

**MODELOS OCULTOS DE MARKOV PARA EL ANÁLISIS DE LA
PRECIPITACIÓN DIARIA EN UNA ESTACIÓN CLIMATOLÓGICA
DEL VALLE DEL CAUCA**

**HAROLD ARMANDO ACHICANOY ESTRELLA
MICHAEL STEVEN LÓPEZ CORREA**

**UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍA
ESCUELA DE ESTADÍSTICA
SANTIAGO DE CALI
2013**

**MODELOS OCULTOS DE MARKOV PARA EL ANÁLISIS DE LA
PRECIPITACIÓN DIARIA EN UNA ESTACIÓN CLIMATOLÓGICA
DEL VALLE DEL CAUCA**

**HAROLD ARMANDO ACHICANOY ESTRELLA
MICHAEL STEVEN LÓPEZ CORREA**

**Trabajo de Grado presentado como requisito parcial para optar al título
de ESTADÍSTICO**

**Director
GABRIEL CONDE ARANGO
Estadístico MSc.**

**Codirector
MERCEDES ANDRADE BEJARANO
Estadística Ph.D.**

**UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍA
ESCUELA DE ESTADÍSTICA
SANTIAGO DE CALI
2013**

Resumen

En diversas áreas de la actividad humana resulta de gran interés analizar y describir el comportamiento de elementos climáticos que ayuden a la toma de decisiones para así obtener los beneficios que el clima proporciona. Bajo esta premisa, el objetivo de este proyecto consistió en analizar el comportamiento de la precipitación diaria durante la primera temporada lluviosa del año (periodo comprendido entre los meses de marzo a mayo) en una estación climatológica ubicada en el valle geográfico del Río Cauca (departamento del Valle del Cauca) entre los años 1994-2011, mediante la determinación de patrones latentes que inciden en la ocurrencia de dicho evento climático, usando un modelo oculto de Markov. Para este propósito, se discretizaron los registros de precipitación definiendo días húmedos y secos; posteriormente se ajustaron seis modelos diferentes, a partir de los cuales, se identificó que el mejor modelo oculto de Markov presentó dos estados. Entre los principales hallazgos se destaca un estado oculto con alta probabilidad de lluvia (0.766), teniendo en cuenta que aproximadamente el 88 % del total de registros en dicho estado presentaron niveles de precipitación superiores a 0.1 mm, mientras el otro estado subyacente se caracterizó por una baja probabilidad de precipitación (0.127), determinando condiciones secas y una baja frecuencia en términos del número de días de permanencia sobre el mismo. Estos resultados indicaron que los estados ocultos se encuentran asociados a patrones climáticos que generan fluctuaciones de la precipitación a escala intraestacional desarrollados dentro del periodo de estudio.

Palabras claves: escalas de variabilidad climática, precipitación, proceso observable, proceso oculto, temporadas lluviosas.

Tabla de Contenido

Resumen	I
1. Introducción	1
2. Formulación del Problema	3
2.1. Planteamiento	3
2.2. Justificación	4
2.3. Objetivos	5
2.3.1. Objetivo General	5
2.3.2. Objetivos Específicos	5
3. Antecedentes	6
4. Marco Teórico	8
4.1. Marco Teórico Conceptual	8
4.1.1. Tiempo y clima	8
4.1.2. Variabilidad climática	9
4.1.3. Factores que inciden sobre el clima de la región	10
4.1.3.1. Zona de Convergencia Intertropical	10
4.1.3.2. Fenómeno El Niño/Oscilación del Sur	11
4.1.4. Precipitación atmosférica	12
4.2. Marco Teórico Estadístico	13
4.2.1. Procesos estocásticos	13
4.2.2. Modelos de Markov	14
4.2.2.1. Ecuaciones de Chapman-Kolmogorov	15
4.2.2.2. Distribuciones marginales	16
4.2.2.3. Propiedades	17
4.2.3. Modelos ocultos de Markov	24
4.2.3.1. Evaluación	27
4.2.3.2. Decodificación	30
4.2.3.3. Entrenamiento	32
4.2.3.4. Consideraciones a la hora de ajustar un HMM	35
4.2.3.5. Selección del modelo	36
4.2.3.6. Inferencia del modelo	38
4.2.4. Estimación de registros faltantes: método de Cressman	39

5. Metodología	40
5.1. Revisión de literatura	40
5.2. Datos del estudio	40
5.3. Validación e imputación de registros faltantes	41
5.4. Transformación de datos	42
5.5. Análisis exploratorio de datos	42
5.6. Ajuste del modelo	42
5.6.1. Supuestos	42
5.6.2. Elementos del modelo	43
5.6.3. Proceso oculto	43
5.6.4. Proceso observable	44
5.6.5. Valores iniciales de los parámetros	44
5.6.6. Estimación de parámetros	45
5.6.7. Selección del modelo	45
5.6.8. Estimación de la secuencia de estados ocultos	46
5.6.9. Caracterización del modelo	46
6. Resultados	47
6.1. Imputación de registros faltantes	47
6.2. Análisis exploratorio de datos	48
6.3. Caracterización de la serie de precipitación a través del HMM	52
6.3.1. Dinámica del proceso oculto	54
6.3.2. Descripción y caracterización de los estados ocultos	57
7. Conclusiones	59
8. Recomendaciones	60
Bibliografía	62

Índice de Tablas

4.1. Clasificación de las fases del fenómeno ENSO para el periodo 1993-2011 según índices macroclimáticos. Fuente, IDEAM (2010)	12
5.1. Estaciones climatológicas auxiliares empleadas en el análisis	41
6.1. Indicadores descriptivos asociados a la distribución de precipitación diaria primera temporada lluviosa	50
6.2. Comparación de modelos mediante criterios de información	52
6.3. Indicadores descriptivos de precipitación para los estados identificados . . .	53
6.4. Coeficientes estimados del modelo ajustado	54
6.5. Indicadores descriptivos del tiempo de permanencia por estado oculto . . .	56
6.6. Indicadores descriptivos por estado oculto	58

Índice de Figuras

4.1. Fotografía satelital ZCIT	10
4.2. Representación esquemática de un HMM	25
6.1. Distribución de datos faltantes diarios para el periodo 1994-2011	47
6.2. Ciclo medio estacional de la precipitación a lo largo del año	48
6.3. Serie de precipitación diaria durante la primera temporada lluviosa del año, periodo 1994-2011	49
6.4. Distribución de precipitación diaria primera temporada lluviosa	50
6.5. Distribución acumulada de la precipitación diaria durante la primera tem- porada lluviosa	51
6.6. Selección del modelo por criterios AIC y BIC	53
6.7. Distribución del tiempo de permanencia por estado oculto	55
6.8. Serie de precipitación diaria durante la primera temporada lluviosa del año junto con la secuencia más probable de estados ocultos	57
6.9. Comparación de la distribución acumulada de precipitación diaria por es- tados ocultos	58

Capítulo 1

Introducción

Colombia es un país diverso, esto en términos culturales, ecosistémicos, geográficos y climáticos. Su posición estratégica en el trópico, lo hace partícipe de las mayores proporciones de energía que el Sol le transmite a la tierra, constituyéndose de esta manera en un epicentro de la circulación atmosférica (IDEAM 2005); sin embargo, su clima a nivel general es heterogéneo debido a las variadas condiciones geográficas del país. En particular, el departamento del Valle del Cauca por su cercanía al océano Pacífico es una región que sufre con mayor intensidad los efectos derivados de eventos océano-atmosféricos de compleja compresión, tales como el fenómeno el Niño Oscilación del Sur (ENSO) en sus diferentes fases (Carvajal et al. 1998), repercutiendo en una gran cantidad de sectores socioeconómicos a nivel departamental.

Un ejemplo de esto es el sector agrícola que constituye uno de los sectores más representativos en la economía de la región, pero que debido a su fuerte dependencia del estado del tiempo experimenta los efectos derivados de la variabilidad climática contribuyendo positiva o negativamente en su desarrollo; de aquí, la importancia de entender y caracterizar el comportamiento de elementos climáticos, con la finalidad de aportar en la toma asertiva de decisiones en la planeación de actividades del sector agrícola (Carvajal et al. 1998).

El clima se encuentra caracterizado por las condiciones predominantes en la atmósfera y en general se describe a partir de variables atmosféricas, entre las que se destacan: la temperatura, la humedad, la presión atmosférica, la velocidad y dirección del viento, y la precipitación. Esta última constituye uno de los elementos climatológicos de mayor importancia para el desarrollo de diversas actividades socioeconómicas, entre ellas la agricultura ya que es fuente del recurso hídrico indispensable para el desarrollo de siembras (IDEAM 2001).

La precipitación atmosférica varía de manera natural en periodos cortos de tiempo, caracterizando fluctuaciones a diferentes escalas entre las que se destacan: la interanual, cuyas alteraciones se presentan de año a año como ocurre con el fenómeno ENSO; la estacional, en la cual se enmarcan variaciones del clima a nivel mensual y sirve de base para la determinación del ciclo anual de precipitación, y la intraestacional, donde se presentan fluctuaciones en periodos cortos de tiempo que van de días a unos cuantos meses

(Martinez 2011).

El estudio de la precipitación atmosférica constituye un pilar importante en la toma de decisiones a nivel económico, político y social, esto debido a que los efectos derivados de la variabilidad inherente de dicho elemento climático, afectan las diversas esferas sociales en las que nos desenvolvemos (ACR 2009). El comportamiento bimodal del ciclo anual de la precipitación en la región plana del Valle del Cauca se traduce en dos temporadas lluviosas que tienen cabida durante los trimestres marzo-mayo y septiembre-noviembre de cada año. Dicho comportamiento, se enmarca en la variabilidad a escala estacional y constituye la unidad temporal de interés en el estudio, seleccionando únicamente la primera temporada lluviosa del año, esto con el fin de identificar los posibles patrones generadores de la precipitación en dicha temporada, lo cual constituiría una información de gran importancia que puede ser utilizada como insumo en actividades de planeación en el sector agrícola.

Por los motivos expuestos anteriormente, el objetivo de esta investigación consiste en analizar el comportamiento de la precipitación atmosférica diaria durante la primera temporada lluviosa del año (trimestre marzo-mayo) para el periodo 1994-2011, mediante un modelo oculto de Markov, en busca de identificar los posibles patrones generadores de la precipitación. Para este fin, se clasifican los registros de precipitación diarios para la temporada de interés en días húmedos y días secos, definiendo como umbral el valor 0.1 mm. Esta discretización de los registros de precipitación permitió modelar el comportamiento de la serie temporal mediante un modelo oculto de Markov con observaciones emitidas por distribuciones de probabilidad Bernoulli por estado oculto.

De esta manera, a partir del ajuste del modelo se logró identificar dos estados ocultos relacionados con patrones de variación a escala intraestacional, uno de ellos asociado a condiciones húmedas con alta probabilidad de lluvia, mientras el otro se relaciona con las condiciones secas dentro de la temporada en estudio y una baja frecuencia en función del número de días en que se permanece en dicho estado. Estos hechos permitieron describir e interpretar los estados ocultos como indicadores descriptivos utilizados para caracterizar los patrones que inciden sobre la generación de precipitación durante la primera temporada lluviosa del año.

El orden con que se abordan los temas del presente trabajo es el siguiente: el Capítulo 2 presenta la formulación del problema que se busca resolver exponiendo: el problema operativo, su respectiva justificación y los objetivos que se pretende alcanzar con la investigación. El tercer capítulo describe los principales antecedentes que se hallaron en la literatura en los cuales se destaca la utilización de los modelos ocultos de Markov en la modelación de la precipitación diaria y los principales resultados alcanzados. A continuación, se presenta el Marco Teórico donde se exponen los principales conceptos que servirán de soporte para la ejecución del análisis. En los Capítulos 4 y 5, se presenta la metodología con la cual se guió la investigación y los principales resultados obtenidos, respectivamente. Finalmente se exponen las conclusiones y recomendaciones con las que se pretende resaltar los principales hallazgos obtenidos por medio de la investigación y brindar un amplio panorama para la realización de investigaciones futuras.

Capítulo 2

Formulación del Problema

2.1. Planteamiento

Desde la antigüedad el hombre ha buscado predecir el estado del tiempo (Lezaun 2002). La incidencia del clima en el desarrollo de actividades socioeconómicas como la producción alimentaria, la generación de energía, el turismo y el transporte, entre otras directamente relacionadas con el estado del tiempo es variada; sus efectos pueden llegar a contribuir en la mejora de las mismas o por el contrario repercutir negativamente en su desarrollo (Mesa et al. 1997). Desde este punto de vista, el propósito fundamental de efectuar predicciones climáticas se ve orientado al desarrollo de actividades de planeación que permitan contrarrestar efectos perjudiciales, aprovechando los beneficios que el clima proporciona (OMM 2001).

La dinámica de la atmósfera y su interacción con los componentes del sistema climático, entre ellos la superficie terrestre y los océanos, genera variaciones espacio-temporales de elementos climáticos como la precipitación, determinando condiciones lluviosas o secas en una región y tiempo específicos (IDEAM 2005). En particular, la región del valle geográfico del río Cauca en el departamento del Valle presenta un régimen de precipitación bimodal¹ a lo largo del año con dos trimestres secos (junio-agosto; diciembre-febrero) y dos lluviosos (marzo-mayo, septiembre-noviembre) como resultado de la actividad de la Zona de Convergencia Intertropical y la influencia de los vientos alisios (Carvajal et al. 1998), (IDEAM 2001), (Puertas et al. 2011).

El conocimiento de este comportamiento bimodal sobre la región sirve como base para direccionar las decisiones a tomar sobre la planeación de siembras y cosechas en el sector agrícola. Sin embargo, las condiciones climáticas del departamento por su ubicación en la Costa Pacífica de Colombia se ven influenciadas por el desarrollo de fenómenos océano-atmosféricos como el fenómeno El Niño/Oscilación del Sur (ENSO, por sus siglas en inglés), cuya fase cálida conocida comúnmente como El Niño trae consigo incrementos de la temperatura superficial del mar, produciendo alteraciones en los patrones de circulación del viento, presiones atmosféricas superficiales y precipitaciones sobre el océano Pacífico (IDEAM 2005), perturbando significativamente el cronograma de siembras, ciclo de vida de plagas y enfermedades, reducción de rendimiento y productividad, entre otras

¹Dos temporadas con altos registros de precipitación durante el año.

consecuencias (IDEAM 2010), (MinAgricultura 2012), (ICA 2012).

El desarrollo del fenómeno ENSO se produce a escala interanual de forma cuasi-periódica y por tanto dependiendo de la temporada en que se desarrolle durante el año, intensifica y/o aplaca los eventos climáticos observados durante dichas temporadas, generando alteraciones sobre las condiciones normales del ciclo. La comprensión de estas particularidades, sin contar con el efecto producido por actividades de origen antropogénico que inciden en alteraciones de la variabilidad natural del clima (Vide 2008), permite hacerse a la idea de la complejidad que subyace al comportamiento temporal de un elemento climático tan esencial como lo es la precipitación en una región propicia para el desarrollo de actividades productivas como la agricultura.

Bajo el conocimiento de estas premisas, el aporte desde el campo estadístico se ve orientado a la modelación del comportamiento temporal de factores climáticos (precipitación, temperatura, radiación solar, etc.) mediante la construcción de modelos pertinentes que permitan describir de la mejor manera, la naturaleza y las características propias de dichos elementos; en este sentido, el problema a trabajar consiste en modelar el comportamiento de la precipitación atmosférica diaria durante la primera temporada lluviosa del año (periodo comprendido entre marzo y mayo) en la estación climatológica² San Marcos ubicada en el centro del departamento del Valle del Cauca, mediante la construcción de un modelo oculto de Markov homogéneo³, utilizando información proporcionada por el Centro de Investigación de la Caña de Azúcar (Cenicaña) para el marco temporal 1994-2011.

2.2. Justificación

Gran variedad de sectores pero en especial la producción agrícola son muy dependientes del estado del tiempo, principalmente en la disponibilidad del recurso hídrico, y se ven afectados negativamente por desastres relacionados con condiciones climáticas cambiantes. Estas variaciones generan desequilibrios en las precipitaciones y las necesidades de los cultivos, ocasionando el fracaso en las cosechas, inseguridad alimenticia, pérdidas de propiedades y vidas, migraciones masivas, y un decrecimiento de la economía nacional (Sivakumar 2004). En este sentido, la comprensión de la variabilidad climática es fundamental para prevenir y mitigar los efectos adversos ocasionados por la ocurrencia de eventos extremos como sequías e inundaciones, esto con el fin de obtener los mayores beneficios en la planeación de actividades tales como la producción de alimentos, ligada de forma esencial a los factores climáticos y fundamentalmente a la precipitación, a partir de una correcta predicción de variables meteorológicas (Mesa et al. 1997).

La ubicación de Colombia en la zona intertropical, hace que su territorio sea partícipe de las mayores proporciones de energía que el Sol le transfiere a la Tierra, esto hace que la dinámica de los factores climáticos se vea afectada por el desarrollo de procesos

²Se hará uso de los términos climatológica o meteorológica, indistintamente.

³El término **homogéneo** hace alusión a un modelo en el cual no se incluyen covariables en el ajuste del mismo; de igual modo, se referirá indistintamente a los términos **modelo oculto de Markov** y **modelo oculto de Markov homogéneo**.

y fenómenos particulares que actúan como elementos importantes en la determinación de patrones característicos a diferentes escalas de variabilidad sobre el territorio nacional (IDEAM 2001). La necesidad de ajustar un modelo que permita capturar la incidencia de los principales procesos que afectan la dinámica del clima y en especial de la precipitación atmosférica durante la primera temporada lluviosa del año en la zona plana del departamento del Valle radica en el hecho de extender el conocimiento sobre el clima, permitiendo mejorar las capacidades de predicción para facilitar el desarrollo de información climática relevante para aplicaciones en agricultura.

La utilidad de los modelos ocultos de Markov en la modelación de la precipitación atmosférica a escala diaria ha mostrado resultados importantes, tal y como se evidencia en los estudios de Hughes, Guttorp & Charles (1999), Robertson, Kirshner & Smyth (2003) y (Tan et al. 2012). Asumir que la dinámica del clima se determina a partir de un conjunto reducido de estados ocultos, permite relacionar la ocurrencia de procesos locales de precipitación con la dinámica de los patrones atmosféricos a diferentes escalas de variabilidad (Robertson et al. 2003). En este sentido, los modelos ocultos de Markov constituyen una alternativa de modelación de series temporales que combina el uso de distribuciones paramétricas, con la dinámica de un modelo de Markov de primer orden, permitiendo manejar la posible correlación serial entre los datos observados. De entrada, los estados del modelo de Markov no se conocen y se identifican a partir de los datos observados; este hecho permite señalar patrones o cambios asociados a factores que inciden en la evolución de las series en estudio, lo que constituye un referente importante a la hora de estudiar series de tiempo asociadas a variables climáticas.

2.3. Objetivos

2.3.1. Objetivo General

Analizar el comportamiento de la precipitación atmosférica diaria durante la primera temporada lluviosa del año en la estación meteorológica San Marcos mediante la construcción de un modelo oculto de Markov homogéneo para el periodo 1994-2011.

2.3.2. Objetivos Específicos

- Determinar los estados ocultos del modelo que inciden sobre la dinámica de la precipitación durante la primera temporada lluviosa del año en la región de estudio.
- Estimar la secuencia más probable de estados ocultos que generan los registros de precipitación observados durante el periodo en estudio.

Capítulo 3

Antecedentes

Son cuantiosas las investigaciones donde se han usado los modelos ocultos de Markov para el análisis de secuencias de datos generados de acuerdo al tiempo u otra variable de indexación bien sean del orden univariado o multivariado, mostrando la gran versatilidad de los mismos para el manejo de información de diversas áreas del conocimiento tales como el reconocimiento automático de voz, bioinformática, psicología, economía y recientemente al análisis de datos ambientales (Visser 2011). A continuación se presenta la descripción de algunos estudios donde se muestra la aplicación de dichos modelos para el análisis de la precipitación atmosférica a escala diaria.

- **Hughes et al. (1999)**, emplearon un modelo oculto de Markov no homogéneo (NHMM, es decir, empleando covariables) con el propósito de estudiar la relación entre la precipitación diaria de un conjunto de 30 estaciones pluviométricas con patrones de circulación atmosférica a gran escala que pueden clasificarse en un número reducido de estados discretos no observables.

La información utilizada en el estudio corresponde a registros de precipitación diaria en la temporada de invierno (mayo-octubre) durante los años 1978-1992 en el suroeste de Australia, se utilizó además un conjunto de 24 covariables asociadas a la circulación atmosférica, entre ellas la presión atmosférica media a nivel del mar, temperatura del aire, componentes de velocidad del viento, gradientes de presión y altura, esto con el fin de estimar la influencia que tienen dichas covariables en las probabilidades de transición entre los estados ocultos.

La relación espacial entre las estaciones utilizadas para el ajuste del modelo se capturó mediante un modelo auto-logístico multivariado. Mientras la estimación de los parámetros del modelo se efectuó bajo una modificación del algoritmo iterativo de Esperanza-Maximización (algoritmo EM). Los resultados más destacados indican que el modelo que presenta un mejor ajuste, cuenta con seis estados ocultos que se relacionan con las covariables: presión atmosférica media a nivel del mar, el gradiente norte-sur en la presión a nivel del mar y el gradiente este-oeste de altura geopotencial a 850 mb, considerando la dependencia espacial como una función de la distancia y la dirección entre las estaciones estimada a través de un modelo no lineal.

Entre los estados identificados, se destaca el *Estado 2* donde se presentan altas probabilidades de precipitación en todas las estaciones indicando un fuerte frente

frío que cruza la zona de estudio, el *Estado 5* con bajas probabilidades de lluvia corresponde a un sistema dominante de alta presión en la Gran Bahía Australiana y los cuatro estados restantes están caracterizados por lluvias en áreas particulares de la región.

- **Robertson et al. (2003)**, el primer objetivo de este trabajo fue examinar la distribución de probabilidad de la ocurrencia de precipitación diaria en escalas temporales intraestacional e interanual, para identificar las relaciones con los grandes patrones de circulación atmosférica. Se utilizó un modelo oculto de Markov para describir la ocurrencia de precipitación diaria mediante información obtenida en 10 estaciones del estado de Ceará en el noreste de Brasil durante la temporada lluviosa de febrero-abril en el periodo anual (1975-2002).

Se identifican cuatro estados ocultos, donde dos de ellos muestran estar asociados con los grandes desplazamientos de la Zona de Confluencia Intertropical ZCIT, el fenómeno El Niño/Oscilación del Sur ENSO y la Oscilación del Atlántico Norte, mientras un segundo par de estados representan las condiciones húmedas y secas en determinado tiempo para cada estación.

Con el objetivo de reducir la escala de ocurrencia diaria de precipitación se ajusta un modelo oculto de Markov no homogéneo el cual proporciona una herramienta útil para entender las estadísticas de ocurrencia de precipitación diaria en términos de los patrones atmosféricos a gran escala, de este modo se muestra una evolución en el tiempo de las probabilidades de transición entre estados ocultos de acuerdo a un conjunto de elementos atmosféricas que funcionan como variables predictoras.

- **(Tan et al. 2012)**, ajustaron un modelo oculto de Markov homogéneo (HMM) para los registros de precipitación diaria en la región de Malasia Peninsular con base en un conjunto de 40 estaciones para las temporadas del monzón del noreste durante los años 1975 - 2008, siendo esta temporada de gran interés por la frecuencia e intensidad de las lluvias producidas.

Para el ajuste del modelo, se discretizan los registros de precipitación como días lluviosos o secos, tomando lluvias por debajo/por encima de 0.1 mm; los autores asumen que la precipitación producida en un día está influenciada por un número reducido de estados ocultos que actúan como un generador estocástico de las cantidades de lluvia, también se asume que los registros de lluvia en cada una de las estaciones se producen de manera independiente con el propósito de facilitar la estimación de los parámetros del modelo.

Entre los principales resultados obtenidos, se identifica un modelo con 6 estados ocultos relacionados con patrones de circulación atmosférica en la región de estudio, dos de ellos están relacionados con las condiciones secas y húmedas en todas las estaciones, mientras otros dos estados describen las anomalías de los gradientes este-oeste.

Capítulo 4

Marco Teórico

A continuación se presenta una descripción de los elementos conceptuales que orientan y soportan el desarrollo del presente estudio, cuyo objetivo general está centrado en analizar el comportamiento de la precipitación atmosférica a escala diaria durante la primera temporada lluviosa del año. El capítulo se encuentra dividido en dos secciones principales: la primera hace referencia a la terminología y los conceptos asociados a la caracterización del estado del tiempo y el clima centrándose principalmente en el factor climático, *precipitación*; mientras la segunda presenta la teoría estadística de los modelos ocultos de Markov, bajo la cual se guía el desarrollo de la investigación.

4.1. Marco Teórico Conceptual

4.1.1. Tiempo y clima

La composición de la atmósfera¹ y los procesos que en ella se desarrollan, generan variaciones espacio-temporales de elementos climáticos² como la precipitación y la temperatura, los cuales producen determinadas condiciones denominadas estados del tiempo (Montealegre 2009). El tiempo corresponde al estado de la atmósfera en un lugar y momento determinados, representado por el conjunto de elementos climáticos, mientras el clima es el conjunto fluctuante de las condiciones atmosféricas, caracterizado por los estados y evoluciones del tiempo en un lugar determinado; este depende del equilibrio radiativo que está controlado por los denominados factores forzantes (factores de forzamiento radiativo: radiación solar, efecto invernadero, entre otros), factores determinantes (condiciones físicas y geográficas: altitud, latitud, distancia al mar, entre otros) y por la interacción entre los diferentes componentes del sistema climático: la atmósfera, la superficie terrestre, los océanos, las áreas terrestres cubiertas de hielo, la biosfera y la antropósfera (IDEAM 2005).

El comportamiento de las variables climáticas en la región del valle geográfico del río Cauca se ve influenciado principalmente por la ocurrencia del fenómeno El Niño/Oscilación del Sur así como por la dinámica de la Zona de Convergencia Intertropical que origina un

¹Capa gaseosa que rodea la Tierra, fundamental para la existencia de todo ser vivo, regula la entrada y salida de energía solar al planeta.

²Condición de la atmósfera, cuyo conjunto define el estado físico del tiempo o del clima en un lugar determinado; entre ellos se destacan: la precipitación, la humedad, la temperatura del aire y la presión atmosférica.

régimen de precipitación bimodal en el año. La ocurrencia de estos dos eventos característicos se presenta a diferentes escalas de tiempo, influyendo en la variabilidad climática de la zona. La definición del concepto de variabilidad climática se presenta en el numeral siguiente.

4.1.2. Variabilidad climática

Las fluctuaciones observadas en el clima en periodos de tiempo parcialmente cortos, originadas por modificaciones en la interacción entre los distintos componentes del sistema climático y por cambios en los factores radiativos forzantes, constituyen la variabilidad climática. Para definir y comparar el clima generalmente se utiliza el valor promedio de una serie continua de mediciones de una variable climatológica durante un período de por lo menos 30 años conocido con el nombre de Normal Climatológica o valor normal (IDEAM 2005). De esta manera a la secuencia de oscilaciones alrededor de los valores normales, se le conoce como variabilidad climática y su valoración se logra mediante la determinación de las anomalías, definidas como la diferencia entre el valor registrado de la variable y su correspondiente valor normal (Montealegre & Pabon 2000).

Naturalmente las variaciones que presenta el clima se enmarcan dentro de los componentes espacial y temporal, sin embargo al restringir la variación en la componente temporal, las escalas de mayor trascendencia en la determinación y modulación de procesos atmosféricos son:

- **Interdecadal**, escala en donde se manifiestan las fluctuaciones del clima a nivel de décadas, por tanto su variación es menos pronunciada.
- **Interanual**, esta corresponde a las variaciones que se presentan en las variables climatológicas de año en año. La variabilidad climática, dentro de esta escala, podría estar relacionada con alteraciones en el balance global de radiación.
- **Estacional**, en esta escala se enmarca la fluctuación del clima a nivel mensual, esta a su vez sirve de base para la determinación del ciclo anual; particularmente en países ubicados en latitudes tropicales como Colombia se produce la alternancia entre temporadas lluviosas y temporadas secas.
- **Intraestacional**, en la escala Intraestacional la variabilidad es menos notoria y menos estudiada, en ella se presentan oscilaciones que determinan las condiciones de tiempo durante decenas de días o de uno a dos meses.

La variabilidad climática en el departamento del Valle, y particularmente hablando de la precipitación como elemento climático en estudio, se ve alterada principalmente por fenómenos característicos que se presentan a diferentes escalas de variabilidad, entre ellos se destaca la actividad de la Zona de Convergencia Intertropical (ZCIT) a escala estacional y el fenómeno El Niño/Oscilación del Sur (ENSO) cuya incidencia se presenta a escala interanual. Dichos fenómenos serán descritos a continuación.

4.1.3. Factores que inciden sobre el clima de la región

4.1.3.1. Zona de Convergencia Intertropical

La Zona de Convergencia Intertropical (ZCIT) es una franja ubicada cerca de la línea ecuatorial, donde confluyen los vientos alisios que provienen del noreste y del sureste. Debido a las altas temperaturas las masas de aire son forzadas a ascender y junto a la interacción de otros factores, favorecen la formación abundante de nubosidad en la zona ecuatorial (ver Figura 4.1), lo que trae consigo fuertes precipitaciones. La localización de esta región varía latitudinalmente³ con el ciclo estacional siguiendo la posición del Sol con un rezago aproximado de dos meses, lo que se refleja en estados del tiempo húmedo y seco, en lugar de estaciones frías y cálidas como en las altas latitudes. La intensidad de la ZCIT es muy variable, tanto en el espacio como en el tiempo (NASA 2000).

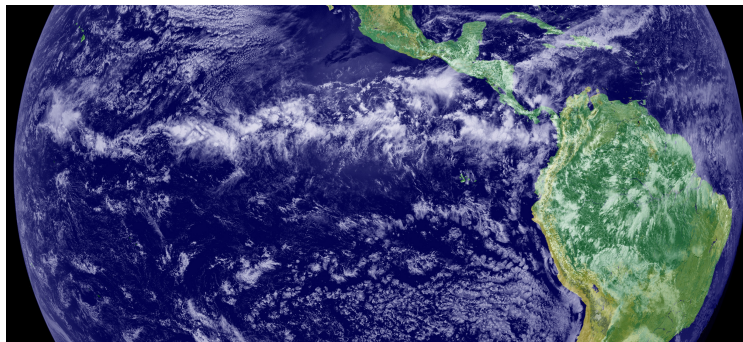


Figura 4.1: Fotografía satelital ZCIT

El desplazamiento de la ZCIT sobre el territorio colombiano determina las temporadas lluviosas en el país (Leon, Zea & Eslava 2000). Sobre Colombia y regiones vecinas, la ZCIT tiende a desplazarse de manera diferente en el océano y en la zona continental. De esta manera, entre enero y febrero en el segmento continental la ZCIT aparece fraccionado y se ubica entre los 5 y 10 grados de latitud sur, mientras que entre marzo y abril la rama continental se conecta con el segmento del océano Atlántico formando un único sistema que se localiza entre los 5 grados de latitud sur y 1 grado de latitud norte al Oriente del país.

Por otro lado, entre los meses de junio y agosto el segmento continental presenta una inclinación Suroeste-Noreste sobre el Oriente del territorio nacional, desplazándose también hacia el norte y pasando del Ecuador a los 8 grados de latitud norte. Luego, entre septiembre y noviembre el segmento del Pacífico comienza su desplazamiento hacia el sur y se registran posiciones desde 11 a 7 grados de latitud norte; la rama continental también inicia su recorrido hacia el sur, moviéndose de los 8 grados de latitud norte al Ecuador sobre la Orinoquía y Amazonía.

³La latitud mide la distancia angular entre la línea del Ecuador y cualquier punto sobre la Tierra ubicado al norte o al sur de la misma, se encuentra expresada en grados sexagesimales.

4.1.3.2. Fenómeno El Niño/Oscilación del Sur

El Niño/Oscilación del Sur (ENSO), corresponde a un evento climático natural que hace referencia a una fluctuación periódica de la temperatura superficial del mar (El Niño) y la presión del aire (Oscilación del Sur), cuyo desarrollo se presenta en el Océano Pacífico ecuatorial central con una duración promedio de doce meses y una recurrencia de 2 a 7 años (IDEAM 2005). La Oscilación del Sur describe una variación bimodal en la presión atmosférica⁴ a nivel del mar, entre las estaciones de observación en Darwin (Australia) y Tahití (Polinesia Francesa). Normalmente, al disminuir la presión sobre Darwin y aumentar en Tahití se produce una circulación del aire de Este a Oeste, transportando el agua caliente de la superficie hacia el occidente llevando precipitaciones a Australia y el Pacífico occidental.

Cuando la diferencia de presión se debilita, se inicia el desplazamiento de las aguas cálidas desde el Oeste hacia el Este en forma de una onda que tiene una longitud de varios miles de kilómetros y una amplitud de unos 20 a 30 centímetros. Esta onda cruza el Pacífico tropical, en un lapso de 50 a 60 días, hasta llegar a la costa de Suramérica incrementando el nivel del mar en dicha zona. La presencia de estas aguas anormalmente cálidas en el sector central y oriental del Océano Pacífico tropical es lo que comúnmente se conoce como la fase cálida del ENSO denominada El Niño. Estas condiciones anómalas generan fuertes precipitaciones y cambios notables en el clima, tanto en los países costeros del Pacífico suroriental, como en otras partes del mundo (Maturana et al. 2004).

La fase inversa o fría del ENSO, conocida como La Niña, se caracteriza por presentar temperaturas de la superficie del mar más frías que lo normal, intensificación de los vientos alisios en el este del océano Pacífico y períodos de sequía (Maturana et al. 2004), (IDEAM 2005), (Shein 2012). La determinación de las fases del fenómeno ENSO se realiza a través de índices macroclimáticos, los cuales son índices océano-atmosféricos implementados por agencias climáticas internacionales como la National Oceanic and Atmospheric Administration (NOAA) bajo ciertas metodologías de cálculo cuyas mediciones se toman en el océano Pacífico tropical, entre ellos se destacan:

- **Índice ONI (*Oceanic Niño Index*)**: se define como la media móvil de tres meses de las anomalías en la temperatura superficial del mar en la región Niño 3.4 (IDEAM 2010).
- **Índice SOI (*Southern Oscillation Index*)**: Índice estandarizado basado en las diferencias observadas de presión atmosférica a nivel del mar entre Tahití y Darwin. El SOI es una medida de las fluctuaciones a gran escala en la presión del aire que se produce entre la parte occidental y oriental del Pacífico tropical durante los episodios de El Niño y La Niña (Enloe 2012).
- **Índice MEI (*Multivariate Enso Index*)**: basado en seis variables observadas sobre el Pacífico tropical: presión, componentes de la superficie del viento zonal y meridional, temperatura superficial del mar, temperatura superficial del aire y fracción de nubosidad total del cielo.

⁴Fuerza ejercida por la atmósfera a nivel del mar (elevación cero metros), medida por un barómetro.

Según IDEAM (2010), la clasificación de las fases del fenómeno ENSO para el marco temporal en estudio se muestra en la Tabla 4.1, esto a partir de un consenso entre los tres índices macroclimáticos descritos previamente, tomando en consideración el año hidrológico definido para Colombia como el periodo comprendido entre el comienzo de junio de determinado año hasta el fin de mayo del siguiente.

Tabla 4.1: Clasificación de las fases del fenómeno ENSO para el periodo 1993-2011 según índices macroclimáticos. Fuente, IDEAM (2010)

Año	Fase	Intensidad	Año	Fase	Intensidad
1993 - 94	Normal		2002 - 03	El Niño	Moderado
1994 - 95	El Niño	Moderado	2003 - 04	Normal	
1995 - 96	El Niño	Débil	2004 - 05	El Niño	Débil
1996 - 97	Normal		2005 - 06	Normal	
1997 - 98	El Niño	Fuerte	2006 - 07	El Niño	Débil
1998 - 99	La Niña	Moderado	2007 - 08	La Niña	Moderado
1999 - 00	El Niño	Débil	2008 - 09	El Niño	Débil
2000 - 01	La Niña	Moderado	2009 - 10	El Niño	Moderado
2001 - 02	Normal		2010 - 11	La Niña	Fuerte

4.1.4. Precipitación atmosférica

La precipitación se define como la caída de agua sobre la tierra, esta se puede presentar en forma líquida como la lluvia o llovizna, o también en forma sólida como la nieve o granizo (IDEAM 2005). Del conjunto de elementos climáticos, la precipitación es seguramente el elemento más importante, ya que es el único que proporciona ganancias de agua, la cual es indispensable para la vida en general, y en particular para el desarrollo de las plantas y la producción agrícola. La distribución de las precipitaciones es extremadamente variable dependiendo de la ubicación y las condiciones orográficas⁵ de la zona, llegando a presentarse abundantes lluvias en un determinado lugar y a unos cuantos cientos de metros del mismo, precipitaciones escasas o nulas; esto hace que el estudio de su distribución sea complejo, tanto a nivel espacial como temporal (Campos 1998).

En este sentido, la precipitación atmosférica se define como la cantidad de agua caída durante determinada unidad de tiempo, bien puede ser horaria, diaria, mensual, anual u otra; la cual se expresa en milímetros, siendo 1 milímetro equivalente a 1 litro de agua por metro cuadrado (1 lt/m²) o a 10 metros cúbicos por hectárea (10 m³/ha) (Montealegre & Pabon 2000). Para el desarrollo del estudio, la variable de interés corresponde a la precipitación atmosférica diaria en la estación climatológica San Marcos, ubicada en el centro del departamento del Valle del Cauca cuyos datos fueron colectados durante el periodo 1994 a 2011.

⁵La orografía hace alusión a las elevaciones que puedan existir en una zona particular.

4.2. Marco Teórico Estadístico

4.2.1. Procesos estocásticos

La teoría de los procesos estocásticos se centra en estudiar y modelar sistemas que evolucionan en función de variables como el tiempo o el espacio, de acuerdo a ciertas condiciones aleatorias (Rincón 2012). Un proceso estocástico se define como una sucesión de variables aleatorias X_t indexadas a un conjunto T y definidas sobre un espacio de probabilidad común $(\Omega, \mathcal{F}, P)$:

$$\{X_t : t \in T\}$$

Donde T representa el conjunto índice asociado al proceso, conocido también como *espacio parametral* y donde los posibles valores que puede tomar la variable aleatoria X_t en cada valor de t se encuentran determinados en un conjunto S , llamado *espacio de estados*. Los conjuntos S y T , pueden tomar valores sobre una gran variedad de conjuntos dependiendo del área y el fenómeno en estudio, no obstante, el subíndice t se asocia de manera general con el tiempo, bien sea que se encuentre medido en instantes discretos $T = \{0, 1, \dots\}$ o continuos $T = [0, \infty)$ (Rincón 2012).

De manera general, para poder considerar las variables aleatorias X_t con valores en el espacio de estados S , es necesario asociar a este conjunto un espacio medible σ -álgebra. Tomando S como un subconjunto de $\mathbb{R}$, se puede utilizar la σ -álgebra de Borel de $\mathbb{R}$ restringida a S , tomando $S \cap \mathcal{B}(\mathbb{R})$, con lo cual se garantiza que para cada t los valores de S que puede tomar la variable aleatoria X_t tengan una medida de probabilidad asociada (Rincón 2007). Luego, un proceso estocástico se puede considerar como una función de dos variables, tomando valores en el espacio parametral ($t \in T$) y el espacio muestral ($\omega \in \Omega$), mediante la relación:

$$\begin{aligned} X : T \times \Omega &\rightarrow S \\ (t, \omega) &\rightarrow X_t(\omega) \end{aligned} \tag{4.1}$$

De esta manera, de la ecuación (4.1) para cada valor de t , la función:

$$\omega \rightarrow X_t(\omega)$$

corresponde a una variable aleatoria; mientras para cada $\omega_0 \in \Omega$ fijo, la expresión:

$$t \rightarrow X_t(\omega_0)$$

es una función del tiempo, conocida como trayectoria o realización del proceso. Las leyes de un proceso estocástico se determinan completamente por la distribución de probabilidad conjunta para todos los posibles valores de t , las relaciones de dependencia entre las variables aleatorias y por los distintos valores que se pueden dar en los conjuntos S y T , respectivamente. En este sentido, los procesos Markovianos o modelos de Markov constituyen un tipo de proceso estocástico que cumple con una propiedad de dependencia muy particular, y serán estudiados a continuación en detalle ya que estos constituyen un componente definidor de los modelos ocultos de Markov (Zucchini & MacDonald 2009).

4.2.2. Modelos de Markov

Los modelos de Markov conocidos comúnmente con el nombre de procesos o cadenas de Markov⁶, constituyen un tipo especial de procesos estocásticos que tienen aplicación en la modelación temporal de una amplia gama de fenómenos científicos y sociales. El éxito de estos modelos radica en el hecho de ser lo suficientemente sencillos para ser analizados matemáticamente y lo suficientemente complejos para describir ciertas características no triviales de algunos sistemas (Rincón 2012). Formalmente un modelo de Markov se define a partir de una sucesión de variables aleatorias a tiempo discreto $\{C_t : t \geq 0\}$, las cuales toman valores en el espacio de estados $S = \{1, 2, \dots, m\}$, y cumplen la siguiente condición:

$$P(C_{t+1} = j | C_0 = i_0, C_1 = i_1, \dots, C_t = i) = P(C_{t+1} = j | C_t = i) \quad (4.2)$$

La igualdad expresada en la ecuación (4.2) define la que se conoce comúnmente como la **propiedad de Markov**, la cual establece que la distribución de probabilidad del proceso al tiempo $t + 1$ depende únicamente del estado del proceso al tiempo t y es independiente de los estados que toma el mismo en instantes previos (proceso de Markov de primer orden), para todo $t \geq 0$ y todo $i, j, i_0, \dots, i_{t-1} \in S$. La probabilidad $P(C_{t+1} = j | C_t = i)$ corresponde a la **probabilidad de transición de un paso** desde el estado i en el tiempo t al estado j en el tiempo $t + 1$, y denota por $\gamma_{i,j}(t, t + 1)$.

Según lo anterior, la propiedad de Markov es la característica que señala la evolución del modelo, por lo cual esta se describe mediante las probabilidades de transición $\gamma_{i,j}(t, t + 1)$. Cuando las probabilidades de transición de un paso no dependen del índice t , la cadena de Markov se denomina **homogénea**, por lo tanto la probabilidad de ir del estado i al estado j en un paso no depende del tiempo en el que se encuentra la cadena; en este sentido, las probabilidades de transición se denotan simplemente por $\gamma_{i,j}$ omitiendo el índice asociado al tiempo.

Al variar los subíndices (i, j) sobre los posibles valores que toma el espacio de estados se obtiene una matriz cuadrada de dimensión $m \times m$ denominada **matriz de probabilidades de transición de un paso**, la cual se denota por:

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \dots & \gamma_{1,m} \\ \gamma_{2,1} & \gamma_{2,2} & \dots & \gamma_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m,1} & \gamma_{m,2} & \dots & \gamma_{m,m} \end{pmatrix}_{m \times m}$$

y está sujeta a las siguientes restricciones:

$$i) \quad \gamma_{i,j} \geq 0, \quad \forall i, j = 1, 2, \dots, m$$

$$ii) \quad \sum_{j=1}^m \gamma_{i,j} = 1, \quad \forall i = 1, 2, \dots, m$$

Donde, $i)$ parte del hecho que tales cantidades son probabilidades y por tanto no pueden tomar valores negativos, y $ii)$ determina que con probabilidad uno, el modelo

⁶Se hará uso indistinto de estos términos para hacer referencia al modelo.

efectúa una transición a algún elemento del espacio de estados a partir de cualquier estado i en una unidad de tiempo. En términos generales, el análisis de los modelos de Markov se centra en estudiar la evolución de $\{C_t : t \geq 0\}$ en dos escalas de tiempo diferentes, a corto y largo plazo (Grimmett & Stirzaker 2001). En el corto plazo la evolución del proceso Markoviano se describe por $\mathbf{\Gamma}$, mientras los cambios a largo plazo se describen mediante las probabilidades de transición en t pasos:

$$\gamma_{i,j}^{(t)} = P(C_{t+s} = j | C_s = i) = P(C_t = j | C_0 = i)$$

Lo que indica, la probabilidad de pasar al estado j a partir del estado i en t pasos⁷, bajo la condición de homogeneidad del modelo, por lo que tal probabilidad es independiente del valor que tome s . Este nuevo conjunto de probabilidades puede organizarse en un arreglo matricial variando los subíndices (i, j) sobre todos los posibles valores del espacio de estados, con lo que se obtiene:

$$\mathbf{\Gamma}^{(t)} = \begin{pmatrix} \gamma_{1,1}^{(t)} & \gamma_{1,2}^{(t)} & \cdots & \gamma_{1,m}^{(t)} \\ \gamma_{2,1}^{(t)} & \gamma_{2,2}^{(t)} & \cdots & \gamma_{2,m}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m,1}^{(t)} & \gamma_{m,2}^{(t)} & \cdots & \gamma_{m,m}^{(t)} \end{pmatrix}_{m \times m}$$

$\mathbf{\Gamma}^{(t)}$ es la matriz de probabilidades de transición en t pasos, y constituye el elemento esencial para describir la evolución del proceso en una escala temporal superior a $t = 1$, ya que $\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}$. El cálculo de las probabilidades de transición en t pasos se describe a continuación.

4.2.2.1. Ecuaciones de Chapman-Kolmogorov

En este aparte se definen las ecuaciones de Chapman-Kolmogorov, las cuales son de utilidad para calcular las probabilidades de transición en t pasos y otros resultados importantes en la teoría de las cadenas de Markov. Estas ecuaciones establecen que la probabilidad de pasar del estado i al estado j en t pasos, puede descomponerse en la suma de probabilidades de las posibles rutas que van desde el estado i al estado j y que atraviesan un estado k cualquiera en un tiempo intermedio r , mediante la ecuación (4.3):

$$\gamma_{i,j}^{(t)} = \sum_{k=1}^m \gamma_{i,k}^{(r)} \gamma_{k,j}^{(t-r)}, \quad 0 \leq r \leq t \quad (4.3)$$

al considerar el caso cuando $r = 1$, se tiene:

$$\gamma_{i,j}^{(t)} = \sum_{k=1}^m \gamma_{i,k} \gamma_{k,j}^{(t-1)}$$

que al generalizarse variando los índices (i, j) , en términos matriciales se obtiene:

$$\mathbf{\Gamma}^{(t)} = \mathbf{\Gamma} \mathbf{\Gamma}^{(t-1)}$$

⁷El número de pasos se escribe dentro de un paréntesis con el fin de distinguirlo de algún posible exponente.

luego, cuando $t = 2$

$$\mathbf{\Gamma}^{(2)} = \mathbf{\Gamma}\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}\mathbf{\Gamma}$$

Lo que equivale a efectuar el producto de la matriz de probabilidades de transición de un paso por sí misma. Este resultado se puede extender para todos los valores de t , permitiendo realizar el cálculo de la matriz de probabilidades de transición en t pasos de manera recursiva efectuando el producto matricial de $\mathbf{\Gamma}$, t veces:

$$\mathbf{\Gamma}^{(t)} = \mathbf{\Gamma}^t$$

4.2.2.2. Distribuciones marginales

Otro elemento importante que permite estudiar la relación en el corto y largo plazo en un modelo de Markov, lo constituyen las probabilidades no condicionadas o marginales, denotadas por:

$$u_i^{(t)} = P(C_t = i), \quad \forall t \in T, i \in S$$

$u_i^{(t)}$ expresa la probabilidad que el proceso se encuentre en el estado i al tiempo t , sin tomar en consideración el comportamiento pasado o futuro del sistema. De acuerdo a lo anterior, cada instante t tiene asociado una distribución de probabilidad $\mathbf{u}^{(t)}$ también conocida como distribución marginal al tiempo t , expresada en negrilla por el vector fila:

$$\mathbf{u}^{(t)} = \left(u_1^{(t)}, u_2^{(t)}, \dots, u_m^{(t)} \right)_{1 \times m}$$

la cual se satisface la condición:

$$\sum_{i=1}^m u_i^{(t)} = 1, \quad \forall t$$

La distribución de probabilidad en el instante $t = 0$, $\mathbf{u}^{(0)}$ se conoce como distribución de probabilidad inicial y es de gran utilidad en la determinación de la distribución de probabilidad no condicionada en el instante t , mediante la relación:

$$u_j^{(t)} = \sum_{i=1}^m u_i^{(0)} \gamma_{ij}^{(t)}$$

que en términos matriciales se expresa:

$$\mathbf{u}^{(t)} = \mathbf{u}^{(0)} \mathbf{\Gamma}^{(t)}$$

Llegando a la importante conclusión de que la evolución aleatoria del modelo se determina por la matriz de probabilidades de transición $\mathbf{\Gamma}$ y la distribución de probabilidad inicial $\mathbf{u}^{(0)}$, permitiendo relacionar la evolución del modelo en el corto y el largo plazo, describiendo como C_t depende de la variable inicial C_0 (Grimmett & Stirzaker 2001).

4.2.2.3. Propiedades

Comunicación

Hasta el momento se ha dado una descripción general de la evolución del modelo por medio de las distribuciones de probabilidad marginales y las probabilidades de transición entre estados. A partir del presente literal se busca describir en mayor detalle algunas propiedades y características particulares de los estados de un proceso Markoviano.

De esta manera, a la propiedad en la cual los estados de un modelo de Markov pueden asociarse o establecer una conexión unos con otros en un número finito de transiciones, se le denomina **comunicación**. Dicha propiedad parte del concepto de **accesibilidad**, en donde para dos estados (i, j) cualesquiera, el estado j es accesible desde i si existe un entero $t \geq 0$ tal que $\gamma_{i,j}^{(t)} > 0$, denotado mediante $i \rightarrow j$.

De ahí que cuando $i \rightarrow j$ y $j \rightarrow i$, se dice que los estados son comunicantes y se denota mediante $i \leftrightarrow j$. Cuando $i \leftrightarrow j$, es importante tener en cuenta que la accesibilidad de i a j puede ocurrir un número de pasos distinto que la accesibilidad de j a i (Grimmett & Stirzaker 2001) (Rincón 2012). La comunicación es una relación de equivalencia por lo cual genera una partición del espacio de estados en subconjuntos de estados que se comunican, y cumple las siguientes propiedades:

1. Es reflexiva: $i \leftrightarrow i$
2. Es simétrica: Si $i \leftrightarrow j$, entonces $j \leftrightarrow i$
3. Es transitiva: Si $i \leftrightarrow j$ y $j \leftrightarrow k$, entonces $i \leftrightarrow k$

Cuando todos los elementos del espacio de estados se comunican entre sí, es decir, cuando solo existe una clase de comunicación, la cadena de Markov se denomina **irreducible**.

Periodo

Otra característica de interés en el estudio de los estados de un modelo de Markov lo constituye el periodo. El periodo es un número entero no negativo que se calcula para cada estado del modelo y se define mediante:

$$d(i) = \text{m.c.d.} \{t : \gamma_{i,i}^{(t)} > 0\}$$

Donde $d(i)$ corresponde al máximo común divisor de los tiempos donde es posible retornar al estado i (Grimmett & Stirzaker 2001). El periodo cumple además con el hecho que después de un número suficientemente grande de pasos, toda cadena puede regresar al estado i cada $d(i)$ pasos (Rincón 2012), esto es:

Para cada estado i , existe un entero r tal que para todo $t \geq r$, se cumple $\gamma_{i,i}^{(td(i))} > 0$.

El hecho que el retorno al estado i , se efectuó cada $d(i)$ pasos, constituye la principal razón por la cual dicha cantidad se le llama periodo. Luego, cuando $\gamma_{i,i}^{(t)} = 0$ para todo

$t \geq 1$, $d(i) = 0$, lo que indica que iniciando en el estado i no es posible regresar al mismo en un tiempo posterior. Por otro lado, si $d(i) = 1$ el estado i se llama **aperiódico**, mientras que si $d(i) = k \geq 2$, se dice que i es periódico de periodo k .

El periodo es una propiedad de clase, luego si los estados i y j pertenecen a la misma clase de comunicación, entonces tienen el mismo periodo, por lo tanto las clases de comunicación en el modelo pueden ser clases periódicas o aperiódicas.

Primeras visitas

Los tiempos de primeras visitas son variables aleatorias de utilidad en el estudio de las cadenas de Markov, ya que permiten complementar el análisis de la clasificación de los estados del modelo. El tiempo de primera visita al estado j a partir del estado i se define como:

$$\tau_{ij} = \min \{t \geq 1 : C_t = j | C_0 = i\}$$

Esto es, el primer tiempo en el cual la cadena toma lugar en el estado j dado que su evolución inicia desde el estado i . Si $i = j$, entonces el tiempo de primera visita al estado i se denota simplemente por τ_i . Los tiempos de primera visita τ_{ij} toman valores discretos en el conjunto $\{1, 2, \dots\} \cup \{\infty\}$. Se incluye el valor infinito, ya que puede darse el caso en que el estado j nunca se pueda alcanzar a partir del estado i . La probabilidad de que τ_{ij} tome lugar en el instante t , es decir, $P(\tau_{ij} = t)$ y se define según Rincón (2012) mediante:

$$f_{ij}(t) = P(C_1 \neq j, C_2 \neq j, \dots, C_{t-1} \neq j, C_t = j | C_0 = i)$$

Luego, para cada $t \geq 1$ el valor $f_{ij}(t)$ representa la probabilidad que una cadena que inicia en el estado i , llegue al estado j por primera vez en exactamente t pasos. Dicha cantidad aleatoria sirve de base para calcular la probabilidad de una eventual visita al estado j a partir del estado i , la cual se define como:

$$f_{ij} = \sum_{t=1}^{\infty} f_{ij}(t)$$

Las probabilidades asociadas a los tiempos de primeras visitas y las probabilidades de transición en t pasos, se relacionan mediante (Rincón 2012):

$$\gamma_{i,j}^{(t)} = \sum_{k=1}^t f_{ij}(k) \gamma_{j,j}^{(t-k)} \quad (4.4)$$

Luego, para cada $t \geq 1$ por medio de la ecuación (4.4) se tiene que la probabilidad de visitar el estado j a partir del estado i en t pasos, puede descomponerse en la suma de probabilidades de eventos excluyentes en los cuales se presenta la primera visita, esto puede darse en el primer paso, segundo paso, y así sucesivamente, hasta el último instante posible t .

Recurrencia y transitoriedad

El presente literal describe la clasificación que se puede dar entre los estados de un modelo de Markov, esto en términos de si el modelo es capaz de regresar con certeza al estado de partida. Para este propósito se define:

$$P(C_t = i \text{ para algún } t \geq 1 | C_0 = i) = \begin{cases} 1 \\ < 1 \end{cases} \quad (4.5)$$

como la probabilidad de que iniciando en el estado i se presente un eventual regreso en algún tiempo t al mismo estado. Luego, si dicha probabilidad es igual a uno, el estado i se llama **recurrente**, lo que indica que el modelo es capaz de regresar eventualmente a ese estado, y cuando ello ocurre en algún momento finito, es posible regresar a él una y otra vez con absoluta certeza. Por otro lado, si la probabilidad expresada en la ecuación (4.5) es menor que uno, el estado i se denomina **transitorio** e indica que existe una probabilidad no nula para la cual el proceso iniciando en dicho estado, no regrese nunca a él.

La determinación de la recurrencia o transitoriedad de un estado, se realiza en función de las probabilidades asociadas a los tiempos de primeras visitas y las probabilidades de transición, bajo las siguientes condiciones:

Un estado i es recurrente si:

$$f_{ii} = 1 \text{ ó si } \sum_{t=1}^{\infty} \gamma_{i,i}^{(t)} = \infty$$

Por otro lado, i es transitorio si:

$$f_{ii} < 1 \text{ ó si } \sum_{t=1}^{\infty} \gamma_{i,i}^{(t)} < \infty$$

La recurrencia y la transitoriedad son propiedades de clase, por lo cual:

- a) Si i es recurrente e $i \leftrightarrow j$, entonces j es recurrente.
- b) Si i es transitorio e $i \leftrightarrow j$, entonces j es transitorio.

Esto conduce a que la partición del espacio de estados en una cadena de Markov produzca clases de estados recurrentes o transitorios. De este modo, si una cadena es irreducible y algún estado es recurrente, todos lo son, por tanto la cadena en sí, es recurrente. Adicional a esto, vale la pena resaltar que para toda cadena de Markov con espacio de estados finito, se tiene por lo menos un estado recurrente; en consecuencia, toda cadena finita e irreducible es recurrente.

Tiempo medio de recurrencia

El tiempo medio de recurrencia de un estado i se define como:

$$\mu_i = E(\tau_i) = \begin{cases} \sum_{t=1}^{\infty} t f_{ii}(t) & \text{si } i \text{ es recurrente} \\ \infty & \text{si } i \text{ es transitorio} \end{cases}$$

Siendo τ_i el tiempo de primera visita al estado i , partiendo del mismo estado. El tiempo medio de recurrencia representa el número promedio de pasos que tarda el modelo en retornar al estado i , luego si dicho estado es transitorio existe la posibilidad que nunca se alcance el estado por lo cual tal valor esperado es infinito (Rincón 2012). Por otro lado, si el estado i es recurrente el tiempo medio de recurrencia puede llegar a ser finito o infinito, este resultado es de importancia para realizar una clasificación de los estados recurrentes de un modelo de Markov, y se expondrá a continuación a partir de la definición de la variable aleatoria número de visitas.

Número de visitas

El estudio de la variable aleatoria que registra el número de visitas que un proceso de Markov realiza sobre un estado j desde un estado i , para cualquier tiempo finito t , permite complementar el análisis de los estados del modelo. Este se define a partir de:

$$N_{ij}(t) = \sum_{k=1}^t 1_{(C_k=j)}$$

Siendo $N_{ij}(t)$ el número de visitas al estado j hasta el tiempo t partiendo desde el estado i en $t = 0$. Si $i = j$, el número de visitas al estado i se denota simplemente por $N_i(t)$. De este modo, las variables $0 \leq N_{ij}(1) \leq N_{ij}(2) \leq \dots$, conforman una sucesión monótona creciente de variables aleatorias no negativas que converge con probabilidad 1 a la variable:

$$N_{ij} = \sum_{k=1}^{\infty} 1_{(C_k=j)}$$

Luego, a partir de la variable aleatoria N_{ij} se tiene que:

- Si j es transitorio entonces el número de visitas a dicho estado a partir del estado inicial i es finito, esto es, $P(N_{ij} < \infty) = 1$, y por tanto el número esperado de visitas al estado j es siempre $E(N_{ij}) < \infty$.
- Si j es recurrente y el proceso inicia en dicho estado, entonces con probabilidad 1 la cadena retornará a j una infinidad de veces, por lo tanto el número esperado de visitas a j es infinito. De igual manera, si el proceso inicia en cualquier otro estado i y si j es accesible desde i , entonces el modelo regresará a j una infinidad de veces y por tanto el número esperado de visitas será infinito. Por otro lado, si $i \nrightarrow j$ es decir $f_{ij} = 0$, entonces el número esperado de visitas será 0.

Tomando como base lo anterior, se tiene que todo modelo de Markov irreducible y recurrente, visita cada uno de sus estados una infinidad de veces con probabilidad 1 (Rincón 2012).

Recurrencia positiva y nula

Como se exponía previamente en el apartado **Recurrencia y transitoriedad**, la clasificación de los estados de una cadena de Markov se realiza en términos de la transitoriedad

y la recurrencia. Los estados recurrentes pueden ser a su vez clasificados en función del tiempo promedio de retorno, de la siguiente manera:

Si j un estado recurrente del modelo, entonces:

$$j \text{ es } \begin{cases} \text{recurrente nulo, si} & \mu_j = \infty \\ \text{recurrente positivo, si} & \mu_j < \infty \end{cases}$$

Es decir, si j es un estado recurrente nulo, el tiempo promedio de retorno a él será infinito, mientras que si j es recurrente positivo dicho tiempo será finito. La recurrencia positiva y nula son propiedades de clase, por tanto:

- a) Si i es recurrente positivo e $i \leftrightarrow j$, entonces j es recurrente positivo.
- b) Si i es recurrente nulo e $i \leftrightarrow j$, entonces j es recurrente nulo.

En este sentido, el espacio de estados puede descomponerse en tres subconjuntos de estados, estos son: transitorios, recurrentes nulos y recurrentes positivos. En particular cuando el espacio de estados del proceso es finito, se cumple que al menos un estado es recurrente, luego todos los estados recurrentes del modelo son recurrentes positivos.

Particularmente un estado que es recurrente positivo y aperiódico se denomina **ergódico**. Si adicional a esto, la cadena es irreducible, el modelo en general adquiere tal distinción.

Distribuciones estacionarias

La matriz de probabilidades de transición $\mathbf{\Gamma}$ induce una dinámica en el conjunto de distribuciones de probabilidad marginales $\mathbf{u}^{(t)}$ definidas sobre el espacio de estados del modelo (Rincón 2012), mediante la relación:

$$\mathbf{u}^{(t)} = \mathbf{u}^{(t-1)}\mathbf{\Gamma} = \mathbf{u}^{(0)}\mathbf{\Gamma}^t, \quad t \geq 1 \quad (4.6)$$

De esta manera, se puede obtener una sucesión infinita de distribuciones de probabilidad para cada uno de los instantes de tiempo $t \geq 1$, a partir de la distribución inicial $\mathbf{u}^{(0)}$, mediante la multiplicación de la distribución marginal en el instante $t - 1$ con la matriz de probabilidades de transición $\mathbf{\Gamma}$. Por medio de este resultado es fácil notar que las distribuciones $\mathbf{u}^{(t)}$ evolucionan en función del tiempo.

Por otro lado, cuando una distribución de probabilidad δ definida sobre el espacio de estados S , no cambia respecto al paso del tiempo se dice que es **estacionaria**. En este sentido, una distribución de probabilidad δ es estacionaria o invariante para una cadena de Markov con matriz de probabilidades de transición $\mathbf{\Gamma}$ si cumple:

$$\delta = \delta\mathbf{\Gamma} \Rightarrow \delta_j = \sum_{i=1}^m \delta_i \gamma_{ij} \quad (4.7)$$

La cual está sujeta a las siguientes restricciones:

$$\begin{aligned} i) \quad & \delta_j \geq 0, \quad \forall j \in S \\ ii) \quad & \sum_{j=1}^m \delta_j = 1 \end{aligned}$$

La identidad expresada en la ecuación (4.7) tiene como consecuencia el hecho de que para cualquier tiempo t , se cumple que $\delta = \delta \mathbf{\Gamma}^{(t)}$, lo que demuestra que dicha distribución de probabilidad no cambia a medida que transcurre el tiempo, de ahí la denominación de distribución estacionaria o invariante. Aunque el vector de ceros cumple la condición $\delta = \delta \mathbf{\Gamma}$, no se considera una distribución estacionaria ya que no corresponde a una distribución de probabilidad.

La existencia de la distribución estacionaria para un modelo de Markov, está dada por los siguientes casos:

- No existe ninguna distribución estacionaria.
- Existe una distribución estacionaria y es única.
- Existe una infinidad de distribuciones estacionarias.

De este modo, las condiciones necesarias y suficientes para que una cadena de Markov tenga alguna distribución estacionaria, se determinan a partir de la siguiente definición:

Sea δ una distribución estacionaria para una cadena de Markov. Si j es un estado transitorio o recurrente nulo, entonces $\delta_j = 0$.

La anterior afirmación, indica que la distribución estacionaria únicamente se puede calcular si la cadena posee estados recurrentes positivos y esto sucede en general cuando la cadena es finita. A partir de este resultado toda cadena de Markov que es irreducible y recurrente positiva tiene una única distribución estacionaria dada por:

$$\delta_j = \frac{1}{\mu_j} > 0$$

donde μ_j es el tiempo medio de recurrencia del estado j .

Distribuciones límite

Expresando $\mathbf{u}^{(t)} = \delta^t$ como la distribución marginal sobre S al tiempo t ; a partir de la ecuación (4.6) se puede generar una sucesión infinita de distribuciones de probabilidad $\delta^0, \delta^1, \delta^2, \dots$; tal sucesión de distribuciones bajo ciertas condiciones converge a una distribución de probabilidad límite, esto es:

$$\delta = \lim_{t \rightarrow \infty} \delta^t$$

En el presente numeral se describen las propiedades y condiciones necesarias para la obtención de la distribución límite, permitiendo tener una mejor comprensión de la evolución del modelo en el largo plazo.

La distribución límite de un modelo de Markov con matriz de probabilidades de transición Γ y distribución inicial δ^0 , se define como:

$$\delta = \lim_{t \rightarrow \infty} \delta^0 \Gamma^t = \lim_{t \rightarrow \infty} \sum_{i=1}^m \delta_i^0 \gamma_{i,j}^{(t)}$$

y cuando $t \rightarrow \infty$, cumple:

$$\delta = \delta \Gamma$$

Lo que indica que toda posible distribución límite es una distribución estacionaria (Rincón 2012). Además:

$$\delta = \delta^0 \left(\lim_{t \rightarrow \infty} \Gamma^t \right)$$

Por tanto δ no depende de la distribución de probabilidad inicial, y se obtiene únicamente a partir del límite de las potencias de Γ . Ahora si los límites $\delta_j = \lim_{t \rightarrow \infty} \gamma_{i,j}^{(t)}$ existen para cada j , y no dependen del estado i , entonces se tiene:

$$\begin{aligned} 1. \quad & \sum_{j=1}^m \delta_j \leq 1 \\ 2. \quad & \delta_j = \sum_{i=1}^m \delta_i \gamma_{i,j} \end{aligned}$$

Donde la igualdad para el primer resultado se garantiza, siempre y cuando el espacio de estados del modelo sea finito (Grimmett & Stirzaker 2001). Las condiciones suficientes para que el límite $\lim_{t \rightarrow \infty} \gamma_{i,j}^{(t)} = \delta_j$ exista (convergencia a la distribución estacionaria), están garantizadas si el modelo es:

- a.** Irreducible
- b.** Aperiódico
- c.** Y tiene distribución estacionaria δ

Mientras que la convergencia a la distribución límite para una cadena de Markov, está dada para un modelo que es:

- a.** Irreducible,
- b.** Recurrente positivo, y
- c.** Aperiódico

En este caso, las probabilidades límite $\delta_j = \lim_{t \rightarrow \infty} \gamma_{i,j}^{(t)}$ existen, están dadas por $\delta_j = 1/\mu_j$, y constituyen la única solución al sistema de ecuaciones:

$$\delta_j = \sum_{i=1}^m \delta_i \gamma_{i,j}$$

Sujeto a las restricciones $\delta_j \geq 0$ y $\sum_{j=1}^m \delta_j = 1$.

4.2.3. Modelos ocultos de Markov

Hasta el momento se han considerado los modelos de Markov como un tipo especial de proceso estocástico, donde cada uno de los estados del proceso corresponde a un evento observable o medible, sin embargo, a pesar de su aplicación en la modelación de series temporales, presentan demasiadas restricciones para realizar el ajuste de muchos problemas de interés (Rabiner 1989). En la presente sección se describe una extensión de los modelos de Markov, en los cuales las distribuciones de probabilidad generadoras de la secuencia de registros observados dependen de un conjunto de estados subyacentes que actúan bajo la dinámica de un modelo de Markov, denominados Modelos Ocultos de Markov (*Hidden Markov Models*, HMM).

Los HMM constituyen una valiosa herramienta estadística para la modelación de series de tiempo. Entre las ventajas que presentan, se incluye el potencial para modelar secuencias temporales de datos cuantitativos y categóricos convirtiéndose en un instrumento de gran atractivo en muchas áreas de estudio (Bulla 2006), como:

- Reconocimiento general de patrones.
- Bioinformática.
- Finanzas.
- Ciencias ambientales.

Formalmente un modelo oculto de Markov se define como un doble proceso estocástico $\{(C_t, X_t) : t \geq 1\}$, en el cual $\{C_t : t \geq 1\}$ corresponde al proceso oculto que controla el grado de dependencia de la serie temporal, mientras $\{X_t : t \geq 1\}$ representa el proceso observable del modelo al cual se le asocian las distribuciones de probabilidad que posiblemente generan los registros observados (Zucchini & MacDonald 2009). De manera general, los HMM actúan bajo la influencia de dos procesos aleatorios que interactúan entre sí y sirven de herramienta de propósito general para la representación de distribuciones probabilísticas que evolucionan en función del tiempo o cualquier otra variable de indexación.

Los supuestos inherentes que constituyen la base de la modelación a través de los HMM, se presentan a continuación:

1. El proceso oculto $\{C_t : t \geq 1\}$ cumple con la propiedad de Markov, por lo cual:

$$P(C_{t+1}|C_t, C_{t-1}, \dots, C_1) = P(C_{t+1}|C_t) \quad (4.8)$$

mientras,

2. El proceso observable $\{X_t : t \geq 1\}$ condicionado sobre la secuencia de estados del proceso oculto presenta *independencia condicional*, esto es:

$$P(X_t|X_{t-1}, \dots, X_1, C_t, \dots, C_1) = P(X_t|C_t) \quad (4.9)$$

De este modo, el primer supuesto asociado al modelo permite manejar la dependencia serial de la serie temporal y capturar la dinámica del proceso generador de la serie observada a partir de un proceso Markoviano de primer orden, es decir, un proceso en el cual la ocurrencia de un estado oculto en un tiempo determinado únicamente se ve afectada por la ocurrencia de un estado oculto en el instante previo y es independiente de cualquier información anterior. Por otro lado, el segundo supuesto indica que la distribución de probabilidad de la variable aleatoria X_t observada en algún instante de tiempo t , depende únicamente del estado oculto en el cual se encuentra el proceso Markoviano en el mismo instante temporal, y es independiente de las observaciones y estados ocultos previos.

Los modelos ocultos de Markov tienen la ventaja sobre las cadenas de Markov en que pueden representar dependencias a largo plazo entre las observaciones por medio del proceso Markoviano de primer orden sobre las variables latentes, es decir, no asumen que la propiedad de Markov se presente entre las observaciones (Fraser 2008)(Murphy 2012). La estructura básica de un HMM se ilustra en la Figura 4.2, donde $\{C_1, C_2, \dots, C_T\}$ representa la secuencia de estados del modelo de longitud T , mientras $\{X_1, X_2, \dots, X_T\}$ representa la correspondiente secuencia de observaciones emitidas.

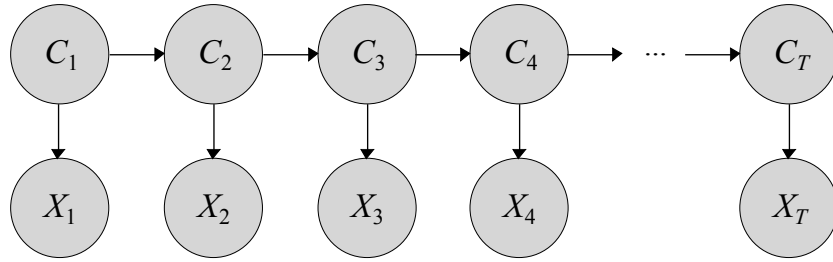


Figura 4.2: Representación esquemática de un HMM

Esta representación en forma de red muestra la influencia que tiene el proceso oculto sobre el proceso observable, indicando que la estructura de dependencia de la serie recae sobre el proceso $\{C_t\}$ y es posible determinarla a partir del proceso $\{X_t\}$. Este hecho es de gran importancia en el estudio de las series temporales ya que permite analizar las transiciones que se pueden dar entre los posibles patrones generadores de las series en estudio.

A continuación se exponen de manera general los elementos básicos de los modelos ocultos de Markov:

- Número de estados ocultos del modelo, m . Aunque los estados son subyacentes (ocultos), para muchas aplicaciones prácticas existe un significado físico conectado a este conjunto. Se denotará el espacio de estados ocultos del modelo por el conjunto $S = \{1, 2, \dots, m\}$.
- Número de observaciones emitidas por estado oculto, n . Estos valores corresponden a las salidas o emisiones del proceso las cuales se encuentran asociadas a la distribución de probabilidad por estado oculto; en el presente estudio solo se consi-

deraran distribuciones de probabilidad discretas, y por consiguiente dichos valores serán identificados por el conjunto $V = \{v_1, v_2, \dots, v_n\}$.

- Matriz de probabilidades de transición de los estados ocultos, $\mathbf{\Gamma} = \{\gamma_{i,j}\}$. Dicha matriz caracteriza la dinámica de la serie temporal y tiene como elemento general:

$$\gamma_{i,j} = P(C_{t+1} = j | C_t = i), \quad 1 \leq i, j \leq m$$

- Distribución de probabilidad de las observaciones emitidas por estado oculto, denotada por:

$$p_i(v_k) = P(X_t = v_k | C_t = i), \quad 1 \leq i \leq m, \quad 1 \leq k \leq n$$

La cual indica la probabilidad de que el proceso en el tiempo t , estando en el estado i , emita la observación v_k . Estas probabilidades se agrupan en una matriz $\mathbf{P} = \{p_i(v_k)\}$, llamada matriz de probabilidades de emisión de tamaño $m \times n$.

$$\mathbf{P} = \begin{pmatrix} p_1(v_1) & p_1(v_2) & \dots & p_1(v_n) \\ p_2(v_1) & p_2(v_2) & \dots & p_2(v_n) \\ \vdots & \vdots & \ddots & \vdots \\ p_m(v_1) & p_m(v_2) & \dots & p_m(v_n) \end{pmatrix}_{m \times n}$$

La cual está sujeta a la siguiente restricción:

$$\sum_{k=1}^n p_i(v_k) = 1, \quad \forall i = 1, 2, \dots, m$$

Se referirá a las m distribuciones $p_i(v_k)$ como distribuciones de probabilidad dependientes del estado i –ésimo del modelo o simplemente probabilidad de que el estado i emita la observación v_k .

- Distribución de probabilidad inicial del proceso oculto, $\delta = (\delta_1, \delta_2, \dots, \delta_m)$. La cual expresa la probabilidad que el proceso se encuentre en algún estado i al tiempo $t = 1$.

$$\delta_i = P(C_1 = i), \quad \forall i = 1, 2, \dots, m$$

Esta distribución de probabilidad puede considerarse estacionaria, o no dependiendo del comportamiento o evolución del proceso oculto en el largo plazo.

Por conveniencia se hará uso de la siguiente notación para simplificar las expresiones que involucren las cantidades expresadas a continuación:

$$\lambda = (\mathbf{\Gamma}, \mathbf{P}, \delta)$$

$$\mathbf{X} = \{X_1, X_2, \dots, X_T\}$$

$$\mathbf{C} = \{C_1, C_2, \dots, C_T\}$$

En el contexto de los HMM, los múltiples usos o aplicaciones que se pueden dar se reducen a solucionar los tres problemas básicos que se presentan a continuación (Rabiner 1989):

Problema 1: Evaluación. Mediante este problema se busca calcular la probabilidad de una secuencia de observaciones $\mathbf{X}$ de longitud T dado un HMM con parámetros λ , es decir, $P(\mathbf{X}|\lambda)$.

Problema 2: Decodificación. Consiste en hallar la secuencia más probable de estados ocultos $\mathbf{C}^*$ asociada a una secuencia de registros observados $\mathbf{X}$, esto es, $P(\mathbf{C}|\mathbf{X}, \lambda)$.

Problema 3: Entrenamiento. Se busca ajustar los parámetros λ del modelo de tal manera que se maximice la probabilidad de la secuencia de observaciones $\mathbf{X}$, es decir, $P(\mathbf{X}|\lambda)$.

El problema 1 se puede resolver usando el algoritmo iterativo Forward-Backward, mientras el problema 2 puede desarrollarse mediante el algoritmo de Viterbi siendo este último un algoritmo de programación dinámica. Por otro lado, el problema 3 se resuelve por el algoritmo EM (Esperanza-Maximización) que en el contexto de los HMM se conoce por el nombre de algoritmo Baum-Welch, este método utiliza las probabilidades Forward y Backward para actualizar iterativamente los parámetros (Zhai 2003). Los problemas 1 y 2 pueden ser vistos como problemas de análisis mientras el problema 3 es un problema de síntesis (o de identificación del modelo) (Dugad & Desai 1996). En las siguientes subsecciones se presenta de manera general la solución de cada uno de los tres problemas fundamentales de los modelos ocultos de Markov.

4.2.3.1. Evaluación

Se desea calcular la probabilidad de una secuencia de observaciones $\mathbf{X}$ dado un modelo HMM con parámetros $\lambda = (\mathbf{\Gamma}, \mathbf{P}, \delta)$, es decir $P(\mathbf{X}|\lambda)$. La manera directa de realizar este cálculo, es mediante la enumeración de todas las posibles secuencias de estados ocultos de longitud T , $\mathbf{C} = \{C_1, C_2, \dots, C_T\}$. En este sentido, por el teorema de Bayes dicha probabilidad se puede descomponer mediante:

$$P(\mathbf{X}|\lambda) = \sum_{c_1, c_2, \dots, c_T} P(\mathbf{X}, \mathbf{C}|\lambda) = \sum_{c_1, c_2, \dots, c_T} P(\mathbf{X}|\mathbf{C}, \lambda) P(\mathbf{C}|\lambda) \quad (4.10)$$

Donde, $P(\mathbf{X}|\mathbf{C}, \lambda)$ corresponde a la probabilidad de la secuencia de observaciones $\mathbf{X}$ dado una secuencia fija de estados $\mathbf{C}$, la cual gracias al supuesto de independencia condicional produce:

$$P(\mathbf{X}|\mathbf{C}, \lambda) = \prod_{t=1}^T P(X_t|C_t, \lambda) = p_{c_1}(x_1) p_{c_2}(x_2) \cdots p_{c_T}(x_T)$$

Mientras, la probabilidad de la secuencia de estados ocultos $P(\mathbf{C}|\lambda)$ corresponde a:

$$P(\mathbf{C}|\lambda) = P(C_1 = c_1, C_2 = c_2, \dots, C_T = c_T|\lambda) = \delta_{c_1} \gamma_{c_1, c_2} \gamma_{c_2, c_3} \cdots \gamma_{c_{T-1}, c_T}$$

De este modo, al efectuar el producto entre $P(\mathbf{X}|\mathbf{C}, \lambda)$ y $P(\mathbf{C}|\lambda)$, y sumar sobre todas las posibles secuencias de estados ocultos, se logra calcular la probabilidad $P(\mathbf{X}|\lambda)$ o función de verosimilitud del proceso mediante:

$$P(\mathbf{X}|\lambda) = \sum_{c_1, c_2, \dots, c_T} \delta_{c_1} p_{c_1}(x_1) \gamma_{c_1, c_2} p_{c_2}(x_2) \cdots \gamma_{c_{T-1}, c_T} p_{c_T}(x_T)$$

Esta probabilidad indica cómo se desarrolla la evolución del proceso desde el primer hasta el último instante observado, en este sentido cuando el proceso inicia en el tiempo $t = 1$ se encuentra en el estado oculto c_1 con probabilidad δ_{c_1} a partir del cual genera o emite la observación x_1 con probabilidad $p_{c_1}(x_1)$, posterior a dicha emisión el proceso realiza la transición al estado oculto c_2 a partir de c_1 con probabilidad γ_{c_1, c_2} y genera la observación x_2 con probabilidad de emisión $p_{c_2}(x_2)$. De esta manera, el proceso continua hasta que en el tiempo T hace la última transición desde el estado c_{T-1} con probabilidad γ_{c_{T-1}, c_T} y genera la observación x_T con probabilidad $p_{c_T}(x_T)$ (Rabiner 1989).

El cálculo de la probabilidad de la secuencia de observaciones $\mathbf{X}$ de forma directa, presentada en la ecuación 4.10 consiste en sumar m^T términos, correspondientes a todas las secuencias posibles de estados, cada uno con $2T$ factores, lo que requiere un orden de $2Tm^T$ operaciones. Por consiguiente el elevado número de operaciones que se deben realizar para resolver dicha probabilidad conduce a la utilización de algoritmos eficientes como el algoritmo Forward-Backward el cual se expone a continuación.

Algoritmo Forward-Backward

Mediante este algoritmo es posible calcular la probabilidad de una secuencia de observaciones dado el modelo $P(\mathbf{X}|\lambda)$, de manera recursiva y eficiente. En el contexto de los modelos ocultos de Markov, el problema de calcular la probabilidad de una secuencia de observaciones se conoce como el *problema de evaluación* (Zhai 2003). Para resolver este problema se define la variable de avance (o variable Forward) como:

$$\alpha_t(i) = P(X_1, X_2, \dots, X_t, C_t = i | \lambda)$$

La cual expresa la probabilidad conjunta de la secuencia parcial de observaciones hasta el tiempo t ($t < T$) y a su vez que el proceso se encuentre en el estado i en el instante t , dado el modelo con parámetros λ . Esta probabilidad puede calcularse de manera inductiva mediante los siguientes pasos:

1. Inicialización:

$$\alpha_1(i) = \delta_i p_i(x_1), \quad i \in S$$

2. Inducción:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^m \alpha_t(i) \gamma_{i,j} \right] p_j(x_{t+1}), \quad 1 \leq t \leq T-1, \quad 1 \leq j \leq m$$

3. Finalización:

$$P(\mathbf{X}|\lambda) = \sum_{i=1}^m \alpha_T(i)$$

De este manera, el paso 1 inicializa la variable de avance como la probabilidad conjunta de que el proceso se encuentre en el estado i y emita la observación x_1 en el instante de tiempo $t = 1$. Posterior a esto, el paso 2 considerado el corazón del algoritmo Forward-Backward, permite calcular de manera recursiva la probabilidad de alcanzar el estado j en el tiempo $t + 1$ desde el estado i en el tiempo t considerando todas las m posibilidades para realizar dicha transición. Como $\alpha_t(i)$ es la probabilidad del evento conjunto de observar $x_1, x_2, \dots, x_t$ y el estado i en el tiempo t , el producto $\alpha_t(i)\gamma_{i,j}$ indica la probabilidad del evento conjunto de observar $x_1, x_2, \dots, x_t$ y que el estado j sea alcanzado en el momento $t + 1$ desde el estado i en el momento t . Sumando este producto sobre todos los m posibles estados i , $1 \leq i \leq m$ en el instante de tiempo t , se obtiene la probabilidad que el proceso esté en el estado j en el tiempo $t + 1$ con todas las observaciones parciales previas. Una vez llevado a cabo lo anterior y el estado j es alcanzado, es fácil incluir en la expresión la probabilidad de emitir la observación x_{t+1} , mediante el producto entre $\alpha_{t+1}(j)$ y $p_j(x_{t+1})$. Este procedimiento de inducción se desarrolla de manera recursiva para todos los posibles estados j , $1 \leq j \leq m$ por instante temporal $t = 1, 2, \dots, T - 1$. Finalmente mediante el paso 3 se obtiene el cálculo deseado de $P(\mathbf{X}|\lambda)$ a través de la suma de las variables terminales $\alpha_T(i)$ para todos los posibles estados $1 \leq i \leq m$ (Rabiner 1989), esto dado que por definición:

$$\alpha_T(i) = P(X_1, X_2, \dots, X_T, C_T = i|\lambda)$$

Examinando el número de cálculos para obtener $\alpha_t(i)$, $1 \leq t \leq T$, $1 \leq i \leq m$, se requiere un orden de m^2T operaciones, de manera precisa es necesario $m(m+1)(T-1) + m$ multiplicaciones y $m(m-1)(T-1)$ adiciones, lo cual es mucho menor comparado con el número de cálculos necesarios mediante la manera directa ($2Tm^T$).

Las variables de avance $\alpha_t(i)$, calculadas mediante el algoritmo Forward son suficientes para calcular la probabilidad $P(\mathbf{X}|\lambda)$ o función de verosimilitud del modelo; no obstante, para resolver el problema de entrenamiento o estimación de parámetros se define otro conjunto de probabilidades las cuales se denominan variables de retroceso (o variables Backward), definidas como (Zhai 2003):

$$\beta_t(i) = P(X_{t+1}, X_{t+2}, \dots, X_T | C_t = i, \lambda)$$

La cual corresponde a la probabilidad de generar la secuencia parcial de observaciones desde el tiempo $t + 1$ hasta el final, condicionada a que el proceso se encuentra en el estado $C_t = i$ en el tiempo t . Equivalentemente, el cálculo de las probabilidades β puede ser resuelto de manera recursiva a partir de los siguientes pasos:

1. Inicialización:

$$\beta_T(i) = 1, \quad i \in S$$

2. Inducción:

$$\beta_t(i) = \sum_{j=1}^m \gamma_{i,j} p_j(x_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \dots, 1, \quad 1 \leq i \leq m$$

El paso de inicialización define arbitrariamente $\beta_T(i)$ como 1 para cada estado $i \in S$, mientras el segundo paso indica cómo alcanzar el estado i en el tiempo t a partir de la información en los tiempos posteriores, haciendo la transición desde el estado i al j considerando todos los posibles estados $j \in S$ en el tiempo $t+1$. Contar con la observación emitida x_{t+1} en el estado j , se define mediante la probabilidad de emisión $p_j(x_{t+1})$ y después contar con el resto de la secuencia de observaciones a partir del estado j , se denota mediante el término $\beta_{t+1}(j)$ (Rabiner 1989). De nuevo el cálculo de $\beta_t(i)$, $1 \leq t \leq T$, $1 \leq i \leq m$ requiere de un orden de m^2T operaciones.

Por medio de las probabilidades de avance y retroceso es posible calcular $P(\mathbf{X}|\lambda)$ mediante:

$$P(\mathbf{X}|\lambda) = \sum_{i=1}^m \alpha_t(i) \beta_t(i), \quad \forall t$$

Recordando que:

$$\alpha_t(i) = P(X_1, X_2, \dots, X_t, C_t = i | \lambda)$$

$$\beta_t(i) = P(X_{t+1}, X_{t+2}, \dots, X_T | C_t = i, \lambda)$$

El cálculo de las variables Forward-Backward al consistir en el producto de un gran número de términos muestra que a medida que T incrementa, el producto de las probabilidades involucradas en el cálculo tiende rápidamente a cero conduciendo a un problema de precisión ya que se pueden obtener valores significativamente pequeños. Para evitar esta situación, en el contexto de los HMM generalmente se aplica un procedimiento de escalamiento, donde las expresiones $\alpha_t(i)$ y $\beta_t(i)$ utilizadas en la estimación son normalizadas realizando un producto por un factor de escala que garantiza que las estimaciones obtenidas varíen en un rango apropiado, evitando de esta manera los inconvenientes de subdesbordamiento numérico (Rabiner 1989).

4.2.3.2. Decodificación

Dado un HMM para una serie de tiempo, es de interés conocer la secuencia de estados ocultos que podría haber generado la serie de tiempo en cuestión, así como identificar los instantes de tiempo donde se ha producido la transición entre estados con el propósito de tener un conocimiento detallado del comportamiento del fenómeno en estudio (Visser 2011). En el contexto de los modelos ocultos de Markov, el problema de calcular la secuencia de estados ocultos de un HMM se conoce como el *problema de decodificación*.

Para resolver el problema de decodificación es necesario calcular la secuencia de estados ocultos $\mathbf{C} = \{C_1, C_2, \dots, C_T\}$ condicionada a los datos observados y los parámetros del modelo, de tal manera que se maximice la probabilidad:

$$P(C_1, C_2, \dots, C_T | X_1, X_2, \dots, X_T, \lambda) = P(\mathbf{C} | \mathbf{X}, \lambda) \quad (4.11)$$

La determinación de la secuencia $\mathbf{C}$ que maximiza la ecuación (4.11) no es una labor simple, ya que a medida que T se incrementa, se hace difícil enumerar todas las posibles secuencias de estados y por consiguiente seleccionar la que tenga mayor probabilidad. Por esta razón se hace necesario utilizar algún procedimiento que permita obtener de manera eficiente la secuencia $\mathbf{C}$ que mejor explique la secuencia $\mathbf{X}$, sin incrementar exponencialmente el número de cálculos involucrados en su determinación. Para efectuar la estimación de la secuencia de estados ocultos más probable se recurre al algoritmo de Viterbi (Viterbi 1967; Forney 1973).

Algoritmo de Viterbi

Este algoritmo de programación dinámica permite calcular la secuencia óptima de estados de manera eficiente y recursiva, maximizando la probabilidad conjunta $P(\mathbf{C}, \mathbf{X})$ en lugar de $P(\mathbf{C} | \mathbf{X})$, ya que:

$$P(\mathbf{C} | \mathbf{X}) = \frac{P(\mathbf{C}, \mathbf{X})}{P(\mathbf{X})}$$

donde $P(\mathbf{X})$ es independiente de la secuencia $\mathbf{C}$ y por lo tanto no afecta su estimación. Básicamente el funcionamiento del algoritmo de Viterbi se resume en dos etapas:

1. Obtener el valor máximo de la probabilidad conjunta $P(\mathbf{C}, \mathbf{X})$ evaluando de forma recursiva todas las posibles rutas de estados ocultos.

Para este propósito se define:

$$\xi_t(i) = \max_{c_1, \dots, c_{t-1}} P(C_1 = c_1, \dots, C_{t-1} = c_{t-1}, C_t = i, X_1 = x_1, \dots, X_t = x_t)$$

como la probabilidad de la mejor ruta hasta el estado i habiendo visto las t primeras observaciones. Esta variable auxiliar $\xi_t(i)$ se calcula para todos los estados $i \in S$ e instantes del tiempo $1 < t < T$, mediante la recursión:

$$\xi_{t+1}(j) = \left[\max_i \xi_t(i) \gamma_{i,j} \right] p_j(x_{t+1}) \quad (4.12)$$

teniendo en cuenta que $\xi_1(i) = \delta_i p_i(x_1)$.

2. Construir un maximizador que permita retornar la secuencia de estados ocultos más probable, es decir, la secuencia de estados que maximice la probabilidad conjunta del paso anterior.

Una vez obtenidas las cantidades $\{\xi_t(i)\}$ se construye una secuencia maximizadora:

$$\mathbf{C}^* = \{i_1^*, i_2^*, \dots, i_T^*\}$$

A partir de un procedimiento de retroceso que se realiza definiendo:

$$i_T^* = \arg \max_{i \in S} \xi_T(i)$$

como el estado oculto al tiempo T que maximiza $\xi_T(i)$; luego para los tiempos restantes $1 \leq t \leq T-1$ los elementos de la secuencia optima $\mathbf{C}$ se obtienen mediante:

$$i_t^* = \arg \max_{i \in S} [\xi_t(i) \gamma_{i, i_{t+1}}], \quad t = T-1, T-2, \dots, 1$$

El algoritmo de Viterbi proporciona una estimación global⁸ de la secuencia de estados ocultos, lo que constituye una característica muy importante; sin embargo, una de las desventajas de este algoritmo es que solo proporciona la estimación de una secuencia de estados ocultos y no provee información sobre secuencias similares que podrían estar próximas a la secuencia óptima (Visser 2011). El procedimiento de decodificación en un HMM asume que el modelo ajustado es apropiado para los datos, esto se logra mediante una estimación adecuada de los parámetros del modelo, tema que se discute en la siguiente subsección.

4.2.3.3. Entrenamiento

El tercer problema de interés en un HMM consiste en determinar el método más apropiado para ajustar los parámetros λ , de tal forma que se logre maximizar la función de verosimilitud del modelo $L = P(\mathbf{X}|\lambda)$. Como se expuso en la solución del problema de evaluación, el algoritmo Forward-Backward proporciona un método recursivo para calcular:

$$L = \sum_{c_1, c_2, \dots, c_T} \delta_{c_1} p_{c_1}(x_1) \gamma_{c_1, c_2} p_{c_2}(x_2) \cdots \gamma_{c_{T-1}, c_T} p_{c_T}(x_T)$$

mediante la utilización de las variables de avance, a través de:

$$P(\mathbf{X}|\lambda) = L = \sum_{i=1}^m \alpha_T(i)$$

No obstante, el cálculo de estas expresiones requiere el desarrollo de un elevado número de productos y sumas de probabilidades, haciendo difícil la maximización de la función de verosimilitud del modelo de manera analítica. Por esta razón, es necesario emplear técnicas numéricas de optimización que permitan estimar los parámetros del modelo de manera apropiada; en este sentido, dos de los principales métodos son: *a)* maximización directa de la función de verosimilitud usando técnicas de gradiente (como el método de Newton o sus variantes) o *b)* algoritmo Esperanza-Maximización (EM), el cual en el contexto de los HMM se le denomina algoritmo Baum-Welch (Rabiner 1989).

Estas dos técnicas proporcionan estimaciones locales de máxima verosimilitud de los parámetros; sin embargo, aunque las técnicas de gradiente convergen mucho más rápido que el algoritmo Baum-Welch en la zona de un máximo, requieren valores de partida

⁸Este término hace referencia al hecho que se calcula la secuencia de estados ocultos de manera conjunta; otros procedimientos de estimación se presentan en Rabiner (1989) y Zucchini & MacDonald (2009).

mucho más precisos que los proporcionados a este último, convirtiendo así al algoritmo Baum-Welch en un método de estimación un poco más robusto (Bulla & Berzel 2008) (Visser 2011). En esta sección se presenta la solución del problema de entrenamiento de un HMM, mediante el algoritmo Baum-Welch, el cual busca realizar una actualización dinámica de los parámetros del modelo en busca de un máximo local de la función de verosimilitud, a partir del siguiente procedimiento:

1. Asignar valores iniciales a los parámetros del modelo $\lambda^{(in)} = (\mathbf{\Gamma}^{(in)}, \mathbf{P}^{(in)}, \delta^{(in)})$, a partir de un procedimiento de selección aleatoria o a través de la incorporación de información a priori de los datos.
2. Calcular la secuencia esperada de estados ocultos a partir de los valores iniciales del paso previo.
3. Reestimar los parámetros del modelo condicionados a la estimación actual de la secuencia de estados ocultos, obteniendo como resultado los parámetros $\lambda^{(out)}$, los cuales cumplen:

$$P(\mathbf{X}|\lambda^{(out)}) \geq P(\mathbf{X}|\lambda^{(in)})$$

4. Repetir los pasos 2 y 3 hasta que la diferencia entre los parámetros reestimados en dos iteraciones sucesivas sea despreciable, de acuerdo a un límite de convergencia establecido.

Este procedimiento busca estimar los parámetros del modelo a través de la maximización de la función de verosimilitud, la cual involucra cantidades no observables, como la secuencia de estados ocultos, posible generadora del proceso observable. En este sentido, el algoritmo Baum-Welch intenta aproximar estos parámetros o cantidades a partir del cálculo de valores esperados (conocido como paso E, del algoritmo EM) intentando maximizar la función de verosimilitud a partir de las reestimaciones de dichos parámetros (paso M). Este proceso se repite de manera iterativa utilizando los parámetros encontrados en el paso M para iniciar el paso E siguiente.

De esta manera, a partir de la secuencia de observaciones $\mathbf{X}$, los valores iniciales $\lambda^{(in)}$ de los parámetros y asumiendo que se conoce la secuencia de estados ocultos del modelo, se construyen las siguientes cantidades:

$$\delta_i^{(out)} = \text{Frecuencia esperada del estado } i \text{ en } t = 1$$

$$\gamma_{i,j}^{(out)} = \frac{E(N_{ij}|X)}{E(N_i|X)}, \quad p_j^{(out)}(v_k) = \frac{E(N_j(v_k)|X)}{E(N_j|X)}$$

Donde N_{ij} corresponde al número de transiciones desde el estado i en el tiempo t al estado j en $t + 1$; N_i el número de veces que se hace transición desde el estado i para algún t y $E(N_j(v_k))$ el número esperado de veces en el estado j en los cuales se emite la observación v_k para algún valor de t . Es claro que las anteriores cantidades aleatorias no se observan de manera directa ya que corresponden a los estados subyacentes, por tanto se define la variable $\xi_t(i, j)$, la cual es de utilidad para el cálculo de los valores esperados de manera eficiente:

$$\xi_t(i, j) = P(C_t = i, C_{t+1} = j | \mathbf{X}, \lambda), \quad \forall i, j \in S, \quad 1 \leq t \leq T$$

Lo que indica la probabilidad de estar en el estado i en el instante de tiempo t y en el estado j en $t + 1$, condicionada a la secuencia de observaciones $\mathbf{X}$:

$$\xi_t(i, j) = P(C_t = i, C_{t+1} = j | \mathbf{X}, \lambda)$$

Usando la definición de probabilidad condicional:

$$\xi_t(i, j) = \frac{P(C_t = i, C_{t+1} = j, \mathbf{X} | \lambda)}{P(\mathbf{X} | \lambda)}$$

La cantidad ξ_t puede ser definida en términos de las variables Forward-Backward de la siguiente manera:

$$\xi_t(i, j) = \frac{\alpha_t(i) \gamma_{i,j} p_j(x_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^m \sum_{j=1}^m \alpha_t(i) \gamma_{i,j} p_j(x_{t+1}) \beta_{t+1}(j)}$$

Similarmente se define $\tau_t(i) = P(C_t = i | \mathbf{X}, \lambda)$ como la probabilidad de estar en el estado i en el tiempo t , dado la secuencia de observaciones y el modelo, por tanto puede relacionarse a $\tau_t(i)$ con $\xi(i, j)$, sumando sobre j de la siguiente manera:

$$\tau(i) = \sum_{j=1}^m \xi_t(i, j)$$

Sumando $\tau_t(i)$ a través del índice de tiempo t , se obtiene una cantidad que puede ser interpretada como el número esperado de veces que el estado i es visitado o equivalentemente, el número esperado de transiciones que se hace desde el estado i . De manera similar la sumatoria de $\xi_t(i, j)$ sobre t (desde $t = 1$ hasta $t = T - 1$) puede ser interpretada como la cantidad esperada de transiciones del estado i al estado j . Esto es:

$$\begin{aligned} \sum_{t=1}^{T-1} \tau_t(i) &= \text{Número esperado de transiciones desde } i \\ \sum_{t=1}^{T-1} \xi_t(i, j) &= \text{Número esperado de transiciones desde } i \text{ a } j \end{aligned}$$

Usando los conceptos anteriores, se proponen las siguientes fórmulas de reestimación del conjunto de parámetros λ de un HMM:

$$\delta_i^{(out)} = \text{Frecuencia esperada del estado } i \text{ en } t = 1$$

$$\delta_i^{(out)} = \tau_1(i)$$

$$\gamma_{i,j}^{(out)} = \frac{\text{Número esperado de transiciones de } i \text{ a } j}{\text{Número esperado de transiciones desde } i}$$

$$\gamma_{i,j}^{(out)} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \tau_t(i)}$$

$$p_j^{(out)}(v_k) = \frac{\text{Número esperado de visitas a } j \text{ y observar } v_k}{\text{número esperado de visitas a } j}$$

$$p_j^{(out)}(v_k) = \frac{\sum_{t=1}^T \tau_t(j)}{\sum_{t=1}^T \tau_t(j)}$$

Mediante las formulas anteriores, utilizando los valores iniciales de los parámetros del modelo $\lambda^{(in)}$ con la finalidad de obtener los parámetros reestimados $\lambda^{(out)}$, se puede probar que el modelo con parámetros $\lambda^{(out)}$ es más probable que el modelo inicial $\lambda^{(in)}$, en el sentido de $P(\mathbf{X}|\lambda^{(out)}) \geq P(\mathbf{X}|\lambda^{(in)})$. Este procedimiento se repite un número suficiente de veces hasta tener garantías de que se ha llegado a las proximidades de un máximo local de la función de verosimilitud.

4.2.3.4. Consideraciones a la hora de ajustar un HMM

La implementación de un HMM requiere tomar en consideración dos aspectos que inciden en los resultados obtenidos durante el ajuste, principalmente en la estabilidad de las estimaciones de los parámetros del modelo. Dichos aspectos se describen a continuación:

■ Valores iniciales de los parámetros

La estimación de los parámetros de un modelo oculto de Markov se realiza a partir del método de máxima verosimilitud. El algoritmo iterativo de Baum-Welch utilizado para este fin proporciona estimaciones que maximizan la función de verosimilitud del modelo solo de manera local dependiendo principalmente de los valores iniciales que se le brinden; en este sentido, proporcionar valores iniciales adecuados mejorará sustancialmente la estimación de los parámetros del modelo.

La elección de un conjunto apropiado de valores iniciales para la estimación de los parámetros tiene como base principal la experiencia o el conocimiento previo que se tenga sobre el comportamiento de los posibles parámetros involucrados en el fenómeno en estudio (principalmente para los parámetros asociados a las distribuciones de probabilidad dependientes de cada estado), no obstante, la literatura presenta algunas recomendaciones para brindar los valores iniciales en el proceso iterativo (Rabiner 1989), (Zucchini & MacDonald 2009):

- Distribución inicial (δ): se recomienda utilizar valores iniciales generados de manera aleatoria o uniformemente distribuidos sobre el número de estados ocultos en consideración, garantizando el cumplimiento de:

$$\sum_{i=1}^m \delta_i = 1$$

- Matriz de probabilidades de transición ($\mathbf{\Gamma}$): entre las estrategias sugeridas por la literatura se encuentran: la asignación de valores aleatorios generados a partir de una distribución uniforme garantizando el cumplimiento de las restricciones asociadas a la matriz de probabilidades de transición, o la utilización de valores iniciales asignando un valor común próximo a cero (por ejemplo, 0.01 o 0.05) a los elementos fuera de la diagonal de la matriz $\mathbf{\Gamma}$, esta última estrategia permite que la correspondiente distribución estacionaria sea uniforme sobre todos los estados del modelo.
- Parámetros asociados a las distribuciones de probabilidad condicionadas a los estados ocultos: dependiendo de la naturaleza de las distribuciones de probabilidad que se asume emiten los registros observados en la serie temporal, existen diferentes estrategias para obtener los valores iniciales; en algunos casos la experiencia con la que se cuente sobre el fenómeno en estudio permitirá brindar conjeturas acertadas para parámetros como proporciones; por otro lado, para parámetros como los promedios, se recomienda usar segmentación manual de la secuencia de observaciones en estados y dentro de cada estado se calculan los promedios muestrales los cuales se toman como valores iniciales. Otras estrategias se exponen en Rabiner (1989).

■ Efectos de una cantidad insuficiente de datos de entrenamiento

Otro de los problemas asociado a la estimación de parámetros a través del algoritmo Baum-Welch corresponde a la cantidad necesaria de datos para efectuar el ajuste del modelo. La utilización de una serie temporal con un número insuficiente de registros, hace que los parámetros estimados sean poco fiables, ya que las estimaciones son calculadas en base a un número reducido de ocurrencias de los eventos de interés dentro de cada estado oculto; luego para identificar un modelo con una mayor cantidad de estados ocultos, así como observaciones emitidas por estado, se necesita una mayor cantidad de registros (Fraser 2008).

Con el propósito de hacer frente a esta dificultad Rabiner (1989), plantea algunas recomendaciones entre las cuales se destacan:

- Incrementar el número de datos de entrenamiento del modelo.
- Reducir el tamaño del modelo, es decir, el número de estados o el número de observaciones por estado (lo que conlleva a replantear la distribución de probabilidad generadora de la secuencia de observaciones).
- Interpolan un conjunto de parámetros estimados mediante un modelo con una cantidad adecuada de datos de entrenamiento.

Estas alternativas en varias circunstancias no son aplicables debido a las restricciones en cuanto a la cantidad de información disponible y a razones conceptuales por las cuales un modelo es ajustado.

4.2.3.5. Selección del modelo

En la modelación por HMM la selección de un modelo apropiado que permita describir el comportamiento de la serie temporal observada viene determinada por el número de estados ocultos. Como se expone en Zucchini & MacDonald (2009), el aumento del número

de estados ocultos mejora el ajuste del modelo, sin embargo, esto se realiza a costa de un incremento cuadrático en el número de parámetros a estimar; es por este motivo que la selección del modelo debe estar determinada por un criterio que permita equilibrar la mejora del ajuste versus la complejidad del modelo (es decir, la cantidad de parámetros a estimar).

La escogencia de un modelo apropiado se ve influenciada por la cantidad de datos de entrenamiento con los que se cuenta en la serie temporal, en este sentido, un estudio de simulación desarrollado por Dugad & Desai (1996) donde se estudió la similaridad entre dos HMM, mediante una medida de distancia basada en la divergencia de Kullback-Leibler, en función de la longitud de la secuencia de observaciones, permitió determinar que la longitud óptima de una serie temporal para realizar la elección adecuada del mejor HMM, debe presentar un orden de aproximadamente $T = 4000$ observaciones.

De este modo, a continuación se presentan dos criterios de selección que permiten balancear las condiciones descritas, permitiendo obtener el modelo más apropiado para el ajuste y posterior inferencia.

Criterio de información de Akaike

El criterio de información de Akaike es un enfoque frecuentista a partir del cual la selección del modelo se realiza a través de la minimización de una medida de distancia entre el modelo ajustado y el que se asume es el verdadero modelo generador de los datos observados. Para este propósito, se recurre a la divergencia de Kullback-Leibler (Zucchini & MacDonald 2009) que bajo ciertas condiciones operativas se simplifica en:

$$AIC = -2 \log L + 2p \quad (4.13)$$

Donde, $\log L$ es el logaritmo de la función de verosimilitud del modelo ajustado y p corresponde al número de parámetros del modelo. El primer término de la expresión (4.13) corresponde a una medida de ajuste del modelo, mientras el segundo es una cantidad de penalización que incrementa cuando se incluyen más estados ocultos, m .

Criterio de información Bayesiano

Desde el punto de vista Bayesiano, la construcción de un criterio que permita realizar la selección del modelo más apropiado para los datos observados viene determinada por la selección del modelo correcto con mayor probabilidad a posteriori. Para esto se define:

$$BIC = -2 \log L + p \log T \quad (4.14)$$

A diferencia del criterio AIC el único término que sufre una modificación es la penalización ($p \log T$), en la cual T es la longitud de la serie temporal.

La diferencia entre los criterios de información AIC y BIC radica en que el criterio BIC seleccionará modelos más parsimoniosos que el AIC, ya que para este último la constante de penalización es $2p$ mientras en el BIC corresponde a $p \log T$. En este sentido, en muestras pequeñas o medianas el criterio AIC tiende a seleccionar modelos con más parámetros de los necesarios (Peña 2002).

4.2.3.6. Inferencia del modelo

La estimación de parámetros por máxima verosimilitud bajo ciertas condiciones de regularidad proporciona estimaciones consistentes, eficientes y asintóticamente normales (Cappé, Moulines & Rydén 2005). Gracias a estas propiedades es posible calcular las varianzas del conjunto de parámetros estimados para realizar contrastes de hipótesis y calcular intervalos de confianza con el fin de estudiar la estabilidad de las estimaciones y examinar posibles asociaciones que se puedan dar entre ellos.

El cálculo de la matriz de varianzas-covarianzas para los parámetros de un HMM puede realizarse a partir de la inversa de la matriz de segundas derivadas parciales de la función de verosimilitud (conocida también como matriz de información de Fisher o matriz Hessiana); sin embargo, dada la complejidad de la función de verosimilitud y la proximidad de las estimaciones de los parámetros a los límites del espacio parametral (probabilidades muy próximas a cero o uno), realizar el cálculo de las segundas derivadas de L se vuelve un asunto de difícil solución (Zucchini & MacDonald 2009).

No obstante, diferentes métodos han sido propuestos para lograr una aproximación asintótica a la matriz de varianzas-covarianzas de los parámetros del modelo (Zucchini & MacDonald 2009). A continuación se describe una prueba de significancia a través del uso de la distribución t -Student con la premisa de normalidad de las estimaciones de los parámetros para una muestra de T observaciones.

Inicialmente se tiene como hipótesis de investigación que los parámetros obtenidos mediante el ajuste del HMM son significativos en el sentido de explicar el conjunto de observaciones obtenidas. Las hipótesis estadísticas a contrastar son las siguientes:

$$H_0 : \gamma = 0 \quad \text{vs} \quad H_a : \gamma \neq 0$$

Suponiendo la veracidad de la hipótesis nula, se define el estadístico de prueba el cual se distribuye t -Student con $T - 1$ grados de libertad, mediante:

$$t_0 = \frac{\hat{\gamma}}{sd(\hat{\gamma})}$$

Donde $sd(\hat{\gamma})$ corresponde a la desviación estándar del parámetro estimado. El procedimiento de prueba calcula t_0 y compara su valor observado con el percentil superior $\alpha/2$ de la distribución t -Student. De este modo, se rechaza la hipótesis nula si:

$$|t_0| > t_{\alpha/2, T-1}$$

También se puede usar el valor $-p$ para tomar la decisión, el cual indica la probabilidad de observar un resultado al menos tan extremo como el que realmente se ha obtenido, de manera que si este es menor o igual al nivel de significancia adoptado se rechaza la hipótesis nula:

$$P(T > |t_0|) = P(T > t_0) + P(T < -t_0)$$

Este test de significancia puede aplicarse para cada uno de los parámetros estimados con el propósito de evaluar su grado de importancia en la construcción del modelo.

4.2.4. Estimación de registros faltantes: método de Cressman

La estimación de los registros diarios faltantes de precipitación en la estación climatológica objeto de estudio se efectúa mediante el método de Cressman (Cressman 1959) descrito en *López, Gutiérrez & de los Santos (1996)* y *Campos (1998)*. Este procedimiento hace uso de los registros de precipitación en estaciones aledañas a la estación de interés durante el mismo día del valor faltante, es decir, tiene en cuenta únicamente el comportamiento espacial de la información. Denotando R_h^j como la precipitación total para el día h en la estación j cuyo registro no es conocido, el cálculo de dicho valor puede realizarse tomando información de n estaciones circundantes de las cuales se disponga de información para el día en cuestión h , ponderando los valores observados en una cantidad igual al recíproco del cuadrado de la distancia entre cada estación vecina y la estación de interés, mediante la ecuación (4.15):

$$R_h^j = \frac{\sum_{i=1}^n R_h^i \left(\frac{1}{d_{ij}^2} \right)}{\sum_{i=1}^n \frac{1}{d_{ij}^2}} \quad (4.15)$$

Dónde:

- R_h^i : corresponde a la precipitación total para el día h (mm) en la estación i , teniendo en cuenta $i \neq j$, $i = 1, 2, \dots, n$.
- d_{ij} : distancia (km) entre cada estación circundante i y la estación j donde se presenta el registro incompleto.

El método puede utilizarse para estimar valores faltantes de orden diario, mensual o anual. Se recomienda utilizar cuatro estaciones auxiliares (las más próximas) de manera que cada una quede localizada en uno de los cuadrantes definidos por los ejes coordenados que pasan por la estación con los registros faltantes; el número mínimo permitido es de dos estaciones.

Capítulo 5

Metodología

En este capítulo se describen las etapas desarrolladas para dar cumplimiento a los objetivos propuestos en la presente investigación, brindando las pautas para el manejo y análisis de datos.

5.1. Revisión de literatura

Se consultó información de fuentes bibliográficas como libros, artículos, documentales, websites y otros, relacionados con la dinámica de la atmósfera y el clima a nivel global y regional; explorando las posibles fuentes de alteración de la variabilidad natural del clima a través de la problemática del cambio climático a nivel mundial y la incidencia sobre nuestro país que se traduce en modificaciones del régimen climático normal esencialmente en la precipitación atmosférica. En segunda instancia, se realizó la revisión de literatura concerniente a la modelación por medio de los procesos estocásticos, en particular mediante los modelos ocultos de Markov, permitiendo obtener las bases teóricas para su aplicación al problema en estudio.

5.2. Datos del estudio

Se dispone de información diaria proporcionada por Cenicaña de 12 estaciones climatológicas pertenecientes a la Red Meteorológica Automatizada (RMA) administrada por dicha entidad, ubicadas en la zona geográfica del valle del río Cauca, cuyo rango de altitud oscila entre los 900 y 1050 msnm¹, para el periodo Enero 1 de 1994 a Diciembre 31 de 2011, con mediciones de variables meteorológicas tales como temperatura (mínima, media, máxima), humedad relativa (mínima, media, máxima), precipitación, radiación solar y evaporación calculada. Para el desarrollo del estudio, la variable seleccionada de esta base de datos correspondió a la precipitación diaria durante la primera temporada lluviosa del año (que comprende los 92 días de los meses de Marzo a Mayo) para la estación San Marcos (Latitud: 3°36'14,75", Longitud: 76°27'15,02", Altitud: 960 msnm) ubicada en el centro del departamento del Valle del Cauca. La longitud de la serie temporal evaluada corresponde a 1656 días (92 días $\times$ 18 años) denotada por T .

¹Metros sobre el nivel del mar

5.3. Validación e imputación de registros faltantes

El objetivo de esta fase se centró en verificar la consistencia de las fechas de los registros de precipitación diaria; esto permitió calcular la cantidad de datos faltantes presentes en la serie temporal en estudio. Los registros faltantes identificados fueron imputados mediante el método de Cressman (Cressman 1959), utilizando la información de la base de datos completa de estaciones cercanas a la estación meteorológica San Marcos. La imputación de los registros faltantes se realizó de la siguiente manera:

- Para el cálculo de un valor faltante en un día determinado, se hizo uso de los registros en el mismo instante temporal de las dos (2) estaciones más próximas. Las estaciones climatológicas más próximas cuyo radio de proximidad a la estación San Marcos no excede los 30 km, se presentan en la Tabla 5.1.

Tabla 5.1: Estaciones climatológicas auxiliares empleadas en el análisis

Estación climatológica	Distancia (km)	Orden de cercanía	Dirección (°)
Guacarí	23.74	4	37.77
Amaime	22.21	2	81.12
Palmira La Rita	19.48	1	99.38
Candelaria	22.81	3	145.26
Meléndez	27.71	5	190.05

En caso que una de las dos estaciones más próximas disponga de un dato faltante para el día en cuestión, se tomó la siguiente estación más próxima comprendida en el límite de distancia considerado. De este modo, la imputación de los registros diarios de precipitación se realizó mediante la expresión:

$$R_h^j = \frac{\sum_{i=1}^2 R_h^i \left(\frac{1}{d_{ij}^2} \right)}{\sum_{i=1}^2 \frac{1}{d_{ij}^2}}$$

Dónde:

R_h^j : precipitación total (mm) durante el día h en la estación climatológica San Marcos.

R_h^i : precipitación total (mm) para el día h en la estación i .

d_{ij}^2 : distancia (km) entre la estación i y la estación San Marcos.

- Cuando no existía información en las estaciones más próximas, la estimación del valor faltante para un día en particular se calculó como el promedio aritmético entre los días anterior y posterior al dato faltante de interés (OMM 2011).

5.4. Transformación de datos

Tomando como referencia el estudio realizado por Tan et al. (2012), los registros de precipitación diaria para el periodo en estudio se clasificaron en: días secos y días húmedos. Un día húmedo (representado por “1”) se definió como un día en que la cantidad total de lluvia es mayor o igual a 0.1 mm, mientras un día seco (“0”) es un día con un registro de precipitación total inferior a 0.1 mm. La discretización de los registros de precipitación permitió modelar el comportamiento de la serie temporal mediante un modelo oculto de Markov con observaciones emitidas por distribuciones de probabilidad Bernoulli por cada estado oculto del modelo.

5.5. Análisis exploratorio de datos

El análisis exploratorio de datos para los registros de precipitación diaria se realizó para la serie temporal sin transformar correspondiente a la primera temporada lluviosa del año para el periodo 1994-2011. Para este propósito se desarrollaron gráficos e indicadores descriptivos, como medidas de centramiento, dispersión y forma de la distribución, esto con el fin de tener un conocimiento previo de la dinámica de la precipitación atmosférica sobre esta temporada lluviosa.

5.6. Ajuste del modelo

El ajuste del modelo que permitió realizar la descripción de los posibles patrones que caracterizan la precipitación diaria durante la primera temporada lluviosa del año, se realizó mediante el software estadístico **R versión 2.15.2** (R Core Team 2012) con la librería **RHmm** (TARAMASCO & Bauer 2013) la cual tiene implementados los algoritmos de estimación descritos en el marco teórico, así como un procedimiento de aleatorización para la elección de los valores iniciales para los parámetros del modelo. En la presente sección se definen los supuestos, elementos y herramientas que fueron utilizadas en el ajuste del modelo oculto de Markov.

5.6.1. Supuestos

Con el propósito de simplificar algunos procedimientos de cálculo, se definen dos supuestos importantes para el ajuste del modelo:

- **No estacionariedad:** se asume que la distribución de probabilidad inicial δ , definida sobre el espacio de estados ocultos $S = \{1, 2, \dots, m\}$ varía respecto al tiempo. Este hecho es de esperar ya que la evolución de la precipitación diaria se ve influenciada por fenómenos que ocurren a distintas escalas de variabilidad temporales y pueden alterar la probabilidad de ocurrencia de los posibles patrones climáticos, por lo tanto sería inadecuado suponer que dicha distribución de probabilidad permanece constante durante el periodo en estudio.
- **Homogeneidad:** considerando los factores a gran escala que inciden sobre la variabilidad climática como lo son: factores antropogénicos, las fases del fenómeno ENSO

y la oscilación de la ZCIT, es de esperar que la matriz de probabilidades de transición entre estados ocultos presente cambios en el transcurso del tiempo, conllevando a la alteración de la probabilidad de lluvia, lo que conduce a pensar que resultaría apropiado la utilización de un proceso de Markov no homogéneo con el objetivo de modelar dicha dinámica. Sin embargo, para el desarrollo de este estudio se consideró un proceso de Markov homogéneo con el fin de simplificar la estimación y el estudio del modelo.

5.6.2. Elementos del modelo

Los elementos que definieron el modelo oculto de Markov ajustado en el presente estudio son:

- Espacio de estados del modelo: $S = \{1, 2, \dots, m\}$
- Observaciones emitidas por estado: $V = \{0, 1\}$
- Probabilidades de transición entre estados γ_{ij} :

$$\gamma_{ij} = P(C_{t+1} = j | C_t = i), \quad i, j \in S, \quad 1 \leq t \leq T$$

Donde, $T = 1656$ corresponde a la longitud de la serie temporal.

- Distribución de probabilidad de las observaciones:

$$p_i(x_t) = P(X_t = x_t | C_t = i), \quad 1 \leq t \leq 1656, \quad i \in S, \quad x_t \in V$$

- Distribución de probabilidad inicial $\delta = (\delta_1, \delta_2, \dots, \delta_m)$ definida sobre el espacio de estados:

$$\delta_i = P(C_1 = i), \quad i \in S$$

Este conjunto de elementos constituyen la base de un HMM, el cual tiene asociados dos procesos estocásticos: uno oculto y otro observable. Tales procesos, en el contexto de modelación se definen a continuación.

5.6.3. Proceso oculto

El proceso oculto $\{C_t : 1 \leq t \leq 1656\}$ que toma valores en el espacio de estados $S = \{1, 2, \dots, m\}$, corresponde a la secuencia de estados del modelo asociados a los posibles *estados* o *patrones del clima* que modifican el régimen de precipitación a escala intraestacional durante la primera temporada lluviosa del año (marzo-mayo) en la zona de estudio. El modelo asume que la dinámica que se presenta entre este conjunto de estados sigue un modelo de Markov de primer orden, por tanto cumple:

$$P(C_t | C_1, \dots, C_{t-1}) = P(C_t | C_{t-1})$$

A partir de este hecho, un modelo de Markov queda descrito completamente por dos elementos esenciales: la distribución de probabilidad inicial y la matriz de probabilidades

de transición. La distribución de probabilidad inicial del modelo δ para un conjunto de m estados ocultos se asume no estacionaria:

$$\delta = (\delta_1, \delta_2, \dots, \delta_m)_{1 \times m}, \quad \delta_i = P(C_1 = i)$$

Mientras, la matriz de probabilidades de transición entre estados se asume homogénea y se denota por:

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \dots & \gamma_{mm} \end{pmatrix}_{m \times m}, \quad \gamma_{ij} = P(C_t = j | C_{t-1} = i), \quad i, j = 1, 2, \dots, m$$

Ya que no se conoce con certeza el número de estados ocultos m que genera la serie de observaciones, se deberán ajustar varios modelos variando el número de estados de 2 a 7, con el propósito estimar el modelo con el número de estados ocultos más apropiado para realizar los análisis posteriores.

5.6.4. Proceso observable

El proceso observable $\{X_t : 1 \leq t \leq 1656\}$ que toma valores en el conjunto $V = \{0, 1\}$, corresponde a la serie de tiempo de los registros transformados (secuencia de días secos y lluviosos) para la primera temporada lluviosa del año durante el periodo 1994-2011. En este sentido, cada una de las variables X_t se define como:

$$X_t = \begin{cases} 1 & \text{si } \text{Día } t \text{ es lluvioso} \\ 0 & \text{si } \text{Día } t \text{ es seco} \end{cases}$$

Por el supuesto de independencia condicional entre las observaciones, se tiene:

$$P(X_t | X_1, \dots, X_{t-1}, C_1, \dots, C_t) = P(X_t | C_t)$$

Además se asume que la distribución de probabilidad asociada a la ocurrencia de lluvia para un día t dado que el proceso se encuentra en el estado i , sigue una distribución de probabilidad Bernoulli con probabilidad de éxito π_i :

$$p_i(x_t) = P(X_t = x_t | C_t = i) = \pi_i^{x_t} (1 - \pi_i)^{1-x_t}, \quad i \in S$$

5.6.5. Valores iniciales de los parámetros

La selección de los valores iniciales para ejecutar el proceso de estimación de parámetros, se realizó mediante la generación de valores aleatorios a partir de una distribución de probabilidad uniforme en el intervalo $[0,1]$, garantizando las restricciones asociadas a cada conjunto de parámetros las cuales presentan a continuación:

Distribución de probabilidad inicial, δ

$$\delta_i \geq 0, \quad \sum_{i=1}^m \delta_i = 1, \quad \forall i, \quad 2 \leq m \leq 7$$

Matriz de probabilidades de transición, $\mathbf{\Gamma}$

$$\gamma_{ij} \geq 0, \quad \sum_{j=1}^m \gamma_{ij} = 1, \quad \forall i, j, \quad 2 \leq m \leq 7$$

Probabilidades de emisión, $\mathbf{P}$

$$p_i(v_k) \geq 0, \quad \forall i$$

5.6.6. Estimación de parámetros

La estimación de los parámetros λ del modelo se realizó mediante la maximización de la función de verosimilitud de los datos observados, $P(\mathbf{X}|\lambda)$, en este sentido se hace uso del algoritmo iterativo Baum-Welch asumiendo no estacionariedad en la distribución de probabilidad inicial del proceso oculto, mediante los siguientes pasos:

1. Asignar los valores iniciales de los parámetros del modelo, denotados mediante $\lambda^{(in)}$, como datos de entrada en el proceso iterativo.
2. Calcular la secuencia esperada de estados ocultos en base a los valores iniciales del paso 1.
3. Reestimar los parámetros del modelo condicionados a la estimación actual de la secuencia de estados ocultos, obteniendo como resultado $\lambda^{(out)}$, los cuales cumplen:

$$P(\mathbf{X}|\lambda^{(out)}) \geq P(\mathbf{X}|\lambda^{(in)})$$

4. Repetir los pasos 2 y 3 hasta que la diferencia entre los parámetros reestimados en dos iteraciones sucesivas sea despreciable, de acuerdo a un límite de convergencia establecido.

Este procedimiento se llevó a cabo variando el número de estados ocultos; iniciando con el ajuste de un HMM con 2 estados ocultos hasta llegar al ajuste de un modelo con 7 estados, esto con la finalidad de seleccionar el modelo que presentara un mejor balance en términos de ajuste y número de parámetros, mediante los criterios de información de Akaike y Bayesiano.

5.6.7. Selección del modelo

La elección del modelo que presentó un balance adecuado entre bondad de ajuste y cantidad de parámetros en el modelo se llevó a cabo a partir de la aplicación de los criterios de información de Akaike (AIC) y Bayesiano (BIC). De este modo, mediante el criterio de información de Akaike se calcula:

$$\text{AIC} = -2 \log L + 2p$$

Donde, $\log L$ corresponde al logaritmo de la verosimilitud obtenida por el modelo ajustado, p es el número de parámetros independientes en el modelo. Mediante la aplicación del criterio de información Bayesiano se realizó el cálculo de:

$$\text{BIC} = -2 \log L + p \log(1656)$$

Este procedimiento se realizó para cada uno de los modelos ajustados variando el número de estados ocultos con lo cual se pudo comparar y seleccionar el modelo más apropiado.

5.6.8. Estimación de la secuencia de estados ocultos

Una vez seleccionado el modelo adecuado para describir el comportamiento de la precipitación diaria durante la primera temporada lluviosa del año, la estimación de la secuencia más probable de estados ocultos $\mathbf{C}^*$ se realizó a partir del algoritmo de Viterbi, buscando maximizar la probabilidad condicional:

$$P(\mathbf{C}|\mathbf{X})$$

A partir de la secuencia estimada de estados ocultos $\mathbf{C}^*$ se buscó clasificar y estudiar los posibles patrones generadores de los registros de precipitación observados. El examen de dicha información se expresa en la siguiente subsección.

5.6.9. Caracterización del modelo

Posterior a la estimación de los parámetros y la secuencia de estados ocultos, el análisis se centró en efectuar la caracterización de los mismos en función de la información disponible. Para este propósito se planteó el desarrollo del análisis desde dos perspectivas complementarias.

1. Análisis del proceso de Markov

La exploración del proceso oculto que describe la evolución de la precipitación diaria durante la primera temporada lluviosa del año es de vital importancia para comprender el comportamiento que presentan los patrones o estados climáticos identificados por el modelo para el periodo en estudio. Para realizar dicho análisis se usaron indicadores descriptivos, utilizando la secuencia de estados ocultos y de pruebas de hipótesis sobre los parámetros estimados.

2. Análisis del proceso observable

La caracterización del proceso observable en función de los estados ocultos identificados por el modelo se realizó a través de indicadores descriptivos utilizando los registros de precipitación con y sin transformación, permitiendo conocer las particularidades que gobiernan el régimen de precipitación sobre la temporada lluviosa de los meses de marzo a mayo para la zona en estudio.

Capítulo 6

Resultados

A continuación se presentan los resultados correspondientes al ajuste de un modelo oculto de Markov con el objetivo de analizar el comportamiento de la precipitación diaria durante la primera temporada lluviosa del año, determinando los estados ocultos que inciden sobre la dinámica de dicho elemento climático, así como también la estimación de la secuencia más probable de estados subyacentes que generan los registros de precipitación observados. Como en todo análisis de tipo estadístico, el presente capítulo inicia explorando la serie temporal a través de indicadores descriptivos mediante el uso de medidas de posición, variabilidad y forma de la distribución de precipitación diaria. Posterior al análisis exploratorio se presenta el ajuste del modelo oculto de Markov, a partir de cual se caracteriza la dinámica de la serie temporal.

6.1. Imputación de registros faltantes

Al examinar la consistencia de las fechas para la serie temporal diaria del total de registros de precipitación durante el periodo 1994-2011 (6574 días), se logra identificar 43 datos faltantes a lo largo de la serie, equivalente al 0.65 % del total. En la Figura 6.1 se presenta la distribución de tales registros faltantes sobre dicha serie temporal.

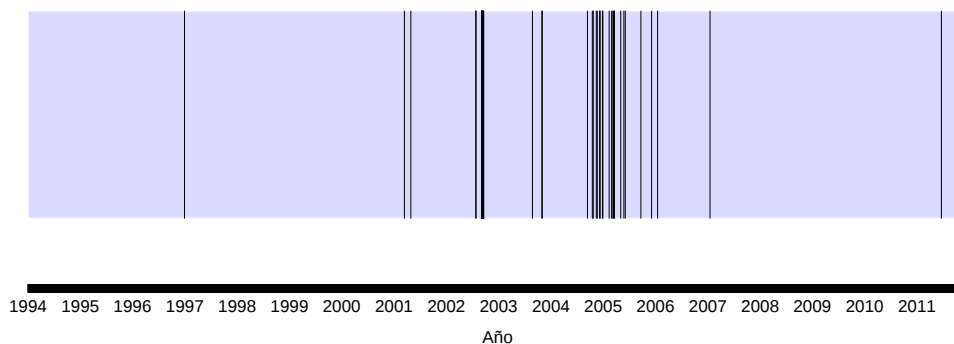


Figura 6.1: Distribución de datos faltantes diarios para el periodo 1994-2011

En dicho gráfico, las líneas verticales corresponden a registros de precipitación faltantes en la estación climatológica San Marcos, evidenciándose para el año 2005 la mayor cantidad de registros faltantes (13 días), mientras los registros restantes se encuentran dispersos principalmente en el periodo 2001 a 2007, con dos excepciones para los años 1996 y 2011. La baja proporción de datos perdidos valida la imputación a través del método de Cressman, permitiendo completar tal porcentaje de registros faltantes para dicha estación.

6.2. Análisis exploratorio de datos

La interacción de la atmósfera con los diferentes componentes del sistema climático genera variaciones espacio-temporales en la dinámica de elementos atmosféricos, como la precipitación, la cual se ve influenciada por el desarrollo de eventos que ocurren a diversas escalas de variabilidad. Tal variabilidad puede verse enmarcada en eventos que se presentan a diferentes intervalos de tiempo; en particular para la región occidental de Colombia, se destaca la ocurrencia del fenómeno ENSO que se presenta a intervalos de años (escala interanual), variaciones a escala intraestacional principalmente por periodos cortos de lluvias (IDEAM 2001), así como fenómenos a gran escala entre los que se destaca el aumento gradual del cambio climático generado por factores antropogénicos (Vide 2008).

Uno de los principales determinantes del ciclo anual de precipitación sobre la región plana del valle geográfico del río Cauca es la variabilidad a escala estacional, la cual se ve influenciada por el desplazamiento de la ZCIT originando un régimen de precipitación bimodal en el transcurso del año, caracterizada por dos trimestres lluviosos correspondientes a los meses de marzo a mayo y de septiembre a noviembre, como se muestra en la Figura 6.2 en donde se presenta el ciclo medio estacional de la precipitación a lo largo del año. Dicho gráfico se calcula a partir de la serie mensual de precipitación acumulada para el periodo 1994-2011, compuesta por un total de 216 meses (12 meses $\times$ 18 años).

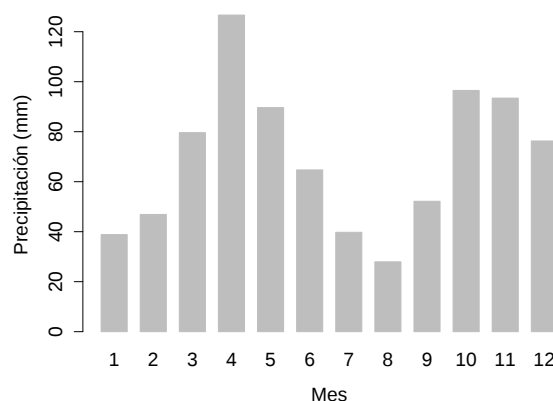


Figura 6.2: Ciclo medio estacional de la precipitación a lo largo del año

En este sentido, si se pretendiera estudiar el comportamiento de la serie temporal de precipitación para todo el periodo en estudio (serie de registros diarios comprendiendo los meses de enero a diciembre para los años 1994-2011), sería necesario incluir en el modelo la

componente de variabilidad estacional que ocurre a lo largo de todos los años. No obstante, con el objetivo de controlar dicha componente de variabilidad se decide estudiar únicamente la primera temporada lluviosa del año, correspondiente a los meses marzo-mayo para el desarrollo del presente estudio. Este hecho permite reducir la complejidad implícita en el ajuste del modelo, capturando un gran porcentaje de la variabilidad observada a escala intraestacional e interanual.

En consecuencia, a partir de este punto se presentan los resultados correspondientes al análisis de la precipitación diaria durante la primera temporada lluviosa para el periodo 1994-2011. En la Figura 6.3 se muestra la serie temporal en estudio, a la cual se sobrepone las fases del fenómeno ENSO¹ con el fin de explorar una posible incidencia de dicho fenómeno sobre la secuencia de observaciones para la temporada estudiada.

De este modo, en contraste a lo esperado no existe una clara demarcación de la incidencia del fenómeno ENSO sobre la precipitación diaria para la primera temporada lluviosa de cada año, ya que en algunos años donde se presentó la fase Niña (años 1999, 2001 y 2008) los datos de precipitación registraron valores inferiores a los que se presentaron en años donde ocurrió la fase Niño del fenómeno (años 2000 y 2007).

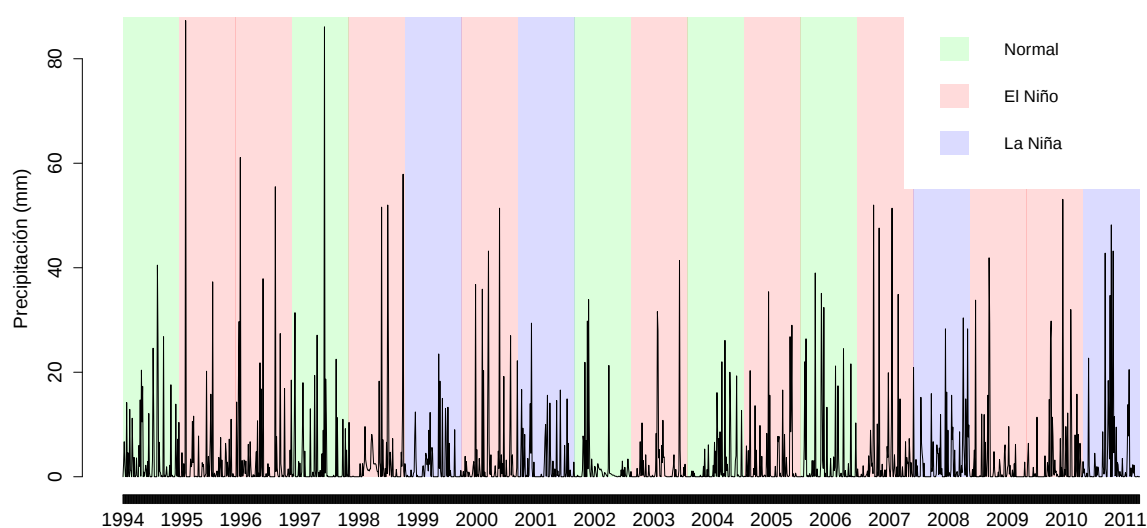


Figura 6.3: Serie de precipitación diaria durante la primera temporada lluviosa del año, periodo 1994-2011

Este hecho se debe a que la intensidad del fenómeno ENSO ocurre de manera variada durante el año en que se presenta, luego la etapa más fuerte del fenómeno bien puede haberse presentado por fuera del periodo marzo-mayo, y en consecuencia los registros de

¹La clasificación de las fases del fenómeno ENSO es realizada por parte del IDEAM mediante un consenso de tres índices macroclimáticos como se expresa en IDEAM (2010) (ver Tabla 4.1 en el Capítulo 4).

precipitación dentro de dicho periodo pueden no capturar esta información. Por otro lado, se destaca el comportamiento no estacionario de la serie temporal de precipitación, es decir, distribución no constante a través del tiempo, exhibiendo transiciones cortas de periodos de lluvia a periodos secos, los cuales se enmarcan en la escala intraestacional.

La distribución de precipitación diaria durante la primera temporada lluviosa de cada año, se presenta en la Figura 6.4, además de las estadísticas descriptivas básicas presentadas en la Tabla 6.1, las cuales ayudan a la descripción del conjunto de datos. El valor mínimo corresponde a 0, indicando la ocurrencia de días secos dentro de la temporada lluviosa, mientras la precipitación máxima es de 87.3 mm.

El nivel promedio de precipitación diario corresponde a 3.22 mm, valor que está alejado por la derecha de la mediana (0.10 mm) indicando asimetría positiva (Coef. Asimetría = 4.44), es decir, la existencia de registros de precipitación elevados o por encima de dicho promedio de nivel de lluvias. Además se presenta un coeficiente de variación de 251.86 % señalando que este elemento climático posee un grado de dispersión significativamente elevado a pesar de la omisión de la variabilidad a escala estacional, por tanto un indicador de tendencia central como la media explicaría muy poco en la descripción de estos datos.

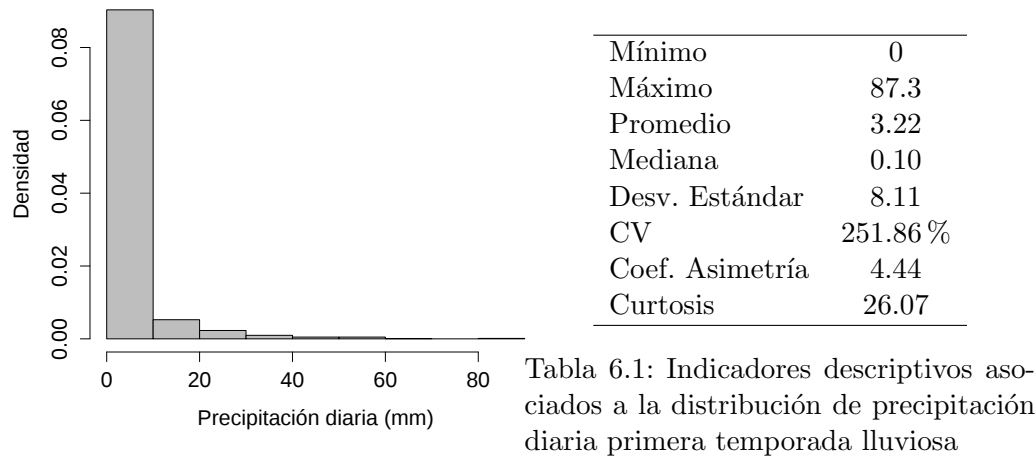


Figura 6.4: Distribución de precipitación diaria primera temporada lluviosa

En la Figura 6.5 se presenta el gráfico correspondiente a la distribución acumulada de precipitación diaria, donde se muestra que el 49 % de registros presentan niveles de cero milímetros, indicando que a pesar de elegir la primera temporada lluviosa, comprendida entre los meses de marzo a mayo, se registran aproximadamente la mitad de días secos, sin embargo, con días lluviosos repartidos a lo largo de toda serie, para el 51 % de casos (Figura 6.3). Además el 90 % de registros corresponden a precipitaciones por debajo de 9.6 mm.

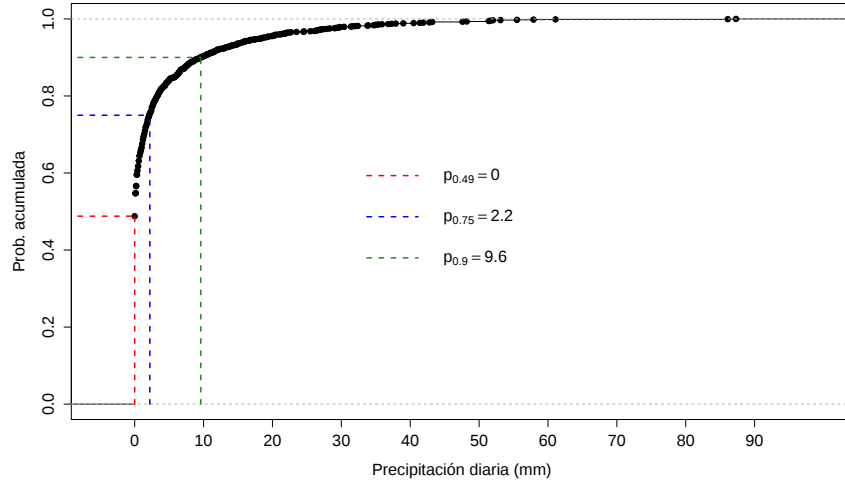


Figura 6.5: Distribución acumulada de la precipitación diaria durante la primera temporada lluviosa

En la siguiente sección se presentan los resultados correspondientes a la caracterización o análisis del comportamiento de la serie de precipitación, mediante el ajuste de un HMM. Previo a esto, se clasificaron los registros de precipitación diaria en días secos y días húmedos. Un día seco (representado por “0”) se definió como un día en el que la cantidad total de lluvia es inferior a 0.1 mm, mientras un día húmedo (“1”) es un día cuyo registro de precipitación es mayor o igual a 0.1 mm. Dicha discretización, permitió modelar el comportamiento de la serie temporal mediante un modelo oculto de Markov con observaciones emitidas por distribuciones de probabilidad Bernoulli por cada estado oculto del modelo.

6.3. Caracterización de la serie de precipitación a través del HMM

Posterior al análisis exploratorio y descriptivo de datos, la caracterización de la serie de precipitación se desarrolló a través de un modelo oculto de Markov con el propósito de identificar y clasificar los posibles estados o patrones climáticos que inciden sobre la dinámica de precipitación, los cuales pueden estar influenciados por fenómenos que actúan a diferentes escalas de variabilidad. Para este fin, se realizó el ajuste de varios modelos variando el número de estados ocultos en un rango de 2 a 7 estados, evaluando su desempeño mediante los criterios de información de Akaike (AIC) y Bayesiano (BIC), respectivamente. Los resultados del análisis se presentan en la Tabla 6.2.

Tabla 6.2: Comparación de modelos mediante criterios de información

Estados ocultos	Número de parámetros	AIC	BIC
2	5	2220.75	2247.81
3	11	2232.70	2292.23
4	19	2242.36	2345.19
5	29	2259.98	2416.94
6	41	2276.30	2498.20
7	55	2304.12	2601.79

Al examinar los resultados asociados a la evaluación de los modelos, se destaca el incremento cuadrático del número de parámetros a medida que se involucran estados ocultos, lo cual es poco conveniente en la determinación de un modelo parsimonioso. Por este motivo se selecciona el modelo con los menores valores AIC y BIC, tal que el balance entre la bondad de ajuste y el número de parámetros del modelo sea apropiado. Se evidencia como los criterios de información señalan la elección de un modelo con dos estados ocultos, destacando que el BIC indica la escogencia de un modelo más parsimonioso, a medida que aumentan el número de estados en comparación al criterio AIC (Figura 6.6), este último según Peña (2002) en muestras pequeñas o medianas selecciona modelos con más parámetros de los necesarios. Con base en la información obtenida, se selecciona un modelo oculto de Markov con dos estados ocultos cuyas parámetros estimados corresponden a:

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0,814 & 0,186 \\ 0,282 & 0,718 \end{pmatrix}$$

$$\hat{\pi}_1 = P(X_t = 1|C_t = 1) = 0,766 \quad \hat{\pi}_2 = P(X_t = 1|C_t = 2) = 0,127$$

Una mirada rápida de las estimaciones de los parámetros del modelo, indica que se cuenta con dos estados ocultos, relacionados con patrones climáticos a escala intraestacional, durante las temporadas de marzo a mayo de cada año, los cuales tienen una dinámica interna que se rige por un proceso de Markov de primer orden con matriz de probabilidades de transición $\hat{\mathbf{\Gamma}}$ y con probabilidades diarias de lluvia del orden de 0.77 y 0.13 para el primer y segundo estado oculto del modelo, respectivamente.

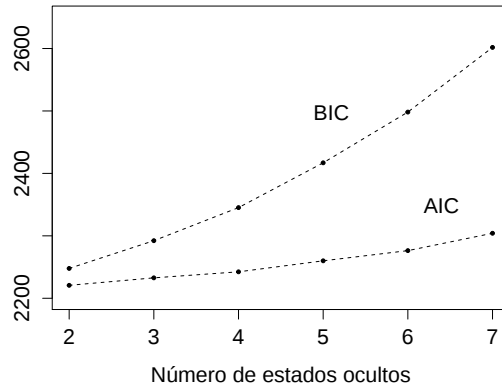


Figura 6.6: Selección del modelo por criterios AIC y BIC

Tabla 6.3: Indicadores descriptivos de precipitación para los estados identificados

Estado oculto	Frecuencia (días)	Frecuencia (%)	Probabilidad de lluvia
1	1033	62.4	0.766
2	623	37.6	0.127

Algunas estadísticas descriptivas básicas se presentan en la Tabla 6.3 como resultado del uso de la secuencia de observaciones transformadas. De este modo, se destaca que el primer estado del modelo, de ahora en adelante etiquetado como *Estado 1*, tiene asociada una alta probabilidad de lluvia (0.766) y es el estado de mayor ocurrencia durante todo el periodo según la secuencia estimada con 1033 días (62.4 %) del periodo en estudio. Por otro lado, el segundo estado identificado, *Estado 2*, muestra una baja probabilidad de lluvia (0.127) y tiene una baja proporción de aparición durante toda la temporada estudiada, con un 37.6 % del total.

Dichas estimaciones representan la dinámica interna del proceso y definen la caracterización del proceso observable a través de los estados o patrones climáticos identificados, es por eso que en la siguiente sección se pretende dar una caracterización del proceso oculto posible generador de la serie temporal en estudio.

6.3.1. Dinámica del proceso oculto

La dinámica del proceso oculto, posible generador de la serie observada, se estudia a través de las propiedades asociadas al proceso Markoviano representado por la distribución inicial y matriz de probabilidades de transición del modelo. En este sentido, ya que se cuenta con una cadena de Markov con espacio de estados finito y probabilidades de transición positivas, la cadena que describe el proceso es ergódica, con distribución límite asociada:

$$\hat{\delta} = (0,603 \quad 0,397)$$

Este resultado indica que a la larga el proceso tiene una probabilidad aproximada de 0.60 de encontrarse en el *Estado 1*, independiente del estado en que se halle el proceso en instantes de tiempo previos, mientras la probabilidad de permanecer en el *Estado 2* (estado con una baja probabilidad de lluvia) oscila alrededor de 0.40. Esta información permite caracterizar el comportamiento del proceso a largo plazo, así como el tiempo medio de recurrencia, es decir, el tiempo promedio necesario para que un estado retorne a si mismo después de haber hecho una transición a otro estado.

El tiempo medio de recurrencia del primer estado oculto del modelo resulto ser aproximadamente de 2 días, es decir, de retornar al estado climático con mayor probabilidad de lluvia, mientras que dicho tiempo para el *Estado 2* se encuentra alrededor de 3 días. Se destaca el hecho que no hay una diferencia marcada entre los tiempos medios de recurrencia para el retorno a un determinado estado del proceso.

Ahora con el fin de explorar la significancia de los parámetros estimados, se presenta en la Tabla 6.4 las estimaciones máximo verosímiles y errores estándar asociados a cada parámetro. Mediante dicha información se efectuaron test de significancia para evaluar el grado de importancia de las estimaciones en la descripción de la serie temporal, utilizando un nivel de significancia $\alpha = 0,05$.

Con base en los resultados obtenidos mediante las pruebas, se puede concluir que todos los parámetros estimados son significativos y permiten caracterizar la serie de tiempo de los registros de precipitación diaria sobre la primera temporada lluviosa del año, a excepción de la probabilidad de lluvia cuando el proceso se encuentra en el *Estado 2*.

Tabla 6.4: Coeficientes estimados del modelo ajustado

Parámetro	Estimación	Error estándar	t	p-value
$\gamma_{1,1}$	0.814	0.061	13.307	$< 2e-16$
$\gamma_{1,2}$	0.186	0.061	3.037	0.0024
$\gamma_{2,1}$	0.282	0.076	3.715	0.0002
$\gamma_{2,2}$	0.718	0.076	9.456	$< 2e-16$
π_1	0.766	0.062	12.344	$< 2e-16$
π_2	0.127	0.079	1.620	0.1053

Por otro lado, el grado de asociación que existe entre las probabilidades de lluvia para cada estado oculto del modelo (π_1, π_2) , corresponde a 0.299, indicando que estos parámetros muestran un leve grado de asociación positiva, siendo esta correlación importante

mediante el test de significancia de Pearson. Este hecho señala que al aumento de la probabilidad de lluvia dado el *Estado 1*, produce un leve aumento de posibilidades de lluvia dado el *Estado 2*, el cual posee baja probabilidad de precipitación.

Además de estudiar el tiempo de recurrencia, se analiza el comportamiento de las distribuciones del tiempo de permanencia en cada estado subyacente, construidas a partir de la secuencia estimada de estados ocultos. Dicha información se presenta en la Figura 6.7 y la Tabla 6.5.

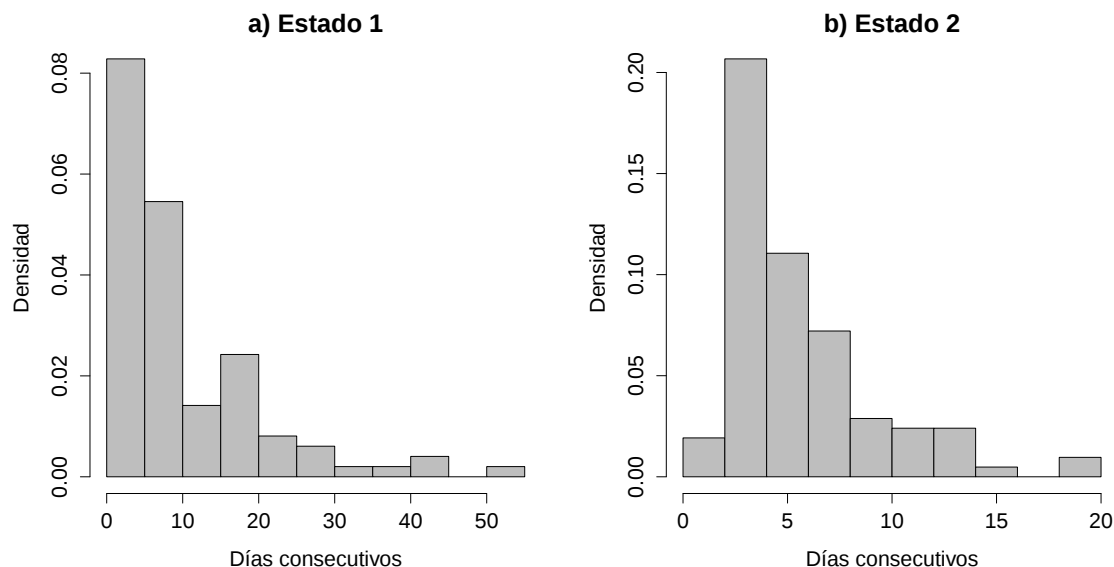


Figura 6.7: Distribución del tiempo de permanencia por estado oculto

Examinando los resultados descriptivos correspondientes de la distribución de tiempos de permanencia por cada estado, el periodo medio de duración en el *Estado 1* es de aproximadamente 11 días, mucho mayor a la permanencia en el *Estado 2* (aproximadamente 6 días), caracterizado por un patrón generador de precipitación con baja probabilidad. Este hecho, destaca la permanencia de días lluviosos dentro de una temporada de intensas precipitaciones, como lo es el periodo de marzo a mayo de cada año.

La distribución del tiempo de permanencia en el *Estado 1* posee un coeficiente de variación del 92.83%, indicando alta variabilidad en términos de días de permanencia, desde un mínimo de 2 hasta un máximo de 51 días en los cuales se presenten precipitaciones con alta probabilidad. Dicha distribución posee asimetría positiva sugiriendo que la curva de valores no está centrada y sus datos están acumulados a la izquierda, es decir mayor frecuencia de días de permanencia alrededor de 7 días, valor correspondiente a la mediana o cuartil 2, el cual indica que el 50% de los días de permanencia es menor o igual a 7 fechas. Por su lado el cuartil 3, indica que el 75% de días de permanencia es menor o igual a 14, en la presencia del *Estado 1*, destacado por un patrón generador de lluvias con alta probabilidad.

Tabla 6.5: Indicadores descriptivos del tiempo de permanencia por estado oculto

	Estado 1	Estado 2
Mínimo	2	1
Máximo	51	19
Promedio	10.43	5.99
Moda	3	3
Total	1033	623
Desv. Estándar	9.68	3.62
CV	92.83 %	60.42 %
Q1	4	3
Q2	7	5
Q3	14	7
Coef. Asimetría	1.87	1.49
Curtosis	3.73	2.04

Por otro lado, el tiempo de duración en el *Estado 2* posee un coeficiente de asimetría de 1.49, indicando que los valores se encuentran acumulados a la izquierda y por tanto que el valor del promedio se ubica a la derecha de la mediana, la cual es de 5 días, correspondiente al tiempo de permanencia dado el *Estado 2*. Además esta distribución presenta menor grado de dispersión (CV=60.42%), en comparación a la distribución del tiempo de permanencia en el *Estado 1*, oscilando entre valores desde 1 hasta 19 días. Es de destacar que el 75 % de los días de permanencia en el *Estado 2* es menor o igual a 7, valor que coincide con la mediana (Q2) de la distribución del tiempo de permanencia en el Estado 1, corroborando su mayor permanencia de días, en comparación al *Estado 2*.

6.3.2. Descripción y caracterización de los estados ocultos

La caracterización de los estados ocultos del modelo ajustado permite describir el comportamiento de la serie temporal de precipitación, identificando los posibles patrones que inciden sobre la generación de lluvias para el periodo marzo-mayo de cada año. De conformidad con esto, en la Figura 6.8 se expone el gráfico correspondiente a la serie temporal de los registros de precipitación, junto con la secuencia de estados ocultos más probable.

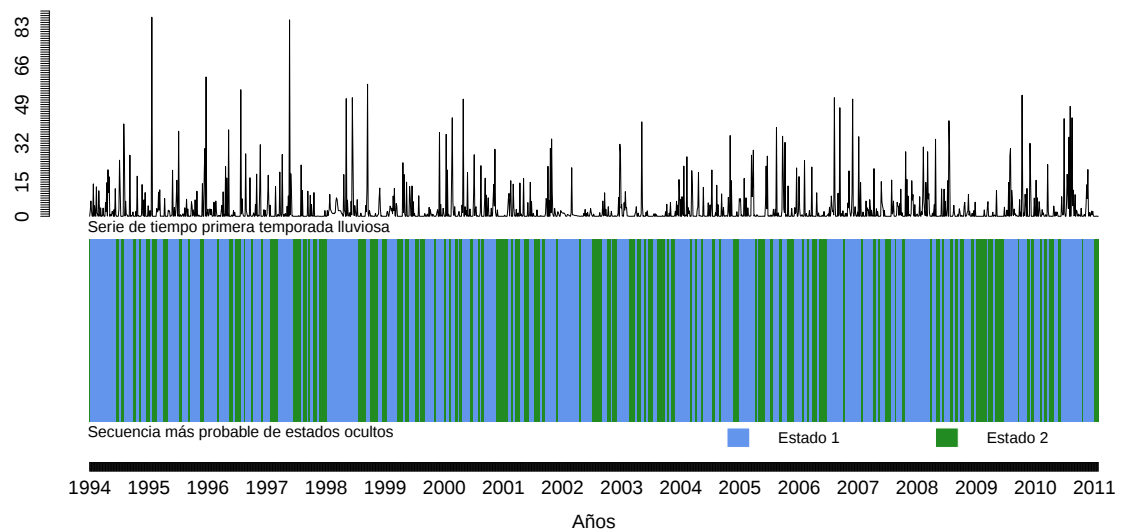


Figura 6.8: Serie de precipitación diaria durante la primera temporada lluviosa del año junto con la secuencia más probable de estados ocultos

Luego al examinar la clasificación de la secuencia de estados más probable (gráfico inferior de la Figura 6.8), se tiene que el *Estado 1* cuya probabilidad de lluvia es de 0.766, se destaca por aparecer en aquellos días donde se presentaron elevados niveles de precipitación, caso contrario a lo que sucede con la ocurrencia del *Estado 2* donde se observan periodos cortos de días con escasas precipitaciones. Este hallazgo indica que los estados ocultos se pueden caracterizar por eventos ocurrentes que generan fluctuaciones de la precipitación a escala intraestacional, esto debido a que no se cuenta con suficiente información sobre la influencia de la variabilidad interanual para el periodo en estudio. En consecuencia, a continuación se presenta una caracterización de los estados identificados a través de estadísticas descriptivas por estado oculto (Tabla 6.6, Figura 6.9).

Estado 1

Este estado se distingue por presentar un rango de variación que va de 0 a 87.3 mm, capturando una proporción de 88 % de días con registros de precipitación no nulos. Los niveles de lluvia en el *Estado 1* presentan un promedio de 5 mm mostrando un elevado grado de variabilidad ($CV=195.15\%$) debido a la fuerte asimetría de la distribución por

la presencia de registros con altos niveles de precipitación. Los resultados indican un total de precipitación de 5168.66 mm durante la primera temporada, en comparación al *Estado 2* donde tan solo se evidencia un total de precipitación 157.6 mm (Tabla 6.6), por la baja probabilidad de lluvia generada por este estado.

Tabla 6.6: Indicadores descriptivos por estado oculto

	Estado 1	Estado 2
Mínimo	0	0
Máximo	87.30	20.90
Promedio	5.00	0.25
Mediana	1.10	0
Total	5168.66	157.60
Desv. Estándar	9.76	1.61
CV	195.15 %	638.16 %
Coef. Asimetría	3.52	8.56
Curtosis	16.31	80.78

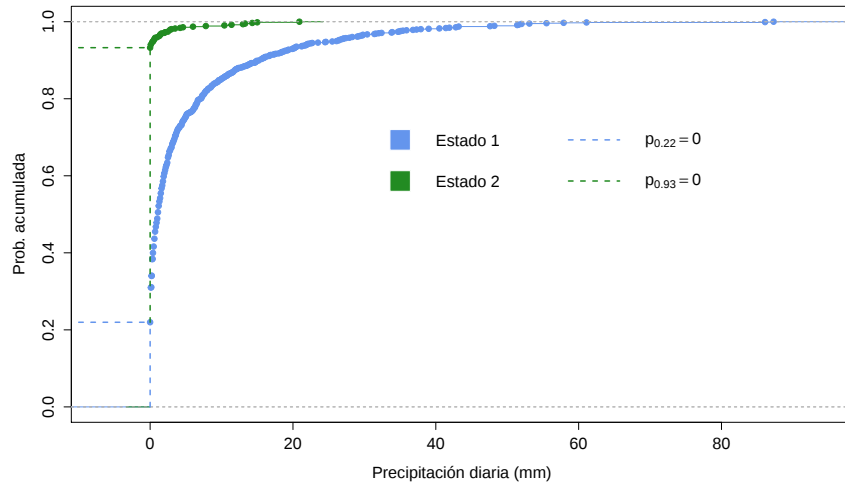


Figura 6.9: Comparación de la distribución acumulada de precipitación diaria por estados ocultos

Estado 2

Los días clasificados en el *Estado 2* se caracterizan por presentar el 93 % de registros sin lluvia, según el percentil 93 (Figura 6.9), corroborando la baja probabilidad de precipitación en la presencia de este estado. Por otro lado, los niveles de lluvia tienen una media de 0.25 mm, valor que esta hacia la derecha de la mediana (0 mm), indicando una pronunciada asimetría positiva (8.56), es decir la mayor concentración de valores alrededor de dicha mediana. También se evidencia un rango de variación que va de 0 mm a 20.9 mm de precipitación por día.

Capítulo 7

Conclusiones

- La distribución de la precipitación diaria durante la primera temporada lluviosa del año (marzo-mayo) presenta una fuerte asimetría positiva (4.44) por la presencia de días sin lluvia, con un elevado grado de variabilidad ($CV=251.86\%$). Este hecho indica que a pesar de controlar la variabilidad a escala estacional tomando la primera temporada lluviosa del año, se deja en evidencia el gran porcentaje de variabilidad implícito a escala intraestacional.
- Los modelos ocultos de Markov tienen una gran utilidad como herramienta descriptiva para la exploración de series de temporales en diversos áreas de estudio, donde se tiene un limitado conocimiento de los componentes generadores de los registros observados. En el presente estudio se mostró una aproximación a la dinámica de la variabilidad a escala intraestacional, permitiendo identificar dos posibles estados generadores de la precipitación diaria durante la primera temporada lluviosa del año.
- El primer estado oculto se encuentra asociado a un régimen de precipitación más pronunciado con alta probabilidad de lluvia (0.766) debido a que del total de registros de precipitación en dicho estado aproximadamente el 88 % de los días se presentaron lluvias por encima de 0.1 mm.
- Por otro lado el segundo estado identificado corresponde a un estado con baja probabilidad de lluvia (0.127), el cual se encuentra caracterizado por presentar una mayor cantidad de días sin precipitación (con un 93 % del total de días en ese estado). Dicho estado es menos frecuente que el estado anterior en términos del número de días de permanencia sobre el mismo (en promedio 6 días).
- Los rangos e interpretación de los estados ocultos son indicadores descriptivos que se utilizan para caracterizar los principales rasgos de los posibles patrones climáticos que gobiernan el régimen de precipitación durante la primera temporada lluviosa del año. Por esta razón deben entenderse como una medida netamente descriptiva y no pueden ser generalizados como una clasificación precisa (Zucchini & MacDonald 2009).

Capítulo 8

Recomendaciones

El campo de acción de los modelos ocultos de Markov en el estudio de variables climatológicas (precipitación, temperatura, radiación solar, etc.) y de una gran variedad de fenómenos es muy amplio; y constituye un largo camino por recorrer, en el cual se pueden alcanzar grandes logros. A continuación se exponen algunas recomendaciones que pueden verse como un punto de inicio para el desarrollo de futuras investigaciones, con el fin ampliar el conocimiento de fenómenos climáticos a partir del uso de los HMM.

- En lugar de discretizar los registros de precipitación atmosférica diaria, se recomienda utilizar distribuciones continuas para efectuar el ajuste del modelo; se espera que esto contribuya a dar una mejor explicación de los estados ocultos identificados y que puedan ser analizados desde una perspectiva mucho más general (Zucchini & MacDonald 2009).
- Los resultados que se presentan deben interpretarse como una metodología apropiada para el análisis de series temporales de datos cuantitativos y categóricos, permitiendo evitar el cumplimiento de supuestos de difícil validación (Visser 2011).
- Para probar el supuesto de homogeneidad de la cadena de Markov sobre el proceso de estados ocultos, se recomienda dividir la serie temporal por partes y ajustar el modelo en cada sub-serie, para verificar si los valores de la matriz de probabilidades de transición no varían significativamente (Mattson & Thorburn 1989).
- Con la finalidad de obtener mejores resultados en términos de la explicación de los estados ocultos identificados por el modelo, se recomienda hacer uso de covariables asociadas a variables que describan los patrones de circulación atmosférica, entre ellos presión atmosférica, temperatura del aire, gradientes de presión y altura, velocidad del viento y otros; como se trata en los estudios de Hughes et al. (1999) y Robertson et al. (2003), utilizando series temporales de varias estaciones en estudio.
- Explorar métodos para dar las semillas o valores iniciales para el proceso de estimación de parámetros, posiblemente explorar otras técnicas de estimación permitiendo utilizar procedimientos de inferencia estadística sobre los parámetros del modelo (Rabiner 1989).
- Explorar otras metodologías que permitan involucrar una relación de dependencia entre los registros observados, es decir las observaciones de la serie temporal, como

se propone en el artículo de Li (2005), donde además de existir dependencia entre los estados ocultos que generan las observaciones también se toma en consideración la relación de dependencia entre las observaciones.

Bibliografía

- ACR (2009), ‘Departamento del Valle del Cauca, diagnóstico socioeconómico’, Alta Consejería para la Reintegración, Presidencia de la República de Colombia.
- Bulla, J. (2006), Application of Hidden Markov Models and Hidden Semi-Markov Models to Financial Time Series, Dissertation, Georg-August-University of Gottingen, Gottingen, Germany.
- Bulla, J. & Berzel, A. (2008), ‘Computational issues in parameter estimation for stationary hidden Markov models’, *Computational Statistics* **23**, 1–18.
- Campos, D. (1998), *Procesos del ciclo hidrológico*, 3 edn, Universidad Autónoma de San Luis de Potosí, México, chapter 4, pp. 4.12–4.14.
- Cappé, O., Moulines, E. & Rydén, T. (2005), *Inference in Hidden Markov Models*, Springer, chapter 12. Statistical properties of the maximum likelihood estimator, pp. 441–469.
- Carvajal, Y., Jiménez, H. & Materón, H. (1998), ‘Incidencia del fenómeno ENSO en la hidroclimatología del valle del Río Cauca - Colombia’, *Bulletin de l’Institut Français d’Etudes Andines* **27**(3), 743–751.
- Cressman, G. (1959), ‘An operational objective analysis system’, *Monthly Weather Review* **87**(10), 367–374.
- Dugad, R. & Desai, U. (1996), A tutorial on Hidden Markov Models, Technical Report 96.1, Indian Institute of Technology, Powai, Mumbai, India.
- Enloe, J. (2012), ‘Southern Oscillation Index (SOI)’, National Oceanic and Atmospheric Administration, NOAA, sitio web, entrada 21 de Agosto de 2012, consultado el 27 de Febrero de 2013. <http://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi.php>.
- Forney, D. (1973), ‘The viterbi algorithm’, *Proceedings of the IEEE* **61**(3), 268–278.
- Fraser, A. (2008), *Hidden Markov models and dynamical systems*, Library of Congress.
- Grimmett, G. & Stirzaker, D. (2001), *Probability and Random Processes*, 3 edn, Oxford University Press, Great Clarendon Street, Oxford, Great Britain, chapter 6, pp. 213–296.
- Hughes, J., Guttorp, P. & Charles, S. (1999), ‘A nonhomogeneous Hidden Markov Model for precipitation occurrence’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48**(1), 15–30.

- ICA (2012), 'Alertas sanitarias y fitosanitarias en función de la predicción climática, recomendaciones de manejo Fenómeno de La Niña', Publicaciones Instituto Colombiano Agropecuario.
- IDEAM (2001), *El medio ambiente en Colombia*, 2 edn, Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá D.C.
- IDEAM (2005), *Atlas Climatológico de Colombia*, Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá D.C.
- IDEAM (2010), *Estudio Nacional del Agua 2010*, Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá D.C.
- Leon, G., Zea, J. & Eslava, J. (2000), 'Circulación general del trópico y la Zona de Confluencia Intertropical en Colombia', *Meteorología Colombiana* (1), 31–38.
- Lezaun, M. (2002), 'Predicciones del tiempo y matemáticas', *Boletín de la Sociedad Española de Matemática Aplicada* (22), 61–100.
- Li, Y. (2005), 'Hidden Markov models with states depending on observations', *Pattern Recognition Letters* **26**, 977–984.
- López, C., Gutiérrez, C. & de los Santos, H. (1996), Comparación de metodologías para la imputación de la lluvia diaria en una pequeña cuenca de latitudes medias, *in* 'IX Congreso Brasileiro de Meteorología', pp. 125–129.
- Martínez, M. (2011), 'Diagnostico situación de riesgos hidrometeorológicos en Colombia y avances en la zonificación de riesgos'. Publicaciones Instituto de Hidrología, Meteorología y Estudios Ambientales.
- Mattson, A. & Thorburn, D. (1989), 'A simple check of the time homogeneity of Markov chains', *Journal of Forecasting* **8**, 65–72.
- Maturana, J., Bello, M. & Manley, M. (2004), 'Antecedentes históricos y descripción del fenómeno El Niño/Oscilación del Sur', Servicio Hidrográfico y Oceanográfico de la Armada de Chile, Departamento de Oceanografía. Valparaíso, Chile.
- Mesa, O., Poveda, G. & Carvajal, L. (1997), *Introducción al clima de Colombia*, 1 edn, Universidad Nacional de Colombia, Medellín, Colombia.
- MinAgricultura (2012), Plan de mitigación de los efectos del Fenómeno El Niño en el sector agropecuario, acuícola y pesquero, Technical Report 2, Ministerio de Agricultura y Desarrollo Rural, Bogotá D.C.
- Montealegre, F. (2009), 'Curso básico de Meteorología y Climatología', Universidad Nacional de Colombia, sede Palmira, sitio web, entrada 30 de Octubre de 2009, consultado el 25 de Febrero de 2013. http://www.virtual.unal.edu.co/cursos/sedes/palmira/5000134/docs_curso/introduccion.html.
- Montealegre, J. & Pabon, J. (2000), 'La variabilidad climática interanual asociada al ciclo El Niño, La Niña, Oscilación del Sur y su efecto en el patrón pluviométrico de Colombia', *Meteorología Colombiana* (2), 7–21.

- Murphy, K. (2012), *Machine learning a probabilistic perspective*, The MIT Press, chapter 17. Markov and hidden Markov models, pp. 589–621.
- NASA (2000), ‘The Intertropical Convergence Zone’, Earth Observatory, National Aeronautics and Space Administration, sitio web, entrada 12 de Julio de 2000, consultado el 1 de Marzo de 2013. <http://earthobservatory.nasa.gov/IOTD/view.php?id=703>.
- OMM (2001), ‘Tiempo, clima y seguridad alimentaria’, *Organización Meteorológica Mundial*.
- OMM (2011), Guide to climatological practices, Technical Report 100, World Meteorological Organization, Geneva, Switzerland.
- Peña, D. (2002), *Análisis de datos multivariantes*, McGraw-Hill, Madrid, España, chapter 11, pp. 311–343.
- Puertas, O., Carvajal, Y. & Quintero, M. (2011), ‘Estudio de tendencias de la precipitación mensual en la cuenca alta-media del río Cauca, Colombia’, *DYNA* (169), 112–120.
- R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>.
- Rabiner, L. (1989), ‘A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition’, *Proceedings of the IEEE* **77**(2), 257–286.
- Rincón, L. (2007), Curso intermedio de probabilidad, Technical report, Universidad Nacional Autónoma de México, Departamento de Matemáticas, México.
- Rincón, L. (2012), Introducción a los procesos estocásticos, Technical report, Universidad Nacional Autónoma de México, Departamento de Matemáticas, México.
- Robertson, A., Kirshner, S. & Smyth, P. (2003), Hidden Markov Models for modeling daily rainfall occurrence over Brazil, Technical report, Information and Computer Science, University of California, Irvine, Irvine, California, United States.
- Shein, K. (2012), ‘ENSO Technical Discussion’, National Oceanic and Atmospheric Administration, NOAA, sitio web, entrada 21 de Agosto de 2012, consultado el 27 de Febrero de 2013. <http://www.ncdc.noaa.gov/teleconnections/enso/enso-tech.php>.
- Sivakumar, M. (2004), Predicción climática y agricultura, in ‘Servicios de Información y Predicción del Clima y Aplicaciones Agrometeorológicas para los países Andinos’, Organización Meteorológica Mundial, Ginebra, Suiza, pp. 29–41. Actas de la Reunión Técnica llevada a cabo en Guayaquil, Ecuador, del 8 al 12 de diciembre de 2003.
- Tan, W., Yusof, F. & Yusop, Z. (2012), Using Hidden Markov Model in analysis of daily rainfall occurrence over Peninsular Malaysia, in ‘2nd International Conference on Water Resources ICWR’, Bayview Hotel, Langkawi Kedah, Malaysia.
- TARAMASCO, O. & Bauer, S. (2013), *RHmm: Hidden Markov Models simulations and estimations*. R package version 2.0.3, <http://CRAN.R-project.org/package=RHmm>.

- Vide, J. M. (2008), ‘La nueva realidad del calentamiento global, un decálogo del cambio climático’, X Coloquio Internacional de Geocrítica, Universidad de Barcelona, España.
- Visser, I. (2011), ‘Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series’, *Journal of Mathematical Psychology* **55**, 403–415.
- Viterbi, A. (1967), ‘Error bounds for convolutional codes and an asymptotically optimal decoding algorithm’, *Information Theory, IEEE Transactions on* **13**(2), 260–269.
- Zhai, C. (2003), ‘A brief note on the Hidden Markov Models’, Department of Computer Science, University of Illinois at Urbana-Champaign.
- Zucchini, W. & MacDonald, I. (2009), *Hidden Markov Models for Time Series an Introduction Using R*, Chapman & Hall/CRC Taylor & Francis Group, United States of America.