2011). Con el fin de evitar soluciones degeneradas, Fruhwirth-Schnatter (2006) y Stephens (1997) proponen que se regularice la matriz de varianzas y covarianzas especificando los hiperparámetros a priori $c_0 = 2.5 + \frac{r-1}{2}$, $g_0 = 0.5 + \frac{r-1}{2}$ y $\boldsymbol{G}_0 = \frac{100 \cdot g_0}{c_0} Diag\left(\frac{1}{R_1^2} \ldots \frac{1}{R_r^2}\right)$. De acuerdo con Fruhwirth-Schnatter (2006, página 193), para mezclas normales bivariadas se puede asumir como: $c_0 = 3$, $g_0 = 0.3$, $\boldsymbol{b}_0$, $\mathbf{B}_0$ y $\boldsymbol{G}_0$ como sigue:

$$\boldsymbol{b}_0 = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \qquad \boldsymbol{B}_0 = \begin{pmatrix} R_1^2 & 0 \\ 0 & R_2^2 \end{pmatrix}, \qquad \boldsymbol{G}_0 = \begin{pmatrix} \frac{100 \cdot g_0}{c_0 \cdot R_1^2} & 0 \\ 0 & \frac{100 \cdot g_0}{c_0 \cdot R_2^2} \end{pmatrix},$$

donde m_l y R_l son el punto medio y la longitud del intervalo de observación del l-ésimo componente de $\boldsymbol{y}_i$.

6.2.2.3. Medida de la distancia entre *DPMM*

En este trabajo de investigación, se utilizó la divergencia de Kullback-Leibler para medir la similitud entre los grupos de los ciclones tropicales para los dos intervalos.

De acuerdo con Hershey y Olsen (2007), la divergencia de Kullback-Leibler (KL), conocida también como la entropía relativa, es la medida de la distancia comúnmente utilizada entre funciones de densidad de probabilidad (fdp). Para dos fdp Gaussianas, $f(y)$ y $g(y)$, la divergencia de KL tiene una expresión en forma cerrada, pero no para dos modelos de mezclas Gaussianas (MMG), y ésta se define como sigue:

$$KL(f \parallel g) \overset{def}{=} \int f(y) \log \frac{f(y)}{g(y)} dy. \tag{6.8}$$

En la estadística, la divergencia de KL se utiliza como una medida de similitud entre dos distribuciones de densidad y satisface las siguientes tres propiedades:

1. Auto similitud: $KL(f \parallel f) = 0$,
2. Auto identificación: $KL\left(\hat{f} \parallel \hat{g}\right) = 0$ sólo si $f = g$, y
3. Positividad: $KL(f \parallel g) \geq 0$ para toda f, g.

Para dos fdp Gaussianas $\hat{f}$ y $\hat{g}$ la divergencia de KL tiene una expresión en forma cerrada:

$$KL\left(\hat{f} \parallel \hat{g}\right) = \frac{1}{2}\left[\log\frac{\left|\boldsymbol{\Sigma}_{\hat{g}}\right|}{\left|\boldsymbol{\Sigma}_{\hat{f}}\right|}\right] + \operatorname{tr}\left[\boldsymbol{\Sigma}_{\hat{g}}^{-1}\boldsymbol{\Sigma}_{\hat{f}}\right] + \left(\boldsymbol{\mu}_{\hat{g}} - \boldsymbol{\mu}_{\hat{f}}\right)^{T}\boldsymbol{\Sigma}_{\hat{g}}^{-1}\left(\boldsymbol{\mu}_{\hat{g}} - \boldsymbol{\mu}_{\hat{f}}\right). \quad (6.9)$$

Ahora, sea f y g *MMG*, las densidades marginales de $\boldsymbol{y} \in R^d$ son:

$$\begin{aligned} f(\boldsymbol{y}) &= \sum_a \pi_a N(\boldsymbol{y}; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \\ g(\boldsymbol{y}) &= \sum_b \omega_b N(\boldsymbol{y}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b), \end{aligned} \quad (6.10)$$

donde π_a es la probabilidad a priori de cada estado, y $N(\boldsymbol{y}; \boldsymbol{\mu}_a; \boldsymbol{\Sigma}_a)$ es una Gaussiana en $\boldsymbol{y}$ con media $\boldsymbol{\mu}_a$ y varianza $\boldsymbol{\Sigma}_a$. Se utiliza con frecuencia la notación abreviada $f_a(y) = N(\boldsymbol{y}; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$ y $g_b(y) = N(\boldsymbol{y}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$. La estimación de $D(f \parallel g)$ hará uso de la divergencia KL entre los componentes individuales, que se escribe por lo tanto como $D(f_a \parallel g_b)$.

Una aproximación utilizada para $D(f \parallel g)$ es la *"Aproximación Gaussiana"*, la cual consiste en reemplazar f y g con densidades Gaussianas, $\hat{f}$ y $\hat{g}$. En una incorporación, uno utiliza Gaussianas cuya media y covarianza coincide con las de f y g. La media y la covarianza de f están dadas por:

$$\begin{aligned} \boldsymbol{\mu}_{\hat{f}} &= \boldsymbol{\Sigma}_a \pi_a \boldsymbol{\mu}_a, \\ \boldsymbol{\Sigma}_{\hat{f}} &= \boldsymbol{\Sigma}_a \pi_a \left(\boldsymbol{\Sigma}_a + \left(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{\hat{f}}\right)\left(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{\hat{f}}\right)^{T}\right). \end{aligned} \quad (6.11)$$

La aproximación $D_{Gaussiana}(f \parallel g)$ es dada por la expresión en forma cerrada, $D_{Gaussiana}(f \parallel g) = D(\hat{f} \parallel \hat{g})$ usando la ecuación 6.9.

6.2.2.4. Estadístico de prueba para la comparación de las fdp de los *DPMM*

Análogamente como en la sección 5.2.2.5 en el Capítulo 5, se aplicó la técnica de re-muestreo paramétrico para estimar el percentil 95 de la distribución empírica del estadístico de prueba que en este caso es la divergencia de Kullback Leibler entre los componentes de los modelos de mezclas Gaussianas para verificar si existen diferencias

estadísticamente significativas entre ellas, ver sección 6.2.2.3.

La regla de decisión es rechazar H_0 si $D(\hat{f} \parallel \hat{g})$ es mayor que el percentil 95 de la distribución muestral de la divergencia de Kullback-Leibler $\hat{D}(\hat{f} \parallel \hat{g})$.

6.3. Resultados

6.3.1. Estimación de la función de densidad de probabilidades

Tal y como ya se mencionó que los modelos de mezclas de procesos *Dirichlet* son algoritmos de clasificación no supervisada. En consecuencia, para el modelo definido en la ecuación 6.5, el nivel de truncamiento para la construcción del *stick-breaking* se estableció en $L = 25$, de acuerdo Gelman *et al.* (2014), quienes mencionan que en la práctica L debe oscilar entre 25 y 50. La incertidumbre en α se determinó con base en el valor esperado del número de grupos a priori $E(K) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$, propuesto por Antoniak (1974), y con base en la distribución espacial de los puntos de ubicación de ciclogénesis por intervalo, ver la Figura 6.4 y 6.5. Luego entonces, se probaron dos a priori para los diferentes intervalos de estudio: $Gama(4, 2)$ y $Gama(2, 4)$. El valor esperado del número de grupos para el intervalo 1951-1975 vs 1976-2013 fue de 2.66 y 1.19 y 4.18 y 1.2, respectivamente. Respecto al número promedio de grupos a priori para el intervalo 1951-1989 vs 1990-2013, éste fue de 3.75 y 1.99 y 3.40 y 1.52, respectivamente.

Con base en lo anterior, se tomó una distribución $Gama(4, 2)$ a priori en α para todos los intervalos.

Los resultados se basan en 4000 iteraciones de muestreador Gibbs por Bloques descrito en la sección anterior. Las 4,000 iteraciones fueron determinadas mediante el paquete CODA (Convergence Diagnosis and Output Analysis) para validar los resultados generados por l os procedimientos MCMC (cadenas de Markov Monte Carlo).

Los valores de los parámetros iniciales para cada uno de los componentes de las mezclas se determinaron directamente en cada intervalo. A continuación se muestran como ejemplo las gráficas de las densidades obtenidas de la inferencia bayesiana basada en la simulación de las muestras para resumir la distribución a posteriori de los parámetros $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, y $\boldsymbol{\Sigma}$ para el grupo uno del intervalo 1951-1975, ver Figura 6.1, 6.2 y 6.3.

En la Figura 6.4 y 6.5 se muestra la función de densidad estimada para cada uno de grupos de los intervalos. Se puede observar que ajustan a la distribución espacial de

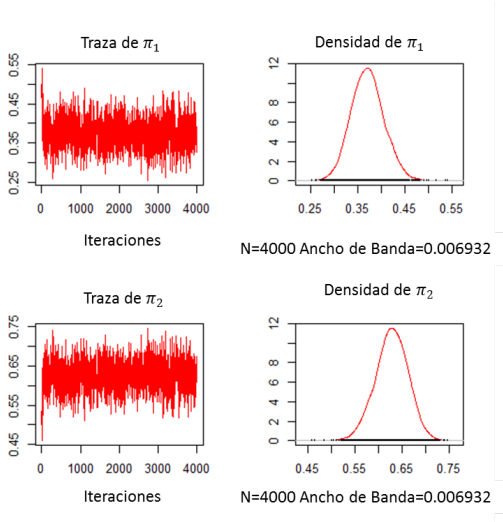


Figura 6.1: Convergencia y densidad a posteriori de la probabilidad de pertenencia (π_1 *y* π_2) para los dos grupos obtenidos en el intervalo 1951-1975. La gráfica muestra una traza perfecta, ya que el centro de la cadena parece estar en torno al valor 0.35 y 0.65, respectivamente, con fluctuaciones muy pequeñas, lo cual indica que el algoritmo converge.

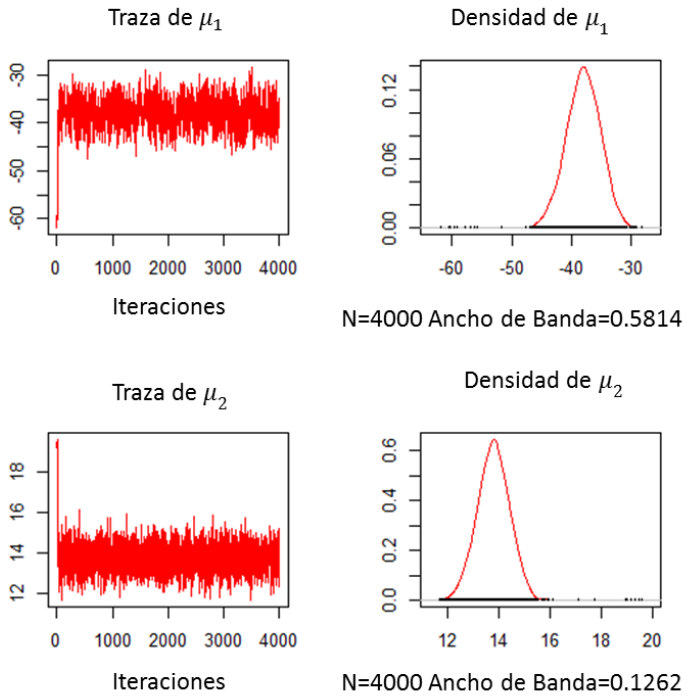
los puntos de ubicación de ocurrencia de los ciclones tropicales.

Adicionalmente a las gráficas de las funciones de densidad para cada uno de los grupos por intervalo, también se graficaron en forma simultánea, con el propósito de comparar la forma y la orientación de los grupos, ver Figura 6.6.

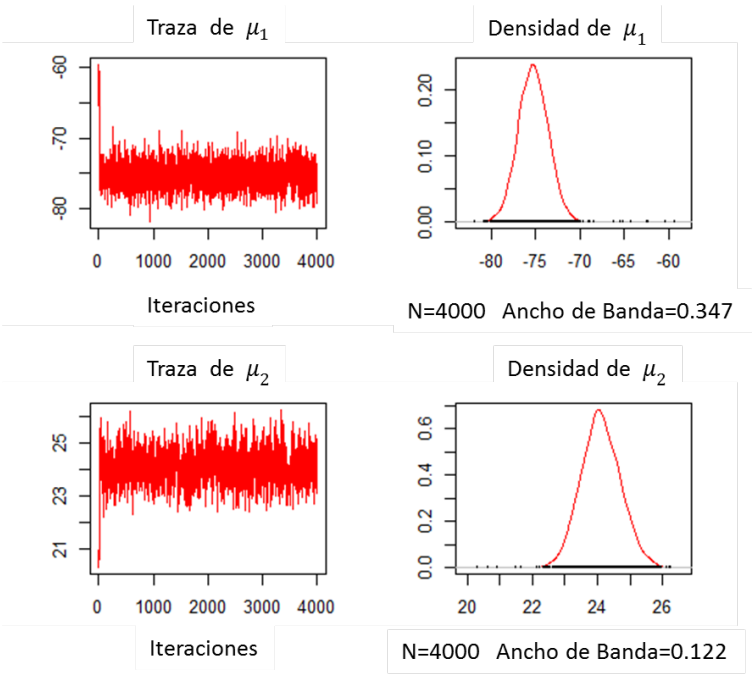
En dicha gráfica se puede observar que las funciones de densidad para el intervalo 1951-1975 vs 1951-1989 tiene diferente forma y aparentemente la misma orientación, esto se debe a que en el primer intervalo se tienen dos grupos y en el segundo se tienen tres grupos ($L = 3$). En contraste, para el intervalo 1976-2013 vs 1990-2013, la función de densidad aparentemente tienen la misma forma y orientación.

6.3.2. Comparación de los vectores de medias de las funciones de densidad

Comparando los centroides estimados de los intervalos 1951-1975 vs 1976-2013, se puede observar en el inciso a) de la Figura 6.7 que el centroide del intervalo 1976-2013 con respecto al centroide del intervalo 1951-1975 ha sufrido cambios en su localización, específicamente moviéndose éste hacia el sur de la cuenca oceánica. Sin embargo; en el intervalo 1976-2013, el centroide que aparecía en color verde en la sección anterior mediante este método se dividió en dos centroides, uno en color rojo y otro en color azul. El centroide de color rojo no se puede comparar con el centroide del intervalo



(a) Grupo 1.



(b) Grupo 2.

Figura 6.2: Convergencia y densidad a posteriori del vector de medias (μ_1 y μ_2) del grupo uno y dos obtenido para el intervalo 1951-1975. La gráfica muestra que la cadena comienza en un valor inicial lejano, pero posteriormente converge a la distribución objetivo.

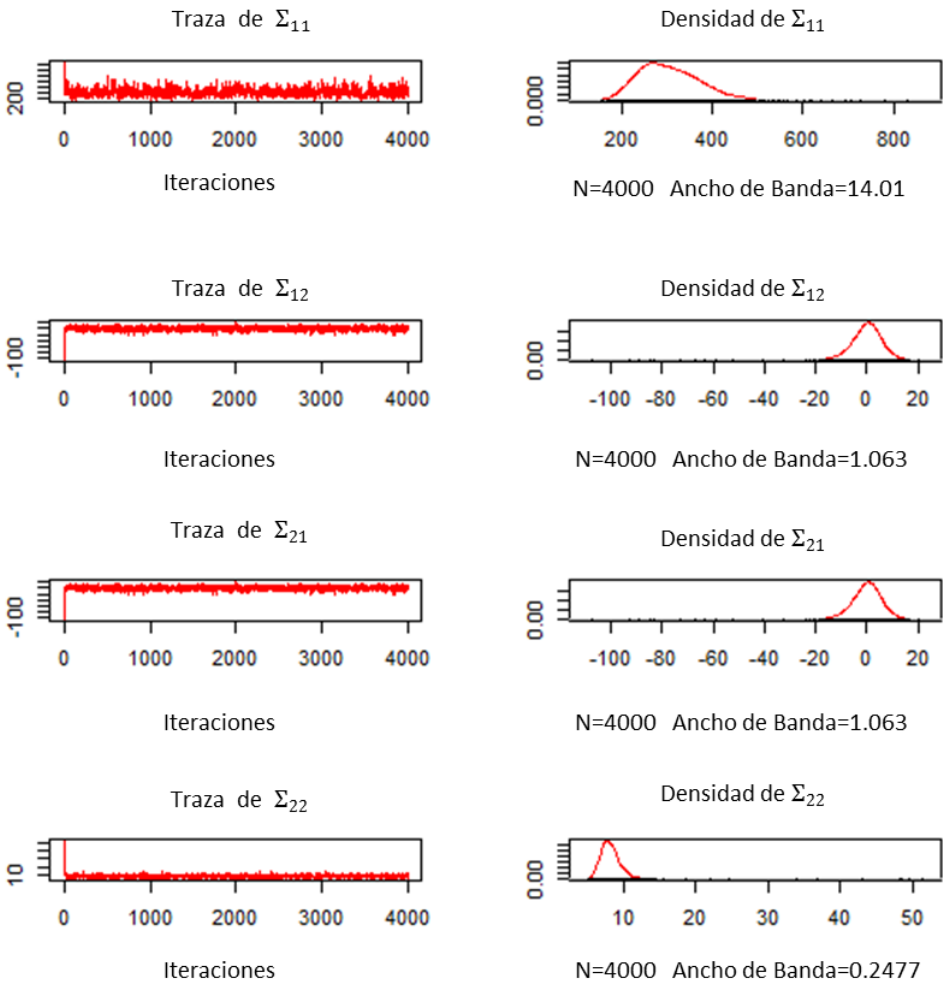
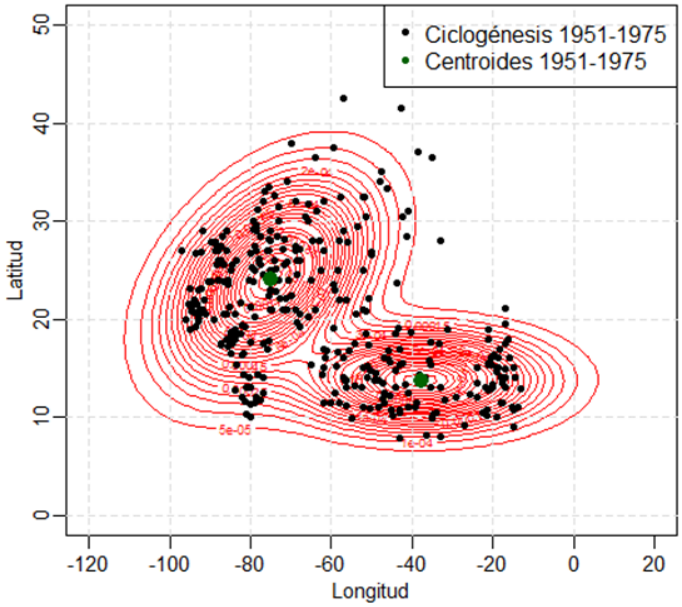


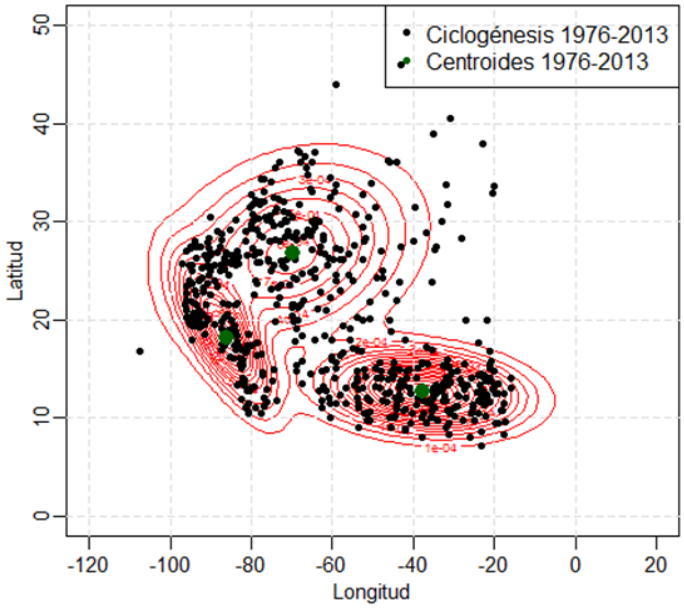
Figura 6.3: Convergencia y densidad a posteriori de la matriz de varianzas y covarianzas ($\Sigma_{11}, \Sigma_{12}, \Sigma_{21}\, y\, \Sigma_{22}$) del grupo uno obtenido para el intervalo 1951-1975. La gráfica muestra que la cadena comienza en un valor inicial lejano, pero posteriormente converge a la distribución objetivo.

1951-1975, ya que en esta región solo se formó un centroide (color azul). Respecto al centroide de color azul este si se comparó, dando como resultado que se está moviendo hacia el nor-este, ver inciso a) en la Figura 6.7.

Respecto al comportamiento de los centroides del intervalo 1951-1989 vs 1990-2013, se puede observar que tienen el mismo comportamiento que los centroides de los intervalo 1951-1975 vs 1976-2013 estimados mediante el *MMG*, con la diferencia que aquí son tres centroides, ver inciso b) en la Figura 6.7.

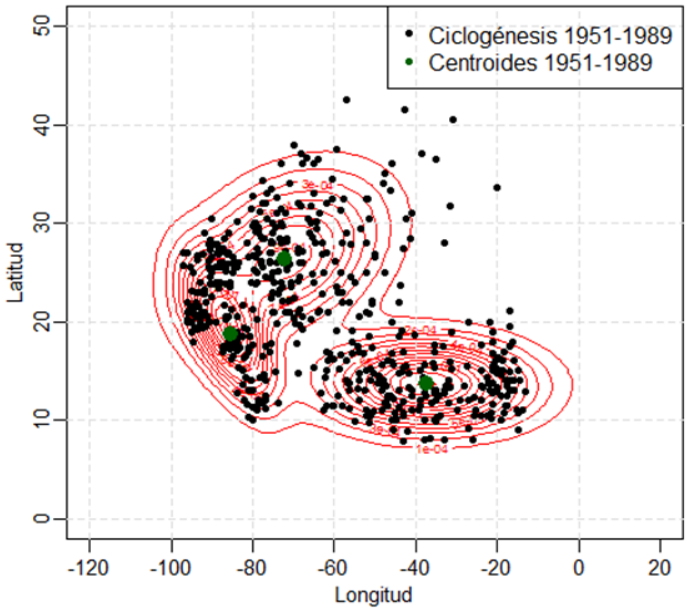


(a) Intervalo 1951-1975.

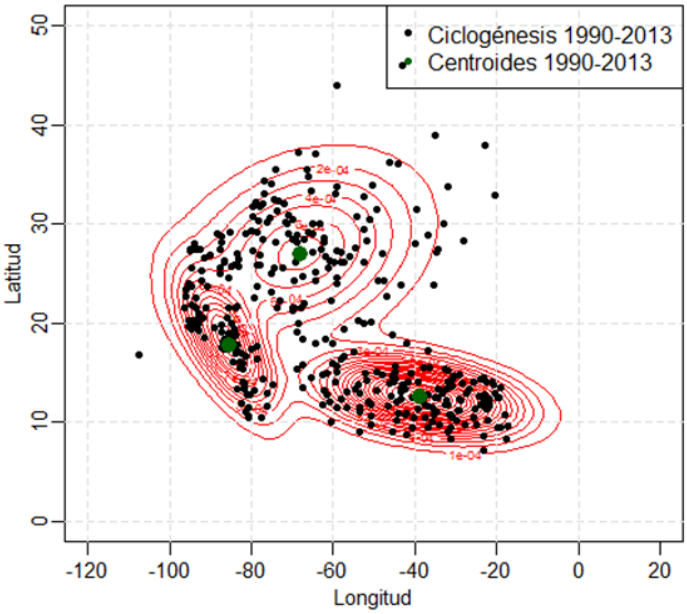


(b) Intervalo 1976-2013.

Figura 6.4: Determinación del número de grupos y de la función de densidad mediante el proceso Dirichlet. (a) Muestra que el número de grupos es dos para el intervalo 1951-1976, en tanto que (b) Muestra que el número de grupos es $L = 3$ para el intervalo 1976-2013. Fuente: Elaboración Propia.

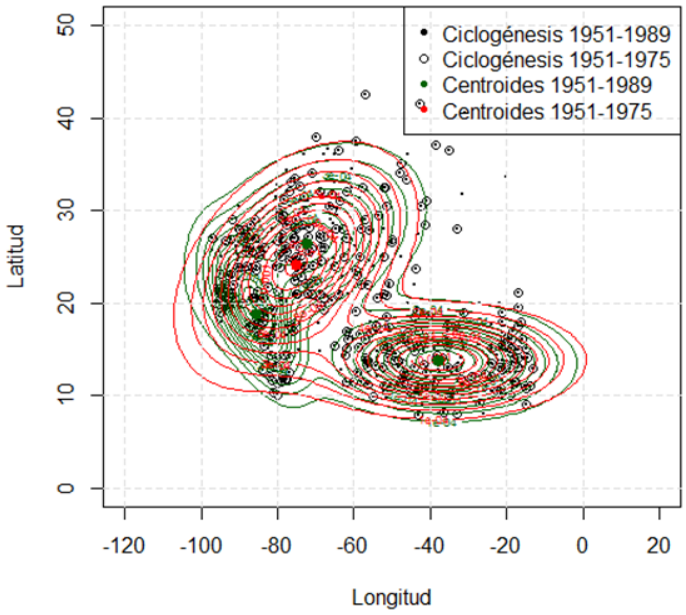


(a) Intervalo 1951-1989.

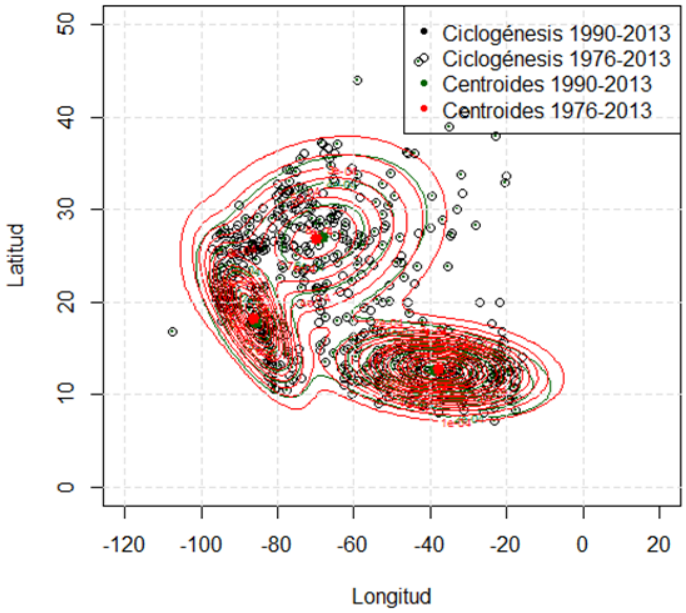


(b) Intervalo 1990-2013.

Figura 6.5: Determinación del número de grupos y de la función de densidad mediante el proceso Dirichlet. (a) y (b) Muestran que el número de grupos es tres para el intervalo 1951-1989 vs 1990-2013, respectivamente. Fuente: Elaboración Propia.

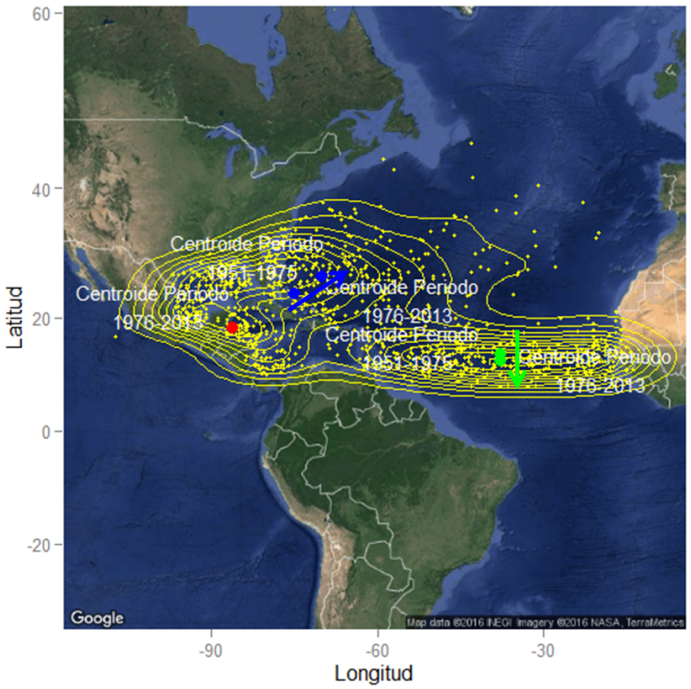


(a) Intervalo 1951-1975 vs 1951-1989.

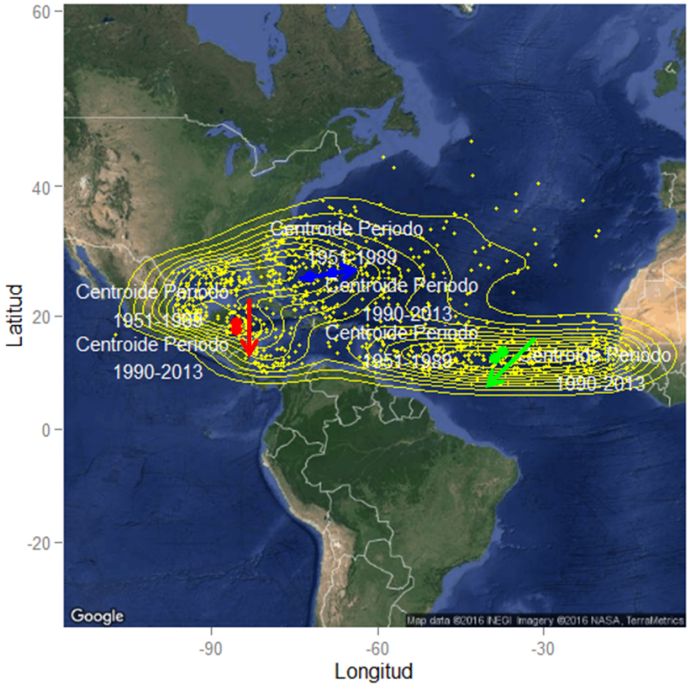


(b) Intervalo 1976-2013 vs 1990-2013.

Figura 6.6: (a) y (b) Muestran la comparación del número de grupos por intervalo (1951-1975 vs 1951-1989 y 1976-2013 vs 1990-2013). En el intervalo 1951-1975 vs 1951-1989, existen diferencias en la ciclogénesis, es decir en el primer período se tienen $L = 2$ grupos en tanto que en el segundo se tienen $L = 3$. Para el intervalo 1976-2013 vs 1990 vs 2013 a diferencia de la comparación anterior aquí coinciden el número de grupos ($L = 3$ grupos) y tienen la misma orientación. Fuente: Elaboración Propia.



(a) 1951-1975 vs 1976-2013.



(b) 1951-1989 vs 1990-2013.

Figura 6.7: Centroides de las regiones ciclogenéticas. Hay dos regiones de génesis de los ciclones tropicales y sus centroides aparentemente se han movido hacia el centro de la cuenca oceánica del Atlántico Norte en los dos intervalos de estudio. Fuente: Elaboración Propia.

6.3.3. Prueba de re-muestreo paramétrica para comparar las fdp

La obtención del percentil 95 de la distribución empírica del estadístico $\hat{D}(\hat{f} \parallel \hat{g})$ aplicó únicamente para el intervalo 1951-1989 vs 1990-2013, ya que en ambos intervalos se tiene el mismo número de componentes ($L = 3$). Sin embargo, para el intervalo 1951-1975 vs 1976-2013 no fue posible aplicar dicha metodología, ya que el número de componentes por intervalo es diferente.

Luego entonces, el percentil 95 del estadístico de prueba (divergencia de Kullback-Leibler) en promedio de los grupos para el intervalo de 1951-1989 vs 1990-2013 fue de 0.057 con un valor critico al 5 %, por lo que se rechaza la hipótesis nula de similitud de las fdp de los grupos de los diferentes modelos de mezclas de procesos Dirichlet, indicando que estadísticamente existen diferencias significativas entre ellas.

6.4. Discusión y conclusiones

Los resultados estadísticos obtenidos para el número de grupos mediante los *DPMM* a diferencia de los obtenidos con el *MMG* es mayor para los dos intervalos (1951-1975 vs 1976-2013 y 1951-1989-2013), esto significa que la región de ciclogénesis del Atlántico Norte se está subdividiendo más que expandiéndose tal y como lo mencionan Henderson-Sellers *et al.* (1998).

Específicamente, en el intervalo 1951-1975 vs 1976-2013 hay dos ($K = 2$) y tres ($K = 3$) sitios de génesis, respectivamente. Mientras que en el intervalo 1951-1989 vs 1990-2013 hay tres ($K = 3$) sitios en ambos intervalos.

Cabe mencionar que los centroides de los sitios de génesis encontrados de manera general han sufrido cambios en su localización para ambos intervalos de estudio, lo cual coincide con las proyecciones hechas por Mori *et al.* (2013) sobre la génesis de los ciclones tropicales para finales del siglo XXI.

Particularmente, el centroide que aparece en color azul en el intervalo 1951-1975, éste se dividió en dos centroides (azul y rojo) en el intervalo 1976-2013, ver inciso a) de la Figura 6.7. Comparando éstos se puede observar que han sufrido cambios en su ubicación, desplazándose específicamente hacia el nor-este (es decir, del Gólfo hacia la Costa-Este) y al sur-oeste (es decir, moviéndose dentro del mismo Atlántico Tropical), respectivamente.

Por lo tanto, es posible suponer que la precipitación inducida por los ciclones tropicales sufrirá cambios en su distribución espacial, en eventos y en cantidad como lo aseveran

Kim *et al.* (2006) y Lau *et al.* (2008). Esto significa que la zona donde se ubica México disminuirá ligeramente su precipitación tal y como lo mencionan Houghton *et al.* (2001) en sus proyecciones de la precipitación para el siglo XXI debidas al cambio climático.

Respecto al comportamiento de la génesis en el intervalo 1951-1989 vs 1990-2013, éste es igual al estimado por el método *MMG*, sólo que a diferencia de que aquí hay tres regiones de génesis ($K = 3$), ver inciso b) de la Figura 6.7.

Capítulo 7

Mezclas Gaussianas de Modelos de Regresión para determinar la influencia del la Temperatura de la Superficie del Mar en la Ciclogénesis en el Atlántico Norte

7.1. Introducción

En las últimas décadas, se ha reportado una relación estrecha entre las tendencias climáticas en la actividad de los ciclones tropicales y la temperatura de la superficie del mar (TSM), dando lugar a un intenso debate sobre si el calentamiento global está aumentando la actividad de los ciclones tropicales (Chan, 2006, Emanuel, 2005, Holland y Webster, 2007, Webster *et al.*, 2005). Emanuel (2005) mostró una relación estrecha entre el Índice de Destrucción Potencial (PDI por sus siglas en inglés, Power Dissipation Index) acumulado anual de los ciclones tropicales en el Atlántico Norte y la TSM subyacente durante el intervalo de 1949 a 2004, definiendo el PDI anual como un efecto colectivo de la intensidad, la duración y la frecuencia anual de los ciclones tropicales. Puntualizando un aumento en el PDI, en el Pacífico Nor-occidental y el Atlántico Norte, lo que indica una mayor duración e intensidad de los ciclones tropicales (Emanuel, 2007, Trenberth *et al.*, 2007). La actividad ciclónica tropical está fuertemente correlacionada con la temperatura de la superficial del mar (TSM) en la región ciclogenética del Atlántico (RCA) (Elsner *et al.*, 2006, Emanuel, 2005, 2007, Holland y Webster, 2007, Saunders y Lea, 2008) y más débilmente en la región ciclogenética del Pacífico (RCP) (Chan y Liu, 2004, Emanuel, 2005, 2007).

La TSM ha aumentado durante las últimas décadas tanto en la zona tropical del Atlántico norte como en zona tropical del Pacífico Nor-occidental (Emanuel, 2005, Trenberth *et al.*, 2007). El calentamiento en el Atlántico Norte tropical se ha asociado con la Oscilación Multidecadal del Atlántico (OMA) (Goldenberg *et al.*, 2001), un modo de variabilidad climática asociada a variaciones en la intensidad de la circulación termohalina(Trenberth *et al.*, 2007), que es el movimiento interno de agua en el océano profundo ocasionado por la diferencia de densidad de las masas de agua que se ordenan las menos densas sobre las más densas (EcuRed, 2010). Sin embargo, la contribución de la OMA a la tendencia al calentamiento del Atlántico Norte es objeto de debate, y Trenberth y Shea (2006) sugieren una definición alterna de la OMA en el que la temperatura media mundial se resta de la TSM del Atlántico. El calentamiento de la RCA se interpreta así como el calentamiento global (Mann y Emanuel, 2006, Trenberth y Shea, 2006). Tales análisis llevaron a Hegerl *et al.* (2007) a concluir que el aumento de las concentraciones de gases de efecto invernadero probablemente han contribuido en el incremento de la TSM en esta región.

Sin embargo, el conocimiento actual sobre el efecto de los factores ambientales, particularmente, la temperatura de la superficie del mar, en la formación de estos hidrometeoros, es en gran medida cualitativo, por lo que es importante también entender el comportamiento de forma cuantitativa. Gaffney *et al.* (2007) utilizaron un modelo de regresión cuadrática de mezclas finitas Gaussianas para describir la difusión en tiempo de la latitud y la longitud de las trayectorias de los ciclones extra-tropicales de invierno con respecto a la TSM en el Atlántico Norte. Análogamente, Camargo *et al.* (2007) implementaron un modelo de regresión cuadrática de mezclas Guassianas para describir las trayectorias de los ciclones tropicales en el Pacífico Nor-occidental.

En este estudio, se explora la dispersión de la TSM de los puntos de ubicación de ocurrencia los ciclones tropicales en la cuenca oceánica del Atlántico mediante la aplicación de una mezcla Gaussiana de modelos de regresión lineal, con el propósito de determinar el número regiones ciclogenéticas y si sus centroides han experimentado algún desplazamiento con respecto a la variación de la TSM.

En la sección 7.2 se presenta la metodología de agrupación que se aplicó a la base de datos de la TSM, la cual se dividió en dos intervalos: el primero de 1951-1975 a 1976-2013 y el segundo de 1951-1989 a 1990-2013. Además, mediante el algoritmo Esperanza-Maximización (EM) se determinó el número de componentes (regiones de génesis) y los parámetros de sus funciones de densidad de probabilidad, específicmanete, las medias como función lineal de la TSM. Posteriormente, se simularon las media con diferentes valores TSM con el propósito de verificar los cambios temporales y espaciales de la ciclogénesis. En la sección 7.3, se muestran los resultados y su relación e interpretación con el fenómeno natural en estudio. Finalmente, en la sección 7.4 se presenta una breve discusión y se presentan las conclusiones.

7.2. Materiales y métodos

7.2.1. Descripción de la Base de Datos de la Ciclogénesis Tropical y su correspondiente Temperatura de la Superficie del Mar (TSM)

Análogamente como en el Capítulo 5 y 6, los datos de ciclones tropicales fueron divididos en dos intervalos: 1951-1975 vs 1976 vs 2013 y 1951-1989 vs 1990-2013, y el conjunto de datos contiene la medición de *Longitud* y *Latitud* de cada punto de génesis de los ciclones tropicales así como su correspondiente temperatura, ver detalle en el Capítulo 4.

7.2.2. Modelos de regresión de mezclas finitas Gaussianas

En muchas áreas de aplicación, las covariables de riesgo relevantes se introducen a los modelos de mezclas finitas, y este modelo se llama modelo de regresión de mezclas finitas. Por lo general, estas covariables están asociados con los parámetros de distribución o incluso con las proporciones de las mezclas en el modelo de mezclas.

De acuerdo con lo anterior y con base en el modelo 3.38, se supone que $\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}$ se distribuye normalmente con media condicional $\boldsymbol{\mu}_{Y}\left(\boldsymbol{x}|\boldsymbol{\beta}_{j}\right)$, dada por una transformación lineal de $\boldsymbol{X}$, y la matriz de covarianza $\boldsymbol{\Sigma}_{\boldsymbol{Y}_{j}}$, $j=1,\ldots,K$. Aquí, $\mu_{\boldsymbol{Y}}\left(\boldsymbol{x}|\boldsymbol{\beta}_{j}\right)=\boldsymbol{\beta}_{j}^{T}\boldsymbol{x}^{*}$ se utiliza cuando $\boldsymbol{B}_{j}\in\mathbb{R}^{(1+q)\times d}$ y $\boldsymbol{x}^{*}=(1,\boldsymbol{x})$. Entonces, el modelo 3.38 se puede reescribir como sigue:

$$p\left(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\phi}\right)=\sum_{j=1}^{K}\pi_{j}\cdot p\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\mu}_{\boldsymbol{Y}}\left(\boldsymbol{x},\boldsymbol{\beta}_{j}\right),\boldsymbol{\Sigma}_{\boldsymbol{Y}_{j}}\right),$$

donde $\boldsymbol{\phi}=(\boldsymbol{\pi}_{j},\boldsymbol{\theta}_{j})$, es decir $\boldsymbol{\pi}_{j}=(\pi_{1},\ldots,\pi_{K-1})$ y $\boldsymbol{\theta}_{j}=(\mathrm{vec}^{T}(B_{1}),\ldots,\mathrm{vec}^{T}(B_{K}),$ $\mathrm{vec}^{T}(\Sigma_{1}),\ldots,\mathrm{vec}^{T}(\Sigma_{K}))^{T}$, con $\pi_{j}>0$ y $\sum_{j=1}^{K}\pi_{j}=1$, para $j=1,\ldots,K$, y $\boldsymbol{\beta}_{j}\in\mathbb{R}^{q\times d}$ es una matriz de coeficientes, para cada j.

En resumen, las mezclas de modelos de regresión lineal se dan como sigue:

$$\boldsymbol{y}_i = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d)^T = \begin{cases} \boldsymbol{x}_{i1}^T \boldsymbol{\beta}_1 + \epsilon_{i1} & \text{con probabilidad } \pi_1 \\ \boldsymbol{x}_{i2}^T \boldsymbol{\beta}_2 + \epsilon_{i2} & \text{con probabilidad } \pi_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ \boldsymbol{x}_{iq}^T \boldsymbol{\beta}_K + \epsilon_{iK} & \text{con probabilidad } \pi_K \end{cases}$$

donde $\boldsymbol{y}_i$ es el vector de la variable respuesta en la i-ésima observación; $\boldsymbol{x}_i^T$ $(i = 1, \ldots, n)$ denota la transpuesta del vector $(q+1)$-dimensional de variables independientes para la i-ésima observación, $\boldsymbol{\beta}_j$ $(j = 1, \ldots, K)$ denota el vector $(q+1)$-dimensional de variables regresoras para el j-ésimo componente, π_j son las probabilidades de las mezclas. Finalmente, ϵ_{ij} son los errores aleatorios; bajo la suposición de normalidad, se tiene que $\epsilon_{ij} \sim N(0, \boldsymbol{\Sigma}_j)$ para toda $i = 1, \ldots, n$ y $j = 1, \ldots, K$ (Dang *et al.*, 2014, Faria y Soromenho, 2010, Huang *et al.*, 2013).

Los parámetros para cada componente de densidad $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\phi})$ pueden estimarse a partir de los datos utilizando el algoritmo Esperanza-Maximización (EM), una técnica ampliamente utilizada para la estimación de parámetros por máxima verosimilitud con modelos de mezclas (Dempster *et al.*, 1977, McLachlan y Krishnan, 1997).

7.2.2.1. Estimación de los parámetros utilizando el algoritmo EM

En este trabajo, la estimación de los parámetros se hizo utilizando técnicas de máxima verosimilitud vía el algoritmo EM (Dang *et al.*, 2014, Dempster *et al.*, 1977). Ahora, de acuerdo con la ecuación 3.39, la verosimilitud condicionada para el conjunto de datos completos es igual a:

$$L_c(\boldsymbol{\phi}|\boldsymbol{S}_c) = \prod_{i=1}^{n} \prod_{j=1}^{K} \left[\pi_j \cdot p\left(\boldsymbol{y}_i|\boldsymbol{\mu}_{\boldsymbol{Y}}\left(\boldsymbol{x}_i, \boldsymbol{\beta}_j\right), \boldsymbol{\Sigma}_{\boldsymbol{Y}_j}\right)\right]^{z_{ij}},$$

y de acuerdo con la ecuación 3.40 su función de log-verosimilitud es igual a:

$$\begin{aligned} l_c(\boldsymbol{\phi}|\boldsymbol{S}_c) &= \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \left[\log \pi_j + \log p\left(\boldsymbol{y}_i|\boldsymbol{\mu}_{\boldsymbol{Y}}\left(\boldsymbol{x}_i, \boldsymbol{\beta}_j\right), \boldsymbol{\Sigma}_{\boldsymbol{Y}_j}\right)\right], \\ &= \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log p\left(\boldsymbol{y}_i|\boldsymbol{\mu}_{\boldsymbol{Y}}\left(\boldsymbol{x}_i, \boldsymbol{\beta}_j\right), \boldsymbol{\Sigma}_{\boldsymbol{Y}_j}\right). \end{aligned} \tag{7.1}$$

El *paso-E* consiste en calcular la esperanza de la log-verosimilitud de los datos completos:

$$\begin{aligned} Q\left(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t-1)}\right) &= E\left\{l_c\left(\boldsymbol{\phi}|\boldsymbol{S}_c\right)\right\}, \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K}\hat{\tau}_{ij}^{(t-1)}\left[Q\left(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_{\boldsymbol{Y}_j}|\boldsymbol{\theta}^{(t-1)}\right) + \log\pi_j^{(t-1)}\right], \end{aligned} \tag{7.2}$$

donde:

$$\hat{\tau}_{ij}^{(t-1)} = E\left\{z_{ij}|\boldsymbol{x}_i, \boldsymbol{y}_i\right\} = \frac{p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\mu}}_{\boldsymbol{Y}}\left(\boldsymbol{x}_i|\hat{\boldsymbol{\beta}}_j^{(t)}\right), \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_j}^{(t-1)}\right)\hat{\pi}_j^{(t-1)}}{\sum_{j=1}^{K} p\left(\boldsymbol{y}_i|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{Y}}}\left(\boldsymbol{x}_i|\hat{\boldsymbol{\beta}}_j^{(t-1)}\right), \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_j}^{(t-1)}\right)\hat{\pi}_j^{(t-1)}}$$

proporciona el valor actual de z_{ij} en la k-ésima iteración y

$$Q\left(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_{\boldsymbol{Y}_j}|\boldsymbol{\phi}^{(t-1)}\right) = \frac{1}{2}\left[-d\log\left(2\pi\right) - \log\left|\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}\right| - \left(\boldsymbol{y}_i - \boldsymbol{\beta}_j^T\boldsymbol{x}_i^*\right)^T\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\beta}_j^T\boldsymbol{x}_i^*\right)^T\right]$$

El *paso-M* consiste en maximizar la esperanza condicional de la log-verosimilitud de los datos completos con respecto a $\boldsymbol{\phi}$. La actualización para $\hat{\pi}_j^{(t)}$ es:

$$\hat{\pi}_j^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_{ij}^{(t-1)}, \tag{7.3}$$

y para $\hat{\boldsymbol{\beta}}_j^{(t)}$ y $\hat{\boldsymbol{\Sigma}}^{(t)}$, $j = 1, \ldots, K$, son:

$$\hat{\boldsymbol{\beta}}_j^{(t)^T} = \left(\sum_{i=1}^{n}\hat{\tau}_{ij}^{(t-1)}\boldsymbol{x}_i\boldsymbol{x}_i^{*T}\right)\left(\sum_{i=1}^{n}\hat{\tau}_{ij}^{(t-1)}\boldsymbol{y}_i\boldsymbol{x}_i^{*T}\right)^{-1}, \tag{7.4}$$

y

$$\hat{\boldsymbol{\Sigma}}^{(t)^T} = \frac{\sum_{i=1}^{n}\hat{\tau}_{ij}^{(t-1)}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\beta}}_j^{(t-1)^T}\boldsymbol{x}_i^*\right)\left(\boldsymbol{y}_i - \hat{\boldsymbol{\beta}}_j^{(t-1)^T}\boldsymbol{x}_i^*\right)^T}{\sum_{i=1}^{n}\hat{\tau}_{ij}^{(t-1)}}. \tag{7.5}$$

Las ecuaciones de la 7.3 a la 7.5 son las actualizaciones de los parámetros para el modelo sin restricciones. El proceso se repite generando valores cada vez más pequeños en la log-verosimilitud de los datos completos con la variable latente incluida.

7.2.2.2. Inicialización en el algoritmo EM

Se ha observado que el algoritmo EM es dependiente de los valores iniciales. Propiedades y convergencia a máximos locales están bien documentados, y los modelos de mezcla Gaussianos también tienen superficies de probabilidad ilimitadas. (Titterington *et al.*, 1985).

Para determinar los valores de los parámetros iniciales que permitan obtener el valor máximo de verosimilitud en el marco de regresión de mezclas Gaussianas multivariadas se utilizó el algoritmo de "Inicialización Aleatoria", el cual consiste en iniciar con diferentes posiciones aleatorias obtenidas de la base de datos (Biernackia *et al.*, 2003, McLachlan y Peel, 2000).

7.2.2.3. Identificación del número óptimo de grupos

La log-verosimilitud se define como el logaritmo de la función de densidad conjunta, la cual se puede ver como un indicador de bondad de ajuste para los modelos probabilísticos.

En este trabajo se determinó el número de componentes con la metodología descrita por Camargo *et al.* (2007) y Gaffney *et al.* (2007), quienes estimaron el número de grupos versus la log-verosimiltud, y el número de grupos versus la suma de cuadrados dentro de grupos.

Para el primer criterio de selección, primero, se fijó el número de grupos y, después se estimaron los valores de la log-verosimilitud mediante el algoritmo EM en modelos de mezclas Gaussinas. Para el segundo criterio, se fijó el número de grupos y después mediante el algoritmo K-medias se estimó la suma de cuadrados dentro de cada grupo (**?**).

Finalmente, se seleccionó el número de grupos para los que la log-verosimilitud es más grande y para los que su suma de cuadrados no cambia radicalmente.

7.3. Resultados

7.3.1. Estimación de la función de densidad de probabilidad

De acuerdo con la Figura 7.1 y 7.2, el número de grupos por intervalo (1976-2013 y 1990-2013) es dos (es decir $K = 2$ componentes). Su log-verosimiltud promedio fue aproximadamente 0.40 para ambos intervalos.

Los valores de los parámetros iniciales para cada uno de los componentes de las mezclas ($K = 2$) se determinaron mediante el algoritmo de inicialización aleatoria. Los parámetros θ se estimaron iterativamente mediante el algoritmo EM.

En las Tablas 7.1, 7.2, 7.3 y 7.4 se muestran los parámetros de los modelos de mezclas ajustados a los datos de génesis de huracanes para los diferentes intervalos de estudio:

Tabla 7.1: Parámetros del modelo de mezclas ajustados a los datos de génesis de ciclones tropicales para el intervalo 1951-1975.

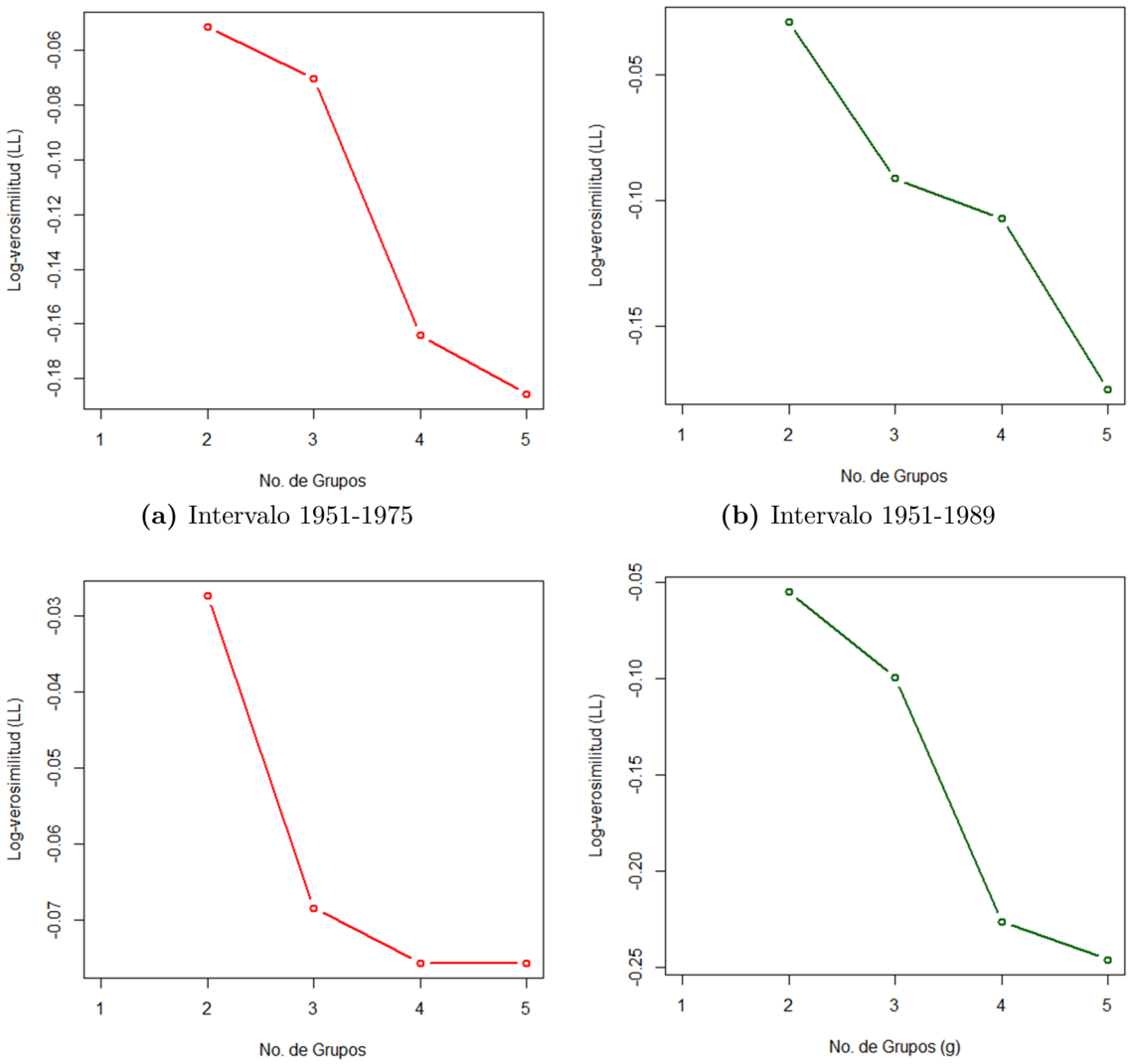
ϕ	**Modelo del grupo 1**	**Modelo del grupo 2**
$\hat{\boldsymbol{\pi}} =$	$\pi_1 = 0.5344$	$\pi_2 = 0.4656$
$\hat{\boldsymbol{\beta}} =$	$\begin{pmatrix} \beta_{10} = 35.7214 & \beta_{20} = 78.2200 \\ \beta_{11} = -3.1249 & \beta_{21} = -2.2669 \end{pmatrix}$	$\begin{pmatrix} \beta_{10} = 52.1107 & \beta_{20} = 92.9000 \\ \beta_{11} = -4.6181 & \beta_{21} = -2.5004 \end{pmatrix}$
$\hat{\boldsymbol{\Sigma}} =$	$\begin{pmatrix} \Sigma_{11} = 671.3671 & \Sigma_{12} = -69.4745 \\ \Sigma_{21} = -69.4745 & \Sigma_{22} = 13.0963 \end{pmatrix}$	$\begin{pmatrix} \Sigma_{11} = 94.0119 & \Sigma_{12} = -12.0057 \\ \Sigma_{21} = -12.0057 & \Sigma_{22} = 3.0502 \end{pmatrix}$

Tabla 7.2: Parámetros del modelo de mezclas ajustados a los datos de génesis de ciclones tropicales para el intervalo 1976-2013.

ϕ	**Modelo del grupo 1**	**Modelo del grupo 2**
$\hat{\boldsymbol{\pi}} =$	$\pi_1 = 0.3589$	$\pi_2 = 0.6411$
$\hat{\boldsymbol{\beta}} =$	$\begin{pmatrix} \beta_{10} = -71.0227 & \beta_{20} = 100.8808 \\ \beta_{11} = 0.0593 & \beta_{21} = -2.9795 \end{pmatrix}$	$\begin{pmatrix} \beta_{10} = -1.9954 & \beta_{20} = 110.8600 \\ \beta_{11} = -1.9788 & \beta_{21} = -3.3457 \end{pmatrix}$
$\hat{\boldsymbol{\Sigma}} =$	$\begin{pmatrix} \Sigma_{11} = 558.5962 & \Sigma_{12} = -61.1992 \\ \Sigma_{21} = -61.1992 & \Sigma_{22} = 17.8760 \end{pmatrix}$	$\begin{pmatrix} \Sigma_{11} = 457.3045 & \Sigma_{12} = -110.8558 \\ \Sigma_{21} = -110.8559 & \Sigma_{22} = 29.9740 \end{pmatrix}$

Tabla 7.3: Parámetros del modelo de mezclas ajustados a los datos de génesis de ciclones tropicales para el intervalo 1951-1989.

ϕ	**Modelo del grupo 1**	**Modelo del grupo 2**
$\hat{\boldsymbol{\pi}} =$	$\pi_1 = 0.5494$	$\pi_2 = 0.4506$
$\hat{\boldsymbol{\beta}} =$	$\begin{pmatrix} \beta_{10} = 1.3827 & \beta_{20} = 92.6383 \\ \beta_{11} = -1.9103 & \beta_{21} = -2.7871 \end{pmatrix}$	$\begin{pmatrix} \beta_{10} = 41.2615 & \beta_{20} = 96.2927 \\ \beta_{11} = -4.2692 & \beta_{21} = -2.6188 \end{pmatrix}$
$\hat{\boldsymbol{\Sigma}} =$	$\begin{pmatrix} \Sigma_{11} = 663.2198 & \Sigma_{12} = -72.8618 \\ \Sigma_{21} = -72.8618 & \Sigma_{22} = 14.0046 \end{pmatrix}$	$\begin{pmatrix} \Sigma_{11} = 96.5979 & \Sigma_{12} = -12.3884 \\ \Sigma_{21} = -12.3884 & \Sigma_{22} = 3.34967 \end{pmatrix}$



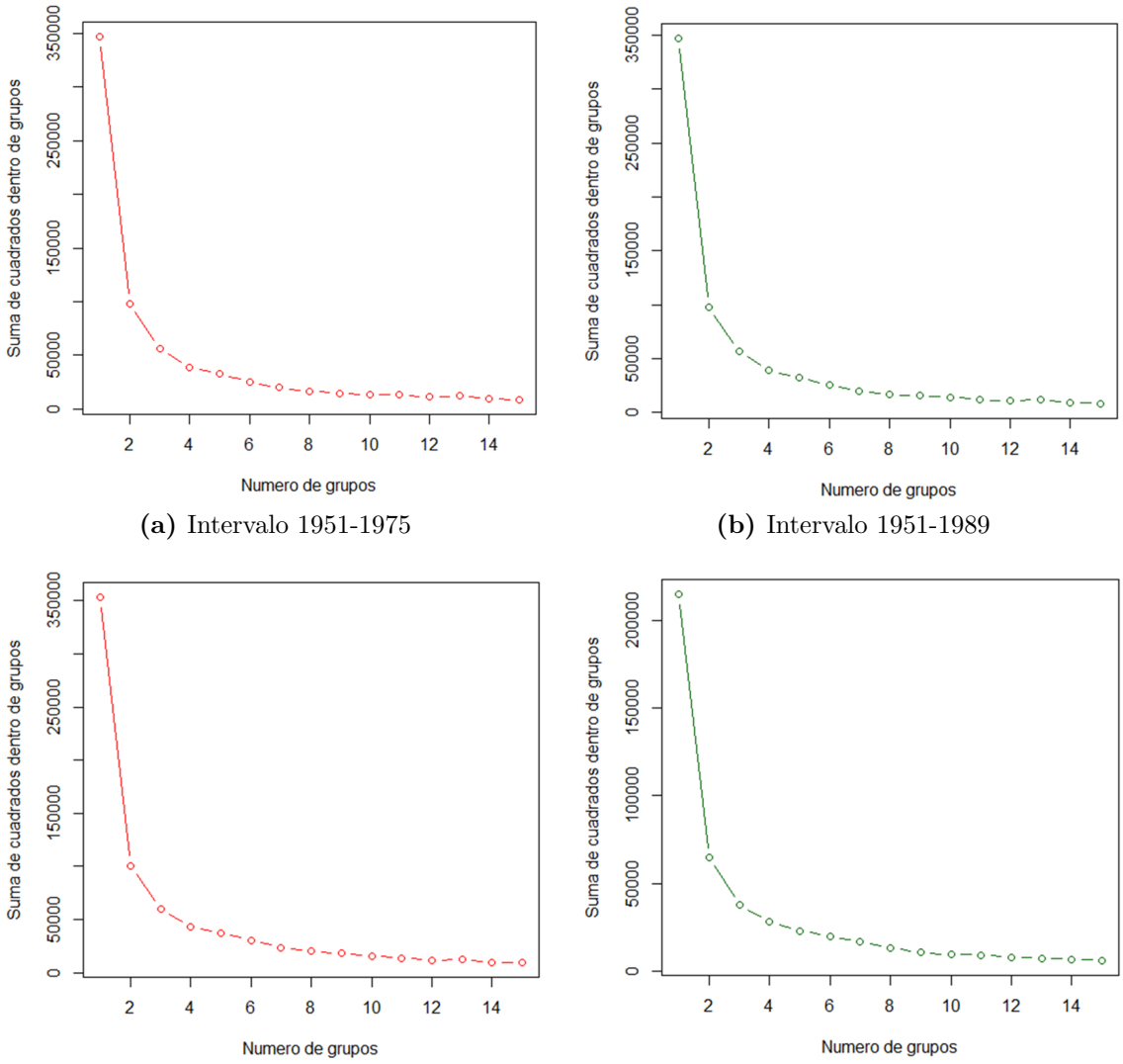
(a) Intervalo 1951-1975

(b) Intervalo 1951-1989

(c) Intervalo 1976-2013

(d) Intervalo 1990-2013

Figura 7.1: La gráfica número de grupos versus log-verosimilitud (LL) muestra que a medida que aumenta el número de grupos por intervalo disminuye la log-verosimilitud en las dos gráficas, lo que significa que la consistencia en su estructura de disminuye.



(a) Intervalo 1951-1975

(b) Intervalo 1951-1989

(c) Intervalo 1976-2013

(d) Intervalo 1990-2013

Figura 7.2: La gráfica número de grupos versus suma de cuadrados dentro del grupo muestra que en general la suma de cuadrados dentro del grupo decrece fuertemente del grupo uno al tres, con poca disminución después, esto significa que un número óptimo de grupos podría ser 2 ó 3.

Tabla 7.4: Parámetros del modelo de mezclas ajustados a los datos de génesis de ciclones tropicales para el intervalo 1990-2013.

$\boldsymbol{\phi}$	**Modelo del grupo 1**	**Modelo del grupo 2**
$\hat{\boldsymbol{\pi}} =$	$\pi_1 = 0.3977$	$\pi_2 = 0.6023$
$\hat{\boldsymbol{\beta}} =$	$\begin{pmatrix} \beta_{10} = -33.7647 & \beta_{20} = 90.5576 \\ \beta_{11} = -1.7707 & \beta_{21} = -2.4877 \end{pmatrix}$	$\begin{pmatrix} \beta_{10} = -47.3257 & \beta_{20} = 124.5170 \\ \beta_{11} = 0.0943 & \beta_{21} = -3.9452 \end{pmatrix}$
$\hat{\boldsymbol{\Sigma}} =$	$\begin{pmatrix} \Sigma_{11} = 78.1543 & \Sigma_{12} = -0.6333 \\ \Sigma_{21} = -0.6333 & \Sigma_{22} = 12.3308 \end{pmatrix}$	$\begin{pmatrix} \Sigma_{11} = 262.6947 & \Sigma_{12} = -56.8318 \\ \Sigma_{21} = -56.8318 & \Sigma_{22} = 17.0725 \end{pmatrix}$

De acuerdo con las Tablas antes mencionadas se puede observar que las probabilidades del grupo 1 son mayores que las del 2 para ambos intervalos de estudio (1951-1975 y 1951-1989). Sin embargo, éstas se comportan de forma inversa para el intervalo 1976-2013 y 1990-2013, es decir las probabilidades del grupo 2 son mayores que las de grupo 1.

7.3.2. Simulación mediante el modelo de regresión de mezclas Gaussianas

Con los parámetros obtenidos de los modelos de mezclas ajustadas a los datos de la temperatura de la superficie del mar TSM se hicieron varias simulaciones para evaluar su efecto versus la ubicación (*Longitud*, *Latitud*) del centroide de los puntos de ocurrencia de los ciclones tropicales. Los datos se generaron a partir de una mezcla bivariada de distribuciones normales mediante el modelo de mezclas de regresión variando la TSM: 27.14°C, 27.50°C, 28.00°C, 28.50°C, 29.00°C, 29.50°C, 30.00°C para los dos intervalos 1976-2013 y 1990-2013. Con los datos generados de la simulación se procedió a estimar sus centroides mediante el algoritmo EM para modelos de mezclas Gaussianas.

A continuación se muestran los modelos de regresión de la génesis de los ciclones tropicales (*Longitud*, *Latitud*) en función de la temperatura de la superficie del mar TSM para los diferentes intervalos de estudio:

Tabla 7.5: Modelo de regresión de la ciclogénesis en función de la TSM para el intervalo 1951-1975.

Modelo del grupo 1	**Modelo del grupo 2**
Longitud $= 35.7214 - 3.1249 \cdot$ **TSM**	**Longitud** $= 78.2200 - 2.2669 \cdot$ **TSM**
Latitud $= 52.1107 - 4.6181 \cdot$ **TSM**	**Latitud** $= 92.9000 - 2.5004 \cdot$ **TSM**

Con los modelos de regresión se simuló la ciclogénesis en función de la TSM, obteniéndose los parámetros para cada uno de los diferentes intervalos.

Tabla 7.6: Modelo de regresión de la ciclogénesis en función de la TSM para el intervalo 1976-2013.

Modelo del grupo 1	Modelo del grupo 2
Longitud $= -71.0227 + 0.0593 \cdot$ **TSM**	**Longitud** $= -2.9795 - 1.9788 \cdot$ **TSM**
Latitud $= 100.8808 - 2.9795 \cdot$ **TSM**	**Latitud** $= 110.8600 - 3.3457 \cdot$ **TSM**

Tabla 7.7: Modelo de regresión de la ciclogénesis en función de la TSM para el intervalo 1951-1989.

Modelo del grupo 1	Modelo del grupo 2
Longitud $= 1.3827 - 1.9103 \cdot$ **TSM**	**Longitud** $= 41.2615 - 4.2692 \cdot$ **TSM**
Latitud $= 92.6383 - 2.7871 \cdot$ **TSM**	**Latitud** $= 96.2927 - 2.6188 \cdot$ **TSM**

Tabla 7.8: Modelo de regresión de la ciclogénesis en función de la TSM para el intervalo 1990-2013.

Modelo del grupo 1	Modelo del grupo 2
Longitud $= -33.7647 - 1.7707 \cdot$ **TSM**	**Longitud** $= -47.3257 + 0.0943 \cdot$ **TSM**
Latitud $= 90.5576 - 2.4877 \cdot$ **TSM**	**Latitud** $= 124.5170 - 3.9452 \cdot$ **TSM**

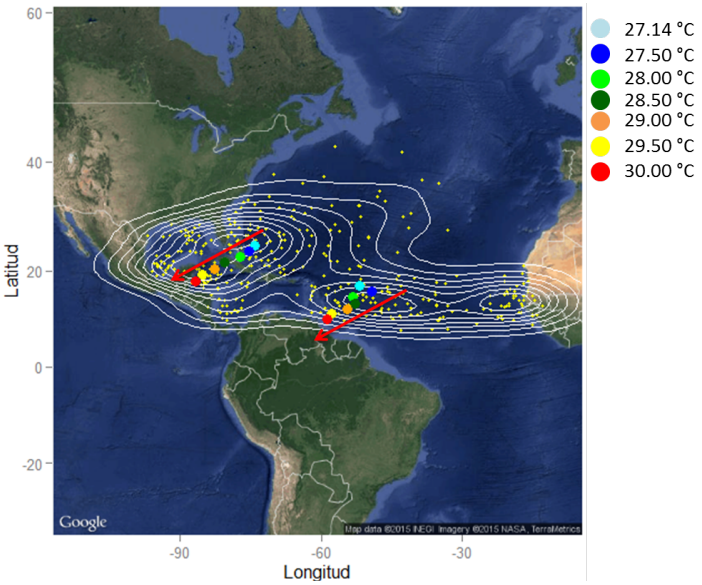
En la Figura 7.3, 7.4, 7.5 y 7.6 se muestra el comportamiento de la posición de los centroides de las regiones ciclogenéticas en función de la TSM.

7.4. Discusión y conclusiones

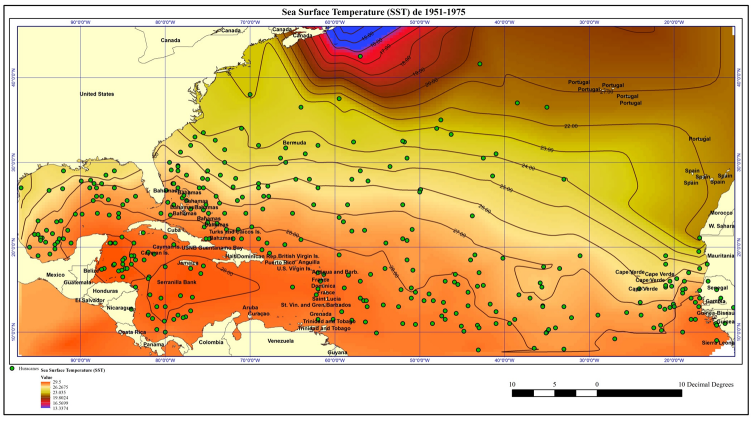
De acuerdo con el criterio estadístico de selección utilizado para determinar el número grupos por intervalo (ver 7.2.2.3), éste indica que solo hay dos, ubicados en la Costa-Este y el Atlántico Tropical de la cuenca oceánica del Atlántico Norte. Por lo tanto se concluye que solamente hay dos regiones que generan los ciclones tropicales desde 1951 hasta el 2013.

Respecto a los modelos de mezclas ajustados, éstos muestran que los centroides de la génesis de ciclones tropicales se desplazan de las zonas de menor a Mayor TSM, comportamiento que es similar en todos los intervalos de estudio. Por lo tanto; se concluye que la TSM influye directamente en el desplazamiento de los centroides, ver Figura 7.3, 7.4, 7.5 y 7.6. Esto fortalece las afirmaciones de Emanuel (2005), Webster *et al.* (2005), Chan (2006), y Holland y Webster (2007) de que la TSM influye en las actividades de los ciclones tropicales, particularmente, la génesis.

Aunado a lo anterior, también se conluye que la génesis de los ciclones tropicales es afectada por el cambio climático de acuerdo con Hegerl *et al.* (2007), quienes mencionan que el aumento de las concentraciones de gases de efecto invernadero han

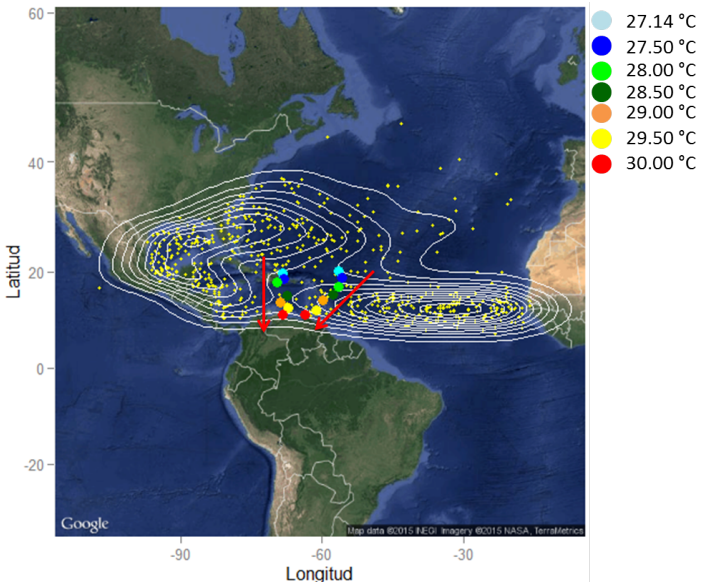


(a) 1951-1975.

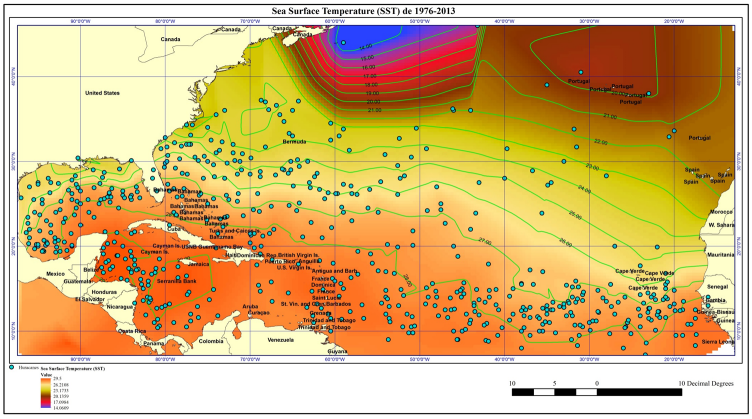


(b) Temperatura promedio de la superficie del mar para el intervalo 1951-1975.

Figura 7.3: (a) Muestra el comportamiento de los centroides de la ciclogénesis respecto a la temperatura de la superficie del mar de acuerdo con los datos del intervalo de 1951-1975. (b) Muestra el comportamiento de la temperatura promedio de la superficie del mar para el intervalo 1951-1975. Fuente: Elaboración Propia con datos de la TSM, generados de la base de datos de la NOAA.

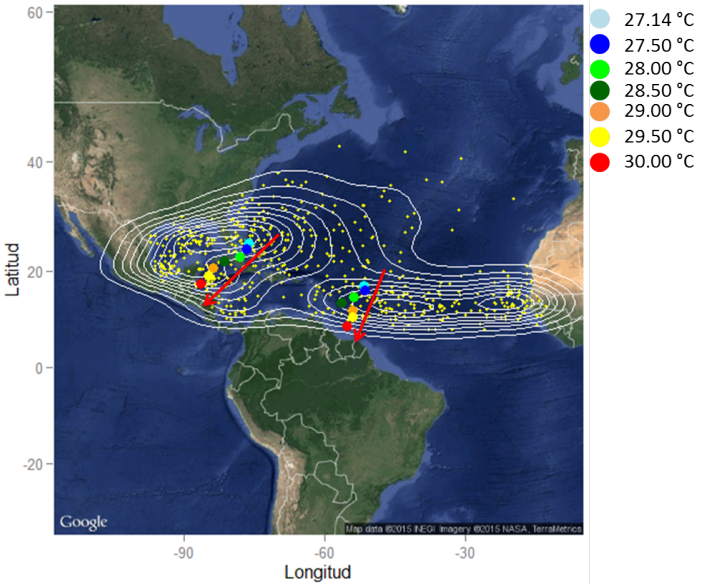


(a) 1976-2013.

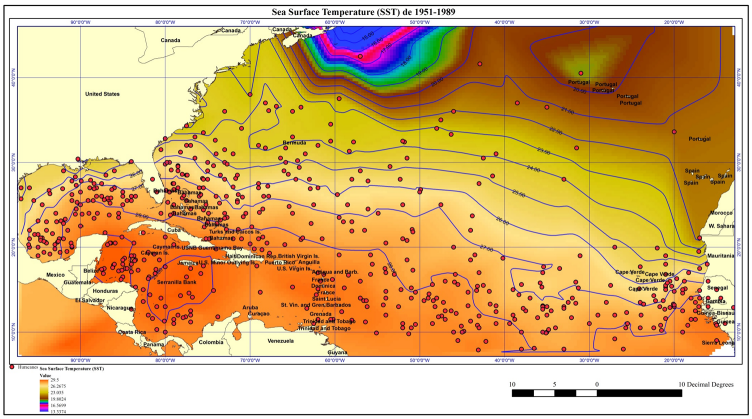


(b) Temperatura promedio de la superficie del mar para el intervalo 1976-2013.

Figura 7.4: (a) Muestra el comportamiento de los centroides de la ciclogénesis respecto a la temperatura de la superficie del mar de acuerdo con los datos del intervalo 1976-2013. (b) Muestra el comportamiento de la temperatura promedio de la superficie del mar para el intervalo 1976-2013. Fuente: Elaboración Propia con datos de la TSM, generados de la base de datos de la NOAA.

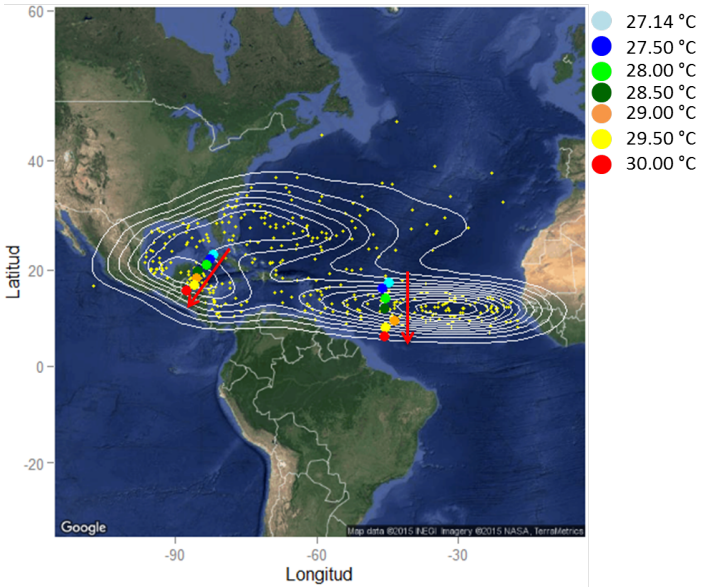


(a) 1951-1989.

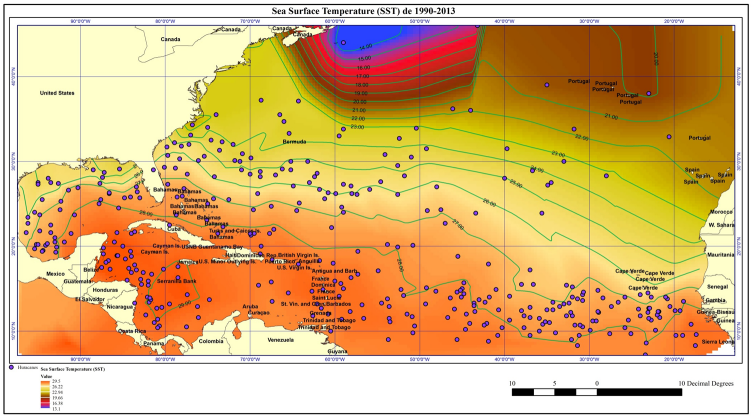


(b) Temperatura promedio de la superficie del mar para el intervalo 1 951-1989.

Figura 7.5: (a) Muestra el comportamiento de los centroides de la ciclogénesis respecto a la temperatura de la superficie del mar de acuerdo con los datos del intervalo 1951-1989. (b) Muestra el comportamiento de la temperatura promedio de la superficie del mar para el intervalo 1951-1989. Fuente: Elaboración Propia con datos de la TSM, generados de la base de datos de la NOAA.



(a) 1990-2013.



(b) Temperatura promedio de la superficie del mar para el intervalo 1990-2013.

Figura 7.6: (a) Muestra el comportamiento de los centroides de la ciclogénesis respecto a la temperatura de la superficie del mar de acuerdo con los datos del intervalo 1990-2013. (b) Muestra el comportamiento de la temperatura promedio de la superficie del mar para el intervalo 1990-2013. Fuente: Elaboración Propia con datos de la TSM, generados de la base de datos de la NOAA.

contribuido en el incremento de la temperatura del mar en la región de génesis del Atlántico Norte.

Respecto al comportamiento de los centroides de las dos regiones ciclogenéticas ubicadas en la Costa-Este de la cuenca Atlántico, éste es el mismo tanto para el intervalo 1951-1976 como para el intervalo 1976-2013; es decir, los centroides se desplazan hacia las zonas de mayor TSM. Con la diferencia de que en el primer intervalo los centroides se desplazan al sur-oeste del Caribe y del Atlántico Tropical (ver Figura 7.3) en tanto que para el intervalo 1976-2013, ambos centroides se desplazan hacia el sur del Atlántico Tropical, ver Figura 7.4.

En el caso del comportamiento de los centroides para el intervalo 1951-1989 vs 1990-2013 es similar al 1951-1975 vs 1976-2013, con la diferencia de que aquí los centroides se desplazaron al sur-oeste del Caribe y del Atlántico Tropical, respectivamente, ver Figuras 7.5 y 7.6.

Capítulo 8

Conclusiones

8.1. Modelos de Mezclas Gaussianas y Procesos Dirichlet

8.1.1. Determinación del número de grupos

El número de grupos determinado para el intervalo 1951-1975 vs 1976-2013 y 1951-1989 vs 1990-2013 mediante el Modelos de Mezclas Gaussianas, *MMG*, fue dos ($K = 2$), respectivamente. En contraste, el número de grupos determinado mediante el Modelo de Mezclas de Procesos Dirichlet, *DPMM*, fue dos ($K = 2$) y tres grupos ($K = 3$) para el intervalo 1951-1975 vs 1976-2013, respectivamente. En tanto que para el intervalo 1951-1989 vs 1990-2013 el número de grupos fue tres ($K = 3$), respectivamente.

En la Tabla 8.1 se muestra el resumen del número de grupos determinado mediante los dos de métodos propuestos. Note que con ambos métodos se obtiene el mismo número de grupos ($K = 2$) para el intervalo 1951-1975. Sin embargo, para el intervalo 1976-2013, 1951-1990 y 1990-2013, el número de grupos cambia dependiendo el método. El número de grupos estimado mediante el método *MMG* es de dos ($K = 2$) en tanto que el número de grupos estimado mediante el *DPMM* es de tres ($K = 3$).

8.1.2. Comparación de las funciones de densidad

En el inciso a) y b) de la Figura 5.6 se muestran los centroides estimados mediante el *MMG*. Los centroides de los intervalo 1951-1975 vs 1976-2013 aparecen en color verde

Tabla 8.1: Número de grupos estimados por los dos métodos aplicados.

Período	Método *MMG*	Método *DPMM*
1951-1975	2	2
1976-2013	2	3
1951-1990	2	3
1990-2013	2	3

y se han desplazado nor-este de la cuenca oceánica, es decir se están moviendo del Golfo hacia la Costa-Este. Respecto a los centroides que aparecen en color rojo, éstos también se han desplazado hacia el sur-este de la cuenca, es decir se están moviendo dentro de la misma zona del Atlántico Tropical. Finalmente, los centroides del intervalo 1951-1989 vs 1990-2013, se puede observar que tienen el mismo comportamiento que el intervalo 1951-1975 vs 1976-2013, ver inciso a) en la Figura 5.6.

Análogamente como en los *MMG*, en el inciso a) y b) de la Figura 6.7 se muestran los centroides estimados mediante el *DPMM*. El centroide del intervalo 1976-2013 con respecto al intervalo 1951-1975 se ha desplazado hacia el sur de la cuenca oceánica. Sin embargo; en el intervalo 1976-2013, el centroide que aparecía en color verde (Figura 5.6) se dividió en dos centroides, uno en color rojo y otro en color azul. El centroide de color rojo no se puede comparar con el el intervalo 1951-1975, ya que en esta región solo se formó solo un centroide (color azul). Respecto al centroide de color azul, ubicado en el Golfo, éste se está moviendo hacia el nor-este, es decir hacia la Costa-Este. Finalmente, los centroides del intervalo 1951-1989 vs 1990-2013 tienen el mismo comportamiento que los centroides del intervalo 1951-1975 vs 1976-2013 estimados mediante el *MMG*, con la diferencia que aquí son tres centroides, ver inciso b) en la Figura 6.7.

8.2. Mezclas Gaussianas de Modelos de Regresión Lineal

8.2.1. Determinación del número de grupos

El número óptimo de grupos determinado mediante la log-verosimlitud y la suma de cuadrados dentro de grupos fue dos ($K = 2$) para cada uno de los intervalos de análisis.

8.2.2. Efecto de la TSM en la ciclogénesis

En las Tablas 7.5, 7.6, 7.7, y 7.8 se muestran los modelos obtenidos mediante mezclas Gaussianas de regresión lineal, con las cuales se predijo la ubicación de los centroides de las regiones de génesis versus diferentes TSM. Las predicciones muestran que la TSM influye en la localización de los centroides de las regiones de génesis, en todos los períodos de estudio, ver Figura 7.3, 7.4 7.5, y 7.6, tal y como mencionan Chan (2006), Emanuel (2005), Holland y Webster (2007), Webster *et al.* (2005), quienes encontraron que la TSM influye en las actividades de los ciclones tropicales, particularmente, las de la génesis. Dichas predicciones también permite concluir que la génesis de los ciclones tropicales es afectada por el calentamiento global, tal y como lo mencionan Hegerl *et al.* (2007), quienes mencionan que el aumento de las concentraciones de gases de efecto invernadero ha contribuido en el incremento de la temperatura de la superficie del mar en la Región Ciclogenética del Atlántico Norte.

8.3. Trabajos Futuros

Como continuación de este trabajo de tesis y como en cualquier otro proyecto de investigación, existen diversas líneas de investigación que quedan abiertas y en las que es posible continuar trabajando. Durante el desarrollo de esta tesis han surgido algunas líneas futuras que se han dejado abiertas y que se esperan atacar en un futuro; algunas de ellas, están más directamente relacionadas con este trabajo de tesis y son el resultado de cuestiones que han ido surgiendo durante la realización de la misma. Otras, son líneas más generales que, sin embargo, no son objeto de esta tesis; estas líneas pueden servir para retomarlas posteriormente o como opción a trabajos futuros para otros investigadores.

A continuación se presentan algunos trabajos futuros que pueden desarrollarse como resultado de esta investigación o que, por exceder el alcance de esta tesis, no han podido ser tratados con la suficiente profundidad. Además, se sugieren algunos desarrollos específicos para apoyar y mejorar el modelo y metodología propuestos. Entre los posibles trabajos futuros se destacan:

- Desarrollar mezclas de proceso Dirichlet de Modelos de Regresión para las covariables Temperatura del Mar, Cizalladura de Viento y Temperatura de la Troposfera.
- Desarrollar modelos lineales dinámicos con mezclas de proceso Dirichlet para datos de génesis de ciclones.
- Modelos de Mezclas de Procesos Dirichlet con datos de Lysis.

- Modelos de Mezclas de Procesos Dirichlet con datos de trayectorias.
- Aplicar los modelos propuesto para diferentes intervalos con el propósito de identificar el efecto del cambio climático.
- Aplicar todas la metodologías propuestas para la cuenca del océano Pacífico.

Referencias

Aldous, D. (1985). Exchangeability and related topics. En *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, 1–198. Springer.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2, 6, 1152–1174.

Archambeau, C. (2008). Lecture 3a: Dirichlet processes. Advanced Topics in Machine Learning (MSc in Intelligent Systems) . http://www0.cs.ucl.ac.uk/staff/C.Archambeau/ATML/atml_files/atml08_lect3_dps.pdf. [Web; accedido el 20-06-2016].

Babu, G. J. y Singh, K. (1983). Inference on means using the bootstrap. *Annals of Statistics*, 11, 338–370.

Bickel, O. y Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9, 1196–1217.

Biernackia, C., Celeuxb, G. y Govaertc, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, 41, 561–575.

Blackwell, D. y MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1, 2, 353–355.

Camargo, S. J., Robertson, A. W., Gaffney, S. J., Smyth, P. y Ghil, M. (2007). Cluster Analysis of Typhoon Tracks. Part I: General Properties. *American Meteorological Society*, 20, 3635–3653.

Chan, J. C. L. (2006). Comment on changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 311, 1713c.

Chan, J. C. L. (2007). Interannual variations of intense typhoon activity. *Tellus*, 59(A), 4, 455–460.

Chan, J. C. L. y Liu, K. S. (2004). Global warming and western North Pacific typhoon activity from an observational perspective. *Journal of Climate*, 17, 4590–4602.

Dang, U. J., Punzoy, A., McNicholasz, P. D., Ingrassiax, S. y Browne, R. P. (2014). Multivariate response and parsimony for Gaussian cluster-weighted models. *arXiv:1411.0560v1 [stat.CO] 3 Nov 2014*, 1–24.

Referencias

Dempster, A. P., Laird, N. y Rubin, D. (1977). Maximum Likelihood for Incomplete Data Via The EM Algorithm. *Journal of the Royal Statistical Society*, 39, 1–38.

DeSarbo, W. S. y Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regressions. *Journal of Classification*, 5, 249–282.

DiCiccio, T. y Tibshirani, R. (1987). Bootstrap confidence intervals and bootstrap approximations. *Journal of American Statistical Association*, 82, 397, 163–170.

EcuRed (2010). Circulación termohalina. http://www.ecured.cu/index.php/Circulaci%C3%B3n_termohalina. [Web; accedido el 03-02-2015].

Efron, B. (1979). Bootstrap Method: another look at the Jackknife . *Annals of Statistics*, 7, 1, 1–26.

Efron, B. y Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York, London.

Elsner, J. B., Kossin, J. P. y Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, 455, 92–95.

Elsner, J. B., Tsonis, A. A. y Jagger, T. H. (2006). High-frequency variability in hurricane power dissipation and its relationship to global temperature. *Journal of Climate*, 87, 6, 763–768.

Emanuel, K. (2005). Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 309, 1844–1846.

Emanuel, K. A. (2007). Environmental factors affecting tropical cyclone power dissipation. *American Meteorological Society*, 20, 5497–5509.

Enfield, D. B., Mestas-Nunez, A. M. y Trimble, P. J. (2001). The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental U.S. *Geophysical Research Letters*, 28, 2077–2080.

Engel, J. (2010). On Teaching Bootstrap Confidence Intervals. *ICOTS8*.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89, 425, 268–277.

Escobar, M. y West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90, 557–588.

Faria, S. y Soromenho, G. (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80, 2, 201–225.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 2, 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, 287–303.

Figueiredo, M. A. T. y Jain, A. K. (2000). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 381–396.

Fox, E. B. (2009). *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Tesis Doctoral, Massachusetts Institute of Technology.

Fraley, C. y Raftery, A. E. (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Inf. Téc. 504, University of Washington, Department of Statistics, Seattle, WA.

Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York, primera edición.

Fukunaga, K. (1993). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Gaffney, S. J., Robertson, A. W., Smyth, P., Camargo, S. J. y Ghil, M. (2007). Bayesian Analysis of U.S. Hurricane Climate. *Springer-Verlag*, 29, 423–440.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B. y Rubin, D. B. (2014). *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, tercera edición.

Ghahramani, Z. (2005). Non-parametric Bayesian Methods. Uncertainty in Artificial Intelligence. Tutorial. http://mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf. [Web; accedido el 20-06-2016].

Goldenberg, S. B., Landsea, C. W., Mestas, A. M. y Gray, W. M. (2001). The recent increase in Atlantic hurricane activity: Causes and implications. *Science*, 293, 474–479.

Green, P. J. y Silverman, B. B. (1994). *Nonparametric regression and generalized linear models*. Chapman and Hall, London.

Gyorfi, L., Kohler, M., Krzyzak, A. y Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Chapman and Hall, Springer, New York.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annal of Statistics*, 16, 3, 927–953.

Hall, T. M. y Jewson, S. (2007). Statistical Modelling of North Atlantic Tropical Cyclone Tracks. *Tellus*, 59A, 486–498.

Hartigan, J. A. y Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28, 100–108.

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, Y., Orsini, J. A. M., Nicholls, N., Penner, J. E. y Stott, P. A. (2007). Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Understanding and Attributing Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Henderson-Sellers, A., Zhang, H., Berz, G., Emanuel, K., Gray, W., Landsea, C., Holland, G., Lighthill, J., Shieh, S.-L., Webster, P. y McGuffie, K. (1998). Tropical Cyclones and Global Climate Change: A Post-IPCC Assessment. *American Meteorological Society*, 79, 1, 19–38.

Henson, B. (2005). Going to extremes. UCAR Quartely. Winter 2004-2005. *Unprinted*.

Hershey, J. R. y Olsen, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. En *IEEE Conf. Acoust. Speech Signal Processing*, 317–320.

Holland, G. J. y Webster, P. J. (2007). Heightened tropical cyclone activity in the North Atlantic: Natural variability or climate trend? *Philosophical Transactions. The Royal Society*, 365, 2695–2716.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., der Linden, P. J. V., Dai, X., Maskell, K. y Johnson, C. A. (2001). Climate Change 2001: The Scientific Basis is the most comprehensive and up-to-date scientific assessment of past, present and future climate change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA*, 1–83.

Huang, J. (2005). Maximum Likelihood Estimation of Dirichlet Distribution Parameters.

Huang, M., Li, R. y Wang, S. (2013). Nonparametric Mixture of Regression Models. *Journal of the American Statistical Association*, 108, 503, 929–941.

Ishwaran, H. y James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 453, 161–173.

Ishwaran, H. y Zarepour, M. (2000). Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, 2, 371–390.

Ishwaran, H. y Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 30, 2, 269–283.

Jones, P. N. y McLachlan, G. J. (1989). Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, 34, 233–240.

Kaufman, L. y Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Kerr, K. A. (2000). A North Atlantic climate pacemaker for the centuries. *Science*, 288, 5473, 1984–1986.

Killick, R. y Eckley, I. A. (2013). changepoint: An R Package for Changepoint Analysis. Inf. téc., Lancaster University.

Kim, J. H., Ho, C. H., Lee, M. H., Jeong, J. H. y Chen, D. (2006). Large increase in heavy rainfall associated with tropical cyclone landfalls in Korea after the late 1970s. *Geophysical Research Letters*, 33, 1–5.

Klotzbach, P. J. (2006). Trends in global tropical cyclone activity over the past twenty years (1986-2005). *Geophysical Research Letters*, 33, 1–4.

Knaff, J. A. y Zehr, R. M. (2007). Reexamination of tropical cyclone wind-pressure relationships. *Weather and Forecasting. American Meteorological Society*, 22, 71–88.

Knapp, K. R., KruK, M. C., Levinson, D. H., Diamond, H. J. y Neumann, C. J. (2010). The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical cyclone Data. *American Meteorological Society*, 366–376.

Knudsen, M. F., Seidenkrantz, M.-S., Jacobsen, B. H. y Kuijpers, A. (2011). Tracking the Atlantic Multidecadal Oscillation through the last 8,000 years. *Nature Communications*, 2, 178, 1–8.

Knutson, T. R. y Tuleya, R. E. (2004). Impact of CO_2 Induced Warming on Simulated Hurricane Intensity and Precipitation: Sensitivity to the Choice of Climate Model and Convective Parameterization. *Journal of Climate*, 17, 18, 3477–3494.

Kossin, J. P., Knapp, K. R., D. J. Vimont, R. J. M. y Harper, B. A. (2007). A globally consistent reanalysis of hurricane variability and trends. *Geophysical Research Letters*, 34, 1–6.

Lau, K. M., Zhou, Y. P. y Wu, H. T. (2008). Have tropical cyclones been feeding more extreme rainfall? *Journal of Geophysical Research*, 113, 1–12.

Lehmann, E. L. y Casella, G. (1998). *Theory of point estimation*. Springer, New York.

MacEachern, S. y Muller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7, 2, 223–238.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297. University of California Press, Berkeley, California.

Maitra, R. (2009). Initializing Partition-Optimization Algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 1, 144–157.

Malsiner-Walli, G., Fruhwirth-Schnatter, S. y Grun, B. (2014). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*.

Mann, M. E. y Emanuel, K. A. (2006). Atlantic hurricane trends linked to climate change. *EOS, Transactions, American Geophysical Union*, 87, 24, 233–244.

McDonnell, K. A. y Holbrook, N. J. (2004). A Poisson Regression Model of Tropical Cyclogenesis for the Australian Southwest Pacific Ocean Region. *Weather and Forecasting-American Meteorological Society*, 19, 440–455.

McLachlan, G. y Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley series in probability and statistics. John Wiley & Sons, New York.

McLachlan, G. y Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

McLachlan, G. J. y Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

Mori, N., Kuniyoshi, S., Nakajo, S., Yasuda, T. y Mase, H. (2013). Projection of Future Tropical Cyclone Activity and Extreme Waves. *Coastal Dynamics*, 1229–1240.

Muller, P., Erkanli, A. y West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83, 1, 67–79.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 2, 249–265.

Nelder, J. A. y Mead, R. (1965). A simplex method for function minimization. *The ComputerJournal*, 7, 308–313.

Orbanz, P. (2014). Lecture Notes on Bayesian Nonparametrics. Inf. téc., Columbia University.

Papaspiliopoulos, O. y Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo for Dirichlet process hierarchical models. *Biometrika*, 95, 169–186.

Pielke Jr., R. A., Landsea, C., Mayfield, M., Laver, J. y Pasch, R. (2005). Hurricanes and Global Warming. *American Meteorological Society*, 1571–1575.

Pitman, J. (2002). Combinatorial stochastic processes. Inf. Téc. 621, U.C. Berkeley Department of Statistics.

Pitman, J. (2006). *Combinatorial stochastic processes, vol. 1875 of Lecture Notes in Mathematics*. Springer Series in Statistics, New York: Springer-Verlag.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67, 306–310.

Ranganathan, A. (2006). The Dirichlet Process Mixture (DPM) Model.

Redner, R. A. y Walker, H. F. (1984). Mixtures densities, Maximum likelihood and The EM Algorithm. *SIAM Review*, 26, 2, 195–239.

Robert, C. P. (2001). *The Bayesian Choice: A Decision Theoretic Motivation*. Springer Verlag, New York, segunda edición.

Rodríguez, A. (2007). *Some Advances in Bayesian Nonparametric Modeling*. Tesis Doctoral, Duke University.

Rumpf, J., Weindl, H., Höpe, P., Rauch, E. y Schmidt, V. (2007). Stochastic Modelling of Tropical Cyclone Tracks. *Mathematical Methods of Operations Research*, 66, 475–490.

Ruppert, D., Wand, M. P. y Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

Saunders, M. A. y Lea, A. S. (2008). Large contribution of sea surface warming to recent increase in Atlantic hurricane activity. *Nature*, 451, 557–560.

Seidel, W., Mosler, K. y Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 58, 3, 481–487.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639–650.

Sfikas, G., Constantinopoulos, C., Likas, A. y Galatsanos, N. (2005). An Analytic Distance Metric for Gaussian Mixture Models with Application in Image Retrieval. *Springer-Verlag Berlin Heidelberg*, 835–840.

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. Tesis Doctoral, University of Oxford.

Sudderth, E. B. (2006). *Graphical Models for Visual Object Recognition and Tracking*. Tesis Doctoral, Massachusetts Institute of Technology.

Teh, Y. W. (2010). Dirichlet processes. *In Encyclopedia of Machine Learning. Springer*, 485–527.

Tippett, M. K., Camargo, S. J. y Sobel, A. H. (2011). A Poisson Regression Index for Tropical Cyclone Genesis and the Role of Large-Scale Vorticity in Genesis. *American Meteorological Society*, 24, 2335–2357.

Titterington, D. M., Smith, A. F. M. y Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester: John Wiley and Sons.

Trenberth, K. E., Jones, P., Ambenje, P., Bojariu, R., Easterling, D., Tank, A. K., Parker, D., Rahimzadeh, F., Renwick, J., Rusticucci, M., Soden, B. y Zhai, P. (2007). Observations: Surface and Atmospheric Climate Change. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Trenberth, K. E. y Shea, D. J. (2006). Atlantic hurricanes and natural variability in 2005. *Geophysical Research*, 33, L12704–1–4.

Villarini, G., Vecchi, G. A. y Smith, J. A. (2011). U.S. Landfalling and North Atlantic Hurricanes: Statistical Modeling of Their Frequencies and Ratios. *Unprinted*, 1–72.

Walker, S. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation*, 36, 1-3, 45–54.

Walker, S. G., Damien, P., Laud, P. W. y Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal Royal Statistical*, 61, 3, 485–527.

Walsh, K. (2004). Tropical cyclones and climate change: unresolved issues. *Climate Research*, 27, 77–83.

Webster, P. J., Holland, G. J., Curry, J. A. y Chang, H. R. (2005). Changes in tropical cyclone number, duration, and intensity in warming environment. *Science*, 406, 686–688.

Werner, A. y Holbrook, N. J. (2011). A Bayesian Forecast Model of Australian Region Tropical Cyclone Formation. *Journal of Climate*, 24, 6114–6131.

Yau, C. y Holmes, C. (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, 6, 2, 329–352.

Yokoi, S. y Takayabu, Y. N. (2009). Multi-model Projection of Global Warming Impact on Tropical Cyclone Genesis Frequency over the Western North Pacific. *Journal of the Meteorological Society of Japan*, 87, 3, 525–538.

Yonekura, E. y Hall, T. M. (2011). A Statistical Model of Tropical Cyclone Tracks in the Western North Pacific with ENSO-Dependent Cyclogenesis. *Journal of Applied Meteorology and Climatology*, 50, 1725–1739.

Apéndices

Apéndice A: Distribución espacial de la génesis de los ciclones tropicales en la región del Golfo en México.

Golfo de México y Océano Pacífico.

En la Figura A.1 se observa la distribución espacial de la ciclogénesis en el Golfo de México y el Océano Pacífico.

Golfo de México.

En la Figura A.2 se observa la distribución espacial de la ciclogénesis en el Golfo de México.

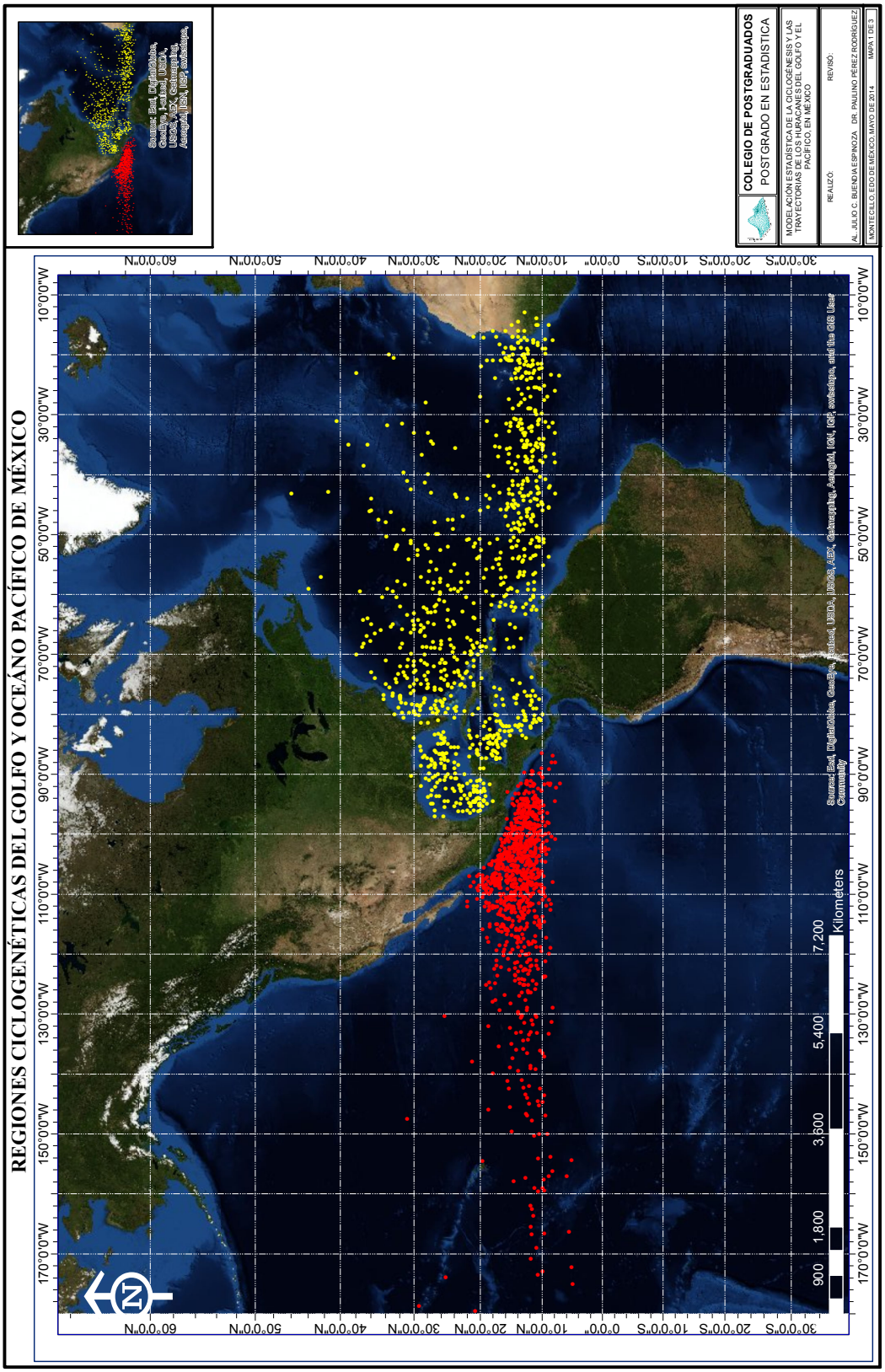


Figura A.1: Distribución espacial de la génesis de los ciclones en las regiones del Golfo y Océano Pacífico en México.

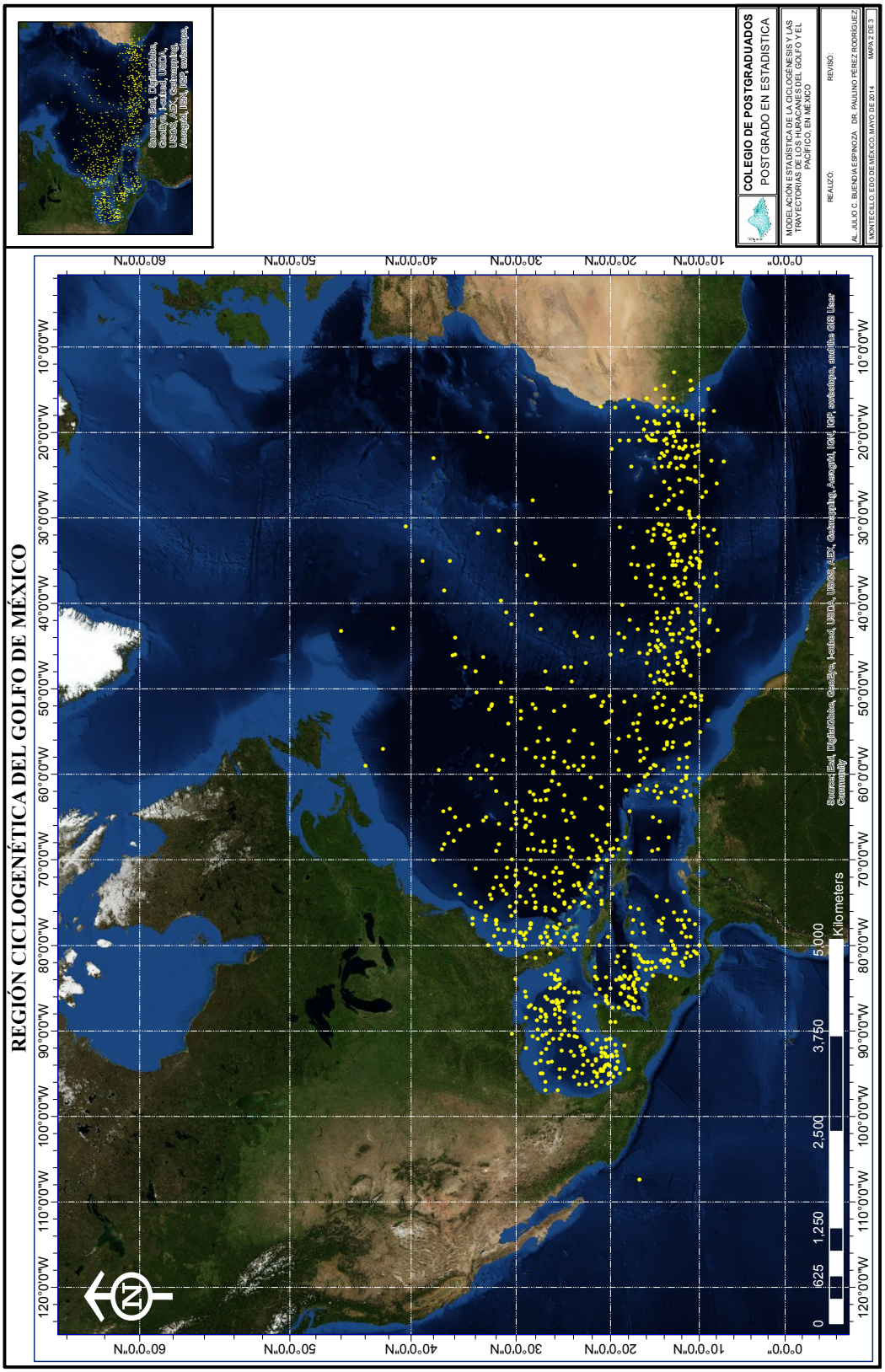


Figura A.2: Distribución espacial de la génesis de los ciclones en la región del Golfo en México.

Apéndice B: Estimación de las condicionales de μ y Σ.

La distribución a posteriori es producida por el producto de la función de verosimilitud y la a priori.

$$p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{Y}\right) = L\left(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot p\left(\boldsymbol{\mu} | \boldsymbol{\Sigma}\right) \cdot p\left(\boldsymbol{\Sigma}\right)$$

Primero se estima la verosimilitud:

$$\begin{aligned} L\left(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) &= \prod_{i=1}^{n} \left((2\pi)^{\frac{-k}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp^{\left[-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})\right]} \right) \\ &\propto \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp^{\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})\right]} \end{aligned}$$

Ahora, se expanden los cuadrados:

$$\begin{aligned} L\left(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) &= \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp^{\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i - 2\boldsymbol{y}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)\right]} \\ &= \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp^{\left[-\frac{1}{2}\left(\sum_{i=1}^{n} \boldsymbol{y}_i^T \Sigma^{-1} \boldsymbol{y}_i - n\boldsymbol{\Sigma}^{-1}\bar{y}^T\bar{y} + n\boldsymbol{\Sigma}^{-1}\bar{y}^T\bar{y} - 2n\bar{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]} \\ &= \left|\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} \exp^{\left[-\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\right)\left(\sum_{i=1}^{n} \boldsymbol{y}_i^T \boldsymbol{y}_i - \bar{y}^T\bar{y}\right) + n(\bar{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{y} - \boldsymbol{\mu})\right)\right]} \end{aligned}$$

dado que: $\sum_{i=1}^{n}\left(\boldsymbol{y}_i^T \boldsymbol{y}_i\right) - \bar{y}^T\bar{y} = \sum_{i=1}^{n}\left(\boldsymbol{y}_i - \bar{y}\right)^T\left(\boldsymbol{y}_i - \bar{y}\right) \equiv S^2$, entonces: $L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{Y}\right)$ es una función de los datos sólo a través de los estadístico suficientes de los dos componentes: $\left(\bar{y}, S^2\right)$. Otras distribuciones a prioris conjugadas para $\boldsymbol{\mu}$ y para $\boldsymbol{\Sigma}$ fueron propuestas por Robert (2001, página 189), a continuación se muestran éstas:

$$\boldsymbol{\mu} | \boldsymbol{\Sigma} \sim N_k\left(m, \frac{1}{n_0}\boldsymbol{\Sigma}\right), \qquad \boldsymbol{\Sigma}^{-1} \sim Wishart\left(\alpha, \beta\right)$$

La distribución Wishart:

$$Wishart\left(\boldsymbol{\Sigma}^{-1} | \alpha, \beta\right) = \frac{\left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{(\alpha - (k+1))}{2}}}{\Gamma_k\left(\alpha\right)\left|\beta\right|^{\frac{\alpha}{2}}} \exp^{-\mathrm{tr}\left(\frac{\beta^{-1}\boldsymbol{\Sigma}^{-1}}{2}\right)}$$

donde:

$$\Gamma_k\left(\alpha\right) = 2^{\frac{\alpha k}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^{k} \Gamma\left(\frac{\alpha + i - 1}{2}\right)$$

$2\alpha > k - 1$, β es no singular.

Esto significa que:

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Y}) &= L\left(Y|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) p\left(\boldsymbol{\mu}|\frac{\boldsymbol{\Sigma}}{n_0}, m\right) p\left(\boldsymbol{\Sigma}|\alpha, \beta\right) \\ &\propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}S^2\right)+n(\bar{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{y}-\boldsymbol{\mu})\right)\right] \\ &\quad \cdot \left|\frac{\boldsymbol{\Sigma}}{n_0}\right|^{-\frac{1}{2}} \exp\left[(\boldsymbol{\mu}-m)^T\left(\frac{1}{n_0}\boldsymbol{\Sigma}\right)^{-1}(\boldsymbol{\mu}-m)\right] \\ &\quad \cdot |\beta|^{-\frac{\alpha}{2}} \left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{(\alpha-(k+1))}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\beta^{-1}\boldsymbol{\Sigma}^{-1}\right)\right] \end{aligned}$$

Ahora, se obtiene la marginal de $\boldsymbol{\mu}$:

$$p(\boldsymbol{\mu}|\boldsymbol{\Sigma}) = \int_0^\infty p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{y}\right) d\boldsymbol{\Sigma}$$

Luego entonces:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim N_k\left(\frac{n_0 m + n\bar{y}}{n_0+n}, \frac{1}{n_0+n}\boldsymbol{\Sigma}\right)$$

El cálculo de la distribución a posteriori marginal para $\boldsymbol{\Sigma}$ es considerablemente menos complicada porque se puede una vez más utilizar la propiedad de probabilidad condicional:

$$p\left(\boldsymbol{\Sigma}|\boldsymbol{y}\right) = \frac{p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{Y}\right)}{p\left(\boldsymbol{\mu}|\boldsymbol{Y}\right)}$$

Esto significa que se puede obtener la marginal a posteriori de $\boldsymbol{\Sigma}$ dividiendo la a posteriori conjunta por la distribución marginal de $\boldsymbol{\mu}$ suponiendo que $\boldsymbol{\Sigma}$ es independiente. Esto funciona muy bien, proporcionando suficiente tamaño de la muestra:

$$\boldsymbol{\Sigma}^{-1} \sim Wishart\left(\alpha+n, \left[\beta^{-1} + S^2 + \frac{n_0 n}{n_0+n}\left(\bar{y}-m\right)\left(\bar{y}-m\right)^T\right]\right)$$

Apéndice C: Modelo de Mezclas de Procesos Dirichlet para determinar las Regiones de Ciclogénesis del Atlántico Norte e identificar sus cambios para la Fase Caliente (1951-1967) vs la Fase Fría (1971-1990).

Las temperaturas de la superficie del Atlántico Norte para 1856-1999 contienen un ciclo de 65-80 años con un rango de 4°C, referido por Kerr (2000) como Oscilación Multidecadal del Atlántico Norte (OMA). Las fases cálidas de la OMA se produjeron aproximadamente durante 1860-1880 y 1940-1960, y las fases frías durante 1905-1925 y 1970-1990 (Enfield *et al.*, 2001).

La Oscilación Multidecadal del Atlántico (OMA) se ha identificado como un patrón coherente de cambios oscilatorios del Atlántico Norte en la temperatura superficial del mar (TSM). Los índices de la OMA se han basado tradicionalmente en el promedio anual de las variaciones de la TSM en la región del Atlántico Norte. Para tener en cuenta la influencia de la reciente tendencia al calentamiento global sobre TSM en el Atlántico Norte, los índices de la OMA se han definido mediante la inclusión de TSM media global (Knudsen *et al.*, 2011).

La idea de que los cambios lentos en circulación oceánica influyen en la variabilidad del clima en el Atlántico Norte se remonta desde hace medio siglo. Actualmente, se tienen varias evidencias que sugieren que las variaciones de la TSM relacionados con la OMA conducen a patrones de precipitación y de clima en América del Norte, sequías en la región del Sahel en África, variabilidad en las precipitaciones del noreste de Brasil, y la frecuencia y la intensidad de los huracanes tropicales.

La OMA oscila entre períodos fríos y calientes de aproximadamente 30 o 35 años cada uno; sin embargo, ésta ha estado en fase de calentamiento desde alrededor de 1995, con consecuencias para el clima en el Atlántico Norte. Esto significa que deberá entrar su fase que puede continuar durante 20 o más años de enfriamiento.

En la Figura A.3 se muestran las tendencias del clima Global y del Atlántico Norte en los últimos 150 años. Cabe mencionar que el índice de la OMA de acuerdo con Trenberth y Shea (2006) se define restando las variaciones de la TSM media global de las variaciones de la TSM del Atlántico Norte. Las líneas negras gruesas con relleno en todos los paneles son el promedio corrido de cinco años.

Bajo este contexto, en este trabajo de investgación se aplicó un modelo de mezclas de Procesos Dirichlet de acuerdo con la sección 6.2, con el propósito por una parte para determinar el número de regiones ciclogenéticas y por otra para determinar los cambios temporales y espaciales de los centroides de dichas regiones de génesis en la cuenca oceánica del Atlántico Norte, para la Fase Caliente (1951-1967) versus la Fase Fría (1971-1990). Se utilizaron los datos de *Longitud* y *Latitud* de los ciclones tropicales de la base de datos de las “mejores

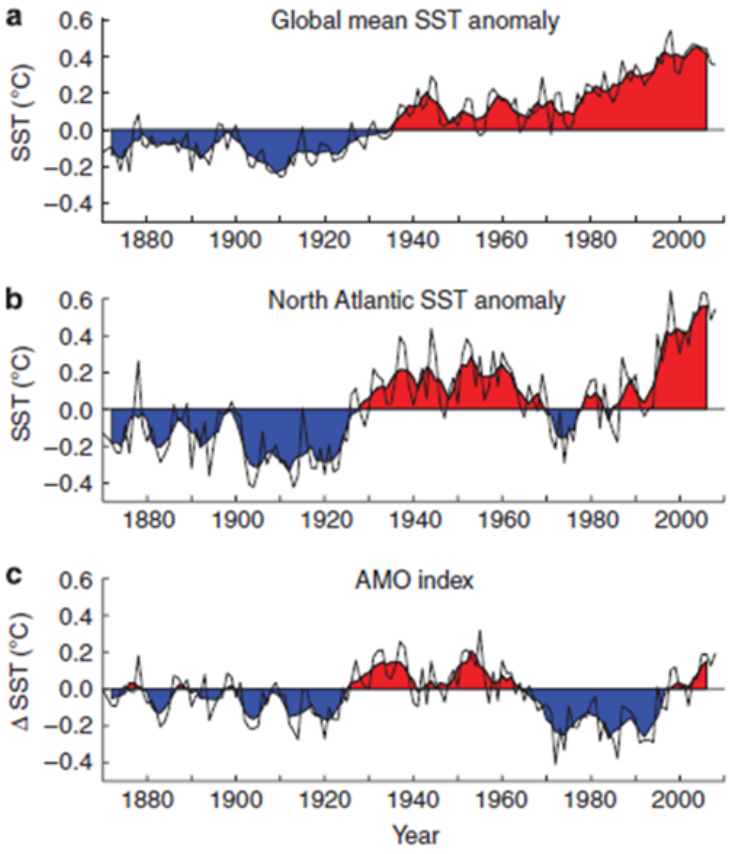
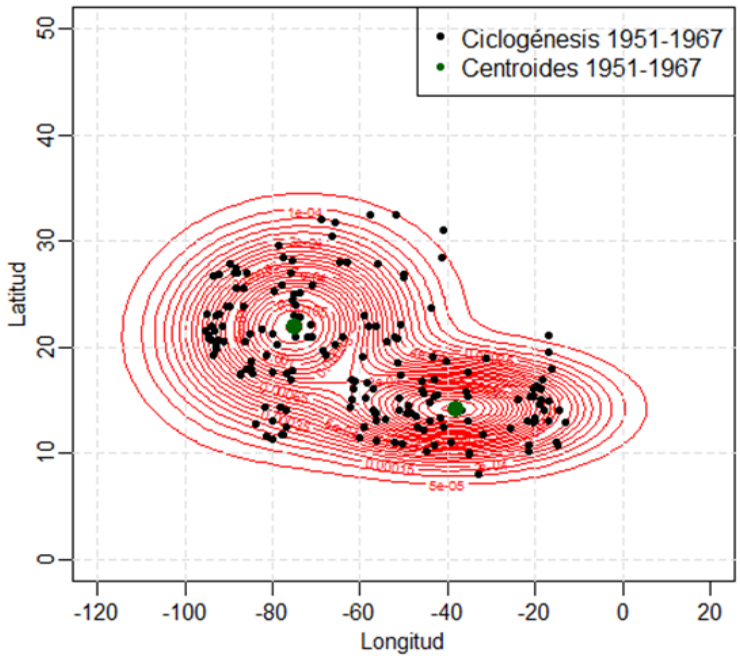


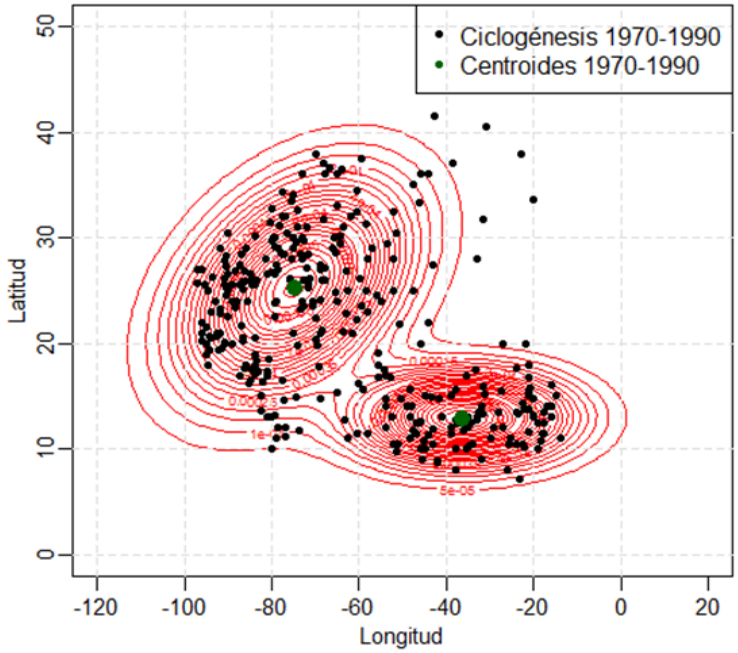
Figura A.3: En el inciso a) se muestran las variaciones globales anuales de la temperatura media de la superficie del mar (TSM) de la HadISST para el período 1870-2008 (línea delgada de color negro). En el inciso b) se muestran las variaciones medias anuales de la TSM del Atlántico Norte para el período 1870-2008 (línea delgada de color negro). Y en el inciso c) se muestra el índice de la OMA para el período 1870-2008. Fuente: Knudsen *et al.* (2011).

trayectorias" o IBTrACS (por siglas en inglés: International Best Track Archive for Climate Stewardship, puede consultarse en: `https://www.ncdc.noaa.gov/ibtracs/index.php?name=ibtracs-data-access`).

Los resultados estadísticos obtenidos para el número de grupos mediante los *DPMM* muestran que hay dos ($K = 2$) sitios de génesis en cada una de las dos fases en la cuenca oceánica del Atlántico Norte, ver Figura A.4. Además, los centroides de dichos sitios se han despalzado de una fase a otra, lo cual coincide con las proyecciones hechas por Mori *et al.* (2013) sobre la génesis de los ciclones tropicales para finales del siglo XXI. Particularmente, comparando el centroide que aparece en color azul en la Fase caliente vs la Fase Fría, éste se ha desplazado hacia el nor-este (es decir, del Gólfo hacia la Costa-Este), A.5. Análogamente, el centroide que aparece en color rojo se ha desplazado hacia el sur-este (es decir, moviéndose dentro del mismo Atlántico Tropical).



(a) Fase Caliente 1951-1967.



(b) Fase Fría 1971-1990.

Figura A.4: (a) Muestra la gráfica de contornos de la longitud y latitud para la Fase Caliente 1951-1967 y, (b) Longitud y la latitud para la Fase Fría 1971-1990. Fuente: Elaboración Propia.

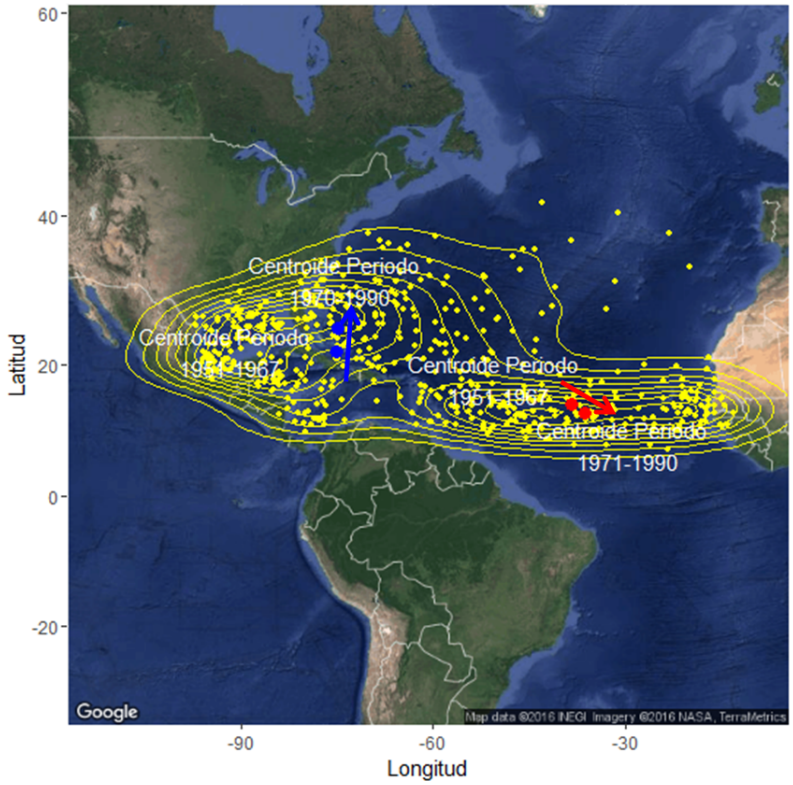


Figura A.5: Distribución espacial de los puntos de ubicación de ocurrencia de los ciclones tropicales en la región del Atlántico Norte para la Fase Caliente (1951-1967) vs la Fase Fría (1971-1990). Fuente: Elaboración Propia con base en los datos del IBTrACS.