TESIS DOCTORAL

Caracterización de la contaminación atmosférica debida a aportes antropogénicos y naturales mediante la aplicación de modelos de mixturas finitas, de Markov homogéneos y otras técnicas de minería de datos

Autor

Álvaro Gómez Losada

Tutora

María Isabel Carretero León

Departamento de Cristalografía,
Mineralogía y Química Agrícola

Director

Rafael Pino Mejías

Departamento de Estadística
e Investigación Operativa

Sevilla, 2016

Caracterización de la contaminación atmosférica debida a aportes antropogénicos y naturales mediante la aplicación de modelos de mixturas finitas, de Markov homogéneos y otras técnicas de minería de datos.

Memoria realizada por Álvaro Gómez Losada, bajo la dirección de Rafael Pino Mejías, profesor del Departamento de Estadística e Investigación Operativa de la Universidad de Sevilla, y la tutoría de María Isabel Carretero León, profesora del Departamento de Cristalografía, Mineralogía y Química Agrícola de la Universidad de Sevilla, para optar al grado de Doctor por la Universidad de Sevilla.

Sevilla, 21 de enero de 2016.

Alumno

Álvaro Gómez Losada

Tutora

Director

Dra. María Isabel Carretero León

Dr. Rafael Pino Mejías

A mi padre

Agradecimientos

Con la defensa de esta tesis se cumplen tres años de tardes, y fines de semana, de trabajo. Publicar sus contenidos ha supuesto, además, un esfuerzo adicional. Rafael Pino, mi director de tesis, ha sido consciente de todo lo anterior. Le agradezco muy sinceramente su apoyo, sus orientaciones y la libertad que me ha dado para desarrollar el trabajo.

Estoy agradecido a mi tutora, María Isabel Carretero, por el recibimiento que ha dado a nuestro proyecto. A Isabel González, del Departamento de Cristalografía, Mineralogía y Química Agrícola, toda mi gratitud.

A José Carlos M. Pires, de la Universidad de Oporto, le agradezco haber aceptado colaborar, compartir su experiencia, así como la revisión tan minuciosa que ha hecho de nuestros trabajos, incluso durante los fines de semana cuando ello fue preciso.

Quiero agradecer a José Fernando Vera, de la Universidad de Granada, el interés en mi trabajo y haberme invitado al congreso de Bolonia. También, el recibimiento que él y José Miguel Angulo me han dado en mis visitas al Departamento de Estadística.

Quiero expresar mi agradecimiento a otras personas, no relacionadas con la tesis, pero que han influido en mi formación: Jesús Martín, profesor de Microbiología en la Universidad de Córdoba, quien dirigió mi tesina y me preparó para la estancia en Londres; y Renato Alvárez, de la Universidad de Sevilla, por su ejemplo.

Gracias, también, a mi amiga Ángela; Luis y Rocío; Ignacio Miró y María José Guerrero. A Maca, Fran y Silvia. A Rocío y Joaquín. A mis compañeros de trabajo Kiko y Alejandra. Y a todas las personas que han mostrado interés por el desarrollo de este trabajo.

Pero, en esencia, gracias a Puri, por su generosidad.

RESUMEN

Son cuantiosos los recursos científicos que se dirigen al estudio de las fuentes de emisión de contaminantes atmosféricos en las áreas urbanas. Este estudio puede ser cuantitativo, determinando la contribución de cada fuente a la contaminación ambiente, o cualitativo, para conocer más sobre la composición de las emisiones que afectan a los residentes en las ciudades. En los países mediterráneos, además, la contaminación causada por fenómenos naturales, como el transporte de polvo desde las regiones áridas del Norte de África, también es de primordial importancia. Este transporte de partículas trasciende de estos ámbitos geográficos al alcanzar las costas americanas atravesando el Océano Atlántico. Entre los instrumentos fundamentales de los que se dispone para medir la contaminación atmosférica, se encuentran las redes de vigilancia de la calidad del aire, integradas por estaciones de medida que se sitúan tanto en ambientes urbanos como en el medio rural, con el fin de determinar e informar sobre la calidad del aire que nos afecta. En las ciudades, algunas de estas estaciones de medida se sitúan en emplazamientos fuera del alcance directo de fuentes de emisión, para determinar la contaminación de fondo urbano, representativa de la exposición a la que la población se expone de forma general.

Esta tesis ha tenido como objetivos los siguientes:

1. La caracterización exhaustiva de la contaminación atmosférica en entornos urbanos y rurales empleando la información obtenida de las redes de vigilancia de la calidad del aire, desarrollando para ello una metodología general para la gestión eficiente de las redes de monitorización.
2. Mejorar la metodología existente para la estimación del aporte de polvo transportado por las masas de aire cálido desde las regiones norteafricanas.
3. Comparar los niveles de contaminación atmosférica entre diferentes redes de monitorización urbanas, sin influencia industrial y localización geográfica distinta, proponiendo para ello una metodología con la que caracterizar la contaminación atmosférica ambiental y de fondo.

Para el desarrollo del primer objetivo se aplicaron, sobre datos de monitorización de contaminantes atmosféricos primarios y secundarios, modelos de mixturas finitas; a partir del cálculo del primer y segundo momento de estas mixturas, se emplearon el análisis clúster jerárquico, la imputación mediante bosques aleatorios y el análisis de componentes principales. Esta aproximación metodológica permitió la detección de duplicidades en los parámetros monitorizados en las estaciones de vigilancia, ofreciendo, por tanto, la posibilidad de reconfigurar estas redes y mejorar el aprovechamiento de los recursos económicos invertidos en ellas.

Con el segundo objetivo se introducen los modelos ocultos de Markov (MOM) y se describen los diferentes regímenes o perfiles de concentración de PM_{10} en algunas de las series temporales (SSTT) estudiadas, permitiendo estimar las contribuciones de cada uno de estos perfiles a la contaminación ambiente. El nuevo método propuesto para la estimación del aporte natural de PM_{10} mejora al de referencia en la Unión Europea (método del percentil 40 medio móvil mensual) en tres aspectos, al evitar su aproximación empírica, utilizar una modelización especialmente orientada al tratamiento de SSTT, y permitir obtener un intervalo de confianza en las estimaciones del aporte de PM_{10}.

Durante el desarrollo del tercer objetivo también se emplearon los MOM; en esta ocasión, para definir y caracterizar, en distintos entornos urbanos de diferentes ciudades, la contaminación de fondo y ambiente por contaminantes atmosféricos primarios. La fracción de fondo de la contaminación ambiente es estimada mediante un procedimiento nuevo basado en el primer perfil de concentraciones definido por los MOM en las SSTT. Se estudiaron también el ratio y la diferencia entre las concentraciones ambiente y de fondo.

Palabras clave: redes de vigilancia de la calidad del aire, intrusión sahariana, percentil 40 medio móvil mensual, contaminación atmosférica, contaminación de fondo, series temporales, modelos de mixturas finitas, modelos ocultos de Markov.

ABSTRACT

A wealth of scientific resources have been dedicated to the study of the sources of pollutant emissions to air in urban areas. Such studies may be quantitative, determining the contribution of each source of environmental pollution, or they may be qualitative, providing insight into the makeup of the emissions that affect a city's inhabitants. In Mediterranean countries, contamination may also be the result of natural phenomenon, such as the flow of dust from the arid regions of North Africa, and are therefore of primary importance as well. The flow of particulate matter transcends these geographic areas, passing over the Atlantic Ocean and reaching the American coasts. Among the fundamental tools available for measuring air pollution are the air-quality monitoring networks, made up of monitoring stations located both in urban areas and rural environments, with the aim of providing information on the air quality that affects us. In cities, some of these monitoring stations are located on sites that are outside of the direct range of emission sources and thus the determination of the urban background pollution, which is indicative of the generalised exposure of the population to air pollution, is possible.

The objectives of this thesis were the following:

1. To exhaustively characterise the air pollutants in urban and rural areas using the information obtained from the air-quality monitoring networks. To this end, a general methodology was developed to efficiently manage the monitoring networks.

2. To improve the existing methodology used to estimate the contribution of dust originating in the North African region that is carried by waves of warm air.

3. To compare the air-pollution levels between the different urban-monitoring networks unaffected by industrial pollution, and between different geographic locations, proposing a methodology that can be used to characterise environmental and background air pollution.

In order to fulfil the first objective, the primary and secondary air-pollution monitoring data were modelled using finite mixture models. Based on the calculation of the first and second moments of these mixtures, hierarchical cluster analysis, imputation using random forests, and principal component analysis were used. This methodological approximation enabled the detection of duplications within the parameters monitored by the monitoring stations, thus allowing these networks to be reconfigured and enabling the economic resources invested in them to be optimised.

For the second objective, hidden Markov models (HMM) were introduced and the different regimes or PM_{10} concentration profiles were described in some of the time series (TS) studied, enabling an estimation of the contribution of each of the profiles to environmental pollution. The new method proposed for estimating the natural contribution of PM_{10} improves upon the reference methodology used in the European Union (monthly moving 40th percentile method) in three ways - it avoids the use of empirical approximation, it applies modelling that is especially designed for the treatment of time-series data, and it allows for obtaining a confidence interval for the contribution estimations for PM_{10}.

For the third objective, hidden Markov models were also used, in this case to define and characterise the environmental and background pollution caused by primary air pollution in different urban areas of different cities. The attributable fraction for background air pollution was estimated using a new procedure based on the first concentration profile defined by the HMMs in the TS. The ratio and difference between environmental and background concentrations were also studied.

Keywords: air quality monitoring networks, Saharan intrusion, monthly moving 40th percentile, air pollution, background pollution, time series, finite mixture models, hidden Markov models.

Índice

ANEXOS

Abreviaturas y Símbolos

Abreviaturas genéricas

ACF	análisis de contribución de fuentes
CO	monóxido de carbono
COVs	compuestos orgánicos volátiles
MP	material particulado
NH_3	amoniaco
NO	óxido de nitrógeno
NO_2	dióxido de nitrógeno
NO_x	óxidos de nitrógeno (NO + NO_2)
O_3	ozono
PM_{10}	material particulado de diámetro aerodinámico 10μm o inferior
$PM_{2.5}$	material particulado de diámetro aerodinámico 2.5μm o inferior
s.n.m.	sobre el nivel del mar
SO_2	dióxido de azufre

ACJ	análisis clúster jerárquico
ACP	análisis de componentes principales
BA	bosques aleatorios
CP	componente principal
EM	esperanza-maximización (algoritmo)
MMF	modelos de mixturas finitas

CM	cadena de Markov
MOM	modelos ocultos de Markov
m.p.t.	matriz de probabilidades de transición
ST y SSTT	serie temporal y series temporales

Notación empleada en los modelos de mixturas finitas

$\mathsf{E}(\cdot)$	operador Esperanza
e.e. SEM	error estándar calculado mediante método SEM
EMV	estimadores de máxima verosimilitud
I_{oc}	matriz de información de los datos completos
$\ell(\Psi\|y)$	función de log-verosimilitud de la mixtura
$\ell(\Psi\|y,z)$	función de log-verosimilitud de los datos completos
$L(\theta\|y)$	función de verosimilitud de una muestra y
$L(\Psi\|y)$	función de verosimilitud de la mixtura
$L(\Psi\|y,z)$	función de verosimilitud de los datos completos
$Q(\Psi\|\Psi^{(h)})$	función Q
$\widehat{se_B}$	estimador *bootstrap* del error estándar
x	vector de datos completos

y	muestra observada
z	vector de variables latentes
μ_m	media de la mixtura
σ_m^2	varianza de la mixtura
cv_m	coeficiente de variación de la mixtura
λ	multiplicador de Lagrange

Notación empleada en los modelos ocultos de Markov

$\mathbf{1}$ y $\mathbf{1}'$	vector fila y columna de unos
C_t	estado ocupado por la cadena de Markov en el instante t
$\mathbf{C}^{(t)}$	$(C_1, C_2, ..., C_t)$
l o l_T	log-verosimilitud
L o L_T	verosimilitud
m	número de estados en la cadena de Markov
p_i	función de densidad en el estado i
$\mathbf{P}(x)$	matriz diagonal con i-ésimo elemento diagonal $p_i(x)$
T	longitud de la serie temporal
$\mathbf{u}(t)$	vector $(\mathrm{P}(C_t{=}1, \ldots, C_t = m))$
w_t	$\boldsymbol{\alpha}_t\ \mathbf{1}' = \sum_i \alpha_t(i)$
X_t	observación en el instante t o la t-ésima observación
$\mathbf{X}^{(t)}$	$(X_1, X_2, ..., X_t)$
$\boldsymbol{\alpha}_t$	vector fila de probabilidades hacia adelante
$\boldsymbol{\beta}_t$	vector fila de probabilidades hacia atrás
$\beta_t(i)$	probabilidad hacia atrás
$\boldsymbol{\delta}$	distribución inicial o estacionaria de la cadena de Markov, o vector de pesos
$\boldsymbol{\Gamma}$	matriz de probabilidades de transición (m.p.t.) de una etapa de la cadena de Markov
γ_{ij}	elemento (i, j) de $\boldsymbol{\Gamma}$
$\boldsymbol{\phi}_t$	vector de probabilidades hacia adelante, normalizado para que sume 1 $(\boldsymbol{\alpha}_t / w_t)$

Notación común a ambos modelos

AIC	criterio de información de Akaike
BIC	criterio de información bayesiano
log	logaritmo natural
$\phi(\cdot \mid \mu, \sigma)$	función de densidad normal de parámetros μ y σ

Preámbulo

Esta tesis aplica, con diferentes fines, modelos bien conocidos en Estadística a datos de la calidad del aire. De forma resumida, los modelos de mixturas finitas (MMF) se emplearon para caracterizar una red de vigilancia de la calidad del aire, y los modelos ocultos de Markov (MOM), para: (i) proponer una nueva metodología con la que estimar la contribución de polvo procedente de los desiertos, y (ii) para caracterizar la contaminación de fondo en diferentes entornos urbanos. Se emplearon, además, otras técnicas estadísticas sobre los resultados obtenidos con los MMF: análisis clúster jerárquico (ACJ), imputación mediante bosques aleatorios (BA) y análisis de componentes principales (ACP). Los MMF y MOM fueron implementados mediante funciones diseñadas para la ocasión en el entorno de programación R, si bien, para las técnicas auxiliares mencionadas, se emplearon la librería randomForest (Breiman y Cutler, 2014), y las funciones hclust (ACJ) y prcomp (ACP) de la librería stats (R Core Team, 2015). A efectos de validación, los resultados de la implementación propia en R de los MMF y MOM fueron posteriormente comparados mediante las librerías mclust (Fraley et al., 2012) y depmixS4 (Visser y Speekenbrink, 2014), respectivamente. Las publicaciones dedicadas a los MOM que forman parte de este trabajo emplean la librería depmixS4 en lugar del código de propio desarrollo, para facilitar su implementación por parte de un público no especialista.

La estructura de la tesis, conforme a lo anterior, se divide en dos partes bien diferenciadas: la primera de ellas, teórica, expone el fundamento de los MMF y MOM, prestando una menor atención a las técnicas auxiliares mencionadas (Capítulos 2, 3 y 4); la segunda, aplicada, emplea estos modelos sobre los datos de la calidad del aire (Capítulos 5, 6 y 7). La implementación computacional en R de los MMF y MOM se explica en los Anexos A y B, respectivamente, al final del documento.

Todo el contenido aplicado de esta tesis ha sido publicado en revistas incluidas en el *Science Citation Index*. Así, los Capítulos 5, 6 y 7 son, respectivamente, una adaptación de las siguientes publicaciones:

- **Gómez-Losada, Á., Lozano-García, A., Pino-Mejías, R., Contreras-González, J.** 2014. Finite mixture models to characterize and refine air quality monitoring networks. *Science of the Total Environment*, 485-486: 292-9. DOI: 10.1016/j.scitotenv.2014.03.091.
- **Gómez-Losada, Á., Pires, J.C.M., Pino-Mejías, R.** 2015. Time series clustering for estimating particulate matter contributions and its use in quantifying impacts from deserts. *Atmospheric Environment*, 117: 271-81. DOI: 10.1016/j.atmosenv.2015.07.027.
- **Gómez-Losada, Á., Pires, J.C.M., Pino-Mejías, R.** 2016. Characterization of background air pollution exposure in urban environments using a metric based on Hidden Markov Models. *Atmospheric Environment*, 127: 255-61. DOI: 10.1016/j.atmosenv.2015.12.046.

Una adaptación del material suplementario de estas tres publicaciones se acompaña en los Anexos C, D y E. El contenido del Capítulo 7 forma parte de un *Abstract* que fue expuesto en Bolonia (Italia), en julio de 2015, en el congreso de la Federación Internacional de Sociedades de Clasificación:

- **Gómez-Losada, A., Vera-Vera, J.F.** 2015. Stability analyses in human exposure to background air pollution in urban environments. Abstract. Conference of the International Federation of Classification Societies Abstracts, 295-6.

Un *Abstract*, en donde se recoge parte del contenido del Capítulo 6, ha sido remitido al octavo Taller Internacional sobre Tormentas de Polvo y Arena y Precipitaciones Asociadas, que se celebrará en Lisboa (Portugal), del 1 al 4 de mayo de 2016:

– **Á. Gómez-Losada, J.C.M. Pires, R. Pino-Mejías** 2016. Time series clustering to estimate particulate matter contributions from deserts. 8th International Workshop on Sand/Duststorms and Associated Dustfall.

El esquema de esta tesis se resume en la Figura 1.

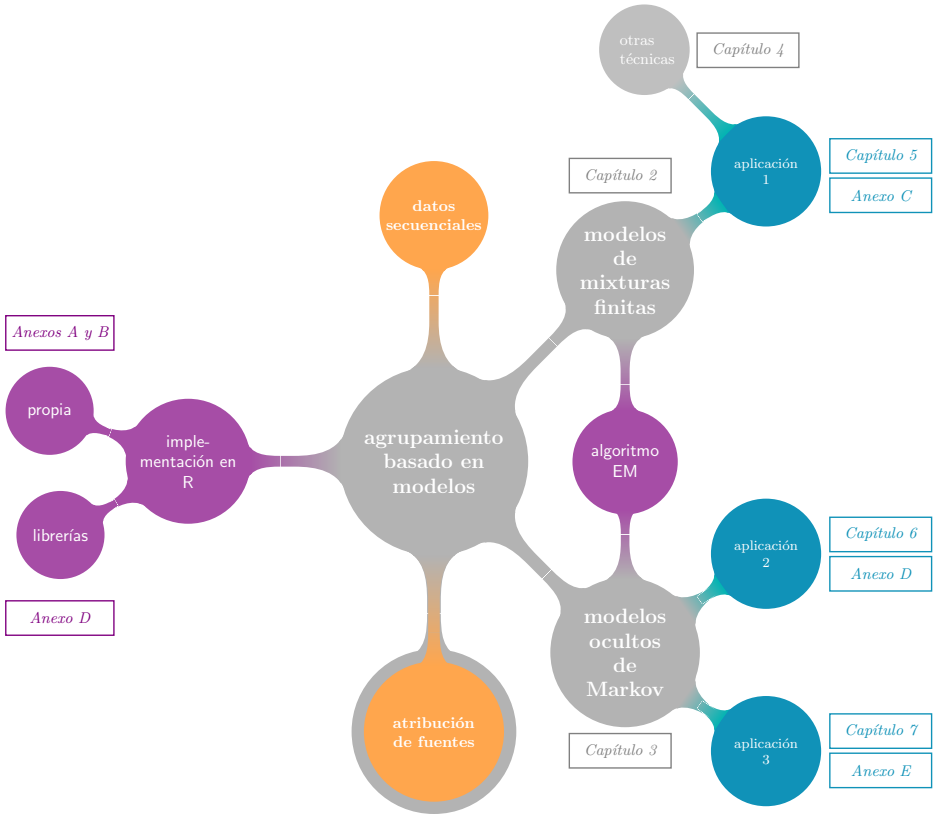


Figura 1 Esquema del contenido de este trabajo, destacándose el papel de los MMF y MOM como técnicas de agrupamiento (*técnicas clúster*). Las dos modelizaciones han compartido el empleo del algoritmo EM.

Los MMF y MOM se han considerado en este trabajo como técnicas de agrupamiento de observaciones (técnicas *clúster*), entendiendo por observaciones datos que representan medias diarias de algún contaminante de interés medido en una estación de vigilancia de la calidad del aire. La segunda de esta modelizaciones, los MOM, ha considerado la existencia, entre estas observaciones, de una dependencia temporal. Así, emplear a ambas modelizaciones con ese fin, la de agrupar observaciones, ha sido, desde un punto de vista ambiental, la principal aportación de este trabajo. Este agrupamiento de observaciones ha permitido proponer, en la mayoría de los casos, una asociación entre cada uno de los grupos resultantes (*clústeres*) y una fuente de emisión contaminante. Esta asociación *grupo de observaciones-fuente de emisión* no había sido empleada, como en esta tesis se hace, en el estudio de la contaminación atmosférica. Este vacío metodológico se debe seguramente a que los equipos científicos no integran personas procedentes de diferentes áreas de conocimiento. La atribución de fuentes no es un mero ejercicio de asociación, sino que trasciende a políticas de reducción de la contaminación y a la toma de medidas concretas para ello por parte de las administraciones competentes.

El diseño experimental de los Capítulos 6 y 7 de esta tesis, que aplica los MOM, se sustenta sobre el del Capítulo 5, basado en los MMF. De él toman prestado el concepto de *grupo de observaciones* y el cálculo

del primer y segundo momento de la mixtura (μ_m y σ_m), que tan útil se ha revelado por su capacidad simplificadora de las parametrizaciones de ambos tipos de modelos. Cabe destacar que en el Capítulo 6 se desarrolla un método para la estimación del material particulado procedente de los desiertos que mejora el que se emplea en la actualidad, por varios motivos: (i) evita el uso de una técnica de alisado de las series temporales (SSTT) y emplea en su lugar una modelización solvente (MOM) para su estudio, (ii) elimina el caracter empírico del método actual, y (iii) proporciona una intervalo de confianza para las estimaciones del material particulado mediante el procedimiento *bootstrap*. No obstante, como se verá, el método propuesto requiere cierto conocimiento de las fuentes de emisión principales en las áreas de estudio.

El Capítulo 7 profundiza en el estudio de la contaminación de fondo en los entornos urbanos, atribuyendo esa fracción de la contaminación al primer régimen de las SSTT modelizadas con MMO. Esta contaminación de fondo tiene un papel relevante para la salud humana, ya que en este trabajo se ha asociado a una contaminación de tipo crónico a la que se expone la población en su vida cotidiana. La notación μ_m y σ_m de los capítulos 5 y 6 se modifica ahora por M y SD, respectivamente, ya que así se ha presentado en el tercer trabajo remitido para publicación, sin ser imprescindible su modificación a la notación anterior. De forma semejante, los errores *bootstrap*, indicados con $\widehat{se_B}$ en el capítulo 5, en el capítulo 6 simplemente se presentan entre paréntesis bajo la estimación del valor de los parámetros (Anexo D). Estos cambios de notación se justifican por dos motivos: (i) porque de forma natural se ha simplificado la notación a medida que se ha ido desarrollando este trabajo, y (ii) porque, desde el primer artículo publicado, se ha intentado emplear una notación sencilla dirigida a un público no especialista, como el medioambiental, que es el usuario final de toda la metodología aquí presentada. Así, este cambio de notación, que obedece a una evolución natural de esta tesis, se ha querido que permanezca inalterado.

Este trabajo pretende ser, antes que nada, útil, en particular para aquellos que no trabajan en el área estadística. Por ello, lector, además de agradecerte encarecidamente que te hayas interesado por esta tesis, te animo, en el caso de que lo consideres oportuno, a que entres en *contacto* conmigo si decides implementar estos modelos, ya que podrían resultarte útiles y enriquecer, de algún modo, tu línea de trabajo.

17 de enero de 2016

1

Introducción

En los apartados siguientes se enmarcan, brevemente, las modelizaciones estadísticas empleadas en esta tesis en su campo de aplicación: la contaminación atmosférica. Para ello, se describe la formación de algunos de los contaminantes de mayor relevancia para la salud humana y de los ecosistemas, se repasa la principal normativa de la calidad del aire en España, y se expone en qué consiste una red de inmisión, y aquella, en particular, dedicada a la vigilancia de la contaminación atmosférica de fondo regional. El fenómeno de la intrusión sahariana se describe al final del capítulo.

1.1. Contaminantes atmosféricos en entornos urbanos

El nexo de unión entre una calidad del aire degradada y los efectos perjudiciales para la salud es ampliamente reconocido (Lim et al., 2012). La contaminación por material particulado (MP) ambiente ocupa, a nivel global, la novena posición como factor de riesgo causante de enfermedades, mientras que el ozono (O_3) aparece en la posición treinta y nueve. Además de los mencionados, entre los contaminantes atmósfericos con distinta repercusión en la atmósfera y, por tanto, en nuestra calidad de vida y de los ecosistemas, se encuentran los óxidos de nitrógeno (NO_x), el dióxido de azufre (SO_2), el monóxido de carbono (CO), así como un elevado número de compuestos orgánicos volátiles (COVs). El medio ambiente urbano adquiere un papel relevante respecto a la incidencia de efectos perjudicales sobre la salud, dado que, como sucede en toda Europa, es en las áreas urbanas donde se concentra la mayor fracción de población, la cual está expuesta a una calidad del aire degradada y en degradación. Si eludimos los factores de contaminación procedentes de fuentes naturales, la contaminación atmosférica en las ciudades es generada por la liberación de contaminantes primarios (aquellos emitidos directamente desde sus fuentes de emisión), pero también por los secundarios, que se forman por transformaciones químicas en las que los contaminantes primarios actúan como precursores. A continuación, se exponen las causas que generan los contaminantes mencionados en entornos urbanos, sugiriéndose para una descripción más detallada, u otros contaminantes de interés, textos especializados como los de Viana (2013), Steyn (2014) o Vallero (2014).

Material Particulado

En las áreas urbanas e industriales, las emisiones gaseosas y particuladas procedentes del tráfico rodado representan una fuente principal de contaminación atmosférica. En el MP generado por el tráfico rodado pueden distinguirse dos categorías de acuerdo con su modo de formación (Karanasiou y Mihalopoulos, 2013): el MP generado por la combustión de los motores, y un conjunto de procesos secundarios de abrasión y corrosión, que puede conllevar la liberación a la atmósfera de MP, entre los que se encuentran el desgaste de neumáticos, embragues y revestimiento de frenos, así como la propia abrasión

de la superficie de las vías de tránsito de los vehículos. Estos procesos de abrasión y corrosión conducen a la deposición de MP en la superficie de las vías. Este MP, que se agrega al existente procedente de los tubos de escape y de actividades no relacionadas propiamente con el transporte (p. ej., MP biogénico y mineral, o procedente de actividades industriales o domésticas), puede ser resuspendido en la atmósfera, como consecuencia de las turbulencias generadas por el tráfico viario o la propia acción del viento. Varios estudios demuestran que las emisiones de MP generadas por los medios "mecánicos" y de resuspensión se equiparan, si no superan, al MP generado por la combustión de los motores de los vehículos (Thorpe y Harrison, 2008; Amato et al., 2009).

La información sobre las emisiones "mecánicas" de MP debidas al tráfico de los vehículos (y no a las emisiones de los tubos de escape) no es tan abundante. Esto se debe, principalmente, a la dificultad encontrada al discriminar fuentes de emisión en ambientes tan complejos como los urbanos. En resumen, Amato et al. (2013) identifica 4 fuentes que originan este polvo depositado en las vías de tránsito: una fuente mineral, que incluye el polvo por el desgaste del pavimento y otras actividades, como la construcción y demolición urbanas; polvo por el desgaste de los frenos de los vehículos; polvo por el desgaste de los neumáticos, y, finalmente, partículas procedentes de los tubos de escape de los vehículos. Comparando las ciudades de Barcelona y Zúrich, las contribuciones medias de cada una de estas fuentes a la carga de polvo en los pavimentos son, respectivamente: 72 % vs 28 %, 17 % vs 33 %, 5 % vs 22 % y 6 % vs 17 %.

Así, las emisiones de MP del tráfico rodado son responsables de una importante proporción de superaciones de los límites de la calidad del aire establecidos para PM_{10} y $PM_{2.5}$ por la Directiva 2008/50/EC, estimándose que representan al menos, respectivamente, el 15-20 % y 10-15 % de las concentraciones de PM_{10} y $PM_{2.5}$ observados a nivel europeo (Hendriks et al., 2013). No obstante, en países como Alemania, el aporte de MP debido a las calefacciones domésticas supera con creces a las del tráfico viario (Brandt et al., 2011). En general, el impacto de las emisiones de polvo es superior en grandes áreas metropolitanas, regiones con alta densidad de población o en países del sur de Europa (condiciones meteorológicas más secas), si bien las emisiones primarias debidas a los tubos de escape (carbón elemental) afectan indistintamente a todas las regiones metropolitanas europeas. Según Amato et al. (2013), el tráfico viario en estos entornos urbanos contribuye a las concentraciones de PM_{10} y $PM_{2.5}$ en un 14-48 % y 9-49 %, respectivamente, mientras que en áreas rurales lo hace en un 1-4 % y 5-7 %.

La composición y cantidad de las emisiones generadas procedentes de los motores de los vehículos dependen de varios factores, los cuales determinan las diferencias en emisiones debidas a esta fuente entre los distintos países europeos. Destacan las características, mantenimiento y condiciones de operación de los propios motores, composición de los combustibles (diésel o gasolina) y sus aditivos, así como los dispositivos de control de emisiones instalados en los vehículos. Por ejemplo, los motores diesel superan en un factor mayor de 10 la emisión en términos de concentración de partículas finas y ultrafinas en comparación con los motores de gasolina. Pero la diferencia del uso de un combustible u otro también está relacionado con la formación de compuestos orgánicos e inorgánicos secundarios. Estudios recientes resaltan la importancia de los motores de gasolina sobre los diesel en la formación de aerosoles orgánicos secundarios (Bahreini et al., 2012), carbono orgánico y elemental. Por su parte, la formación de aerosoles inorgánicos secundarios (iones SO_4^{2-}, NH_4^+ y NO_3^-) debidos a NO_x y SO_2 está determinada también por el NH_3, otro componente atmósferico importante en la calidad del aire urbano. La emisión de NH_3, a su vez, se debe mayoritariamente al tráfico y otras fuentes de emisión fugitivas, como contenedores de residuos municipales y aguas residuales. Este gas alcalino, emitido en escenarios de alta concentración de NO_2, puede potenciar la formación de nitrato amónico, un componente principal de $PM_{2.5}$.

No obstante, además de por el tráfico viario, la concentración de MP en ambientes urbanos también está determinada por otras fuentes. Estas son, principalmente, las emisiones industriales, residenciales y domésticas, las resuspensiones de polvo del entorno regional transportado por el viento, aerosoles

orgánicos (p. ej., polen y esporas), y las intrusiones debidas a fenómenos naturales, como el aerosol marino, erosiones eólicas, volcanes, incendios, o las procedentes de algunos desiertos (p. ej. Sáhara), contribución esta última que, aunque esporádica, eleva puntualmente los niveles de PM_{10} de forma significativa. A escala global, las emisiones de MP procedentes de estas fuentes naturales se estima que superan con un amplio margen a las emisiones procedentes de actividades antropogénicas (Hainsch, 2003, citando a Warneck, 1999). No obstante, esta tendencia se invierte en países con una marcada industrialización y densidad de población (Quass et al., 2013). Otro aporte de MP a la atmósfera en los entornos urbanos costeros lo constituye el tráfico marítimo comercial. En las áreas urbanas situadas en la cuenca mediterránea se estima que estas emisiones son responsables del 2-4 % de la media anual de PM_{10} y del 14 % de la media anual de $PM_{2.5}$.

Óxidos de nitrógeno

El tráfico rodado es, de forma mayoritaria, el principal contribuyente de las emisiones de los óxidos de nitrógeno NO_x (NO + NO_2) en los ambiente urbanos (Nagl, 2013) y, en general, en la Unión Europea (Boulter et al., 2013); implicado en fenómenos como la acidificación y eutrofización, el compuesto de mayor interés en relación con la calidad del aire local y salud humana es el NO_2. Este gas oxidante es precursor de contaminantes atmosféricos secundarios perjudiciales, como el ácido nítrico, la parte nitrato de aerosoles secundarios inorgánicos y foto-oxidantes como el O_3. Los NO_x se forman principalmente cuando el nitrógeno y el oxígeno se combinan a altas temperaturas y presión, condiciones que se alcanzan en los motores de combustión interna, por lo que los modos de transporte motorizados y, en particular, los vehículos por carretera, son los mayores contribuyentes a las emisiones globales de NO_x. En 2008, las emisiones sectoriales en la Unión Europea de NO_x se deben en un 41 % al transporte rodado; en un 28 % son debidas a partes iguales al uso energético en la industria y a los hogares, instituciones y comercios; en un 20 % a la producción y distribución de energía eléctrica, y el resto, a fuentes menores como al transporte no rodado, la agricultura y los procesos industriales (EEA, 2010).

Como sucedía con el MP, las emisiones de NO_x debidas al tráfico están influenciadas por el tipo de vehículo (p. ej., turismo o vehículos de carga), el tipo de combustible, la tecnología del vehículo y, en particular, por su modo de conducción, estimándose que este condiciona las emisiones de NO_x de un vehículo hasta en más de un orden de magnitud.

Ozono

En las capas bajas de la atmósfera, el O_3 se forma principalmente mediante una serie de reacciones químicas complejas iniciadas por la luz solar. Estas reacciones, en las que el NO_x y COVs reaccionan para formar O_3, pueden conllevar horas o días, dependiendo de la concentración de COVs; una vez que el O_3 se ha formado, puede persistir en la atmósfera durante varios días. En consecuencia, el O_3 medido en una localización particular puede haberse formado a partir de emisiones de NO_x y COVs generadas a muchos cientos, o incluso miles, de kilómetros. Cercano a las regiones de emisión (p. ej., áreas urbanas), el NO emitido puede reaccionar con el O_3 para formar NO_2, reduciendo las concentraciones de O_3 localmente. En general, las concentraciones netas de O_3 aumentan a medida que las masas de aire abandonan lon núcleos urbanos.

En la Figura 1.1 se muestra el comportamiento de los niveles de NO_x y O_3 en ambientes rurales, urbanos y a nivel de calle (**A**), y su comportamiento a su paso por áreas de emisión (**B**), como puede ser un área urbana. En estas áreas urbanas, la concentración de NO_x aumenta, lo que causa la disminución, de forma concomitante, de los niveles de O_3. A medida que las masas de aire abandonan las áreas de emisión, la concentración de NO_x disminuye debido a procesos de dilución, mientras que aumenta la de O_3.

En Europa se describe una concentración *base* o *de fondo* de O_3 cuyos valores se consideran generalmente mínimos, y que suelen encontrarse en el rango entre los 40 y 85 $\mu g/m^3$ (Derwent y Hjellbrekke,

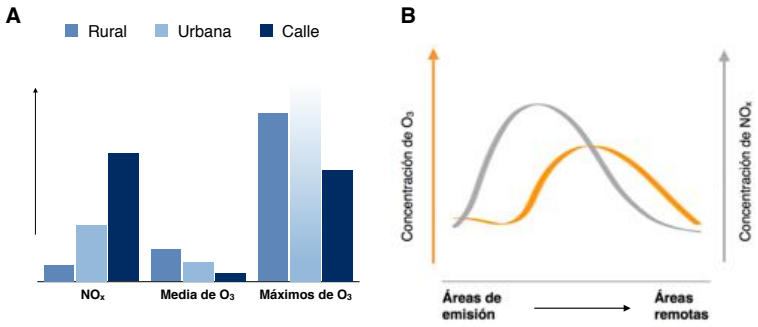


Figura 1.1 A. Comparación cualitativa de los niveles de NO_x y O_3 en ambientes rurales, urbanos y a nivel de calle. **B**. Representación del desarrollo de los niveles de NO_x y y O_3 en el aire a su paso por áreas de emisión (p. ej. una ciudad) (modificado de Ozone Position Paper, 1999).

2013). Este O_3 es transportado desde la estratosfera, pero también desde la troposfera, por un origen fotoquímico este último (Derwent, 2008) en el que participan, en igual proporción, precursores tanto naturales como antropogénicos (Ozone Position Paper, 1999).

Dióxido de azufre y monóxido de carbono

El primero de estos contaminantes es, mayoritariamente, de origen antropogénico, como consecuencia de la combustión de carburantes fósiles que contienen azufre (petróleo, combustibles sólidos), producida sobre todo en los procesos industriales de alta temperatura y de generación eléctrica. Respecto al CO, este contaminante se genera principalmente como consecuencia de las combustiones incompletas de combustibles que contienen carbono (p. ej., carburantes fósiles, madera, biomasa).

Más allá de los procesos de génesis descritos, es necesario destacar, en referencia a las aglomeraciones urbanas, el "efecto fin de semana", que surge como consecuencia de la diferencia en la movilidad de las personas y, por tanto, en el uso de los medios de transporte, según los días de la semana. Estas circunstancias llevan a que las emisiones del tráfico también difieran según el día de la semana, siendo en las zonas urbanas y suburbanas mayores durante los días laborables que en los fines de semana (Figura 1.2).

1.2. Marco normativo de la calidad del aire

La normativa europea sobre calidad del aire actualmente en vigor (octubre 2014) viene representada por las siguientes Directivas (un resumen de los objetos de las aprobaciones de estas Directivas puede consultarse en Análisis de la Calidad del Aire en España -2014-):

1. Directiva 2008/50/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008, relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa, transpuesta en España mediante el Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire.

2. Directiva 2004/107/CE del Parlamento Europeo y del Consejo, de 15 de diciembre de 2004, relativa al arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos en el aire ambiente, transpuesta en España mediante el Real Decreto 812/2007, de 22 de junio, sobre evaluación y gestión de la calidad del aire ambiente en relación con el arsénico (As), el cadmio (Cd), el mercurio (Hg), el níquel (Ni) y los hidrocarburos aromáticos policíclicos (HAP), norma a su vez derogada por el Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire.

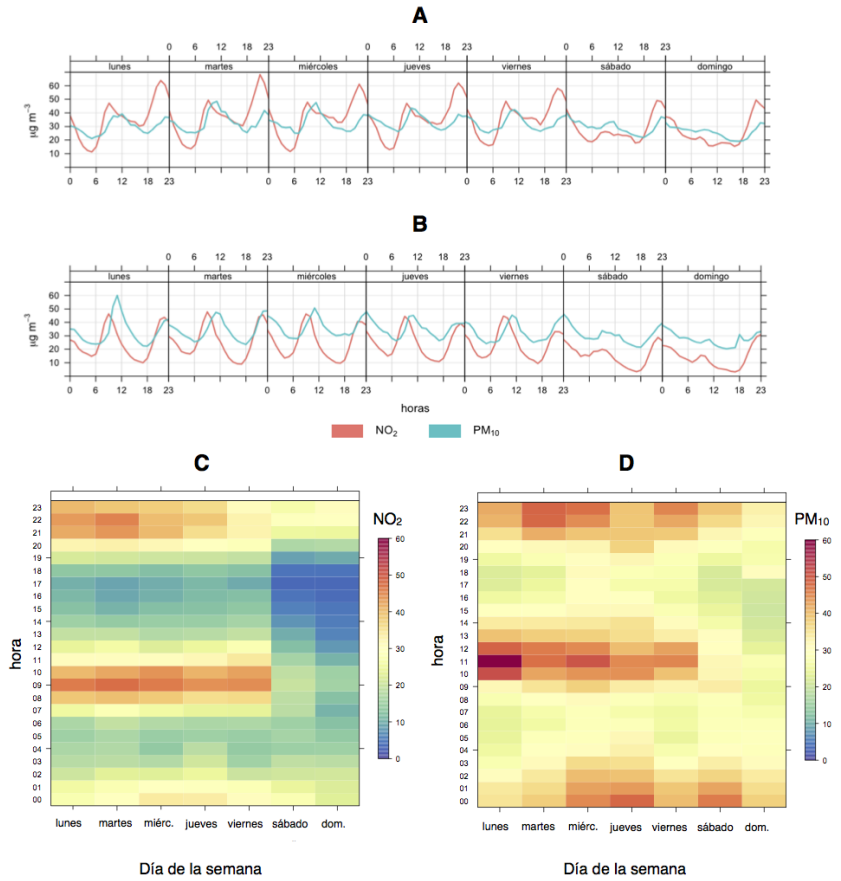


Figura 1.2 Representación gráfica del efecto fin de semana utilizando los contaminantes NO_2 y PM_{10}, medidos en dos estaciones urbanas, de tráfico (**A**) y de fondo (**B**). **C** y **D** muestran una representación alternativa de dicho efecto en la estación de fondo urbano.

La legislación española sobre calidad del aire actualmente en vigor viene representada por las siguientes normas, siendo la segunda de ellas la que resulta esencial en la evaluación y gestión de la calidad del aire en España:

1. La Ley 34/2007, de 15 de noviembre, de calidad del aire y protección de la atmósfera, que actualiza la base legal para los desarrollos relacionados con la evaluación y la gestión de la calidad del aire en España, y tiene como fin último el de alcanzar unos niveles óptimos de calidad del aire para evitar, prevenir o reducir riesgos o efectos negativos sobre la salud humana, el medio ambiente y demás bienes de cualquier naturaleza.

   En esta Ley se establecen los principios esenciales en materia de prevención, vigilancia y reducción de la contaminación atmosférica, destacando los siguientes:

   - ▷ Se desarrollan los fundamentos de la evaluación y gestión de la calidad del aire, basados en tres pilares: los contaminantes a evaluar y sus objetivos de calidad, las obligaciones de la evaluación, y la zonificación del territorio, según los niveles de contaminantes para los que se hayan establecido objetivos de calidad.
   - ▷ La planificación, centrada en la elaboración de planes y programas para la protección de la atmosféra y para minimizar los efectos negativos de la contaminación atmosférica.
   - ▷ El control, la inspección, la vigilancia y seguimiento, recogiendo el deber de las comunidades autónomas y en su caso, entidades locales, de disponer de estaciones, redes y otros sistemas de evaluación de la calidad del aire suficientes para el cumplimiento de sus obligaciones, conforme a lo indicado en la norma.

2. Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire, esencial en la evaluación y gestión de la calidad del aire en España. Se aprueba con la finalizadad de evitar, prevenir y reducir los efectos nocivos de la sustancias dióxido de azufre, dióxido de nitrógeno y óxidos de nitrógeno, partículas, plomo, benceno, monóxido de carbono, ozono, As, Cd, Ni, benzo(a)pireno, HAP, HAP distintos al benzo(a)pireno, Hg y amoniaco. Entre sus principales objetivos se encuentran:

   - ▷ Concretar aspectos relacionados con las mediciones, tales como los criterios de ubicación de los puntos de muestreo o la determinación del número mínimo de estos en medición fija, los objetivos de calidad de los datos o los métodos de referencia para la evaluación; y proporcionar criterios adicionales para las partículas $PM_{2.5}$ y para los metales.
   - ▷ Evaluar la calidad del aire en la relación al ozono y amoniaco.

   Respecto a la gestión de calidad del aire (se destaca un ítem de interés en este trabajo):

   - ▷ Fija diversas obligaciones en lo que respecta a plazos de cumplimiento y a la necesidad de elaborar listados diferenciados por contaminante donde se indiquen los umbrales y límites legislados superados, por zonas y aglomeraciones.
   - ▷ Define un Indicador Medio de Exposición, como "el nivel medio, determinado a partir de las mediciones efectuadas en ubicaciones de fondo urbano de todo el territorio nacional, que refleja la exposición de la población", y que se emplea para calcular el objetivo nacional de reducción de la exposición.
   - ▶ Contempla las aportaciones procedentes de fuentes naturales (art. 22, Anexo IV), en relación con las superaciones de los valores límite imputables a dichas causas, que por ello no se consideran como tales. La demostración y sustracción de los niveles atribuibles a fuentes naturales será conforme a las directrices publicadas por la Comisión Europea y, en su ausencia, a los procedimientos elaborados por el Ministerio de Agricultura, Alimentación y Medio Ambiente en colaboración con las Comunidades Autónomas.
   - ▷ Considera la posibilidad de solicitar prórrogas de los plazos de cumplimiento.
   - ▷ Estipula la elaboración de planes de calidad del aire por las comunidades autónomas cuando en determinadas zonas o aglomeraciones los niveles de contaminantes en el aire ambiente superen cualquier valor límite o valor objetivo, así como el margen de tolerancia correspondiente a cada caso.

Basadas en la evaluación por parte de los expertos de las evidencias científicas del momento, relativas a la contaminación del aire y sus consecuencias para la salud, la Organización Mundial de la Salud (OMS) establece unos valores guía que revisa cada cierto tiempo, para determinados contaminantes del aire. Estos valores guía son más estrictos que los valores límite establecidos por las Directivas Europeas y constituyen un objetivo a perseguir por la legislación europea. Los valores límite de la legislación tienen implicaciones y se fijan en función de que puedan ser cumplidos por los Estados miembros, forzando la aplicación de las Mejores Técnicas Disponibles. La Tabla 1.1 compara valores límites y objetivo de la Directiva 2008/50/EC y la OMS (2000, 2006).

EC Directiva 2008/50/EC				OMS (2000,2006)
Contaminante	Periodo	Valor	Superación	Valor objetivo
SO_2	1 h	350 $\mu g/m^3$	≤ 24 veces/año	
		500 $\mu g/m^3$	Umbral de alerta (3h)	
	24 h	125 $\mu g/m^3$	≤ 3 veces/año	20 $\mu g/m^3$
	[1] Anual	20 $\mu g/m^3$	Protección vegetación	
NO_2	1 h	200 $\mu g/m^3$	≤ 18 veces/año	200 $\mu g/m^3$
	Anual	40 $\mu g/m^3$		40 $\mu g/m^3$
NO_x	Anual	30 $\mu g/m^3$	Como NO_2 Protección vegetación	
PM_{10}	24 h	50 $\mu g/m^3$	≤ 35 veces/año	20 $\mu g/m^3$
	Anual	40 $\mu g/m^3$		
$PM_{2.5}$	Anual	25 $\mu g/m^3$		
CO	[2] 8 h	10 mg/m^3		10 mg/m^3
O_3	1 h	180 $\mu g/m^3$	Umbral de información	
	1 h	240 $\mu g/m^3$	Umbral de alerta 3h consecutivas	
	[2] 8 h	120 $\mu g/m^3$	≤ 25 veces/año Promedio de 3 años	
	[3] AOT40	18000 $\mu g/m^3 h$	Promedio de 5 años	

[1] Y periodo de invierno (1/10-31/3).
[2] Máximo diario de las medias móviles octohorarias (p.ej., para un día en particular [día 1]: el primer periodo de cálculo comprende desde las 17 h del día 0 hasta las 1h del día 1; el último periodo de cálculo del día 1 comprende el periodo a partir de las 16 h hasta las 24 h).
[3] Suma de las diferencias entre las 8AM y 8PM de los valores horarios que excedan 80 $\mu g/m^3$, cada día, entre mayo y julio.

Tabla 1.1 Comparación entre algunos valores límites y objetivo de contaminantes legislados.

1.3. Las redes de inmisión

La calidad del aire viene determinada por la presencia en la atmósfera de sustancias contaminantes, que pueden ser gases o aerosoles. La protección de la atmósfera y de la calidad del aire en cualquier territorio pasa por la prevención, vigilancia y reducción de los efectos nocivos de dichas sustancias contaminantes sobre la salud y el medio ambiente en su conjunto. Para ello, las normativas vigentes en materia de calidad del aire establecen unos objetivos de calidad del aire o niveles de concentración de contaminantes en la atmósfera que no deben sobrepasarse. También ha de cumplirse con el requisito imprescindible de informar a la población y a las organizaciones interesadas.

En España, la Ley 34/2007, de 15 de noviembre, de calidad del aire y protección de la atmósfera, define la evaluación de la calidad del aire como el resultado de aplicar cualquier método que permita medir, calcular, predecir o estimar las emisiones, los niveles o los efectos de la contaminación atmosférica.

Conforme al Real Decreto 102/2011, de 28 de enero, la evaluación de la calidad del aire ambiente se realizará, dependiendo del nivel de los contaminantes con respecto a los umbrales/valores objetivo, mediante mediciones fijas, técnicas de modelización, campañas de mediciones representativas, mediciones indicativas o investigaciones, o una combinación de todos o algunos de estos métodos. Por tanto, los diferentes métodos de evaluación pueden ser los siguientes:

- ▷ Mediciones fijas: mediciones efectuadas en emplazamientos fijos, bien de forma continua (Figura 1.3), bien mediante un muestreo aleatorio, con el propósito de determinar los niveles de conformidad con los objetivos de calidad de los datos establecidos por la legislación (incertidumbre, recogida mínima de datos y cobertura mínima temporal).
- ▷ Mediciones indicativas: mediciones cuyos objetivos de calidad de los datos en cuanto a cobertura temporal mínima son menos estrictos que los exigidos para las mediciones fijas (esto es, se efectúan con una menor frecuencia), pero satisfacen todos los demás objetivos de calidad de los datos establecidos por la legislación.

- ▷ Modelizaciones: herramientas matemáticas que simulan el comportamiento de la atmósfera para determinar los niveles de un determinado contaminante en ella.
- ▷ Una combinación de los anteriores.

Figura 1.3 Estación fija de medida de la calidad del aire en la calle Torneo (Sevilla). En la parte superior de la estación se aprecian los diferentes equipos de medición. Véase detalle de su localización en la Figura 1.4B.

Las estaciones fijas de medida pueden ser de diferentes tipos, según el área en que se localizan y según la principal fuente emisora implicada. Así, las estaciones de vigilancia de la contaminación del aire pueden clasificarse del siguiente modo (Figura 1.4):

1. Según el tipo de área en el que se localizan:
   - ▷ Urbanas: las ubicadas en zonas edificadas de forma continua.
   - ▷ Suburbanas: las que se encuentran en zonas con presencia continuada de edificios, separadas por zonas no urbanizadas (pequeños lagos, bosques, tierras agrícolas).
   - ▷ Rurales: entendidas como las situadas en aquellas zonas que no satisfacen los criterios de las dos categorías anteriores.
2. Según la tipología de la principal fuente de emisión que la influye (que determina unos contaminantes predominantes):
   - ▷ De fondo: estaciones en las que no se manifiesta ninguna fuente de emisión como predominante.
   - ▷ De tráfico: estaciones situadas de tal manera que su nivel de contaminación está determinado principalmente por las emisiones procedentes de los vehículos de una calle o carreteras próximas.
   - ▷ Industriales: estaciones situadas en áreas cuyo nivel de contaminación se debe fundamentalmente a la contribución de fuentes industriales.

1.3.1. Contaminación atmosférica de fondo regional e intrusiones saharianas

La contaminación atmosférica de fondo regional es la que existe en zonas alejadas de focos de emisión directa. Proporciona información acerca de cuál es el nivel de contaminación regional, debida tanto a fuentes antropogénicas, naturales, regionales o transfonterizas. Desde 2006, estos niveles de fondo regional se determinan a partir de mediciones realizadas por las estaciones de la red española EMEP/VAG/CAMP, gestionada por la Agencia Estatal de Meteorología. Esta red pretende satisfacer

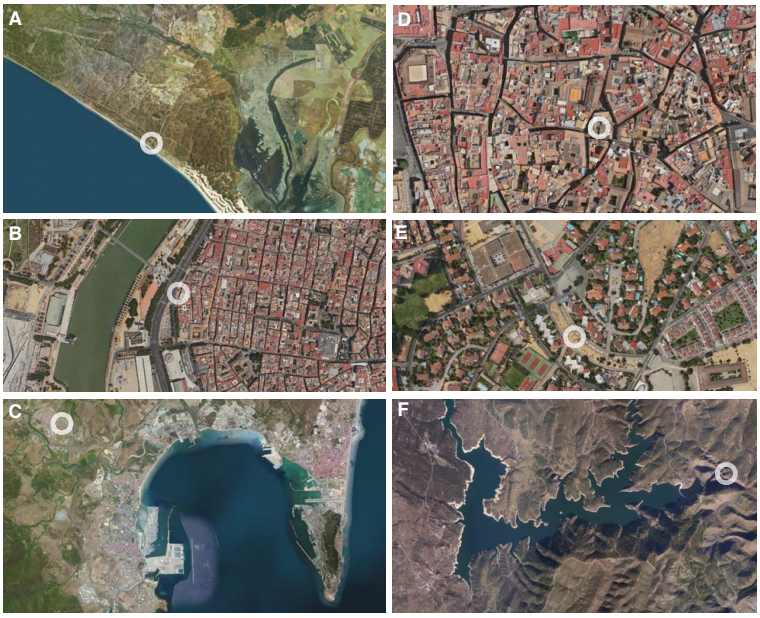


Figura 1.4 Comparación entre ubicaciones (circunferencia) de estaciones fijas de medida, clasificadas según el tipo de fuente principal de emisión y el tipo de área en donde se localizan. Según la fuente principal de emisión: **A**, de fondo (Parque Nacional de Doñana, Huelva); **B**, de tráfico (calle Torneo, Sevilla); **C**, industrial, en el municipio de Los Barrios (Cádiz), junto a la Bahía de Algeciras. Según el área donde se ubica la estación: **D**, urbana (calle Pajaritos, Sevilla); **E**, suburbana (barrio de Santa Clara, Sevilla), y **F**, rural (El Atazar, Madrid).

los compromisos de medición de contaminantes contraídos por España con los programas EMEP (Programa concertado de seguimiento y de evaluación del transporte a larga distancia de los contaminantes atmosféricos en Europa), VAG (Vigilancia Mundial de la Atmósfera) y CAMP (Programa Integral de Control Atmosférico).

Las mediciones obtenidas de las estaciones de dicha red permiten determinar los niveles de contaminación de fondo en una región, así como evaluar el transporte desde fuentes emisoras situadas a grandes distancias de ellas. Por ello son representativas, en cuanto a la calidad del aire, de un área extensa en torno a ellas, por lo que también se utilizan para la verificación de los pronósticos de los modelos de predicción de la calidad del aire. Además, en estas estaciones se determinan los contaminantes regulados por la legislación europea y nacional, por lo que dan apoyo a las redes autonómicas y locales en su evaluación de la calidad del aire. También se determinan una serie de contaminantes distintos a los regulados en dicha legislación (contaminantes sobre los que no se dispone de información acerca de su comportamiento o efectos sobre la salud o vegetación). Ello permite que también sirvan para estudios científicos sobre dichos compuestos, cuyos resultados, a su vez, influyen en la generación de nueva legislación de la calidad del aire.

Respecto al transporte desde la fuentes emisoras lejanas, de acuerdo con Querol et al. (2013a), el impacto del material en suspensión atmosférico africano en la visibilidad y en la composición de la deposición húmeda se conoce desde antiguo (lluvias y nevadas rojas registradas en Europa, o el "Mar oscuro" en el Atlántico Ecuatorial descrito por Ehrenberg en 1862). A escala global, la fracción mineral es el componente mayoritario de los aerosoles atmosféricos. El IPCC (Grupo Intergubernamental de Expertos sobre el Cambio Climático) -2001- estima unas emisiones naturales de partículas crustales (o partículas de origen mineral) del orden de 1500 millones de toneladas anuales. El transporte a largas distancias de material paticulado crustal (Figura 1.5) se produce cuando se generan procesos masivos de resuspensión en zonas áridas como las presentes en el Norte de África, Oriente Próximo o Asia Central. Otras zonas desérticas como Atacama o los desiertos de Australia no generan este tipo de transporte a larga distancia. Las áreas exportadoras de partículas crustales tienen como característica común el ser cuencas en las que se acumula una gran cantidad de este material particulado de granulometría muy fina, debido a la erosión de zonas áridas en épocas de lluvia torrencial. En la época seca, este material fino

queda expuesto a posibles procesos de resuspensión. En la zona del Norte de África, existen infinidad de cuencas endorreicas donde este fino material se puede depositar (desiertos de Sahel, Sáhara Occidental, Argelia, Túnez, Libia y Egipto). En estas áreas norteafricanas, desérticas, las condiciones climatológicas constituyen situaciones muy favorables para la resuspensión masiva de grandes cantidades de material particulado.

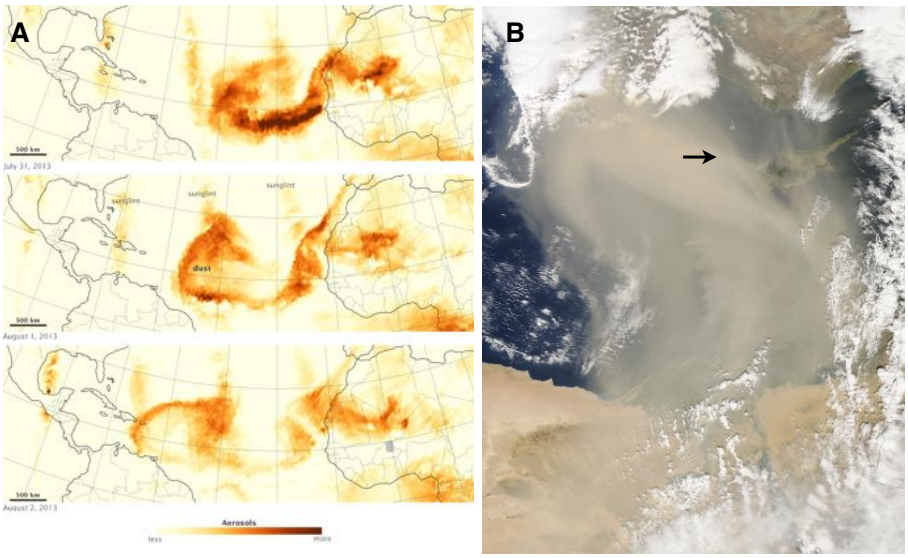


Figura 1.5 Imágenes procedentes de NASA (Earth Observatory) mostrando masas de polvo procedentes del Desierto del Sáhara. **A**. Concentración de partículas aerosoles a través del Océano Atlántico, desde el 31 de julio al 2 de agosto de 2013, detectada mediante información espectral ultravioleta. La luz solar causa un bandeado vertical en las imágenes. **B**. Fotografía tomada el 2 de marzo de 2014 por el satélite Terra sobre el Mar Mediterráneo. La flecha indica la Isla de Chipre.

2

Modelos de mixturas finitas

El presente capítulo comienza con las definiciones básicas relacionadas con los modelos de mixturas finitas y la estimación de máxima verosimilitud (apartado 2.1). A continuación, se presenta el algoritmo EM como un método general de obtención de soluciones para las ecuaciones de verosimilitud en el caso de que no exista una analítica para ellas (apartado 2.2). Seguidamente, se expone el problema de los valores iniciales, un aspecto al que el algoritmo se muestra especialmente sensible. La selección del mejor modelo se trata en el apartado 2.4. Finalmente, el capítulo acaba con los dos métodos utilizados en el trabajo para aproximar el valor de los errores de los estimadores EM: el método SEM y el replicativo *bootstrap*.

2.1. Modelos de mixturas finitas

Las distribuciones mixtas se utilizan para la modelización de datos heterogéneos en multitud de situaciones experimentales, en donde aquellos pueden interpretarse como procedentes de dos o más subpoblaciones (componentes). La obtención de estas componentes conduce a la estimación de los parámetros de la mixtura. Este problema de estimación tiene una larga historia y se remonta a Pearson (1894), quien trabajó con una mixtura de dos componentes con varianzas iguales usando el método de los momentos. Trabajos posteriores que utilizan esta aproximación son los de Charlier (1906), Charlier y Wicksell (1924), Cohen (1967), y Tan y Chang (1972). Rao (1948) y Hasselblad (1966, 1969) utilizaron la estimación de máxima verosimilitud en este contexto.

Un amplio rango de aplicaciones prácticas y un análisis estadístico detallado de las mixturas finitas, considerando diferentes métodos de estimación, fueron presentados por Everitt y Hand (1981), y Titterington et al. (1985). Descripciones más generales fueron publicadas por McLachlan y Basford (1988), McLachlan y Jones (1988), McLachlan y Krishnan (1997), y McLachlan y Peel (2000).

Una buena revisión histórica de estos modelos puede encontrarse en Holgersson y Jorner (1978), Redner y Homer (1984), y Everitt (1996). Algunas aplicaciones en un contexto médico fueron presentadas por Schlattmann (2009) y Frühwirth-Schnatter (2010). El trabajo reciente de Mengerse et al. (2011) muestra la relevancia de los modelos mixtos considerando un esquema bayesiano.

Para el desarrollo conceptual del algoritmo EM, que se verá más adelante, es conveniente proporcionar una formulación paramétrica para la representación del modelo, y se adoptará aquí, en general, la notación de McLachlan y Peel (2000). En lo sucesivo, mediante $Y = (Y_1, Y_2, \ldots, Y_n)$, se denotará a una muestra aleatoria de tamaño n, donde Y_j es un vector aleatorio q-dimensional con función de densidad de probabilidad $f(y_j)$ en $\mathbb{R}^q$. Así, $y = (y_1, y_2, \ldots, y_n)$ representa a una muestra observada o realización de Y, donde y_j constituye un valor observado del vector aleatorio Y_j.

Definición 2.1 *La distribución de una variable aleatoria Y_j cuya función de densidad se escribe*

$$f(y_j|\Psi) = \sum_{i=1}^{g} \pi_i \; f_i(y_j|\theta_i), \quad y_j \in \mathbb{R}^q, \tag{2.1}$$

se denomina distribución de mixtura finita de g componentes, con un vector de parámetros del modelo

$$\Psi = (\pi_1, \ldots, \pi_g, \theta_1, \ldots, \theta_g).$$

Así, $f_i(y_j|\theta_i)_{i=1,\ldots,g}$ denotan las *densidades de las componentes* de la mixtura con parámetros θ_i, y $\pi_1, \ldots, \pi_g$, las *proporciones o pesos*. Mediante la notación $f_i(\cdot|\theta_i)$, se asume que estas densidades pueden pertenecer a diferentes familias paramétricas, agrupándose cada uno de sus vectores parámetricos en θ_i.

Las proporciones de la mixtura representan las probabilidades de que la realización y_j de la variable aleatoria haya sido generada por las g diferentes densidades y, como probabilidades que son, están sujetas a las restricciones

$$0 \leqslant \pi_i \leqslant 1 \qquad i = 1, \ldots, g \tag{2.2}$$

y

$$\sum_{i=1}^{g} \pi_i = 1, \tag{2.3}$$

por lo que uno de los pesos resulta redundante. La Figura 2.1 muestra un ejemplo de mixturas de distribuciones gaussianas con dos componentes y diferente parametrización, creadas a partir de muestras sintéticas generadas mediante una misma semilla aleatoria.

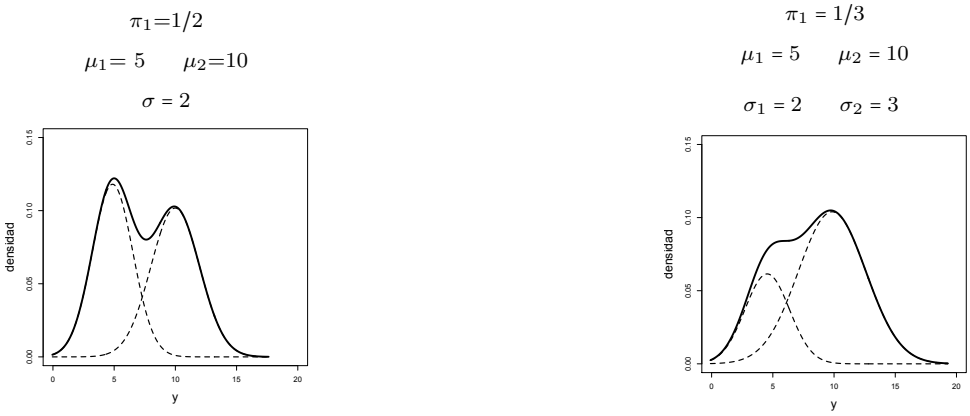


Figura 2.1 Las líneas discontinuas muestran la densidad de cada componente. A la izquierda, mixtura homocedástica.

Observación 2.1 Para el desarrollo experimental de este trabajo, las densidades de las componentes han sido del tipo gaussianas univariantes heterocedásticas, por lo que en (2.1) estas podrían haberse representado mediante $f(y_j|\theta_i)$, con $\theta_i = (\mu_i, \sigma_i)$, o bien $\phi(y_j|\mu_i, \sigma_i)$. Para mixturas cuyas componentes pertenecen a otras familias de densidades, puede consultarse Simar (1976), y Moharir (1992) para mixturas de distribuciones Poisson; Falls (1970), para mixturas Weibull; o Blischke (1962) y Medgyessy (1961), para mixturas binomiales.

Como se decía, dado que las distribuciones empleadas en este trabajo han sido las gaussianas, procede a continuación indicar su función de densidad.

Definición 2.2 *Se dice que una variable aleatoria Y sigue una distribución normal o gaussiana si su función de densidad puede escribirse como*

$$f(y|\theta) = \phi(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ exp\ (-\frac{(y-\mu)^2}{2\sigma^2}) \qquad -\infty < y < \infty \tag{2.4}$$

con $-\infty < \mu < \infty$, $\sigma^2 > 0$ *y* $\theta = (\mu, \sigma^2)$ *los parámetros de la distribución.*

2.1.1. Identificabilidad

Para un conjunto de observaciones $y_1, \ldots, y_n$ se persigue su ajuste a una distribución mixta mediante la estimación de todos los parámetros Ψ, como se definió en (2.1). La estimación de Ψ en función de las observaciones y_j solo tiene sentido si Ψ es identificable, lo que hace referencia a la existencia de un única caracterización para un modelo de mixtura. En general, una familia paramétrica de densidades $f(y_j|\Psi)$ es identificable si valores diferentes del parámetro Ψ determinan miembros distintos de la familia de densidades. Esto es

$$f(y_j|\Psi) = f(y_j|\Psi^*), \tag{2.5}$$

Si y solo si

$$\Psi = \Psi^*.$$

En las mixturas de distribuciones la identificabilidad es algo diferente. Si $f(y_j|\Psi)$ posee dos componentes con densidades $f_i(y_j|\theta_i)$ y $f_h(y_j|\theta_h)$ que pertenecen a la misma familia paramétrica, entonces (2.5) solo es cierta si los índices de las componentes i y h se intercambian en Ψ. Aunque esta clase de mixturas pueda ser identificable, Ψ no lo es. Por tanto, la identificabilidad en el caso de mixturas finitas se interpreta como sigue:

Definición 2.3 *Sean*

$$f(y_j|\Psi) = \sum_{i=1}^{g} \pi_i\ f_i(y_j|\theta_i) \qquad y \qquad f(y_j|\Psi^*) = \sum_{i=1}^{g^*} \pi_i^*\ f_i(y_j|\theta_i^*)$$

dos miembros cualesquiera de una familia paramétrica de mixturas. Esta clase de mixturas finitas se dice que es identificable para Ψ si

$$f(y_j|\Psi) \equiv f(y_j|\Psi^*)$$

Si y solo si $g = g^*$ y las etiquetas de las componentes pueden permutarse tal que

$$\pi_i = \pi_i^* \qquad y \qquad f(y_j|\theta_i) \equiv f(y_j|\theta_i^*) \qquad i = 1, \ldots, g.$$

Con esta definición, las mixturas de distribuciones normales son identificables si pueden permutarse los índices de las componentes. Para solventar este intercambio, pueden imponerse restricciones a la solución, como, por ejemplo, $\mu_1 < \mu_1 < \ldots < \mu_k$ (Aitkin y Rubin, 1985). Por tanto, en la práctica, todos los parámetros pueden ser determinados con exactitud. Para una descripción detallada del concepto de identificabilidad, puede consultarse Titterington et al. (1985), Teicher (1961, 1963), Frühwirth-Schnatter (2010), y Yakowitz y Spragins (1968). La pérdida de identificabilidad no supone un inconveniente en el ajuste mediante mixturas de distribuciones por el método de la máxima verosimilitud, como es el caso en el empleo del algoritmo EM (McLachlan y Peel, 2000).

2.1.2. Mixtura de 3 componentes gaussianas

Como podrá comprobarse en el Capítulo 5 (Tabla 5.2), las mixturas finitas obtenidas para la modelización de las distribuciones de datos experimentales no presentaron en ningún caso más de tres componentes ($g \leq 3$). Por ello cobra sentido el presente apartado, que contiene la definición de este tipo de mixtura y donde se expone la utilidad del cálculo de sus momentos.

Definición 2.4 *Una distribución cuya función de densidad es*

$$g(y|\Psi) = \pi_1\ \phi(y|\mu_1, \sigma_1^2) + \pi_2\ \phi(y|\mu_2, \sigma_2^2) + \pi_3\ \phi(y|\mu_3, \sigma_3^2) \tag{2.6}$$

se dice que sigue una mixtura de tres componentes gaussianas, donde $\phi(\cdot)$ es la función de densidad gaussiana, como se definió en (2.4) *y*

$$\Psi = (\pi_1,\ \pi_2,\ \mu_1,\ \mu_2,\ \mu_3,\ \sigma_1^2,\ \sigma_2^2,\ \sigma_3^3).$$

Para asegurar la identificabilidad de Ψ, se asume que las medias de las componentes se encuentran en orden ascendente,

$$\mu_1\ < \mu_2\ < \mu_3.$$

En la Figura 2.2 se muestran algunos ejemplos de mixturas de distribuciones con 3 componentes gaussianas con diferentes parametrizaciones. Como en la Figura 2.1, para su representación se crearon muestras artificiales de 3 componentes, a partir de una misma semilla, estimándose posteriormente los parámetros mediante el algoritmo EM especialmente diseñado para este trabajo.

Observación 2.2 El solapamiento de las componentes en mixturas gaussianas puede determinarse cuantitativamente a través de la diferencia de cada una de las medias de las componentes. No obstante, es necesario suponer una condición de homocedasticidad entre las componentes ($\sigma = \sigma_i\,, i = 1, \ldots, g$). Dado que no es el caso en este trabajo, su desarrollo se omite, aunque se remite a McLachlan y Peel (2000, cap.2) para su consulta.

Lema 2.1 Suponiendo que Y es una variable aleatoria que sigue una distribución mixta de tres componentes gaussianas definida como en (2.6), y que los momentos de primer y segundo orden de las mismas existan, entonces, la media μ_m y varianza σ_m^2 de la mixtura son:

$$\mu_m = \pi_1\mu_1 + \pi_2\mu_2 + \pi_3\mu_3.$$
$$\sigma_m^2 = \pi_1(\sigma_1^2 + \mu_1^2) + \pi_2(\sigma_2^2 + \mu_2^2) + \pi_3(\sigma_3^2 + \mu_3^2) - \mu_m^2.$$

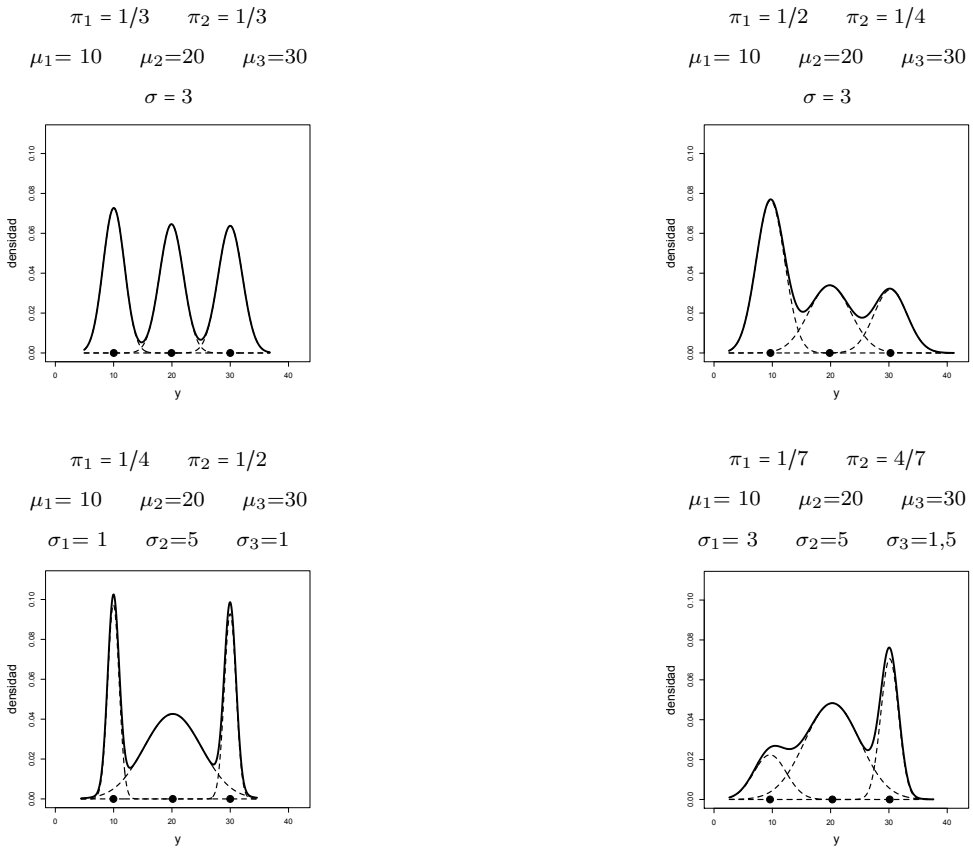


Figura 2.2 Ejemplos de mixturas con 3 componentes gaussianas, homocedásticas en la fila superior. Los puntos en el eje de abcisas representan la media de cada componente.

Demostración 2.1 Utilizando el operador E como esperanza, la media y la varianza de la mixtura se calculan

$$\begin{aligned}
\mu_m &= \mathsf{E}[Y|\Psi] = \\
&= \int_{-\infty}^{\infty} y\ g(y|\Psi)\ dy \\
&= \int_{-\infty}^{\infty} y\ \pi_1\ f(y|\theta_1)\ dy + \int_{-\infty}^{\infty} y\ \pi_2\ f(y|\theta_2)\ dy + \int_{-\infty}^{\infty} y\ \pi_3\ f(y|\theta_3)\ dy \\
&= \pi_1\mu_1 + \pi_2\mu_2 + \pi_3\mu_3.
\end{aligned}$$

$$\begin{aligned}
\sigma_m^2 &= \mathsf{E}[Y^2|\Psi] - \mathsf{E}[Y|\Psi]^2 = \\
&= \int_{-\infty}^{\infty} y^2\ g(y|\Psi)\ dy\ - \mu_m^2 \\
&= \pi_1 \int_{-\infty}^{\infty} y^2\ f(y|\theta_1)\ dy + \pi_2 \int_{-\infty}^{\infty} y^2\ f(y|\theta_2)\ dy + \pi_3 \int_{-\infty}^{\infty} y^2\ f(y|\theta_3)\ dy - \mu_m^2 \\
&= \pi_1(\sigma_1^2 + \mu_1^2) + \pi_2(\sigma_2^2 + \mu_2^2) + \pi_3(\sigma_3^2 + \mu_3^2) - \mu_m^2.
\end{aligned}$$

Expresiones que pueden generalizarse mediante:

$$\mu_m = \sum_{i=1}^{g} \pi_i \mu_i \qquad \sigma_m^2 = \sum_{i=1}^{g} \pi_i(\mu_i^2 + \sigma_i^2) - \mu_m^2. \tag{2.7}$$

La ventaja de la utilización de estos momentos reside en que permite resumir la parametrización completa de una mixtura mediante dos simples números, μ_m y σ_m^2. Por simplicidad, para describir la variabilidad de la mixtura, se utilizará $\sigma_m = \sqrt{\sigma_m^2}$.

2.1.3. Estimación mediante máxima verosimilitud

Existen numerosos métodos de estimación puntuales, entre los que se incluyen el método de los momentos, procedimientos gráficos, bayesiana, mínimo cuadrática, mínimo χ^2 o máxima verosimilitud. La elección del método más adecuado para la estimación de los parámetros en las mixturas ha sido motivo de controversia y una respuesta parcial puede encontrarse en Tan y Chang (1972). Estos autores realizaron una comparación entre el método de los momentos y el de máxima verosimilitud, demostrando que el segundo es superior. Holgerson y Jorner (1978) compararon igualmente varios métodos de estimación, llegando igualmente a semejante conclusión.

Finalmente, Day (1969) mencionó las ventajas de estimación mediante máxima verosimilitud sobre el mínimo χ^2 y los estimadores bayesianos. Por tanto, la estimación mediante máxima verosimilitud es la utilizada en este trabajo, lo que conlleva la maximización de la función de verosimilitud, o equivalentemente, la maximización de la función de log-verosimilitud, la cual es descrita, por ejemplo, en Little y Rubin (2002).

Definición 2.5 *Sea $y = (y_1, y_2, \ldots, y_n)$ observaciones independientes de una variable aleatoria Y con función de densidad $f(y|\theta)$, donde θ es el vector de parámetros desconocidos que queremos estimar; entonces, la función de densidad conjunta de y se escribe*

$$f(y|\theta) = \prod_{j=1}^{n} f(y_j|\theta) = L(\theta|y) \tag{2.8}$$

donde $L(\theta|y)$ representa la función de verosimilitud y se considera una función de θ.

Observación 2.3 Para resaltar el hecho de que la función de verosimilitud es una función de θ, se denota también como f_θ. El proceso de estimación está basado en la asignación de un valor al parámetro desconocido θ que caracteriza una población, y que es observada a través de la muestra aleatoria Y. La idea que subyace en el método de máxima verosimilitud es dar como estimación del parámetro aquel valor, de entre los posibles, que haga máxima la probabilidad de la muestra observada. Así, se considera que es preciso ajustar el valor de θ, permaneciendo fijos los valores de la muestra.

Como el máximo de una función y el de su logaritmo se alcanzan en el mismo valor de θ, habitualmente resulta más simple emplear su transformación logarítmica:

$$\ell(\theta|y) = \log L(\theta|y) = \log \prod_{j=1}^{n} f(y_j|\theta) = \sum_{j=1}^{n} \log f(y_j|\theta)$$

Definición 2.6 *Un estimador de máxima verosimilitud $\hat{\theta}$ de θ es un valor que maximiza $L(\theta|y)$, es decir,*

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

donde Θ representa el espacio paramétrico.

Esta definición permite la posibilidad de obtener más de un estimador de máxima verosimilitud, como puede ocurrir en aproximaciones experimentales donde se detectan múltiples máximos. No obstante, para muchos modelos destacables "el estimador de máxima verosimilitud es único, y más aún, la función de verosimilitud es diferenciable y acotada superiormente" (Little y Rubin, 2002, p. 81). En tales casos, puede encontrarse una solución mediante la resolución de las correspondientes ecuaciones normales.

Definición 2.7 *Las ecuaciones normales o de verosimilitud vienen dadas por*

$$S(y|\theta) = \frac{\partial \ell(\theta|y)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, k.$$

en el supuesto de que $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ sea un parámetro k dimensional.

Definición 2.8 *Sea*

$$I(\theta|y) = -\frac{\partial^2}{\partial\theta\partial\theta^T}\ \ell(\theta|y)$$

las derivadas segundas parciales negativas de la función de log-verosimilitud con respecto a θ, donde θ^T denota el vector transpuesto de θ. Entonces, $I(\theta|y)$ se denomina la matriz de información observada. La matriz de información esperada viene dada, bajo condiciones de regularidad, por

$$\mathcal{I}(\theta|y) = \mathsf{E}[S(y|\theta)\, S^T(y|\theta)] = \mathsf{E}[I(\theta|y)].$$

Observación 2.4 Esta definición es aplicable en el caso de la mixtura sustituyendo θ por Ψ.

Ejemplo 2.1 Sea $(Y_1, Y_2, \ldots, Y_n)$ una muestra aleatoria simple de una población $N(\mu, \sigma^2)$, como se definió en (2.4), siendo $\theta = (\mu, \sigma^2)$ dos parámetros desconocidos. Entonces, las funciones de verosimilitud y log-verosimilitud son

$$L(\theta|y) = f_\theta(y_1, y_2, \ldots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\}.$$

$$\begin{aligned}\ell(\theta|y) &= -\frac{n}{2}\ \log\,\sigma^2 + \log\left(\frac{1}{\sqrt{2\pi}}\right)^n - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2 \\ &= -\frac{n}{2}\ \log\,\sigma^2 - \frac{n}{2}\ \log(2\pi) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2.\end{aligned}$$

La derivada de $\ell(\theta|y)$ respecto a cada uno de los parámetros μ y σ^2 permite obtener las siguientes ecuaciones de verosimilitud:

$$\frac{\partial}{\partial\mu}\,\ell(\theta|y) = \frac{1}{2\sigma^2}\sum_{t=1}^{n} 2\,(y_i-\mu) = \frac{1}{\sigma^2}\sum_{j=1}^{n}(y_i-\mu) = 0,$$

$$\frac{\partial}{\partial\sigma^2}\,\ell(\theta|y) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{2\,\sum_{j=1}^{n}(y_j-\mu)^2}{4(\sigma^2)^2} = -\frac{1}{2\sigma^2}\left(-n + \frac{\sum_{j=1}^{n}(y_j-\mu)^2}{\sigma^2}\right) = 0,$$

representando un sistema de dos ecuaciones con dos incógnitas, que tiene como soluciones la media y la varianza muestral:

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \qquad\qquad \hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j-\bar{y})^2$$

Para la comprobación de que se trata de un máximo, se calculan las segundas derivadas en $\ell(\theta|y)$:

$$\frac{\partial^2}{\partial\theta\partial\theta^T}\,\ell(\theta|y) = -\begin{pmatrix} \dfrac{n}{\sigma^2} & \dfrac{1}{\sigma^4}\displaystyle\sum_{j=1}^{n}(y_j-\mu) \\[2ex] \dfrac{1}{\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu) & \dfrac{1}{\sigma^6}\displaystyle\sum_{j=1}^{n}(y_j-\mu)^2 - \dfrac{n}{2\sigma^4} \end{pmatrix} = I(\theta|y),$$

matriz que es definida negativa cuando $\hat{\mu} = \bar{y}$ y $\hat{\sigma}^2 = s^2$.

La matriz de información esperada, entonces:

$$\mathcal{I}(\theta|y) = \mathsf{E}[I(\theta|y)] = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\[2ex] 0 & \dfrac{n\sigma^2}{\sigma^6} - \dfrac{n}{2\sigma^4} \end{pmatrix} = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\[2ex] 0 & \dfrac{n}{2\sigma^4} \end{pmatrix},$$

que tiene por inversa:

$$\mathcal{I}(\theta|y)^{-1} = \begin{pmatrix} \dfrac{\sigma^2}{n} & 0 \\[2ex] 0 & \dfrac{2\sigma^4}{n} \end{pmatrix}.$$

Definición 2.9 *Sea $y = (y_1, y_2, \ldots, y_n)$ observaciones independientes de una variable aleatoria Y con función de densidad $f(y|\Psi)$, donde Ψ es el vector de parámetros desconocidos que queremos estimar, entonces,*

$$L(\Psi|y) = \prod_{j=1}^{n} f(y_j|\Psi) = \prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i\, f_i(y_j|\theta_i), \tag{2.9}$$

recibe el nombre de función de verosimilitud de la mixtura, que, tomando logaritmos en $L(\Psi|y)$, conduce a su función de log-verosimilitud:

$$\ell(\Psi|y) = log\; L(\Psi|y) = log\;\prod_{j=1}^{n}\left\{\sum_{i=1}^{g}\pi_i\, f_i(y_j|\theta_i)\right\} = \sum_{j=1}^{n} log\left\{\sum_{i=1}^{g}\pi_i\, f_i(y_j|\theta_i)\right\}, \tag{2.10}$$

cuya correspondiente ecuación de verosimilitud es

$$\frac{\partial}{\partial \Psi} \sum_{j=1}^{n} log \left\{ \sum_{i=1}^{g} \pi_i \ f_i(y_j|\theta_i) \right\} = 0. \tag{2.11}$$

Esta expresión, debido a la presencia del logaritmo dentro de una suma, es de difícil resolución y requiere, en el mejor de los casos, procedimientos iterativos para ser resuelta. En el caso de que $f(\cdot)$ presente una forma compleja, puede llegar a no tener una solución analítica (McLachlan y Basford, 1988, cap.2).

Entre estos procedimientos, los más ampliamente utilizados son el método de Newton-Raphson (NR) y *scoring* de Fisher. El algoritmo EM se ha convertido en otra técnica estándar para el cálculo de los EMV. Este algoritmo representa en sí mismo un esquema de trabajo para la obtención de estimaciones en problemas de datos incompletos. La idea que en él subyace es resolver problemas de datos incompletos de cierta complejidad abordando de forma repetida una situación de datos completos de resolución más asequible. Para ello es necesario considerar que la población en estudio, aunque sin datos faltantes, los posee. Así, se asigna a cada dato observado una etiqueta que indique su pertenencia a una u otra subpoblación en la muestra de partida, asumiendo que estas existen. De este modo, la asignación de un conjunto de datos a una u otra componente de la mixtura conduce de forma natural al agrupamiento de estos o a la formación de *clústeres*.

El valor de estas etiquetas o indicadores de pertenencia a un clúster es, a priori, desconocido. Así, bajo una distribución mixta, la estimación de su valor puede considerarse como un problema de datos faltantes, y el algoritmo EM puede utilizarse. A diferencia del método scoring de Fisher, EM no requiere del cálculo de una matriz Hessiana en cada iteración, en particular la inversa de la matriz de información, lo que supone una ventaja si esta es compleja de calcular. Esta matriz Hessiana puede ser aproximada numéricamente mediante simulación empleando métodos de Montecarlo, aunque el coste computacional puede ser elevado si se requieren numerosas iteraciones para su obtención.

2.2. El algoritmo EM

2.2.1. Introducción al algoritmo

El algoritmo EM fue descrito y presentado por Dempster et al. (1977) para la obtención de un máximo de la función de verosimilitud cuando su cálculo no es posible. Además de problemas con datos censurados o perdidos, como se ha mencionado anteriormente, el algoritmo puede utilizarse también en otras situaciones en donde no se considera aparentemente la ausencia de observaciones, como es el caso de las mixturas finitas. Bajo este último enfoque, se plantea la necesidad de formular el conjunto de datos observados como un caso de datos incompletos. El extenso ámbito de aplicación de este algoritmo en diferentes campos de aplicación lo ha hecho ciertamente popular. En 1992, Meng y Pedlow publicaron una relación bibliográfica con más de 1000 artículos relacionados con el algoritmo EM. McLachlan y Khirshan (1997) estimaron al menos 1700 publicaciones.

La idea básica del algoritmo fue anterior a Dempster et al. (1977). La primera referencia en la literatura sobre un algoritmo tipo EM pertenece a Newcomb (1886), quien consideró la estimación de los parámetros de una mixtura gaussiana de dos componentes. Este trabajo fue seguido por muchos otros, como McKendrick (1926), quien presentó una aplicación en un contexto médico, y Healy y Westmacott (1956), quienes propusieron un ejemplo del algoritmo EM en un diseño por bloques completamente aleatorizado. Baum et al. (1970) utilizaron este algoritmo conjuntamente con modelos de Markov, y Orchad y Woodbuty (1972) trabajaron en un algoritmo similar denominado "principio de información perdida". Un resumen estructurado de la historia del algoritmo EM puede encontrarse en McLachlan y Krishnan (1997), o en Redner y Hoomer (1984). Otro resumen interesante al respecto puede encontrarse en McLachlan y Jones (1988), y en Little y Rubin (2002).

El algoritmo EM posee ciertas ventajas atractivas comparado con otros algoritmos iterativos, como el algoritmo de NR o método scoring de Fisher. Por ejemplo, la economía de almacenamiento, la facilidad de implementación y la estabilidad numérica. Más aún, en la mayoría de situaciones prácticas, el algoritmo EM converge en condiciones de regularidad a un máximo local, concepto que se desarrolla al final de esta sección.

No obstante, la utilización de este algoritmo en muchas situaciones estadísticas revela su limitación. Por ello, se han desarrollado numerosas modificaciones y extensiones. Especialmente, el hecho de que en ciertas situaciones el algoritmo converge muy lentamente ha llevado al desarrollo de modificaciones de su aceleración. Por ejemplo, Redner y Homer (1984) recomendaron utilizar el algoritmo EM de forma conjunta con un algoritmo de tipo Newton, donde las buenas propiedades de convergencia del algoritmo EM se sumarán a la rápida convergencia local del método de Newton. Otro algoritmo híbrido, denominado EM/GN, fue introducido por Aitkin y Aitkin (1996), donde el algoritmo EM es combinado con el algoritmo Gauss-Newton (GN). Igualmente, Du (2002) propuso la combinación del algoritmo EM con el NR. Para más detalles sobre algoritmos EM modificados y algoritmos híbridos, se recomienda dirigirse a McLachlan y Krishnan (1997), y Little y Rubin (2002).

Datos incompletos y completos

Definición 2.10 *Sea $y = (y_1, y_2, \ldots, y_n)$ una muestra observada de tamaño n, a la que denominaremos vector de datos incompleto, correspondiente a una realización de Y, con función de densidad $f(y|\Psi)$, donde Ψ es el vector de parámetros que se desea estimar. Sea igualmente una variable $Z = (Z_1, Z_2, \ldots, Z_n)$, que denominaremos latente, que representará a los datos no observados, y cuya realización es $z = (z_1, z_2, \ldots, z_n)$. El vector aleatorio $X = (Y, Z)$ recibe el nombre de vector de datos completos y sus realizaciones serán $x_1 = (y_1, z_1)$, $x_2 = (y_2, z_2), \ldots,$ $x_n = (y_n, z_n)$, de tal forma que a una realización y_j le corresponde siempre otra z_j.*

Observación 2.5 La presencia de estas nuevas variables asociadas a unas observaciones, ahora consideradas incompletas, se entiende como un *aumento* de los datos, denominación que surge naturalmente en aplicaciones que presentan datos faltantes (Tanner, 1987). Las funciones de verosimilitud y log-verosimilitud de la mixtura dadas en (2.9) y (2.10) se denominan entonces, cada una de ellas, *incompletas*.

En este contexto, Z_j representa a una variable indicadora binaria g-dimensional cuyo elemento i-ésimo, Z_{ij}, indica la pertenencia de la observación y_j a la componente i-ésima de la mixtura ($i = 1, \ldots, g$; $j = 1, \ldots, n$). Es decir, $z_{ij} \in \{0, 1\}$ y:

$$Z_{ij} = \begin{cases} 1 & \text{si la observación } y_j \text{ proviene de la componente } i\text{-ésima} \\ 0 & c.c. \end{cases}$$

La representación matricial del vector de datos incompletos (y) y completos (x) es:

$$y' = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad x' = \begin{pmatrix} y' & z' \end{pmatrix} = \begin{pmatrix} y_1 & z_1 \\ y_2 & z_2 \\ \vdots & \vdots \\ y_n & z_n \end{pmatrix} = \begin{pmatrix} y_1 & z_{11} & \cdots & z_{1g} \\ y_2 & z_{21} & \cdots & z_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & z_{n1} & \cdots & z_{ng} \end{pmatrix}.$$

Observación 2.6 Cada variable Z_j toma $(g-1)$ valores 0 y un único valor 1. Por su parte, las variables Z_{ij} son referidas en la literatura como $I_{\{Z_j=i\}}$, $\mathbf{1}_{\{Z_j=i\}}$ u otra representación adecuada de variable indicadora.

Dada la naturaleza categórica de las variables Z_{ij} al indicar la pertenencia ("*labelling*") de los puntos muestrales a una componente u otra de la mixtura, puede asumirse que Z_j sigue una distribución

multinomial de solo una realización sobre g categorías con probabilidades $\pi = (\pi_1, \ldots, \pi_g)$, es decir, la función de probabilidad de Z_j será:

$$Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} \text{Mult}_g(1, \pi),$$

$$P(Z_j = z_j) = \binom{1}{z_{j1}\ z_{j2}\ \cdots\ z_{jg}} \pi_1^{z_{j1}}\ \pi_2^{z_{j2}}\ \cdots\ \pi_g^{z_{jg}} = \prod_{i=1}^{g} \pi_i^{z_{ji}}, \tag{2.12}$$

verificándose de nuevo las restricciones (2.2) y (2.3), y además :

$$\sum_{i=1}^{g} z_{ij} = 1 \qquad \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij} = n.$$

Observación 2.7 Titterington (1990) denomina a este modelo multinomial oculto.

Función de log-verosimilitud de los datos completos

Teniendo en cuenta la relación entre la densidad marginal y condicional de una observación, la función de densidad conjunta de una observación, que denominamos *completa*, es:

$$f(x_j) = f(y_j, z_j) = f(y_j|z_j)\ p(z_j). \tag{2.13}$$

Las variables $Z_1, \ldots, Z_n$, con $Z_j = \{Z_{j1}, \ldots, Z_{jg}\}$, están relacionadas con la observación muestral Y_j condicionalmente, siendo esta la única información de la que se dispone de la distribución de Z_j:

$$f_i(Y_j|Z_{ji} = 1) \sim f_i(y_j|\theta_i). \tag{2.14}$$

Desarrollando (2.13), la distribución conjunta de Y_j y todos los estados posibles de Z_j es:

$$\begin{aligned} f_i(Y_j, Z_j) &= f_i(Y_j = y_j, Z_{j1} = z_{j1}, \ldots, Z_{jg} = z_{jg}) = \\ &= f_i(Y_j = y_j \mid Z_{j1} = z_{j1}, \ldots, Z_{jg} = z_{jg}) \times P(Z_{j1} = z_{j1}, \ldots, Z_{jg} = z_{jg}) \\ &= \left\{\prod_{i=1}^{g}\left[f_i(y_j|\theta_i)^{z_{ji}}\right]\right\} \times \left\{\prod_{i=1}^{g}\pi_i^{z_{ji}}\right\} = \prod_{i=1}^{g}\left[\pi_i f_i(y_j|\theta_i)\right]^{z_{ji}}. \end{aligned}$$

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

$$L(\Psi|y, z) = \prod_{j=1}^{n}\prod_{i=1}^{g}\left[\pi_i f_i(y_j|\theta_i)\right]^{z_{ij}},$$

que tras aplicar la función logaritmo, se obtiene la función de log-verosimilitud de los datos completos:

$$\ell(\Psi|y,z) = \log\, L(\Psi|y,z) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\,[\,\pi_i f_i(y_j|\theta_i)\,] = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ [\log\ \pi_i\ +\ \log\ f_i(y_j|\theta_i)\,] =$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\ \pi_i\ + \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\ f_i(y_j|\theta_i) \qquad \Psi = (\pi_i,\theta_i)_{i=1,\ldots,g}. \tag{2.15}$$

Observación 2.8 Si se compara esta expresión con la (2.10), se observa que la función log no precede a una suma, sino que actúa directamente sobre una función de densidad, que, al ser miembro de la familia exponencial, favorece la operatoria de maximización.

La función de log-verosimilitud de una mixtura gaussiana, con $\theta_i = (\mu_i, \sigma_i^2)$, considerando la forma de (2.15) es:

$$\begin{aligned}
\ell(\Psi|y,z) &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\pi_i\ +\ \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\,\phi(y_j|\mu_i,\sigma_i^2) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\pi_i\ +\ \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \left[-\frac{1}{2}\log 2\pi - \log\sigma_i - \frac{(y_j-\mu_i)^2}{2\sigma_i^2}\right] \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\pi_i\ -\ \frac{1}{2}\,n\,\log 2\pi\ -\ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\left[2\log\sigma_i + \frac{(y_j-\mu_i)^2}{\sigma_i^2}\right] \\
&= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\pi_i\ -\ \frac{n}{2}\,\log 2\pi\ -\ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\left[\log\sigma_i^2 + \frac{(y_j-\mu_i)^2}{\sigma_i^2}\right].
\end{aligned} \tag{2.16}$$

Agrupamiento (clustering) mediante modelos de mixturas

Una vez definidas las variables Z_{ij}, puede introducirse ahora el concepto de agrupamiento sobre los datos observados. Uno de los propósitos de los modelos mixtos es el de proporcionar una partición de los datos en g grupos, siendo g un número previamente establecido. La i-ésima proporción de la mixtura ($\pi_i,\ i = 1,\ldots,g$) puede interpretarse como la probabilidad *a priori* de que una observación muestral pertenezca a la población g, así:

$$P(Z_{ij} = 1) = \pi_i \qquad i = 1,\ldots,g$$

Bajo la especificación de los datos completos, el procedimiento de agrupamiento tiene como objetivo asociar cada una de las variables $z_1,\ldots,z_n$ con los datos observados $y_1,\ldots,y_n$. Una vez que el modelo de mixtura ha sido ajustado y su parámetro Ψ estimado, se proporciona un agrupamiento probabilístico de las observaciones en términos de las probabilidades posteriores (Teorema de Bayes) de pertenencia a uno u otro clúster:

$$\hat{\tau}_{ij} = P\{Z_{ij} = 1|Y_j = y_j\} = \frac{\hat{\pi}_i f_i(y_j|\hat{\theta}_i)}{\sum_{\ell=1}^{g}\hat{\pi}_\ell f_\ell(y_j|\hat{\theta}_\ell)} \qquad \ell = 1,\ldots,g \qquad j = 1,\ldots,n \tag{2.17}$$

Por tanto, $\hat{\tau}_{1j},\hat{\tau}_{2j},\ldots,\hat{\tau}_{gj}$, representan las probabilidades (posteriores) de que el punto y_j pertenezca a la g-ésima componente de la mixtura. Finalmente, la asignación de una observación a uno u otro clúster se decide mediante la mayor de estas probabilidades:

$$\hat{z}_{ij} = \begin{cases} 1 & \text{si} \quad i = \arg\max_{\ell} \hat{\tau}_{\ell j} \\ 0 & c.c. \end{cases} \qquad i = 1, \ldots, g \qquad j = 1, \ldots, n \tag{2.18}$$

2.2.2. Formulación del algoritmo

A continuación, se presenta la formulación del algoritmo como en McLachlan y Krishnan (1997). Cada iteración del algoritmo consta de dos etapas, la etapa de esperanza y la de maximización, nombres que son abreviados con la denominación paso E y paso M, debiendo esta terminología a Dempster et al. (1977). Como ya se dijo, la idea básica es asociar a un problema de datos incompletos dado otro de datos completos para el cual la estimación de máxima verosimilitud es más manejable.

De forma resumida, el paso E adecúa los datos completos, lo cual incluye el cálculo de la función de verosimilitud del conjunto de datos completos. Como esta función de verosimilitud está basada parcialmente en datos no observados, es reemplazada por su esperanza condicional, dados los datos observados, utilizando el valor $\Psi^{(t)}$. Finalmente, el paso M maximiza esta función de verosimilitud de datos completos sobre Ψ. Comenzando con un conjunto de valores de parámetros iniciales adecuados, el algoritmo itera hasta la convergencia:

$$\Psi^{(0)} \longrightarrow \Psi^{(1)} \longrightarrow \cdots \longrightarrow \Psi^{(t)} \longrightarrow \Psi^{(t+1)} \longrightarrow \cdots \longrightarrow \Psi^{(\infty)} = \hat{\Psi}$$

Observación 2.9 A continuación se describe el paso E y M para la primera iteración, siendo válido para cualquier par de iteraciones consecutivas t y $t+1$.

– **Paso E**. Sea $\Psi^{(0)}$ algún valor inicial de Ψ. Entonces, en la primera iteración, el paso E requiere el cálculo de la esperanza condicional de la función de log-verosimilitud de los datos completos, dado el dato observado y, y empleando el valor inicial $\Psi^{(0)}$, que se representa:

$$\mathsf{E}\left[\ \ell(\Psi|y, Z) \mid Y = y,\ \Psi^{(0)}\ \right] := Q(\ \Psi \mid \Psi^{(0)}\ ) \tag{2.19}$$

Observación 2.10 Esta esperanza, que utiliza el valor de Ψ, se denota habitualmente como E_{Ψ}.

Desarrollando (2.19), y teniendo en cuenta la linealidad de $\mathsf{E}(\cdot)$ sobre los datos no observados Z_{ij}:

$$\begin{aligned} Q(\ \Psi \mid \Psi^{(0)}\ ) &= \mathsf{E}\ \left[\ \ell(\Psi|y, Z) \mid Y = y,\ \Psi^{(0)}\ \right] \\ &= \mathsf{E}\ \left[\ \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\ \log\left[\pi_i f_i(y_j|\theta_i)\right] \mid Y = y,\ \Psi^{(0)}\ \right] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} \mathsf{E}\left[\ z_{ij} \mid Y_j = y_j, \Psi^{(0)}\right]\left[\ \log\pi_j\ + \log f_i(y_j|\theta_i)\ \right] \end{aligned} \tag{2.20}$$

Por lo tanto, el paso E solo requiere el cálculo del primer factor en (2.20):

$$\mathsf{E}\left[\ Z_{ij} \mid Y_j = y_j, \Psi^{(0)}\right] = P(Z_{ij} = 1 | Y_j = y_j, \Psi^{(0)})$$

$$= \left. \frac{f_i(Y_j = y_j | Z_{ij} = 1) P(Z_{ij} = 1)}{\sum_{\ell=1}^{g} f_i(Y_j = y_j | Z_{\ell j} = 1) P(Z_{\ell j} = 1)} \right|_{\Psi^{(0)}}$$

$$= \left. \frac{\hat{\pi}_i \ f_i(y_j | \hat{\theta}_i)}{\sum_{i=1}^{g} \hat{\pi}_i \ f_i(y_j | \hat{\theta}_i)} \right|_{\Psi^{(0)}} := \ \hat{\tau}_{ij}^{(0)} \tag{2.21}$$

Por lo que la expresión (2.20) puede reescribirse:

$$Q(\ \Psi \mid \Psi^{(0)}\ ) = \sum_{j=1}^{n} \sum_{i=1}^{g} \ \hat{\tau}_{ij}^{(0)} \left[\ \log \pi_i \ + \ \log f_i(y_j | \theta_i)\ \right]$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{g} \ \hat{\tau}_{ij}^{(0)} \log \pi_i + \sum_{j=1}^{n} \sum_{i=1}^{g} \ \hat{\tau}_{ij}^{(0)} \log f_i(y_j | \theta_i) \tag{2.22}$$

Así, $\hat{\tau}_{ij}$ en (2.17) es reemplazado por $\hat{\tau}_{ij}^{(0)}$. Estas $\hat{\tau}_{ij}^{(0)}$ representan probabilidades estimadas de que el punto y_j pertenezca a las componentes $i = 1, \ldots, g$ de la mixtura, siendo la mayor probabilidad entre ellas la que asigne el mismo a una u otra componente, es decir:

$$\hat{z}_{ij}^{(0)} = \begin{cases} 1 & \text{si} \quad i = \arg\max_j \hat{\tau}_{ij}^{(0)}(y_j | \Psi^{(0)}) \\ 0 & c.c. \end{cases} \qquad i = 1, \ldots, g \qquad j = 1, \ldots, n$$

Observación 2.11 Estas son las probabilidades posteriores ("responsabilities"). Con frecuencia, a los valores de estas probabilidades se les denomina asignaciones *soft*, dado que toman valores en [0,1], en contraste con las probabilidades de asignación a un clúster que se obtienen, por ejemplo, en el algoritmo *k-means*, que son referidas como *hard* al tomar valores en {0,1} o en $\{1,\ldots,g\}$.

– **Paso M**. Este paso requiere a continuación la maximización de la función Q con respecto a Ψ. Dado que π_i aparece únicamente en el primer término de (2.22) y que θ_i lo hace en el segundo, esta maximización puede realizarse independientemente. Comenzando con la maximización del primer término, es necesario resolver:

$$\frac{\partial}{\partial \pi_i} \left(\sum_{j=1}^{n} \sum_{i=1}^{g} \ \hat{\tau}_{ij}^{(0)} \ \log \pi_i \ + \lambda \left[\sum_{i=1}^{g} \pi_i - 1 \right] \right) = 0$$

en donde se ha introducido una restricción con un multiplicador de Lagrange (λ). Así:

$$\sum_{j=1}^{n} \hat{\tau}_{ij}^{(0)} \frac{1}{\pi_i} + \lambda = 0 \tag{2.23}$$

$$\sum_{j=1}^{n} \hat{\tau}_{ij}^{(0)} = -\lambda \, \pi_i \tag{2.24}$$

Sumando sobre g, en ambos términos de (2.24) obtenemos:

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)} = n$$
$$\sum_{i=1}^{g} -\lambda\,\pi_i = -\lambda\sum_{i=1}^{g}\pi_i = -\lambda$$

Por lo que:

$$-\lambda = n$$

Sustituyendo en (2.23) obtenemos un estimador iterativo para π_i :

$$\hat{\pi}_i^{(1)} = \frac{1}{n}\sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}. \tag{2.25}$$

La maximización de (2.22) respecto a θ_i depende de la función de densidad $f_i(y_j|\theta_i)$, extremo que se desarrolla a continuación para el caso particular de las distribuciones gaussianas:

$$\log f_i(y_j|\theta_i) = \log\phi(y_j|\mu_i,\sigma_i^2)$$
$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\frac{1}{2}(y_j-\mu_i)^2}{\sigma_i^2} = -\frac{1}{2}\log(2\pi) - \log\,\sigma_i - \frac{(y_j-\mu_i)^2}{2\,\sigma_i^2}.$$

Comenzando con μ_i, la maximización incluye el cálculo de:

$$\frac{\partial}{\partial\mu_i}\sum_{j=1}^{n}\sum_{i=1}^{g}\hat{\tau}_{ij}^{(0)}\left(-\frac{1}{2}\log(2\pi) - \log\,\sigma_i - \frac{(y_j-\mu_i)^2}{2\,\sigma_i^2}\right) = 0, \tag{2.26}$$

Obteniendo:

$$2\sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}\left(\frac{y_j-\mu_i}{2\,\sigma_i^2}\right) = 0 \quad\Longrightarrow \sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}\,y_j = \sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}\,\mu_i \tag{2.27}$$

Que da como resultado un estimador iterativo para μ_i:

$$\hat{\mu}_i^{(1)} = \frac{\sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}\,y_j}{\sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}}. \tag{2.28}$$

Para obtener el estimador de σ_i^2, procedemos como con μ_i, derivando con respecto a σ_i^2 en (2.26), en lugar de μ_i:

$$-\sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}\frac{1}{\sigma_i} + \sum_{j=1}^{n}\hat{\tau}_{ij}^{(0)}(y_j-\mu_i)^2\,\frac{1}{\sigma_i^3} = 0$$

Obteniendo:

$$\hat{\sigma}_i^{2\,(1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(0)} (y_j - \hat{\mu}_i^{(1)})^2}{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(0)}}. \tag{2.29}$$

2.2.3. Criterio de parada

Los pasos E y M son repetidos alternativamente hasta que se satisface un criterio de parada adecuado. En este proceso se va generando una secuencia de valores de la función de verosimilitud observada, ya definida en (2.10). Para detener la iteración pueden considerarse la diferencia absoluta

$$|\ell(\Psi^{(t+1)}|y) - \ell(\Psi^{(t)}|y)|,$$

o la diferencia relativa

$$\frac{|\ell(\Psi^{(t+1)}|y) - \ell(\Psi^{(t)}|y)|}{|\ell(\Psi^{(t)}|y)|} \tag{2.30}$$

en la mencionada secuencia, aunque también puede utilizarse la magnitud del cambio entre los parámetros estimados en cada iteración

$$|\Psi^{(t+1)} - \Psi^{(t)}|.$$

Si la diferencia utilizada es menor que un valor ϵ escogido previamente, el algoritmo finaliza. Si esto sucede en la iteración $(t+1)$, la estimación de Ψ es $\hat{\Psi} = \Psi^{(t+1)}$.

Seidel et al. (2000) demostraron que los resultados de la estimación dependen fuertemente de esta implementación y de la selección de los parámetros iniciales. Aunque no existe un consenso extendido sobre qué criterio de parada utilizar, los más frecuentemente usados son los basados en la verosimilitud. En el **Capítulo 5**, donde se desarrolla la parte aplicada de los MMF, se utilizó el de la diferencia relativa, por su adimensionalidad, y en cuanto al valor máximo de tal diferencia, $\epsilon = 1 \cdot 10^{-6}$. En general, este valor puede hacerse más pequeño aún, si bien se encuentra limitado por la precisión computacional y generalmente solo redunda en un aumento del número de iteraciones obtenidas hasta la convergencia.

2.2.4. Propiedades de convergencia

Una vez elegido el criterio de parada apropiado, la convergencia del algoritmo es especialmente relevante y, en definitiva, sobre la que descansa la esencia de este. Una descripción detallada puede encontrarse en Wu (1983) o Dempster et al. (1977). Este último muestra la monotonía de la secuencia de valores de log-verosimilitud.

Lema 2.2 La secuencia de log-verosimilitud no decrece tras una iteración EM, esto es

$$\ell(\Psi^{t+1}) \geq \ell(\Psi^t) \qquad \forall t$$

Demostración 2.2 Ver Dempster et al. (1977).

De acuerdo con este resultado, y si los valores de log-verosimilitud se encuentran acotados superiormente, la secuencia converge de forma monótona a algún $L^* = L(\theta^*)$ (ver Wu, 1983). Este autor formuló condiciones de regularidad; entre otras, la compacidad del espacio parámetrico, bajo la cual cualquier secuencia de verosimilitud se encuentra acotada. Más aun, bajo condiciones débiles, por ejemplo, si $Q(\Psi, \Psi^{(t)})$ es continua en Ψ y $\Psi^{(t)}$, L^* representa un punto estacionario.

Como la verosimilitud puede tener varios puntos estacionarios, la convergencia a un máximo, como se adelantó, depende de los valores iniciales. No obstante, no puede garantizarse la convergencia a un máximo local o global. Únicamente en el caso de que la verosimilitud sea unimodal, cualquier secuencia EM converge a un único estimador de máxima verosimilitud, independientemente del valor inicial (Wu, 1983). Además, este autor demostró que si la secuencia de verosimilitud no queda atrapada en un *punto de silla*, el punto estacionario representa un máximo local. Sin embargo, en la práctica, esta condición es limitada y de difícil verificación. Por tanto, este autor recomendó probar con varios valores iniciales en la implementación del algoritmo, ya que una pequeña perturbación en los mencionados puntos silla provoca que el algoritmo diverja (McLachlan y Basford, 1988).

No obstante, pueden encontrarse ejemplos en donde el algoritmo EM converge a un punto de silla y no a un máximo local, como se recoge en McLachlan y Krishnan (1997), donde se discute un ejemplo adaptado de Murray (1977). En consecuencia, dado que no puede garantizarse esta convergencia a un máximo global, la recomendación de Wu sobre probar con diferentes valores iniciales cobra sentido.

La Figura 2.3 muestra, para 4 experiencias distintas empleando un mismo conjunto de datos, el incremento de la log-verosimilitud a lo largo de las primeras 150 iteraciones del algoritmo EM. En cada gráfico, A, B, C y D, el incremento de la log-verosimilitud se corresponde con 20 valores iniciales diferentes del algoritmo EM (cada una de las líneas), empleando una mixtura de dos componentes gaussianas.

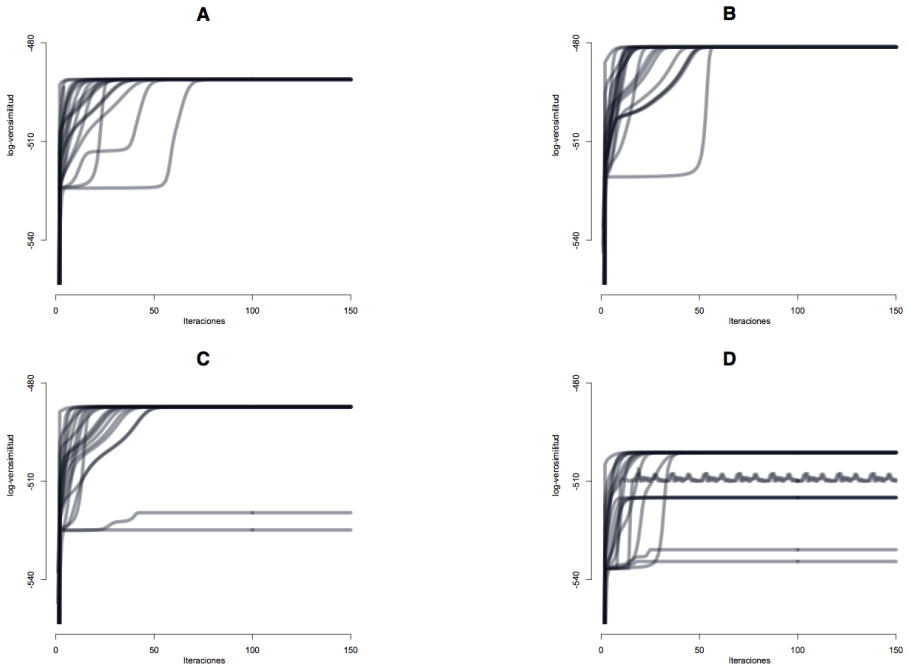


Figura 2.3 Incremento de la log-verosimilitud a lo largo de las 150 primeras iteraciones del algoritmo EM, tras proporcionarle 20 valores iniciales distintos (líneas en color), escogidos aleatoriamente. En **A** y **B**, los diferentes valores iniciales conducen al mismo valor de la log-verosimlitud, mientras que en **C** y **D**, a diferentes en contadas ocasiones.

2.3. El problema de los valores iniciales

El algoritmo EM es altamente dependiente del valor escogido para los parámetros iniciales, ya que influye sobre la velocidad de convergencia y la capacidad para alcanzar un máximo global (Karlis y Xekalaki, 2003). Aunque se han propuesto numerosas estrategias para su inicialización, no existe un consenso extendido al respecto. En general, como se mencionaba en el apartado anterior, se recomienda probar varias estrategias de inicialización y seleccionar aquella que mejores resultados ofrezca. La Figura 2.4 muestra los diferentes resultados del algoritmo EM mediante dos estrategias de inicialización sobre unos datos experimentales.

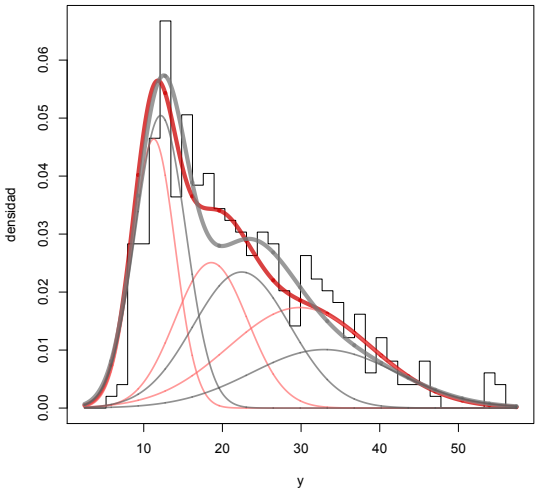


Figura 2.4 Efecto del valor de inicialización de $\hat{\Psi}$ sobre la estimación de la densidad final (línea gruesa) de la mixtura empleando el algoritmo EM. La densidad estimada de la mixtura mediante 3 componentes aparece sobreimpuesta al histograma de las observaciones. La línea gris representa la inicialización mediante el algoritmo *k-medias* (en gris) y la utilizada en este trabajo (en rojo).

El problema de los valores iniciales fue considerado por McLachlan (1988) para el caso de datos multivariantes. Este autor propuso la utilización de un diagrama de puntos en dos dimensiones como una exploración previa para detectar la presencia de clústeres en la distribución en estudio. Se puede utilizar un procedimiento similar para el caso univariante.

Se pueden encontrar estrategias para este último caso, como el que aquí se trata, en Karlis y Xekalaki (2003), donde además se incluye una revisión de trabajos previos al respecto. La utilizada en este trabajo reproduce la de Finch et al. (1989), solo para el cálculo de las medias de cada componente, por la cual se divide la muestra en g particiones y, sobre cada una de ellas, se calcula la media de las observaciones que contiene. Estos valores representan $\mu_1^{(0)}, \ldots, \mu_g^{(0)}$, es decir, la media de cada componente sobre las que el algoritmo comienza a iterar. Para las varianzas, Finch et al. (1989) obtuvieron una común, $\sigma^{2\,(0)}$, ponderada según $\mu_1^{(0)}, \ldots, \mu_g^{(0)}$; y en cuanto a las proporciones iniciales, utilizaron g números aleatorios extraídos de una distribución $U(0,1)$. En este trabajo, para el cálculo de la desviación típica se procedió como con las medias, calculando la correspondiente para cada partición, y respecto a las proporciones, todas semejantes, según $\pi_1^{(0)}, \ldots, \pi_g^{(0)} = 1/g$.

Ejemplo 2.2 Para una mixtura de dos componentes ($g = 2$) gaussianas de tamaño n_1 y n_2 sobre una muestra de tamaño $n = n_1 + n_2$, el cálculo de los valores iniciales se obtendría:

$$\pi_1^{(0)} = \pi_2^{(0)} = 1/2$$
$$\mu_1^{(0)} = \frac{1}{n_1}\sum_{s=1}^{n_1} y_s$$
$$\mu_2^{(0)} = \frac{1}{n_2}\sum_{s=n_1+1}^{n} y_s$$
$$\sigma_1^{(0)} = \left(\frac{1}{n_1-1}\sum_{s=1}^{n_1}(y_s-\mu_1^{(0)})^2\right)^{1/2}$$
$$\sigma_2^{(0)} = \left(\frac{1}{n_2-1}\sum_{s=n_1+1}^{n}(y_s-\mu_2^{(0)})^2\right)^{1/2}.$$

2.4. Selección del mejor modelo

La elección del número de componentes que constituyen un modelo de mixtura merece considerarse con detenimiento y ha sido objeto de estudio en las últimas decadas. Esta elección depende de varios factores, aunque son relevantes la distribución de los datos que son modelizados y la forma de las componentes. En un contexto de discriminación, puede obtenerse dicho número empleando diferentes técnicas de agrupamiento, cada una utilizando un diferente número de componentes, y consensuar, a través de un criterio adecuado, cuál es la cantidad de componentes que mejor estructura el conjunto de datos inicial.

Otra alternativa es el empleo de los denominados "criterios de información". Existen varios de estos criterios y se puede encontrar un estudio comparativo de ellos en Bozdogan (1993), quien utilizó muestras artificiales compuestas por diferentes clústeres, más o menos solapados, con diferentes formas y compacidad.

En este trabajo se utilizaron cuatro de estos criterios para consensuar el número "apropiado" de componentes. El primero de ellos, de muy extendida utilización durante las últimas decadas, ha sido el debido a Akaike (AIC) (Akaike, 1974, 1978), dado por la expresión

$$AIC = -2\log L(\Psi) + 2k,$$

donde $L(\Psi)$ es la función de verosimilitud del modelo, como en la ecuación (2.9), y k, el número de parámetros independientes de la mixtura según el número de componentes propuesto $(3g-1)$. Cuando se selecciona entre los diferentes modelos, cada uno definido por un número de componentes distintos, se selecciona aquel que menor AIC presenta. El AIC penaliza el sobreajuste que se presenta con modelos grandes (elevado g) a través de k (principio de parsimonia), siendo esta característica extensible a todos los criterios aquí expuestos. No obstante, se considera que el AIC sobrestima el número de componentes (McLachlan y Peel, 2000).

Otro criterio muy utilizado es el bayesiano (BIC), propuesto por Schwarz (1978), que responde a la expresión

$$BIC = -2\log L(\Psi) + k\log(n),$$

semejante a AIC excepto por el término de penalización, que incluye ahora el número de observaciones independientes de la muestra univariante (n). Para $n \geq 8$ este término de penalización es mayor que el de AIC, por lo que BIC es menos susceptible de sobrestimar el número de componentes.

Con el fin de comparar los resultados de estos dos criterios, también se utilizaron para la determinación del "mejor" g otros dos más recientes que los anteriores.

El criterio de verosimilitud completa integrada (ICL, *integrated complete likelihood*) fue propuesto por Biernacky et al. (2010) y es similar al BIC, excepto que añade un nuevo término de penalización. Este término, denominado entropía media, es

$$Ent(g) = -\sum_{i=1}^{g}\sum_{j=1}^{n}\hat{\tau}_{ij}\log\hat{\tau}_{ij} \geq 0$$

y representa la capacidad del modelo de mixtura para dar una partición relevante de la distribución, de tal forma que si las componentes están bien separadas, este término tiende a 0. Como en (2.17), $\hat{\tau}_{ij}$ denota la probabilidad condicional de que la observación y_j pertenezca a la g-ésima componente de la mixtura . Así, el número de clúster g' estimado por ICL es siempre menor o igual que el de BIC, debido a este término de penalización adicional que incorpora.

El cuarto criterio escogido se debe a Hurvich y Tsai (1989), el cual modifica a AIC mediante un nuevo término (AIC_c, AIC corregido), tomando la expresión

$$AIC_c = -2\log\, L(\Psi) + 2\,k + \frac{2\,k(k+1)}{n-k-1}.$$

A medida que el tamaño de la muestra se incrementa, el último término se aproxima a 0, llegando a las mismas conclusiones que con AIC.

La elección del criterio que se utilizó en este trabajo fue finalmente el bayesiano (BIC). Para ello influyeron varios motivos. Según se observa en el Anexo I, en donde se incluyen los resultados del número de clúster tras utilizar cada uno de estos criterios, el AIC y AIC_c tienden a sobrestimar el número de componentes, mientras que el ICL parece subestimarlos. Esta apreciación, subjetiva, se corroboró mediante pruebas gráficas, enfrentando el histograma de la distribución de los datos observados con el número de componentes propuesto por estos criterios. El mayor respaldo en la literatura hacia el criterio BIC, junto con la confirmación de las pruebas gráficas, sugirieron su elección.

2.5. Obtención de errores en los estimadores

2.5.1. Método SEM

Este procedimiento, descrito por Meng y Rubin (1972), obtiene unos estimadores estables de la matriz de varianzas-covarianzas de los parámetros obtenidos mediante el algoritmo EM. Para su cálculo, emplea todas las iteraciones obtenidas hasta la obtención de estos, motivo por el que se le considera un producto derivado del algoritmo EM, y la matriz de información de los datos completos. Su uso no está extendido debido a ciertas peculiaridades:

1. Es susceptible de presentar inexactitudes numéricas e inestabilidad, especialmente en configuraciones de alta dimensionalidad (Baker, 1992; McCulloch, 1998; Segal et al., 1994).

2. Requiere obtener la matriz de información de datos completos, que, en el caso de modelos complejos, no siempre es posible (Baker, 1992).

3. Puede ser mucho más costoso (computacionalmente) (Belin y Rubin, 1995).

No obstante, estos inconvenientes pueden ser parcialmente obviados para el caso de este trabajo, ya que, aplicado para mixturas normales univariantes, su asumible complejidad y reducida dimensión han permitido su implementación sin demasiada dificultad. En particular, el coste computacional no ha sido tal, como lo demuestran los tiempos de proceso obtenido en los cálculos (resultados no mostrados) al compararlos con la aproximación replicativa (*bootstrap*) que en el siguiente apartado se trata.

A continuación, se interpreta el algoritmo EM como una sucesión de iteraciones, que alternan entre un paso E y otro M, y que constituye una aplicación tal que:

$$\begin{aligned} M &: \Psi \to \Psi \\ \Psi^{(t+1)} &= M(\Psi^{(t)}) \qquad t = 0, 1, \ldots \end{aligned}$$

Así, si $\Psi^{(t)}$ converge a algún punto Ψ^*, entonces Ψ^* satisface

$$\Psi^* = M(\Psi^*).$$

Mediante este planteamiento, se consigue obtener lo que los autores denominaron la tasa de convergencia del algoritmo EM, que viene dada por

$$DM = \left(\frac{\partial M_j(\Psi)}{\partial \Psi_i} \right) \Bigg|_{\Psi = \Psi^*}$$

que representa la matriz jacobiana $(d \times d)$ de $M(\Psi)$ evaluada en Ψ^*, siendo d el número de parámetros del modelo de mixtura.

Por su parte, el planteamiento de los datos completos introducido en el apartado 2.2.1 permite introducir a su matriz de información (I_{oc}), que viene dada según la expresión

$$I_{oc} = \mathsf{E}[I_o(\Psi|X)|Y, \Psi] \Big|_{\Psi = \Psi^*}$$

$$I_o(\Psi|X) = -\frac{\partial^2 \ell(\Psi|X)}{\partial \Psi^2}$$

siendo I_o la matriz de información de los datos observados.

La esencia del método SEM se debe a la siguiente expresión, debida a Meng y Rubin (1977):

$$V = I_{oc}^{-1} + I_{oc}^{-1} DM \, (I - DM)^{-1}, \tag{2.31}$$

siendo V la matriz de varianzas-covarianzas de los valores de los parámetros utilizando los datos completos, e $I_{d \times d}$ la matriz identidad. Los errores SEM se obtienen aplicando la raíz cuadrada a la diagonal de la matriz V, lo que equivale a los errores estándares de los estimadores EM.

El algoritmo SEM consta fundamentalmente de tres partes bien diferenciadas: (1) La evaluación de I_{oc}^{-1}, (2) la de DM y (3) la obtención final de la estimación de la matriz V.

En el caso de las mixturas gaussianas, para obtener I_{oc} se recurre a la obtención del hessiano de $Q(\Psi \,|\, \Psi^{(t)})$, lo que no es difícil analíticamente debido a la forma de esta distribución. Por su parte, la

matriz DM requiere un cálculo numérico basado en todas las iteraciones del algoritmo EM hasta que este ha obtenido su convergencia.

Obtención de la matriz I_{oc}

Como se mencionaba, la obtención de I_{oc} implica la obtención del hessiano de $Q(\Psi \mid \Psi^{(t)})$. Con los mismos resultados, puede operarse en su lugar sobre la ecuación (2.16) y sustituir z_{ij} por $\hat{\tau}_{ij}$, multiplicando finalmente por (-1):

- Cálculo de $\frac{\partial \ell(\Psi|x)}{\partial \Psi}$:

$$\frac{\partial \ell(\Psi|y,z)}{\partial \pi_i} = \frac{1}{\pi_i} \sum_{j=1}^{n} z_{ij}$$

$$\frac{\partial \ell(\Psi|y,z)}{\partial \mu_i} = \frac{1}{\sigma_i^2} \sum_{j=1}^{n} z_{ij} \, (y_j - \mu_i)$$

$$\frac{\partial \ell(\Psi|y,z)}{\partial \sigma_i} = \sum_{j=1}^{n} z_{ij} \left(\frac{(y_j - \mu_i)^2}{\sigma_i^3} - \frac{1}{\sigma_i} \right)$$

- Cálculo de $\frac{\partial^2 \ell(\Psi|x)}{\partial \Psi \partial \Psi^T}$:

$$\frac{\partial^2 \ell(\Psi|y,z)}{\partial \pi_m \partial \pi_i} = \frac{\partial^2 \ell(\Psi|y,z)}{\partial \pi_i \partial \pi_m} = \begin{cases} -\dfrac{1}{\pi_i^2} \displaystyle\sum_{j=1}^{n} z_{ij} & \text{si} \quad i = m \\ 0 & \text{si} \quad i \neq m \end{cases}$$

$$\frac{\partial^2 \ell(\Psi|y,z)}{\partial \mu_m \partial \mu_i} = \frac{\partial^2 \ell(\Psi|y,z)}{\partial \mu_i \partial \mu_m} = \begin{cases} -\dfrac{1}{\sigma_i^2} \displaystyle\sum_{j=1}^{n} z_{ij} & \text{si} \quad i = m \\ 0 & \text{si} \quad i \neq m \end{cases}$$

$$\frac{\partial^2 \ell(\Psi|y,z)}{\partial \sigma_m \partial \sigma_i} = \frac{\partial^2 \ell(\Psi|y,z)}{\partial \sigma_i \partial \sigma_m} = \begin{cases} \displaystyle\sum_{j=1}^{n} z_{ij} \left[\frac{-3(y_j - \mu_i)^2}{\sigma_i^4} + \frac{1}{\sigma_i^2} \right] & \text{si} \quad i = m \\ 0 & \text{si} \quad i \neq m \end{cases}$$

$$\frac{\partial^2 \ell(\Psi|y,z)}{\partial \sigma_m \partial \mu_i} = \frac{\partial^2 \ell(\Psi|y,z)}{\partial \sigma_i \partial \mu_m} = \begin{cases} -\dfrac{2}{\sigma_i^3} \displaystyle\sum_{j=1}^{n} z_{ij}(y_j - \mu_i) & \text{si} \quad i = m \\ 0 & \text{si} \quad i \neq m \end{cases}$$

$$\frac{\partial^2 \ell(\Psi|y,z)}{\partial \pi_m \partial \mu_i} = \frac{\partial^2 \ell(\Psi|y,z)}{\partial \pi_i \partial \mu_m} = 0 \qquad \forall j,m \qquad\qquad \frac{\partial^2 \ell(\Psi|y,z)}{\partial \pi_m \partial \sigma_i} = \frac{\partial^2 \ell(\Psi|y,z)}{\partial \pi_i \partial \sigma_m} = 0 \qquad \forall j,m$$

Obtención de la matriz DM

Siendo r_{ij} el elemento (i,j)-ésimo de la matriz DM, se define

$$\Psi_{(i)}^{(k)} = (\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(k)}, \theta_{i+1}^*, \ldots, \theta_d^*) \tag{2.32}$$

como el conjunto de parámetros obtenidos en una iteración EM, (k), siendo d el número de parámetros del modelo. Solo el parámetro $\theta_i^{(k)}$ es distinto al resto, ya que los demás coinciden con los valores de Ψ^*, es decir, con los valores de los estimadores EM, una vez que el algoritmo ha convergido. El cálculo de r_{ij} se obtiene

$$\begin{aligned} r_{ij} = \left(\frac{\partial M_j(\Psi)}{\partial \theta_i} \right) \Bigg|_{\Psi=\Psi^*} &= \lim_{\theta_i \to \theta_i^*} \frac{M_j(\theta_i^*, \ldots, \theta_{i-1}^*, \theta_i, \theta_{i+1}^*, \ldots, \theta_d^*) - M_j(\Psi^*)}{\theta_i - \theta_i^*} \\ &= \lim_{k\to\infty} \frac{M_j(\Psi_{(i)}^{(k)}) - \theta_j^*}{\theta_i^{(k)} - \theta_i^*} = \lim_{k\to\infty} r_{ij}^{(k)}. \end{aligned} \tag{2.33}$$

Para llevar a cabo este procedimiento, es necesario asumir que el algoritmo EM converge en K iteraciones y que todas ellas son conocidas. Así, el algoritmo definido por Meng y Rubin comprende:

1. Fijar un $i = 1$ y obtener $\Psi_{(i)}^{(k)} = (\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(k)}, \theta_{i+1}^*, \ldots, \theta_d^*)$ y evaluar $M(\Psi_{(i)}^{(k)})$.
2. Calcular

$$\frac{M_j(\Psi_{(i)}^{(k)}) - \theta_j^*}{\theta_i^{(k)} - \theta_i^*}$$

   para $j = 1, \ldots, d$.
3. Repetir los pasos 1 y 2 para $i = 2, \ldots, d$.

Obteniendo la matriz DM, compuesta por r_{ij}, $i,j = 1, \ldots, d$. El elemento r_{ij} se obtiene cuando la secuencia $r_{ij}^{(k)}, r_{ij}^{(k+1)}, r_{ij}^{(k+2)}, \ldots$ es estable para algún k. El criterio de parada que se aplica suele ser más laxo que el que se aplica en las iteraciones EM y puede llegar a ser diferente según el parámetro del que se trate en Ψ, generalmente la raíz cuadrada del ϵ utilizado en el algoritmo EM (Jamshidian y Jennrich, 2000, p. 260):

$$| r_{ij}^{(k+1)} - r_{ij}^{(k)} | \approx \sqrt{\epsilon} \tag{2.34}$$

Tanner (1996) recomienda evitar los errores de redondeo que puedan cometerse debido al cálculo de esta diferencia. Igualmente, advierte que la matriz $\hat{V}$ puede no ser simétrica, por lo que sugiere en su lugar el cálculo de $\frac{1}{2}(\hat{V} + \hat{V}^T)$.

En este trabajo, en lugar de (2.34), se prefirió como criterio de parada a aquella iteración que presentara la mínima diferencia entre dos iteraciones sucesivas, representando una modificación al método habitual. Con esta modificación se evitó tener que considerar un ϵ distinto según el estimador del que se tratase.

El único inconveniente encontrado al obtener los resultados de este método está relacionado con la interrupción del algoritmo. Como describió Jamshidian y Jennrich (2000), el algoritmo se interrumpe al hacerse 0 la diferencia $\theta^{(k)} - \theta^*$ de algún componente.

2.5.2. Método *bootstrap*

Para la estimación de los errores mediante este método replicativo, se ha seguido el procedimiento descrito en Basford et al. (1988) adaptado a mixturas univariantes. En esencia, consiste en generar a partir del conjunto de datos observados, y, B muestras bootstrap, $y_1^*, y_2^*, \ldots, y_B^*$, y obtener la parametrización de la mixtura mediante el algoritmo EM para cada una de ellas. Con posterioridad, se calcula el estimador bootstrap del error de muestreo ($\widehat{se_B}$) sobre cada uno de los estimadores EM de las B mixturas. Esquemáticamente:

1. A partir de la muestra de datos observados, y, generar muestras bootstrap, y_b^* $(b = 1, \ldots, B)$, de tamaño n, con B=1000.
2. Aplicar el algoritmo EM sobre cada y_i^* para obtener los estimadores EM de cada una de las mixturas obtenidas $\hat{\Psi}_1^*, \ldots, \hat{\Psi}_B^*$.
3. Sobre cada uno de los parámetros $\hat{\theta}_{bj}^*$ $(b = 1, \ldots, B\,;\, j = 1, \ldots, 3g-1)$ que componen $\hat{\Psi}_b^*$, calcular a continuación el estimador bootstrap del error estándar:

$$\widehat{se_B(\theta_j^*)} = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_j^*(b) - \bar{\hat{\theta}}_j) \right)^{1/2}$$

donde

$$\bar{\hat{\theta}}_j = \left(\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_j^*(i) \right).$$

3

Modelos ocultos de Markov

Dada la naturaleza temporal de los datos secuenciales objeto de esta tesis, una alternativa a los modelos de mixtura consiste en el tratamiento explícito de la dependencia en el tiempo. Los modelos ocultos de Markov (MOM) permiten modelizar secuencias de observaciones con la posibilidad de tratar de forma eficiente esta evolución temporal. El propósito de este capítulo es introducir los MOM y algunas de sus propiedades, y su utilidad en el estudio de las series temporales (SSTT), univariantes y estacionarias, en las que se apoya la segunda parte experimental de este trabajo (Capítulos 6 y 7).

Dado que las Cadenas de Markov (CM) representan la estructura subyacente de los MOM, aquellas se describen en la Sección 3.1. Los tres problemas clásicos con los que se aborda el estudio de los MOM (evaluación, decodificación y aprendizaje) en el trabajo de Rabiner (1989) son de nuevo presentados en esta ocasión, dada su utilidad para exponer estos modelos. Más alla de este préstamo, se ha preferido emplear para su descripción el enfoque del texto, más actual, de Zucchini y MacDonald (2009), motivada esta elección por su claridad conceptual y simplicidad en la notación, lo que ha permitido acompañar ejemplos a lo largo del texto del capítulo. Especial atención se ha dedicado a la implementación computacional de los MOM, y en particular, al escalado de las probabilidades hacia adelante y hacia atrás (Sección 3.2.1), con los que se evita el desbordamiento de sus valores. Para evitar reiteración, el algoritmo EM es ligeramente descrito en la Sección 3.2.3. Otros aspectos ya abordados en el Capítulo 2, como el problema de la identificabilidad de las mixturas, la elección del número de componentes en el modelo, o la estimación de los errores bootstrap en los estimadores EM, son brevemente reseñados al final del capítulo, dado que atañen a los MOM de forma semejante.

3.1. Cadenas de Markov

En este sección se expone de forma básica el fundamento necesario de las CM para la definición y desarrollo de los MOM. Referencias clásicas de CM son aquellas de Feller (1968), Freedman (1975), Isaacson y Madsen (1976), y Grimmett y Stirzaker (2001).

Un proceso estocástico describe la evolución de un fenómeno o sistema sujeto a perturbaciones aleatorias independientes en diferentes instantes de tiempo. Más formalmente, es un conjunto de variables aleatorias $\{C_t : t \in \mathbb{T}\}$. Mediante $\mathbb{T}$ se representa a un conjunto de parámetros cuyos elementos generalmente son instantes de tiempo t. Los posibles valores de la variable aleatoria C_t se interpretan como los posibles *estados* del proceso en el instante t, y al conjunto de todos los estados, el *espacio de estados* $\mathcal{S}$. Si $\mathbb{T}$ toma valores es un intervalo en la recta real, al proceso se le denomina en tiempo *continuo*, mientras que si lo hace en un conjunto numerable, en tiempo *discreto*. Por su parte, si $\mathcal{S}$ toma valores

en un conjunto numerable, al proceso se le denomina de espacio de estados *discreto*, mientras que toma valores en un conjunto infinito, recibe el nombre de espacio de estados *continuo*. Cada una de las variables aleatorias tiene su propia función de distribución y estas pueden ser incluso independientes unas de otras. Las CM son un tipo de proceso estocástico en el que el estado del sistema en un instante depende únicamente del estado en el instante anterior, motivo por el que se le denomina "sin memoria". Cuando este estado es el más reciente, la CM es de *primer orden*. Las CM descritas en esta sección serán de primer orden, con espacio de estados y tiempo discretos.

Definición 3.1 *Una sucesión de variables aleatorias discretas* $\{C_t : t \in \mathbb{N}\}$ *se denomina una CM en tiempo discreto si* $\forall\, t \in \mathbb{N}$*, satisface la propiedad de Markov*

$$P(C_{t+1}|C_t, \ldots, C_1) = P(C_{t+1}|C_t).$$

Observación 3.1 Como se muestra en la Figura 3.1, la evolución del proceso hasta el instante de tiempo t depende entonces únicamente del valor más reciente $C(t)$.

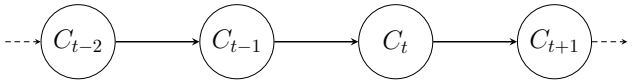


Figura 3.1 Dependencia de los estados en una CM.

Por simplicidad en la notación, esta evolución del proceso $(C_1, C_2, ..., C_t)$ se representará como $\mathbf{C}^{(t)}$, por lo que la CM puede representarse

$$P(C_{t+1}|\,\mathbf{C}^{(t)}) = P(C_{t+1}|C_t).$$

Observación 3.2 Esta propiedad de Markov representa la primera relajación del supuesto de independencia entre variables aleatorias.

La CM se inicia en un estado y se mueve sucesivamente entre instantes de tiempo de un estado a otro hasta finalizar, a través de diferentes etapas. La probabilidad de cambio de la CM de un estado i a otro j después de haber transcurrido t-etapas queda definido por unas probabilidades condicionadas, denominadas *de transición*:

$$\gamma_{ij}(t) = P(C_{s+t} = j|C_s = i) \qquad \forall s, t \in \mathbb{N} \qquad \forall i, j \in \mathcal{S}.$$

Estas probabilidades, en caso de ser independientes del tiempo, definen a una CM *homogénea*.

Definición 3.2 *Una CM se denomomina homogénea si*

$$P(C_{s+t} = j|C_s = i) \qquad \forall s, t \in \mathbb{N}$$

es independiente de s *para cualquier* i, $j \in \mathcal{S}$.

Todas las posibles probabilidades de transición entre estados de una cadena son elementos de una matriz de transición.

Definición 3.3 *Sea una CM homogénea. Entonces* $\mathbf{\Gamma}(t) = (\gamma_{ij}(t))$ *se denomina la matriz de transición de* t*-etapas o de probabilidades de transición (m.p.t.), de dimensiones* $m \times m$*, donde* m *es finito e indica*

el número de estados de $\mathcal{S}$. Mediante $\mathbf{\Gamma}(1)$ se denota la matriz de probabilidades de transición de una etapa.

Observación 3.3 $\mathbf{\Gamma}(t)$ es una matriz estocástica. Por tanto, los elementos de $\mathbf{\Gamma}(t)$ verifican:

$$\gamma_{ij}(t) \geq 0, \quad \forall i, j \in \mathcal{S} \tag{3.1}$$

$$\sum_{j=1}^{m} \gamma_{ij}(t) = 1 \quad \forall i \tag{3.2}$$

Ambas propiedades indican que cada fila de $\mathbf{\Gamma}(t)$ representa una distribución de probabilidad. En terminos de $\mathbf{\Gamma}(t)$, la segunda propiedad puede representarse mediante $\mathbf{\Gamma}(t)\ \mathbf{1}' = \mathbf{1}'$, siendo $\mathbf{1}'$ un vector columna de dimensiones $m \times 1$ cuyos elementos son todos unos.

Una importante propiedad de las CM homogéneas de espacio de estados finito es que satisfacen las ecuaciones de Chapman-Kolmogorov:

$$\mathbf{\Gamma}(t+u) = \mathbf{\Gamma}(t)\,\mathbf{\Gamma}(u).$$

Estas ecuaciones implican, que

$$\mathbf{\Gamma}(t) = \mathbf{\Gamma}(1)^t \quad \forall\, t > 1,$$

indicando que la matriz de transición de t-etapas es la t-ésima potencia de la matriz de transición de una etapa. Por simplicidad en la notación, en adelante $\mathbf{\Gamma}(1)$ se denotará simplemente como $\mathbf{\Gamma}$.

Proposición 3.1 *Cualquier potencia de $\mathbf{\Gamma}$ es una matriz estocástica.*

Las distribución marginal $P(C_t = j)$, $j \in \mathcal{S}$, de una CM indica la probabilidad de la cadena de estar en el estado j en la etapa t. Así, para cada etapa se tendrá un vector fila

$$\mathbf{u}(t) = (P(C_t = 1), \ldots, P(C_t = m)), \quad t \in \mathbb{N} \tag{3.3}$$

que representa la probabilidad de que la cadena se encuentra en cada uno de los posibles estados de la etapa t. Estas distribuciones de probabilidad para la primera etapa $t = 1$ recibe el nombre de *distribución inicial* de la CM, permitiendo conocer el punto de partida del proceso:

$$\mathbf{u}(1) = (P(C_1 = 1), \ldots, P(C_1 = m)).$$

Por ser distribución de probabilidad, esta distribución inicial cumple, $\forall j \in \mathcal{S}$:

$$P(C_1 = j) \geq 0,$$

$$\sum_{j=1}^{m} P(C_1 = j) = 1.$$

La distribución de la CM en la etapa $t+1$ a partir de la distribución en la etapa t se deduce multiplicando por la matriz de transición $\mathbf{\Gamma}$:

$$\mathbf{u}(t+1) = \mathbf{u}(t)\,\mathbf{\Gamma}, \tag{3.4}$$

pudiéndose demostrar por inducción que

$$\mathbf{u}(t) = \mathbf{u}(1)\,\mathbf{\Gamma}^t. \tag{3.5}$$

Observación 3.4 Por tanto, una CM en tiempo discreto queda completamente determinada por la distribución inicial $\mathbf{u}(1)$ y por las probabilidades de transición $\gamma_{ij}(t)$.

Clasificación de los estados en las CCMM

Las CCMM pueden clasificarse de diferentes maneras según se clasifican sus estados. Para un desarrollo de esta clasificación puede consultarse, por ejemplo, Grimmet y Stirzaker (2001). En este apartado se exponen dos condiciones en la teoría de las CCMM, que desempeñan un papel fundamental en el estudio de las distribuciones estacionarias de estas cadenas, la irreducibilidad y aperiodicidad.

Definición 3.4 *El periodo $d(i)$ de un estado $i \in \mathcal{S}$ se determina como $d(i) = mcd\{t : \gamma_{ii}(t) > 0\}$, el máximo común divisor de los intervalos de tiempo que transcurren hasta que la cadena regresa al estado i, cuando este regreso es posible. El estado i es periódico si $d(i) > 1$, y aperiódico cuando $d(i) = 1$. Una CM es aperiódica si todos sus estados son aperiódicos.*

El periodo $d > 1$ del estado $i \in \mathcal{S}$ se refiere al menor número tal que la duración de todas las secuencias de transiciones que parten del estado i y regresan al estado i son múltiplos de d; o de otra forma, que $\gamma_{ii}(t) = 0$ a menos que t sea un múltiplo de $d(i)$, y $d(i)$ es el máximo con esta propiedad.

Definición 3.5 *Un estado $i \in \mathcal{S}$ comunica con un estado $j \in \mathcal{S}$ si $\gamma_{ij}(t) > 0$, e intercomunican si además $\gamma_{ji}(t) > 0$. Se dice que una CM es irreducible si todos sus estados comunican entre sí.*

Distribuciones estacionarias

Para cada CM finita, aperiódica e irreducible existe una única *distribución estacionaria*. Los valores de la distribución estacionaria se pueden interpretar como la proporción final de tiempo que la cadena ha pasado en cada estado a lo largo de su evolución, con probabilidad 1. Se puede decir que es también la proporción, a largo plazo, de etapas en las que la cadena se encuentra en el estado i a lo largo de su evolución, si ha partido de i o de otro estado recurrente que intercomunica con i.

Ejemplo 3.1 Sea la siguiente realización de una CM, con un espacio de estados finito $\mathcal{S} = \{1, 2\}$. La secuencia de estados observada, representada por filas, es:

1 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 1 1 2 2 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2

Una de las formas posibles de estimar la matriz de transición dada una realización es mediante el recuento de transiciones posibles de un estado a otro. Representando por f_{ij} el número de transiciones del estado i al j, obtenemos la matriz de frecuencia de transiciones de una etapa

$$(f_{ij}) = \begin{pmatrix} 9 & 9 \\ 8 & 72 \end{pmatrix}.$$

Dado que el número de transiciones del estado 1 al estado 2 es 9, y que el número total de transiciones desde el estado 1 es 18, puede obtenerse $\hat{\gamma}_{12}(1) = 9/18$. Así, puede estimarse la matriz de transición

$$\mathbf{\Gamma} = \begin{pmatrix} 1/2 & 1/2 \\ 1/10 & 9/10 \end{pmatrix},$$

que verifica las condiciones (3.1) y (3.2). Se deduce que la CM es irreducible, al comunicar todos los estados entre sí, y que es aperiódica, ya que $d(1) = d(2) = 1$.
Si consideramos que la distribución inicial de esta cadena es $\mathbf{u}(1) = (1, 0)$, entonces, por (3.4) y (3.5):

$$\begin{aligned}
\mathbf{u}(2) &= \mathbf{u}(1)\,\mathbf{\Gamma} &&= (0.3, 0.7) \\
\mathbf{u}(3) &= \mathbf{u}(2)\,\mathbf{\Gamma} &&= (0.22, 0.78) \\
\mathbf{u}(4) &= \mathbf{u}(1)\,\mathbf{\Gamma}^4 &&= (0.188, 0.812) \\
\mathbf{u}(8) &= \mathbf{u}(1)\,\mathbf{\Gamma}^8 &&= (0.1672128, 0.8327872) \\
\mathbf{u}(16) &= \mathbf{u}(1)\,\mathbf{\Gamma}^{16} &&= (0.1666670, 0.8333330) \\
\mathbf{u}(32) &= \mathbf{u}(1)\,\mathbf{\Gamma}^{32} &&= (0.1666667, 0.8333333) \\
\mathbf{u}(64) &= \mathbf{u}(1)\,\mathbf{\Gamma}^{64} &&= (0.1666667, 0.8333333) = (1/6, 5/6).
\end{aligned}$$

Resultado que se obtiene también, por ejemplo, para el valor de $\mathbf{u}(1) = (0, 1)$, o cualquier otro. Suponiendo esta vez un valor de $\mathbf{u}(1) = (1/6, 5/6)$, se obtiene:

$$\begin{aligned}
\mathbf{u}(2) &= \mathbf{u}(1)\,\mathbf{\Gamma} &&= (1/6, 5/6) \\
\mathbf{u}(32) &= \mathbf{u}(1)\,\mathbf{\Gamma}^{32} &&= (1/6, 5/6) \\
\mathbf{u}(64) &= \mathbf{u}(1)\,\mathbf{\Gamma}^{64} &&= (1/6, 5/6),
\end{aligned}$$

resultado únicamente alcanzable con el valor de $\mathbf{u}(1)$ supuesto. Observando la secuencia de estados al principio del ejemplo, la frecuencia con la que la CM ha transcurrido por el estado 1 es 2/11, aproximándose al valor 1/6 obtenido. La distribución inicial $\mathbf{u}(1)$ recibe el nombre de *distribución estacionaria* de la CM y se denotará por $\boldsymbol{\delta}$.

Proposición 3.2 *Un resultado general que puede ser convenientemente utilizado a efectos del cálculo de la distribución estacionaria establece que*

$$\boldsymbol{\delta}(\mathbf{I}_m - \mathbf{\Gamma} + \mathbf{U}) = \mathbf{1}, \tag{3.6}$$

siendo $\mathbf{I}_m$ la matriz identidad $m \times m$ y $\mathbf{U}$ la matriz $m \times m$ de unos.

Así, puede comprobarse que $\mathbf{u}(1) = (1/6, 5/6)$ es distribución estacionaria de esta CM:

$$(1/6, 5/6)\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1/2 & 1/2 \\ 1/10 & 9/10 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right) = (1, 1).$$

Observación 3.5 $\boldsymbol{\delta}$ es también denominada en la literatura como distribución *invariante* o de *equilibrio*.

El resultado de este ejemplo permite enunciar en términos de $\mathbf{\Gamma}$:

Definición 3.6 *Una CM, irreducible y aperiódica, con matriz de transición de probabilidad $\mathbf{\Gamma}$ se dice que tiene una distribución estacionaria $\boldsymbol{\delta}$ (un vector fila con elementos no negativos) si $\boldsymbol{\delta}\,\mathbf{\Gamma} = \boldsymbol{\delta}$ y $\boldsymbol{\delta}\,\mathbf{1}' = 1$. Esta distribución estacionaria es única.*

El primero de los requerimientos expresa la condición de estacionariedad, y la segunda, que $\boldsymbol{\delta}$ es en efecto una distribución de probabilidad. Por (3.4), una CM que comienza en su distribución estacionaria

continuará teniendo esta distribución en todos los instantes de tiempo posteriores, denominándose una CM *estacionaria*[1]. Aunque la condición de irreducibilidad subyace de forma necesaria en la Definición 3.6, la de aperiodicidad no (ver Grimmet y Stirzaker [2001], Lema 6.3.5 y Teorema 6.4.3).

3.2. Modelos ocultos de Markov

Un modelo oculto de Markov es un tipo particular de mixtura dependiente. En los MOM la distribución estadística que genera una observación depende del *estado* de un proceso de Markov no observado. Estos modelos, tradicionalmente utilizados en las últimas décadas para el procesado de señales (reconocimiento automático del habla y facial, escritura) (Miller, 1952; Rabiner, 1989; Baumm y Petri, 1966, y Jelinek, 1998), estudio del aprendizaje (Wickens, 1982) y sociología (Langeheine y Van de Pol, 1990), han extendido su uso a áreas como la bioinformática (Krogh, 1998; Etheridge et al., 2008), economía (Kim, 1994; Mamon, 2007), biofísica (Blanco et al., 2013) y al aprendizaje automático y minería de datos (Ghahramani y Jordan, 1997). Su utilidad para el modelado de largas SSTT es destacable, por su flexibilidad para la captura de la estacionalidad y tendencia de estas, a la vez que permite incluir covariables de otras SSTT (Zucchini y MacDonald, 2009). Los HMMs poseen atractivas características, como su versatilidad, tratabilidad matemática y asequible implementación computacional. En particular, el cálculo de sus momentos, de su función de verosimilitud y su distribución marginal. En la mayoría de los casos estos modelos son fácilmente interpretables.

Aunque el estudio de los MOM comenzó con la publicación en 1966 del trabajo de Baum y Petri (1966), varios aspectos de estos modelos han sido desarrollados posteriormente, como la estimación de su orden (Rydén, 1995a; Csiszár y Shields, 2000) o las propiedades asintóticas de los estimadores de máxima verosimilitud (Bickel et al., 1998; Douc y Matias, 2001; Rydén, 1995b). Una completa revisión histórica de estos modelos puede encontrarse en Ephraim y Merhav (2002), Cappé et al. (2005) y Rydén et al. (1998).

Definición 3.7 *Sea* $\{X_t : t \in \mathbb{N}\}$ *un MOM. Representando mediante* $\mathbf{X}^{(t)} = \{X_t : t = 1, 2, ...\}$ *y* $\mathbf{C}^{(t)} = \{C_t : t = 1, 2, ...\}$ *las variables aleatorias que describen las evoluciones del proceso desde el instante 1 al instante t, entonces, un modelo de este tipo se resume mediante las dos siguientes propiedades:*

$$P(C_t|\,\mathbf{C}^{(t-1)}) = P(C_t|\,C_{t-1}),\ t = 2, 3, ...$$
$$P(X_t|\,\mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = P(X_t|\,C_t),\ t \in \mathbb{N}.$$

El modelo consiste en dos partes. En primer lugar un proceso de parámetros $\{C_t : t = 1, 2, ...\}$, oculto al observador, que satisface la propiedad de Markov, y en segundo lugar, un proceso dependiente de estado $\{X_t : t = 1, 2, ...\}$, observable, tal que, cuando C_t es conocido, la distribución de X_t depende sólo del estado actual C_t y no de observaciones o estados anteriores (Figura 3.2). Si la CM $\{C_t\}$ tiene m estados, al MOM $\{C_t\}$ se le denomina según m (un MOM de m estados). Cuando $m = 1$ supone un caso degenerado de un MOM, representando una secuencia de variables aleatorias independientes entre sí.

Observación 3.6 La denominación de los MOM es diversa en la literatura: procesos de Markov ocultos, mixturas dependientes de Markov, modelos latentes de Markov, modelo de mixtura de Markov, modelos sujetos a un régimen de Markov, modelos de cambio de Markov, o en inglés, hidden Markov models.

En cada ocasión de medida, el proceso en el estado C_t emite una observación X_t, constituyendo el proceso de generación de las observaciones, denominadas también *símbolos* o *símbolos de observación.*

[1] No todos los autores utilizan esta terminología. McLachlan y Peel (2000, p. 328) se refieren a una CM estacionaria a la que en este trabajo se denomina homogénea.

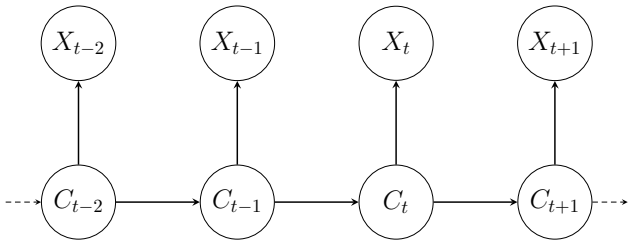


Figura 3.2 Relaciones de independencia condicional en un MOM.

Estas observaciones son emitidas siguiendo una particular distribución, discreta o continua. En el caso de este trabajo, la empleada es una mixtura de distribuciones Gaussianas. Siendo p_i la función de densidad de probabilidad de X_t si la CM se encuentra en el estado i en el intante t, se denomina a las m distribuciones (condicionales) p_i como *distribuciones dependientes del estado* del modelo o *de emisión*, representándose como $p_i(x_t) \sim f_i(X_t|C_t = i)$, $i = 1, ..., m$, con $f_i \sim \phi_i(\cdot|\mu_i, \sigma_i)$. En forma matricial, puede escribirse para los m estados,

$$\mathbf{P}(x_t) = \begin{pmatrix} p_1(x_t) & & 0 \\ & \ddots & \\ 0 & & p_m(x_t) \end{pmatrix},$$

siendo $\mathbf{P}(x_t)$ la matriz diagonal con i elementos $p_i(x_t)$. Se definen a continuación $u_i(t) = P(C_t = i)$, $t = 1, 2, ...$, como la probabilidad de que la CM se encuentre en el instante t en el estado i, y al vector resultante para todos los posibles estados $\mathbf{u}(t) = \{u_1(t), ..., u_m(t)\}$, como en (3.3). Se tiene entonces

$$p(x_t) = \sum_{i=1}^{m} u_i(t) p_i(x_t),$$

o en forma matricial,

$$p(x_t) = (u_1(t), \ldots, u_m(t)) \begin{pmatrix} p_1(x_t) & & 0 \\ & \ddots & \\ 0 & & p_m(x_t) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{u}(t)\, \mathbf{P}(x_t)\, \mathbf{1}'.$$

De la ecuación (3.4), se deduce que $\mathbf{u}(t) = \mathbf{u}(1)\, \mathbf{\Gamma}^{t-1}$, obteniéndose

$$p(x_t) = \mathbf{u}(1)\, \mathbf{\Gamma}^{t-1}\, \mathbf{P}(x_t)\, \mathbf{1}'. \tag{3.7}$$

La ecuación (3.7) asume que la CM es homogénea, pero no necesariamente estacionaria. Admitiendo esta última condición, con distribución estacionaria $\boldsymbol{\delta}$, entonces $\boldsymbol{\delta}\, \mathbf{\Gamma}^{t-1} = \boldsymbol{\delta}$, $\forall t \in \mathbb{N}$. Actuando $\boldsymbol{\delta}$ como *peso*, se obtiene la distribución marginal que relaciona los MOM con los MMF:

$$p(x_t) = \boldsymbol{\delta}\, \mathbf{P}(x_t)\, \mathbf{1}', \tag{3.8}$$

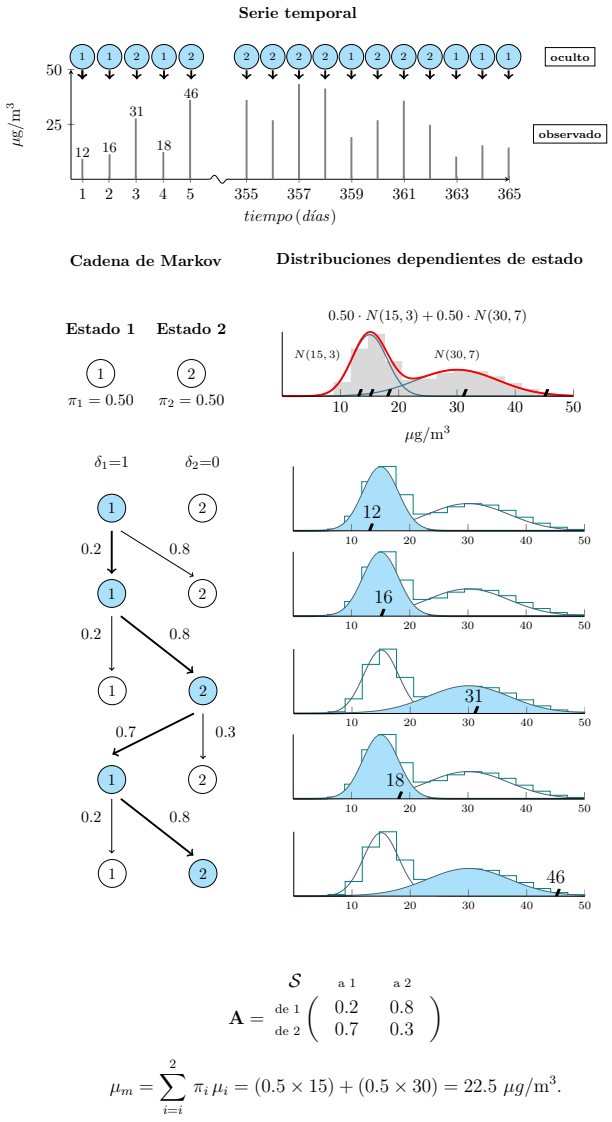
en donde $\boldsymbol{\delta}$, por ser distribución estacionaria, verifica $\boldsymbol{\delta}\, \mathbf{\Gamma} = \boldsymbol{\delta}$ y $\boldsymbol{\delta}\ \mathbf{1}\ ' = 1$.

Observación 3.7 Por tanto, cada MOM, queda determinado por tres elementos: la m.p.t. $\mathbf{\Gamma}$, las distribuciones dependientes de estado $\mathbf{P}(x_t)$, y la distribución inicial $\boldsymbol{\delta}$.

Ejemplo 3.2 La Figura 3.3 muestra una CM de 2 estados, a la izquierda, que genera las cinco primeras observaciones de una ST a partir de una mixtura de distribuciones con dos componentes. La CM con una distribución estacionaria $\boldsymbol{\delta}$=(1,0), alterna entre los dos estados según la m.p.t.

$$\Gamma = \begin{pmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{pmatrix}.$$

Las distribuciones dependientes del estado vienen representadas por dos distribuciones gaussianas, $p_1(x) \in \phi(x|15,3)$ y $p_2(x) \in \phi(x|30,7)$, con pesos $\boldsymbol{\pi} = (0.5, 0.5)$, que forman una mixtura.



$$\mathbf{A} = \begin{array}{c} \mathcal{S} \\ \text{de 1} \\ \text{de 2} \end{array} \begin{array}{c} \begin{array}{cc} \text{a 1} & \text{a 2} \end{array} \\ \begin{pmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{pmatrix} \end{array}$$

$$\mu_m = \sum_{i=i}^{2} \pi_i \mu_i = (0.5 \times 15) + (0.5 \times 30) = 22.5 \ \mu g/\mathrm{m}^3.$$

Figura 3.3 Modelo oculto de Markov de 2 estados, $\mathcal{S} = \{1, 2\}$, y proceso de generación de las cinco primeras observaciones $\mathbf{X}$={12,16,31,18,46} de una ST, anual, de un contaminante atmosférico. En la mixtura de distribuciones, las áreas coloreadas en azul indican la distribución activa, representándose su densidad mediante una línea roja en la parte superior. A la izquierda, la ruta que genera las observaciones se indica mediante flechas resaltadas, junto a sus probabilidades de transición, y el estado que ocupa la cadena en un determinado instante, mediante círculos azules. La cadena es inicializada mediante $\boldsymbol{\delta}$, comenzando en el estado 1. La m.p.t. se denota mediante $\mathbf{A}$, y con μ_m, el cálculo del primer momento de la mixtura.

El estudio de los MOM se aborda de forma clásica mediante la resolución de sus denominados *tres problemas básicos*, presentados por Ferguson (1980), y que fueron popularizados por Poritz (1988) y más tarde por Rabiner (1989). Resumidamente, los dos primeros problemas consisten, dado un MOM conocido, en obtener la probabilidad de las observaciones y la secuencia de estados ocultos, y el tercero, en estimar los parámetros de un MOM dada una secuencia de observaciones. Esta perspectiva no pretende sino caracterizar una secuencia de observaciones desde varios puntos de vista, no siendo nece-

sariamente todos de interés para el investigador. Aunque no es preciso desarrollar el fundamento de los MOM considerando estos tres problemas, se procederá de esta forma, dado que serán de utilidad para articular el resto del contenido del capítulo; no obstante, es el segundo y tercero de estos problemas el que más aplicabilidad obtiene en este trabajo en consonancia con su propósito del estudio. Se enumeran a continuación los tres problemas, que pretenden:

1. Calcular la probabilidad con la que puede obtenerse una secuencia de observaciones $x_1, x_2, ..., x_T$, asumiendo que esta ha sido generada por un MOM de m estados. Recibe el nombre del problema de *evaluación*. Se utilizan las denominadas *probabilidades hacia adelante*, definidas recursivamente, como una solución eficiente para su cálculo, al reutilizar los cálculos efectuados para obtener estas probabilidades.

2. Determinar la secuencia de estados ocultos más probable del MOM que ha generado la secuencia de observaciones de una serie temporal, y en qué instantes de tiempo, los cambios de uno a otro estado entre los que componen $\mathcal{S}$ han ocurrido. Se le denomina el problema de la *decodificación*, distinguiéndose entre *local*, si se determina cada estado de forma individual para cada instante de tiempo, o *global*, si se efectúa dicho análisis de forma global para el conjunto de instantes. En este último caso, se utiliza generalmente el algoritmo de Viterbi (Viterbi, 1967; Forney, 1973) para su resolución.

3. Estimar los parámetros del MOM con los que, con mayor probabilidad, se ha obtenido una secuencia de observaciones concreta. Este aspecto es referido como el problema del *aprendizaje* (o del *entrenamiento*, según el autor, dado que al conjunto de observaciones de la serie temporal se le denomina datos de aprendizaje o de entrenamiento). Los parámetros de un MOM pueden estimarse mediante métodos de optimización numérica de la función de log-verosimilitud y también mediante métodos iterativos (algoritmo EM). Como en el caso de los MMF, el algoritmo EM fue el utilizado, y así, todas las limitaciones de este algoritmo son igualmente trasladables al caso de los MOM. Las *probabilidades hacia adelante*, mencionadas en el problema uno, junto con unas nuevas, las probabilidades condicionadas denominadas *hacia atrás*, serán necesarias para la implementación del algoritmo EM. En el contexto de los MOM, es denominado algoritmo Baum-Welch (Baum et al., 1970; Baum, 1972; Welch, 2003), considerándose un caso especial de aquel.

El algoritmo de Viterbi y el algoritmo EM, utilizados, respectivamente, para la resolución de los problemas dos y tres, han sido implementados computacionalmente para este trabajo, acompañándose su código R y explicación del mismo más adelante, en el Anexo B.

3.2.1. Evaluación

En este apartado, mediante las probabilidades *hacia adelante*, se expone un procedimiento computacionalmente eficiente para estimar la verosimilitud de una secuencia de observaciones, la cual se define a continuación.

Definición 3.8 *Sea una secuencia de observaciones $x_1, x_2, ..., x_T$ generada por un MOM $\{X_t : t \in \mathbb{N}\}$ de m estados, con distribución inicial $\boldsymbol{\delta}$ y m.p.t. $\boldsymbol{\Gamma}$ para la CM, y matriz diagonal $\mathbf{P}(x_t)$, $t = 1, ..., T$, de dimensiones $m \times m$, con el i-ésimo elemento diagonal la función de densidad de probabilidad dependientes de estado p_i. Entonces, la función de verosimilitud del conjunto de observaciones se escribe*

$$\begin{aligned} L_T &= P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \cdots, X_T = x_T) \\ &= \boldsymbol{\delta}\,\mathbf{P}(x_1)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_2)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\,\mathbf{P}(x_T)\,\mathbf{1}'. \end{aligned} \tag{3.9}$$

Si $\boldsymbol{\delta}$, la distribución de C_1, es además la distribución estacionaria de la CM, entonces,

			1	2	3	4	5	6	
i	j	k	$p_i(1)$	$p_j(2)$	$p_k(5)$	δ_i	γ_{ij}	γ_{jk}	producto
1	1	1	$2.42\ 10^{-1}$	$3.98\ 10^{-1}$	$4.43\ 10^{-3}$	0.4	0.3	0.3	$1.54\ 10^{-5}$
1	1	2	$2.42\ 10^{-1}$	$3.98\ 10^{-1}$	$1.76\ 10^{-1}$	0.4	0.3	0.7	$\mathbf{1.43\ 10^{-3}}$
1	2	1	$2.42\ 10^{-1}$	$1.21\ 10^{-1}$	$4.43\ 10^{-3}$	0.4	0.7	0.5	$1.82\ 10^{-5}$
1	2	2	$2.42\ 10^{-1}$	$1.21\ 10^{-1}$	$1.76\ 10^{-1}$	0.4	0.7	0.5	$7.22\ 10^{-4}$
2	1	1	$6.47\ 10^{-2}$	$3.98\ 10^{-1}$	$4.43\ 10^{-3}$	0.6	0.5	0.3	$1.03\ 10^{-5}$
2	1	2	$6.47\ 10^{-2}$	$3.98\ 10^{-1}$	$1.76\ 10^{-1}$	0.6	0.5	0.7	$9.55\ 10^{-4}$
2	2	1	$6.47\ 10^{-2}$	$1.21\ 10^{-1}$	$4.43\ 10^{-3}$	0.6	0.5	0.5	$5.21\ 10^{-6}$
2	2	2	$6.47\ 10^{-2}$	$1.21\ 10^{-1}$	$1.76\ 10^{-1}$	0.6	0.5	0.5	$2.07\ 10^{-6}$
								suma :	$3.36\ 10^{-3}$

Tabla 3.1 Ejemplo del cálculo de la verosimilitud como una suma para secuencias de observaciones. En la última columna se resalta el mayor producto entre los ocho calculados.

$$L_T = \boldsymbol{\delta}\,\boldsymbol{\Gamma}\,\mathbf{P}(x_1)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_2)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\,\mathbf{P}(x_T)\,\mathbf{1}'. \tag{3.10}$$

Ejemplo 3.3 Se considera un MOM de 2 estados, con m.p.t.

$$\Gamma = \begin{pmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix},$$

distribuciones dependientes de estado y distribución inicial, no estacionaria,

$$p_i(x) \in \phi_i(x|2i, i), \quad i = 1, 2 \quad \text{y} \quad \boldsymbol{\delta} = (0.4, 0.6).$$

Se calcula, a continuación, la probabilidad de obtener la secuencia de observaciones $X_1 = 1$, $X_2 = 2$, $X_3 = 5$, y la secuencia de estados más probables, que la ha generado.

Utilizando la expresión (3.9) el valor de $L_3 = \boldsymbol{\delta}\,\mathbf{P}(1)\,\boldsymbol{\Gamma}\,\mathbf{P}(2)\,\boldsymbol{\Gamma}\,\mathbf{P}(5)\,\mathbf{1}'$ o, en forma escalar,

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \delta_i\, p_i(1)\, \gamma_{ij}\, p_j(2)\, \gamma_{jk}\, p_k(5), \tag{3.11}$$

que persigue obtener la combinación de estados que haga máxima la triple suma.

En la Tabla 3.1 se evalúa la probabilidad de obtener la secuencia de interés mediante la combinación de los diferentes valores de i, j, k. La suma en la última columna obtiene tal probabilidad ($3.36\ 10^{-3}$). La triple suma en (3.11) requiere de m^T términos (las 8 filas de la tabla), y cada uno de los términos es el producto de $2T = 2 \times 3$ factores (las 6 columnas numeradas de la tabla), por lo que el cálculo mediante este procedimiento precisa de $2T \cdot m^T$ operaciones. El mayor producto de cada una de las filas de la tabla (resaltado) es el que corresponde a los valores $i = 1, j = 1, k = 2$, siendo por tanto la secuencia de estados 1 1 2 la que máximiza L_3, y por tanto, la secuencia de estados buscada. Este es un caso de *decodificación global*, que será tratado más adelante (Ejemplo 3.9, página 53). El requerimiento computacional observado para el cálculo de L_T para secuencias de observaciones con un T elevado motiva el desarrollo de las siguientes probabilidades.

Probabilidades hacia adelante

El cálculo de L_T puede exponerse mediante la forma de un algoritmo, denominado *hacia adelante*. Para ello se define un vector fila $\boldsymbol{\alpha}_t$, $t = 1, ..., T$,

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\,\mathbf{P}(x_1)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_2)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\,\mathbf{P}(x_t) = \boldsymbol{\delta}\,\mathbf{P}(x_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\,\mathbf{P}(x_s). \tag{3.12}$$

De esta definición se deduce que

$$L_T = \boldsymbol{\alpha}_T\,\mathbf{1}' \qquad \text{y} \qquad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\,\boldsymbol{\Gamma}\,\mathbf{P}(x_t) \quad t \geq 2,$$

y la ecuación (3.9) puede estructurarse entonces en tres etapas, mediante una inicialización, seguida de inducción y una suma final:

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\,\mathbf{P}(x_1)\,, \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\,\boldsymbol{\Gamma}\,\mathbf{P}(x_t) \quad t = 2,3,...,T\,, \\ L_T &= \boldsymbol{\alpha}_T\,\mathbf{1}' \end{aligned} \tag{3.13}$$

A efectos de comparación de este algoritmo con el de Viterbi, que se verá más adelante, se expresa (3.13) de forma escalar:

$$\alpha_1(j) = \delta_j\, p_j(x_1)\,, \tag{3.14}$$

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{m}\alpha_t(i)\gamma_{ij}\right) p_j(x_{t+1}) \quad t = 1,2,...,T-1 \quad j = 1,...,m, \tag{3.15}$$

Si $\boldsymbol{\delta}$, la distribución de C_1, es la distribución estacionaria de la CM, entonces las tres etapas anteriores se expresan:

$$\begin{aligned} \boldsymbol{\alpha}_0 &= \boldsymbol{\delta}\,; \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\,\boldsymbol{\Gamma}\,\mathbf{P}(x_t) \quad t = 1,2,...,T\,; \\ L_T &= \boldsymbol{\alpha}_T\,\mathbf{1}'. \end{aligned}$$

Los elementos del vector fila $\boldsymbol{\alpha}_t$ se denominan probabilidades *hacia adelante*, y son, cada una, una probabilidad conjunta. Así, el j-ésimo elemento de $\boldsymbol{\alpha}_t$,

$$\alpha_t(j) = P(X_1 = x_1,\ldots,X_t = x_t, C_t = j) = P(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j), \quad j = 1,\ldots,m \quad t = 1,\ldots,T,$$

representa la probabilidad de que la cadena se encuentre en el estado j en el instante t después de haber sido observada la secuencia $x_1, x_2, \ldots, x_T$.

Ejemplo 3.4 Utilizando el Ejemplo 3.3 se calcula L_3 mediante el algoritmo (3.13). Se acompaña, con mayor precisión, el valor de $l_3 = \log(L_3)$.

$$
\begin{aligned}
\boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\,\mathbf{P}(x_1) = (0.4, 0.6)\begin{pmatrix} 2.42\ 10^{-1} & 0 \\ 0 & 6.47\ 10^{-2} \end{pmatrix} = (9.67\,10^{-2}, 3.88\,10^{-2}),\\
\boldsymbol{\alpha}_2 &= \boldsymbol{\alpha}_1\,\boldsymbol{\Gamma}\mathbf{P}(x_2) = (9.67\,10^{-2}, 3.88\,10^{-2})\begin{pmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix}\begin{pmatrix} 3.98\ 10^{-1} & 0 \\ 0 & 1.20\ 10^{-1} \end{pmatrix}\\
&= (1.93\,10^{-2}, 1.05\,10^{-2}),\\
\boldsymbol{\alpha}_3 &= \boldsymbol{\alpha}_2\,\boldsymbol{\Gamma}\mathbf{P}(x_3) = (1.93\,10^{-2}, 1.05\,10^{-2})\begin{pmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix}\begin{pmatrix} 4.43\ 10^{-3} & 0 \\ 0 & 1.76\ 10^{-1} \end{pmatrix}\\
&= (4.91\,10^{-5}, 3.31\,10^{-3}).\\
L_3 &= \boldsymbol{\alpha}_3\,\mathbf{1}' = (4.91\,10^{-5}, 3.31\,10^{-3})\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3.36\ 10^{-3},\\
l_3 &= \log(L_3) = -5.6958.
\end{aligned}
$$

resultado L_3 que coincide con el del Ejemplo 3.3, aunque han sido necesarias sólo $T \cdot m^2$ en lugar de $2T \cdot m^T$ operaciones para su obtención.

Observable en este ejemplo, y en tan solo 3 instantes de tiempo, las numerosas multiplicaciones matriciales traen como consecuencia que los elementos de $\boldsymbol{\alpha}_t$, tiendan a hacerse exponencialmente menores a medida que t aumenta, convergiendo finalmente a 0 y resultando en un desbordamiento, lo que conlleva la imposibilidad del cálculo de L_T para secuencias de observaciones incluso de moderada longitud. Esta circunstancia, que hace que la computación del valor de la función de verosimilitud en los MMF sea más sencilla que en los MOM, ha propiciado que se propongan varias soluciones (Durbin et al., 1998, p.78; Rabiner, 1989, p. 282-3). Una de ellas es la de Zucchini y MacDonald (2009, p. 46-47), que será la que se emplee en este trabajo, y que se expone a continuación.

Escalado del vector $\boldsymbol{\alpha}_t$

Este procedimiento permite obtener $\log L_T$ utilizando un escalado de los elementos de $\boldsymbol{\alpha}_t$, a medida que t progresa, además de que, a la vez, se previene el desbordamiento de los mismos. Se define para $t = 1, ..., T$, la matriz $\mathbf{B}_t = \boldsymbol{\Gamma}\mathbf{P}(x_t)$, y el vector

$$\boldsymbol{\phi}_t = \boldsymbol{\alpha}_t / w_t, \quad \text{con} \quad w_t = \sum_i \alpha_t(i) = \boldsymbol{\alpha}_t\,\mathbf{1}'.$$

Tras la definición de $\boldsymbol{\phi}_t$ y w_t se deduce:

$$
\begin{aligned}
w_0 = \boldsymbol{\alpha}_0\,\mathbf{1}' &= \boldsymbol{\delta}\,\mathbf{1}' = 1\,;\\
\boldsymbol{\phi}_0 &= \boldsymbol{\delta}\,;\\
w_t\,\boldsymbol{\phi}_t &= w_{t-1}\,\boldsymbol{\phi}_{t-1}\,\mathbf{B}_t\,;\\
L_T = \boldsymbol{\alpha}_T\,\mathbf{1}' &= w_T\,(\boldsymbol{\phi}_T\,\mathbf{1}') = w_T.
\end{aligned}
\tag{3.16}
$$

Por tanto, $L_T = w_T = \prod_{t=1}^{T}(w_t / w_{t-1})$. De (3.16) sigue que

$$w_t = w_{t-1}(\boldsymbol{\phi}_{t-1}\,\mathbf{B}_t\ \ \mathbf{1}'),$$

obteniéndose

$$\log L_T = \sum_{t=1}^{T} \log(w_t / w_{t-1}) = \sum_{t=1}^{T} \log(\boldsymbol{\phi}_{t-1}\,\mathbf{B}_t\ \ \mathbf{1}').$$

El cálculo de la log-verosimilitud se resume a continuación en forma de algoritmo, en donde $\mathbf{\Gamma}$ y $\mathbf{P}(x_t)$ son matrices de dimensiones $m \times m$, $\mathbf{v}$ y $\boldsymbol{\phi}_t$ son vectores de longitud m, u es un escalar, y l es el escalar en donde se acumula la log-verosimilitud.

$$
\begin{aligned}
\text{asignar} \quad & \boldsymbol{\phi}_0 \leftarrow \boldsymbol{\delta} \ \text{ y } \ l \leftarrow 0 \\
\text{para} \quad & t = 1, 2, ..., T \quad \text{hacer} \\
& \mathbf{v} \leftarrow \boldsymbol{\phi}_{t-1} \mathbf{\Gamma} \mathbf{P}(x_t) \\
& u \leftarrow \mathbf{v}\, \mathbf{1}' \\
& l \leftarrow l + \log u \\
& \boldsymbol{\phi}_t \leftarrow \mathbf{v}/u \\
\text{devolver} \quad & l
\end{aligned}
\tag{3.17}
$$

El valor final requerido de la log-verosimilitud, $\log L_T$, viene representado por l.

Ejemplo 3.5 Se muestran los resultados de la implementación en R del algoritmo (3.17) mediante la función `gauss.HMM.lalpha` (Anexo B.1). Esta función calcula el valor de l_T a partir de los valores escalados de $\boldsymbol{\alpha}_t$. Se acompaña un ejemplo de su uso y se obtiene el resultado de l_3 para su comparación con el del Ejemplo 3.4 (pág. 45).

La función `gauss.HMM.lalpha` emplea como argumentos la secuencia de observaciones y la parametrización del modelo indicadas en el Ejemplo 3.3 (pág. 44):

```
datos<-c(1,2,5)
gamma<-matrix(c(0.3,0.7,0.5,0.5), ncol=2, byrow=T)
delta<-c(0.4, 0.6)
mu<-c(2,4)
sigma<-c(1,2)
```

Los resultados obtenidos al aplicar la función son devueltos en un formato de lista, cuyos componentes, `lalpha` y `l`, representan los valores de $\boldsymbol{\alpha}_t$ escalados y l, respectivamente, en los instantes de tiempo $t = 1, 2, 3$, representadas por columnas de izquierda a derecha. La primera y segunda fila en `lalpha` representan el estado 1 y 2 de la CM, respectivamente.

```
> gauss.HMM.lalpha(datos,m=2,mu=mu, sigma=sigma, gamma=gamma, delta=delta)
$lalpha
          [,1]      [,2]      [,3]
[1,] -2.335229 -3.945870 -9.922091
[2,] -3.247911 -4.551872 -5.710570

$l
[1] -1.997725 -3.510505 -5.695855
```

Probabilidades hacia atrás

Una vez definidas las probabilidades hacia adelante, con las que ha podido resolverse eficientemente el cálculo de L_T dado un MOM de m estados, procede a continuación definir el vector de probabilidades *hacia atrás*. No obstante, no se requieren para la resolución del problema uno de los MOM, sino que participan, en su lugar, en la resolución de los problemas dos y tres.

El vector de estas probabilidades se denota mediante $\boldsymbol{\beta}_t$, tal que para un conjunto de $t = 1, \ldots, T$ instantes de tiempo,

$$\boldsymbol{\beta}'_t = \boldsymbol{\Gamma}\,\mathbf{P}(x_{t+1})\,\boldsymbol{\Gamma}\,\mathbf{P}(x_{t+2})\,\boldsymbol{\Gamma}\,\mathbf{P}(x_{t+3})\cdots\boldsymbol{\Gamma}\,\mathbf{P}(x_T)\,\mathbf{1}' = \left(\prod_{s=t+1}^{T}\boldsymbol{\Gamma}\,\mathbf{P}(x_s)\right)\mathbf{1}', \tag{3.18}$$

con la convención de que en el instante $t = T$, $\boldsymbol{\beta}_T = \mathbf{1}$. El j-ésimo elemento de $\boldsymbol{\beta}_t$, $\beta_t(j)$, se identifica como la probabilidad condicional

$$\beta_t(j) = P(X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \ldots, X_T = x_T \,|\, C_t = j), \tag{3.19}$$

suponiendo que $P(C_t = j) > 0$, representando la probabilidad de la secuencia (parcial) de observaciones $x_{t+1}, x_{t+2}, \ldots, x_{T-1}, x_T$, dado que la CM se encuentre en el estado j en el instante t. Como sucedía con las probabilidades hacia adelante, $\boldsymbol{\beta}_t$ puede resolverse mediante una inicialización y una inducción, siendo necesarias igualmente para su obtención $T \cdot m^2$ operaciones:

$$\begin{aligned} \boldsymbol{\beta}_T &= \mathbf{1}; \\ \boldsymbol{\beta}'_t &= \mathbf{B}_{t+1}\,\boldsymbol{\beta}'_{t+1} = \boldsymbol{\Gamma}\,\mathbf{P}(x_{t+1})\,\boldsymbol{\beta}'_{t+1} \qquad t = 1, 2, ..., T-1. \end{aligned} \tag{3.20}$$

Ejemplo 3.6 Utilizando el Ejemplo 3.3, se calcula $\boldsymbol{\beta}_t$, $t = 1, 2, 3$, mediante el algoritmo recursivo (3.20).

$$\begin{aligned} \boldsymbol{\beta}'_3 &= \mathbf{1}, \\ \boldsymbol{\beta}'_2 &= \boldsymbol{\Gamma}\mathbf{P}(x_3)\,\mathbf{1} = \begin{pmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix}\begin{pmatrix} 4.43\ 10^{-3} & 0 \\ 0 & 1.76\ 10^{-1} \end{pmatrix}\begin{pmatrix} 1 & 1 \end{pmatrix} = \begin{pmatrix} 1.24\ 10^{-1} \\ 9.02\ 10^{-2} \end{pmatrix}, \\ \boldsymbol{\beta}'_1 &= \boldsymbol{\Gamma}\mathbf{P}(x_2)\,\boldsymbol{\beta}'_2 = \begin{pmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix}\begin{pmatrix} 3.98\ 10^{-1} & 0 \\ 0 & 1.21\ 10^{-1} \end{pmatrix}\begin{pmatrix} 1.24\ 10^{-1} \\ 9.02\ 10^{-2} \end{pmatrix} = \begin{pmatrix} 2.25\ 10^{-2} \\ 3.03\ 10^{-2} \end{pmatrix}. \end{aligned}$$

Las probabilidades hacia atrás son escaladas de forma similar a como lo son las probabilidades hacia adelante, dado que, igualmente, están sujetas al problema de desbordamiento.

Escalado del vector $\boldsymbol{\beta}_t$

Se define el vector ψ_t, $t = 1, 2, \ldots, T$, de dimensiones $1 \times m$, resultante del escalado del vector $\boldsymbol{\beta}_t$, tal que

$$\psi_t\,\mathbf{1} = 1.$$

Así, para $t = 1, 2, \ldots, T$,

$$\psi_t = \frac{\boldsymbol{\beta}_t}{s_t}, \tag{3.21}$$

donde,

$$s_t = \boldsymbol{\beta}_t\,\mathbf{1}'. \tag{3.22}$$

En particular, con $\boldsymbol{\beta}_T = \mathbf{1}$ y utilizando (3.22),

$$s_T = \boldsymbol{\beta}_T \mathbf{1}' = \mathbf{1}\mathbf{1}' = m.$$

Sabiendo que $\boldsymbol{\beta}'_t = \mathbf{B}_{t+1}\boldsymbol{\beta}'_{t+1}$ y empleando la expresión (3.21), se obtiene

$$\boldsymbol{\beta}'_t = s_{t+1}\mathbf{B}_{t+1}\boldsymbol{\psi}'_{t+1}, \tag{3.23}$$

Así,

$$\log \boldsymbol{\beta}'_t = \log(\mathbf{B}_{t+1}\boldsymbol{\psi}'_{t+1}) + \log s_{t+1}. \tag{3.24}$$

Multiplicando (3.23) por $\mathbf{1}$ y utilizando (3.22),

$$s_t = s_{t+1}\mathbf{1}\mathbf{B}_{t+1}\boldsymbol{\psi}'_{t+1},$$

se obtiene

$$\log s_t = \log s_{t+1} + \log(\mathbf{1}\mathbf{B}_{t+1}\boldsymbol{\psi}'_{t+1}). \tag{3.25}$$

Las ecuaciones (3.24) y (3.25) proporcionan un algoritmo para el cálculo de los logaritmos naturales de las probabilidades hacia atrás, $\log \boldsymbol{\beta}_t$, semejante a como ocurría con las probabilidades $\boldsymbol{\alpha}_t$. Este algoritmo es el siguiente:

$$\begin{aligned}
&\text{asignar} \quad l\,\boldsymbol{\beta}_T \leftarrow \mathbf{0},\ \ \boldsymbol{\psi}_T \leftarrow \mathbf{1}/m \ \text{ y } \ \log s_T \leftarrow \log m \\
&\text{para} \quad t = T-1, T-2, ..., 1 \quad \text{hacer} \\
&\qquad \mathbf{v}' \leftarrow \boldsymbol{\Gamma}\mathbf{P}(x_{t+1})\boldsymbol{\psi}'_{t+1} \\
&\qquad \log \boldsymbol{\beta}_t \leftarrow \log \mathbf{v} + \log s_{t+1} \\
&\qquad u \leftarrow \mathbf{1}\mathbf{v}' \\
&\qquad \boldsymbol{\psi}_t \leftarrow \mathbf{v}/u \\
&\qquad \log s_t \leftarrow \log s_{t+1} + \log u \\
&\text{devolver} \quad \log \boldsymbol{\beta}_t
\end{aligned} \tag{3.26}$$

Ejemplo 3.7 Se muestran los resultados de la implementación en R de la función `gauss.HMM.lalphabeta` (Anexo B.2), la cual obtiene los valores escalados de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$ mediante los algoritmos (3.17) y (3.26) expuestos.

Los resultados obtenidos al aplicar la función son devueltos en un formato de lista, cuyos componentes, `lalpha` y `lbeta`, contienen, respectivamente, los valores de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$ en los instantes de tiempo $t = 1, 2, 3$

para la CM de 2 estados, representados por columnas de izquierda a derecha. La primera y segunda fila de `lalpha` y `lbeta` representan el estado 1 y 2 de la CM, respectivamente.

```
> gauss.HMM.lalphabeta(datos,m=2,mu=mu, sigma=sigma, gamma=gamma, delta=delta)
$lalpha
          [,1]      [,2]      [,3]
[1,] -2.335229 -3.945870 -9.922091
[2,] -3.247911 -4.551872 -5.710570

$lbeta
          [,1]      [,2] [,3]
[1,] -3.792086 -2.083029    0
[2,] -3.496508 -2.405368    0
```

Un resumen de los resultados se muestra en la Tabla 3.2.

t	$\boldsymbol{\alpha}_t$ no escalado	$\boldsymbol{\alpha}_t$ escalado
1	$(9.67\,10^{-2},\ 3.88\,10^{-2})$	(-2.335, -3.248)
2	$(1.93\,10^{-2},\ 1.05\,10^{-2})$	(-3.946, -4.552)
3	$(4.91\,10^{-5},\ 3.31\,10^{-3})$	(-9.922, -5.711)

t	$\boldsymbol{\beta}_t$ no escalado	$\boldsymbol{\beta}_t$ escalado
1	$(2.25\,10^{-2},\ 3.03\,10^{-2})$	(-3.792, -3.497)
2	$(1.24\,10^{-1},\ 9.02\,10^{-2})$	(-2.083, -2.405)
3	(1, 1)	(0, 0)

Tabla 3.2 Comparación de los valores no escalados y escalados de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$.

Comprobándose que mediante el escalado, los valores de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$ adquieren semejante orden de magnitud, previniendo el desbordamiento a medida que t aumenta.

3.2.2. Decodificación

Se presentan a continuación unos resultados que relacionan las probabilidades hacia delante $\alpha_t(i)$ y hacia atrás $\beta_t(i)$ con las probabilidades $P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i)$. Estos resultados serán utilizados en los denominados problemas de *decodificación local* (Proposición 3.3), así como en la aplicación del algoritmo EM a los MOM (Proposición 3.4), pudiendo consultarse cada una de sus demostraciones en Zucchini y MacDonald (2009). Se expone, además, el problema de la *decodificación global*, por el que se determina la secuencia de estados más probable que ha dado lugar a un conjunto de observaciones.

Proposición 3.3 *Para* $t = 1, 2, \ldots, T \quad y \quad i = 1, 2, \ldots, m,$

$$\alpha_t(i)\,\beta_t(i) = P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i), \tag{3.27}$$

y, en consecuencia,

$$\boldsymbol{\alpha}_t\ \boldsymbol{\beta}_t{}' = P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = L_T,\ \textit{para cada } t. \tag{3.28}$$

Observación 3.8 La igualdad (3.27) se refiere a que la probabilidad conjunta de las observaciones y $C = i$, en el instante t, se obtiene mediante producto de las probabilidades hacia adelante y atrás, en el instante t. La expresión (3.28) sigue a (3.27), teniendo en cuenta la expresión matricial de probabilidad y las definiciones de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$:

$$L_T = (\boldsymbol{\delta}\,\mathbf{P}(x_1)\,\boldsymbol{\Gamma}\,\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\,\mathbf{P}(x_t))\;(\boldsymbol{\Gamma}\,\mathbf{P}(x_{t+1})\cdots\boldsymbol{\Gamma}\,\mathbf{P}(x_T)\,\mathbf{1}') = \boldsymbol{\alpha}_t\;\boldsymbol{\beta}_t{}'.$$

La expresión (3.28) permite calcular L_T mediante T rutas, una por cada valor de t.

Proposición 3.4 *Para* $t = 1, 2, \ldots, T$,

$$P(C_t = j \,|\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{P(C_t = j\,,\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)})}{P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)})} = \frac{\alpha_t(j)\,\beta_t(j)}{L_T}; \tag{3.29}$$

y, además, para $t = 2, 3, \ldots, T$,

$$P(C_{t-1} = j, C_t = k \,|\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \frac{\alpha_{t-1}(j)\;\gamma_{jk}\;p_k(x_t)\;\beta_t(k)}{L_T}. \tag{3.30}$$

Observación 3.9 La expresión (3.29), la distribución condicional de C_t dado el conjunto de observaciones, indica la probabilidad de que la CM se encuentre en el estado j en el instante t. Esta probabilidad se obtiene gracias al denominador L_T, que asegura la normalización del producto $\alpha_t(j)\,\beta_t(j)$.

Observación 3.10 Las expresiones (3.29) y (3.30) serán posteriormente utilizadas en el paso E del algoritmo EM, para la resolución del problema de aprendizaje de los MOM.

Para cada instante de tiempo $t \in \{1, \ldots, T\}$ puede determinarse la distibución del estado C_t, dada las observaciones $\mathbf{x}^{(T)}$, la cual, para m estados, es una distribución de probabilidad discreta con soporte $\{1, \ldots, m\}$. Para cada $t \in \{1, \ldots, T\}$ el estado más probable i_t^*, dada la secuencia de observaciones, se define como

$$i_t^* = \underset{i=1,\ldots,m}{\arg\max}\, P(C_t = i \,|\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}). \tag{3.31}$$

Esta aproximación determina el estado más probable de forma separada para cada instante de tiempo t mediante la maximización de la probabilidad condicional $P(C_t = i \,|\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$, motivo por el que se la denomina *decodificación local*.

Ejemplo 3.8 Se determina la secuencia de estados más probable utilizando la *decodificación local*, empleando los valores de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$ de los ejemplos anteriores. Por (3.27):

$$\begin{aligned}
P(X_1 = 1, X_2 = 2, X_3 = 5, C_1 = i) &= (2.18\;10^{-3}, 1.17\;10^{-3}), \quad \text{para } i{=}1,\,2.\\
P(X_1 = 1, X_2 = 2, X_3 = 5, C_2 = i) &= (2.41\;10^{-3}, 9.51\;10^{-4}), \quad \text{para } i{=}1,\,2.\\
P(X_1 = 1, X_2 = 2, X_3 = 5, C_3 = i) &= (4.91\;10^{-5}, 3.31\;10^{-3}), \quad \text{para } i{=}1,\,2.
\end{aligned}$$

Mediante (3.28) se calcula el valor de L_T, para $t = 1, 2, 3$:

$$L_3 = \boldsymbol{\alpha}_1\;\boldsymbol{\beta}'_1 = \boldsymbol{\alpha}_2\;\boldsymbol{\beta}'_2 = \boldsymbol{\alpha}_3\;\boldsymbol{\beta}'_3 = 3.36\;10^{-3}.$$

Observación 3.11 Se prefiere realizar el cálculo de L_3 a partir de $\boldsymbol{\alpha}_3\,\mathbf{1}'$ ($\boldsymbol{\alpha}_3\;\boldsymbol{\beta}'_3$, con $\boldsymbol{\beta}_3 = \mathbf{1}'$), como en el Ejemplo 3.4, ya que requiere solo del cálculo de probabilidades hacia adelante y de un único recorrido (hacia adelante) a través de la cadena.

Entonces, por (3.29),

$$\begin{array}{lcll} P(C_1 = i \,|\, X_1 = 1, X_2 = 2, X_3 = 5) & = & (\mathbf{0.6496}, 0.3504), & \text{para } i\text{=1, 2.} \\ P(C_2 = i \,|\, X_1 = 1, X_2 = 2, X_3 = 5) & = & (\mathbf{0.7167}, 0.2833), & \text{para } i\text{=1, 2.} \\ P(C_3 = i \,|\, X_1 = 1, X_2 = 2, X_3 = 5) & = & (0.0146, \mathbf{0.9854}), & \text{para } i\text{=1, 2.} \end{array} \tag{3.32}$$

Mediante (3.31) se obtiene que la secuencia de estados más probable (probabilidades resaltadas) es $i_1^* = 1$, $i_2^* = 1$, $i_3^* = 2$. Estos últimos resultados se relacionan por estados consecutivos mediante (3.30), resaltándose igualmente los de mayor probabilidad. Se obtiene, para $t = 2$ y

$$\begin{array}{rcl} j = 1, k = 1 & : & P(C_1 = 1, C_2 = 1 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_1(1)\,\gamma_{11}\,p_1(x_2)\,\beta_2(1)/L_3 = \mathbf{0.42942}. \\ j = 1, k = 2 & : & P(C_1 = 1, C_2 = 2 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_1(1)\,\gamma_{12}\,p_2(x_2)\,\beta_2(2)/L_3 = 0.22014. \\ j = 2, k = 1 & : & P(C_1 = 2, C_2 = 1 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_1(2)\,\gamma_{21}\,p_1(x_2)\,\beta_2(1)/L_3 = 0.28732. \\ j = 2, k = 2 & : & P(C_1 = 2, C_2 = 2 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_1(2)\,\gamma_{22}\,p_2(x_2)\,\beta_2(2)/L_3 = 0.06312. \end{array}$$

y para $t = 3$ y

$$\begin{array}{rcl} j = 1, k = 1 & : & P(C_2 = 1, C_3 = 1 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_2(1)\,\gamma_{11}\,p_1(x_3)\,\beta_3(1)/L_3 = 0.00765. \\ j = 1, k = 2 & : & P(C_2 = 1, C_3 = 2 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_2(1)\,\gamma_{12}\,p_2(x_3)\,\beta_3(2)/L_3 = \mathbf{0.70909}. \\ j = 2, k = 1 & : & P(C_2 = 2, C_3 = 1 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_2(2)\,\gamma_{21}\,p_1(x_3)\,\beta_3(1)/L_3 = 0.00696. \\ j = 2, k = 2 & : & P(C_2 = 2, C_3 = 2 \,|\, X_1 = 1, X_2 = 2, X_3 = 5) = \alpha_2(2)\,\gamma_{22}\,p_2(x_3)\,\beta_3(2)/L_3 = 0.27630. \end{array}$$

Resultado que es consistente con los obtenidos en (3.32).

En ocasiones, el interés de la aplicación de los MOM no se centra en conocer el estado más probable que ha generado una observación en un instante de tiempo concreto, como en el caso de la decodificación local. Por el contrario, el interés se centra en la *secuencia* de estados más probable que ha generado las observaciones, tratándose por tanto de un problema denominado *decodificación global*. En lugar de maximizar $P(C_t = i \,|\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$ en i para cada t, como en (3.31), se busca la secuencia de estados $c_1, c_2, \ldots, c_T$ que maximiza la probabilidad condicional

$$P(\mathbf{C}^{(T)} = \mathbf{c}^{(T)} \,|\, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}); \tag{3.33}$$

o de forma equivalente, la probabilidad conjunta

$$P(\mathbf{C}^{(T)}, \mathbf{X}^{(T)}) = \delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1}, c_t} \prod_{t=1}^{T} p_{c_t}(x_t). \tag{3.34}$$

Los resultados obtenidos mediante la decodificación local y global son con frecuencia similares, pero no idénticos (Zucchini y MacDonald, 2005). Maximizar (3.33) sobre todas las posibles secuencias de estados $c_1, c_2, \ldots, c_T$ implicaría la evaluación de m^T funciones, como en el caso del Ejemplo 3.3 en la página 44. El algoritmo de Viterbi, mediante programación dinámica, determina la secuencia más probable de estados de forma eficiente, creciendo el coste computacional de esta estimación de forma proporcional a la longitud de la cadena (lineal en T).

Inicialmente, se define

$$\xi_{1i} = P(C_1 = i, X_1 = x_1) = \delta_i \, p_i(x_1), \tag{3.35}$$

y, para $t = 2, 3, \ldots, T$,

$$\xi_{ti} = \max_{c_1, c_2, \ldots, c_{t-1}} P(\mathbf{C}^{(t-1)} = \mathbf{c}^{(t-1)}, C_t = i, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}).$$

Las probabilidades ξ_{tj} satisfacen la siguiente recursión, para $t = 1, 2, \ldots, T-1$ e $i = 1, 2, \ldots, m$:

$$\xi_{t+1\ j} = \left[\max_i (\xi_{ti} \, \gamma_{ij}) \right] p_j(x_{t+1}), \tag{3.36}$$

representando una forma eficiente de obtener la matriz de valores ξ_{tj}, de dimensiones $T \times m$. La secuencia de estados requerida, $\widehat{i}_1, \widehat{i}_2, \ldots, \widehat{i}_T$, que da lugar a la secuencia de observaciones, puede determinarse recursivamente mediante la relación

$$\widehat{i}_T = \arg\max_{i=1,\ldots,m} \xi_{Ti} \tag{3.37}$$

y, para $t = T-1, T-2, \ldots, 1$, mediante

$$\widehat{i}_t = \arg\max_{i=1,\ldots,m} (\xi_{ti} \, \gamma_{i\, \widehat{i}_{t+1}}). \tag{3.38}$$

El algoritmo de Viterbi es aplicable a CM estacionarias y no estacionarias, sin ser necesario asumir que la distribución inicial $\boldsymbol{\delta}$ es la distribución estacionaria. Si se comparan las expresiones (3.35) con (3.14), y (3.36) con (3.15), se percibe la similitud en la implementación del algoritmo *hacia adelante* y el de Viterbi, sustituyendo la suma del primero por una maximización en el segundo.

Ejemplo 3.9 Se considera la información del Ejemplo 3.3 (página 44) para estimar la secuencia de estados más probable que ha generado las observaciones $X_1 = 1$, $X_2 = 2$, $X_3 = 5$, utilizando el algoritmo de Viterbi. Se valida posteriormente la implementación en R de este algoritmo mediante la función `viterbi.Gauss` (Sección B.3).

Se calculan, en primer lugar, las probabilidades $\{\xi_{ti}\}_{t=1,2,3\ i=1,2}$:

Para $t = 1$,

$$\begin{aligned} \xi_{11} &= \delta_1 \, p_1(x_1) = 9.67\,10^{-2}, \\ \xi_{12} &= \delta_2 \, p_2(x_1) = 3.88\,10^{-2}. \end{aligned}$$

Para $t = 2$,

$$\begin{aligned} \xi_{21} &= [\, \max(\xi_{11}\,\gamma_{11}, \xi_{12}\,\gamma_{21}) \,]\, p_1(x_2) = 1.15\,10^{-2}, \\ \xi_{22} &= [\, \max(\xi_{11}\,\gamma_{12}, \xi_{12}\,\gamma_{22}) \,]\, p_2(x_2) = 8.19\,10^{-3}. \end{aligned}$$

Para $t = 3$,

$$\begin{aligned}\xi_{31} &= [\,\text{máx}(\xi_{21}\,\gamma_{11}, \xi_{22}\,\gamma_{21})\,]\; p_1(x_3) = 1.81\,10^{-5},\\ \xi_{32} &= [\,\text{máx}(\xi_{21}\,\gamma_{12}, \xi_{22}\,\gamma_{22})\,]\; p_2(x_3) = \mathbf{1.43\,10^{-3}}.\end{aligned}$$

Así, aplicando (3.37) para $t = 3$ y (3.38) para $t = 1, 2$:

$$\begin{aligned}\widehat{i}_3 &= \arg\max_{i=1,2}(\xi_{31}, \xi_{32}) = 2.\\ \widehat{i}_2 &= \arg\max_{i=1,2}(\xi_{21}\gamma_{12}, \xi_{22}\gamma_{22}) = 1.\\ \widehat{i}_1 &= \arg\max_{i=1,2}(\xi_{11}\gamma_{11}, \xi_{12}\gamma_{21}) = 1,\end{aligned}$$

obteniéndose que la secuencia de estados más probable es $\widehat{i}_1 = 1, \widehat{i}_2 = 1, \widehat{i}_3 = 2$, resáltandose la probabilidad de obtener esa secuencia, $1.43\,10^{-3}$, que coincide con la obtenida en el Ejercicio 3.3. El procedimiento de escalado puede ser igualmente aplicado para evitar el desbordamiento en el cálculo de $\{\xi_{tj}\}$. La implementación del algoritmo en R rinde semejantes resultados, los cuales se devuelven en formato de lista, de componentes `xi` ($\{\xi_{ti}\}_{t=1,2,3\ i=1,2}$), `x` (observaciones) y `iv` (secuencia de estados más probable):

```
gamma<-matrix(c(0.3,0.7,0.5,0.5), ncol=2, byrow=2)
delta<-c(0.4,0.6)
x<-c(1,2,5)
mu<-c(2,4)
sd<-c(1,2)

> viterbi.Gauss(x=x,mu=mu,sd=sd, gamma=gamma, delta=delta)
$xi
         [,1]     [,2]
[1,] 0.096788 0.038855
[2,] 0.011584 0.008197
[3,] 0.000018 0.001427

$x
[1] 1 2 5

$iv
[1] 1 1 2
```

En dicha implementación, las T filas de la matriz (ξ_{tj}) pueden normalizarse para que su suma sea 1. En tal caso el resultado es el siguiente:

```
> viterbi.Gauss(x=x,mu=mu,sd=sd, gamma=gamma, delta=delta)
$xi
         [,1]     [,2]
[1,] 0.713549 0.286451
[2,] 0.585611 0.414389
[3,] 0.012565 0.987435

$x
[1] 1 2 5

$iv
[1] 1 1 2
```

Estos últimos resultados escalados (matriz `xi`) pueden compararse con los resultantes de la decodificación local (3.32).

No obstante, el algoritmo de Viterbi proporciona una secuencia única de estados, considerada una solución óptima en función del criterio de la máxima probabilidad, sin aportar información sobre otras

secuencias probables y posiblemente cercanas a esa solución (ver Gedon, 2007, para una discusión sobre otras posibilidades de estimación). La codificación local, en comparación con la global, aporta resultados con errores más pequeños, en particular cuando los diferentes estados ocultos no se encuentran bien separados (Bulla, 2011), si bien la decodificación local puede rendir secuencias de estados inadmisibles, por ejemplo, aquellas que para dos estados cualesquiera $i, j \in \mathcal{S}$ tengan definidas $\gamma_{ij} = 0$, suponiendo que exista alguna probabilidad nula en la m.p.t. del MOM en estudio. No obstante, cuando ese no es el caso y los estados se consideran, además, estables (elementos de la diagonal principal de la m.p.t. próximos a 1), la diferencias entre los resultados de la codificación local y global suelen ser mínimas (Visser et al., 2011). En la práctica, la decodificación local resulta de escaso interés.

3.2.3. Aprendizaje

El algoritmo EM puede aplicarse para la maximización de la verosimilitud de un MOM de una forma parecida a como se hizo en el caso de los MMF, aunque, no obstante, se presentará esta vez de forma más resumida. Es conveniente representar la secuencia de estados $c_1, c_2, \ldots, c_T$ mediante las variables aleatorias indicadoras definidas como sigue:

$$\begin{aligned} u_j(t) &= 1 \quad \text{si y sólo si} \quad c_t = j, \quad (t = 1, 2, \ldots, T) \\ v_{jk}(t) &= 1 \quad \text{si y sólo si} \quad c_{t-1} = j \quad c_t = k, \quad (t = 2, 3, \ldots, T) \end{aligned}$$

Con esta notación, y por (3.34), la función de log-verosimilitud de datos completos[2] de un MOM viene dada por

$$\log\Big(P(\mathbf{x}^{(T)}, \mathbf{c}^{(T)})\Big) = \log\left(\delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1},c_t} \prod_{t=1}^{T} p_{c_t}(x_t)\right) = \log \delta_{c_1} + \sum_{t=2}^{T} \log \gamma_{c_{t-1},c_t} + \sum_{t=1}^{T} \log p_{c_t}(x_t).$$

Así, esta función se obtiene como la suma de tres términos:

$$\begin{aligned} \log\big(P(\mathbf{x}^{(T)}, \mathbf{c}^{(T)})\big) &= \sum_{j=1}^{m} u_j(1) \log \delta_j + \sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{t=2}^{T} v_{jk}(t)\right) \log \gamma_{jk} + \sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log p_j(x_t) \qquad (3.39) \\ &= \textit{término 1} + \textit{término 2} + \textit{término 3}, \end{aligned}$$

siendo $p_j(x)$ la densidad Gaussiana como se definió en (2.4), y $\boldsymbol{\delta}$, la distribución inicial de la CM (la distribución de C_1), no necesariamente la distribución estacionaria. El algoritmo EM para el caso de lo MOM procede como sigue:

– **Paso E**. Sustituir las cantidades $v_{jk}(t)$ y $u_j(t)$ por su esperanza condicional, dada las observaciones $\mathbf{x}^{(T)}$ y las estimaciones de los parámetros actuales:

[2] Denotada como $\ell(\Psi|y, z)$ en los MMF.

$$\begin{aligned}\hat{u}_j(t) &= \mathsf{E}\left[\, u_j(t) \,|\, x_1, x_2, \ldots, x_T \,\right] = P(C_t = j \,|\, \mathbf{x}^{(T)}) = \alpha_t(j)\ \beta_t(j)\ /\ L_T\ ; \\ \hat{v}_{jk}(t) &= \mathsf{E}\left[\, v_{jk}(t) \,|\, x_1, x_2, \ldots, x_T \,\right] = P(C_{t-1} = j, C_t = k \,|\, \mathbf{x}^{(T)}) \\ &= \alpha_{t-1}(j)\ \delta_{jk}\ p_k(x_t)\ \beta_t(k)\ /\ L_T.\end{aligned}$$

Las probabilidades hacia adelante y hacia atrás son necesarias para el cálculo de estas probabilidades condicionadas.

– **Paso M**. Habiendo reemplazado $v_{jk}(t)$ y $u_j(t)$ por $\hat{v}_{jk}(t)$ y $\hat{u}_j(t)$, maximizar (3.39) con respecto a los tres conjuntos de parámetros: la distribución inicial $\boldsymbol{\delta}$, la m.p.t. $\boldsymbol{\Gamma}$, y los parámetros de la distribuciones dependientes de estados (Gaussianas).

Cada uno de los sumandos en (3.39) pueden ser maximizados en el paso M independientemente, dado que el *término 1* depende solo de la distribución inicial $\boldsymbol{\delta}$, el *término 2* de $\boldsymbol{\Gamma}$, y el *término 3*, de los "parámetros dependientes de estado". Por tanto, ha de maximizarse:

1. $\sum_{j=1}^{m} \hat{u}_j(1) \log \delta_j$ con respecto a $\boldsymbol{\delta}$;
2. $\sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{t=2}^{T} \hat{v}_{jk}(t) \right) \log \gamma_{jk}$ con respecto a $\boldsymbol{\Gamma}$; y
3. $\sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log p_j(x_t)$, con respecto a μ_j y σ_j^2.

Omitiendo el desarrollo algebráico de las maximizaciones, su resultado es:

1. $\hat{\delta}_j = \hat{u}_j(1) / \sum_{j=1}^{m} \hat{u}_j(1) = \hat{u}_j(1)$;
2. $\hat{\gamma}_{jk} = \sum_{t=2}^{T} \hat{v}_{jk}(t) / \sum_{k=1}^{m} \sum_{t=2}^{T} \hat{v}_{jk}(t)$; y
3. $\hat{\mu}_j = \sum_{t=1}^{T} \hat{u}_j\, x_t / \sum_{t=1}^{T} \hat{u}_j$; $\hat{\sigma}_j^2 = \sum_{t=1}^{T} \hat{u}_j\, (x_t - \hat{\mu}_j)^2 / \sum_{t=1}^{T} \hat{u}_j$.

La estacionariedad en un MOM es una propiedad deseable cuando estos modelos se aplican para la caracterización de SSTT. Según Bulla y Berzel (2008), la aplicación del algoritmo EM, tal como se ha expuesto, se indica para la estimación de los parámetros de MOM homogéneos, pero no estacionarios. Estos autores señalan que el tratamiento del *término 1* y *término 2* en el paso M como sumandos individuales de (3.39) (componentes de inicio y de transición) conlleva que no se verifiquen las condiciones indicadas en la Definición 3.6 ($\hat{\boldsymbol{\delta}}$ no es distribución estacionaria). Así, estos autores plantean un algoritmo EM modificado en su paso M, basado en la maximización conjunta respecto a $\boldsymbol{\Gamma}$ de la suma del *término 1* y *término 2* de la expresión (3.39). No obstante, esta salvedad no ha sido considerada trascendente como para ser implementada en este trabajo y llevar a cabo esta modificación del algoritmo EM, dado que los propios autores no encontraron diferencias significativas frente al algoritmo EM convencional, más allá de su rapidez. Este último factor, que no es relevante aquí, ya que la longitud de las SSTT estudiadas ha sido de pequeño tamaño, junto a la complejidad que hubiera supuesto implementar este nuevo algoritmo EM, hace que excluir la adopción de este nuevo método parezca entonces justificada.

3.3. Otras consideraciones

Varios aspectos que se abordaron en el Capítulo 2 referente a MMF no lo han sido en este. Estos son el problema de la identificabilidad, el cálculo de los momentos de un MOM, el criterio de parada en

el algoritmo EM, la selección del mejor modelo y la estimación de los errores de los parámetros de un MOM.

La identificabilidad de un MOM descansa sobre dos condiciones necesarias. La primera de ellas responde a lo expuesto en la Sección 2.1.1 sobre la definición de identificabilidad; la segunda, hace referencia a que la CM que gobierna el MOM sea irreducible y aperiódica. No obstante, la identificabilidad en los MOM es un problema que parece no generar inconvenientes técnicos en la práctica, dado que solo ha sido descrita para conjuntos de datos muy particulares; de hecho, este problema no es mencionado en, por ejemplo, el texto de Zucchini y MacDonald (2009). No obstante, puede consultarse Leroux (1992) y Frühwirth-Schnatter (2010, pág. 313-4) para una breve descripción más detallada.

Respecto al cálculo de los momentos y a la estimación de los errores bootstrap en un MOM, son aplicables las soluciones dadas en el capítulo anterior. Timmermann (2000, Corolario 1) ha desarrollado una expresión nueva para la varianza de un MOM de dos componentes gaussianas:

$$\sigma^2 = \pi_1 \sigma_1^2 + \pi_2 \sigma_2^2 + \pi_1 \pi_2 (\mu_1 - \mu_2)^2,$$

si bien esta expresión no se ha empleado en este trabajo, ya que únicamente es aplicable cuando se detectan dos componentes en la mixtura.

En el **Capítulo 6**, donde se desarrolla la parte aplicada de los MOM, respecto al criterio de parada, y siguiendo la referencia de Zucchini y MacDonald (2009), se optó por la suma de las diferencias absolutas por grupos de parámetros en lugar de la diferencia relativa. Se adoptó un valor $\epsilon = 1 \cdot 10^{-6}$, como en el caso de los MMF. También comparten los MOM y MMF el método para generar los valores iniciales del algoritmo EM.

En relación con la selección del mejor modelo, esto es, el número más apropiado de estados[3] en el MOM, igualmente son aplicables los criterios de información expuestos en la Sección 2.4. No obstante, como se procedió con los MMF, se empleó el criterio *BIC* para la selección del "mejor" MOM, si bien, como ya se mencionó, favorece los modelos con menor número de parámetros en comparación con el criterio *AIC*.

[3] En ocasiones este número recibe el nombre de "orden" del MOM, pudiendo confundirse con el término que relaciona la dependencia entre estados en un MOM.

4

Otras técnicas de minería de datos utilizadas

Las siguientes técnicas auxiliares se emplearon sobre los resultados experimentales de los MMF, en particular, a partir de la información proporcionada por el primer y segundo momento de las mixturas. En el Capítulo 5 se justifica su empleo, mientras que a continuación se repasa su fundamento de forma breve.

4.1. Análisis clúster jerárquico

Con está técnica multivariante se ha pretendido agrupar a las estaciones de medida (Tabla 5.1 - Estación) en función del grado de homogeneidad de la calidad del aire que en ellas se registra (Tabla 5.1 - Contaminantes estudiados). Así, estaciones con un alto grado de homogeneidad interna (niveles de calidad del aire asumiblemente similares) se situarán próximas en la representación gráfica del análisis (Figura 5.2). La métrica escogida fue la distancia euclídea, indicando las distancias más pequeñas una mayor similitud. Por ello, antes de realizar el análisis propiamente dicho, es necesario obtener una matriz de distancias, que fue calculada mediante la función dist de la librería stats de R.

El algoritmo empleado (función hclust, librería stats) asigna inicialmente a cada objeto un clúster propio, e identifica las dos observaciones más parecidas (cercanas) que no estén en el mismo clúster y las combina. Así, las iteraciones comienzan con cada observación en su propio clúster, combinando dos conglomerados a un tiempo, hasta que todas las observaciones se reúnen en un única *solución* clúster en función de su homogeneidad. El método aglomerativo utilizado fue el asociado al argumento single de hclust. La implementación del algoritmo se muestra en el Anexo C.2.1.

4.2. Imputación mediante bosques aleatorios

La denominación de bosques aleatorios responde a que, para la implementación del algoritmo, se utiliza un número determinado de árboles de decisión. En problemas de clasificación, se elige la clase más votada entre los distintos árboles, y en los de regresión, se calcula la media de todas las predicciones. El algoritmo de bosques aleatorios proporciona muy buenos resultados en estudios empíricos y se ejecuta de forma eficiente sobre grandes bases de datos, pudiéndose procesar miles de variables sin tener que eliminar ninguna y proporcionando estimaciones de la importancia de cada variable.

La función empleada (rfImpute, librería randomForest) comienza calculando la mediana de cada columna de la matriz de datos como valor inicial para la imputación de los datos faltantes en ellas, para posteriormente emplear el algoritmo Random Forest sobre la matriz de datos completada. rfImpute precisa de los argumentos iter y ntree, con los que se especifica el número de iteraciones que se desea

implementar y el número de árboles de decisión empleado en cada iteración. Estos valores fueron de 250 y 2 500, respectivamente (ver Anexo C.2.2).

El resultado de la implementación del algoritmo en R es una matriz de datos completa, en la que la variable respuesta viene representada por el tipo de estación en el que se han tomado las observaciones (Rural-R; S-Suburbana; U-Urbana), la cual es incluida en la primera columna de la matriz de datos en estudio. Las variables predictoras, continuas, se asocian a los contaminantes estudiados (p = 5: CO, NO_2, O_3, PM_{10}, SO_2). Al ser las variables predictoras continuas, rfImpute efectúa una regresión para la estimación de los valores imputados.

La aleatoriedad del algoritmo Random Forest se basa en dos aspectos: 1) cada árbol de decisión, entre los ntree, se crea a partir de un conjunto de observaciones de la muestra original; y 2) cada nodo de división en cada uno de estos árboles de decisión se crea a partir de un número de variables predictoras candidatas ($mtry < p$), seleccionadas aleatoriamente. En Random Forest, el término OOB ("out of the bag") se refiere al conjunto de observaciones de la muestra original que no ha sido incluida en la muestra bootstrap. Este término está asociado al error de los árboles generados en cada iteración. La predicción final con la que se obtienen los valores imputados finales se basa en el promedio de las predicciones realizadas por cada árbol. El algoritmo *bagging* es un caso especial de Random Forest cuando $mtry = p$. En la Figura 4.1 se muestra un esquema del algoritmo Random Forest.

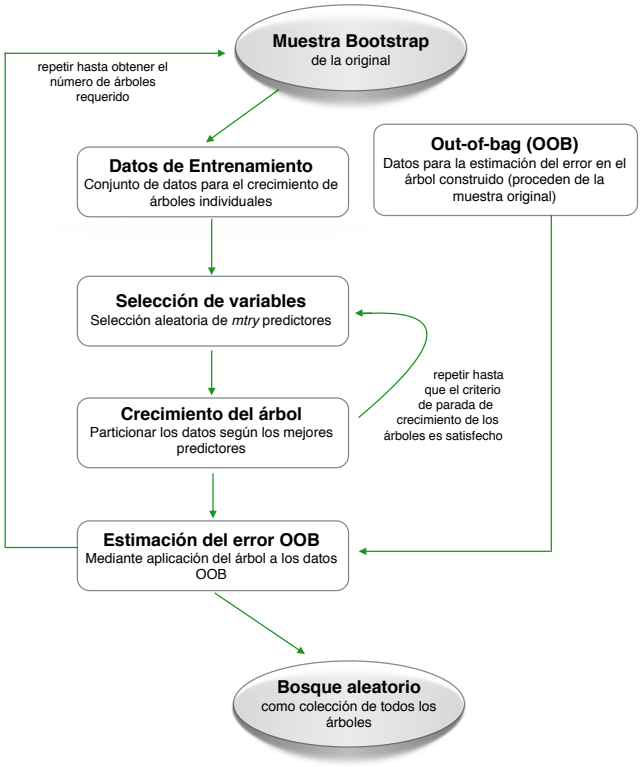


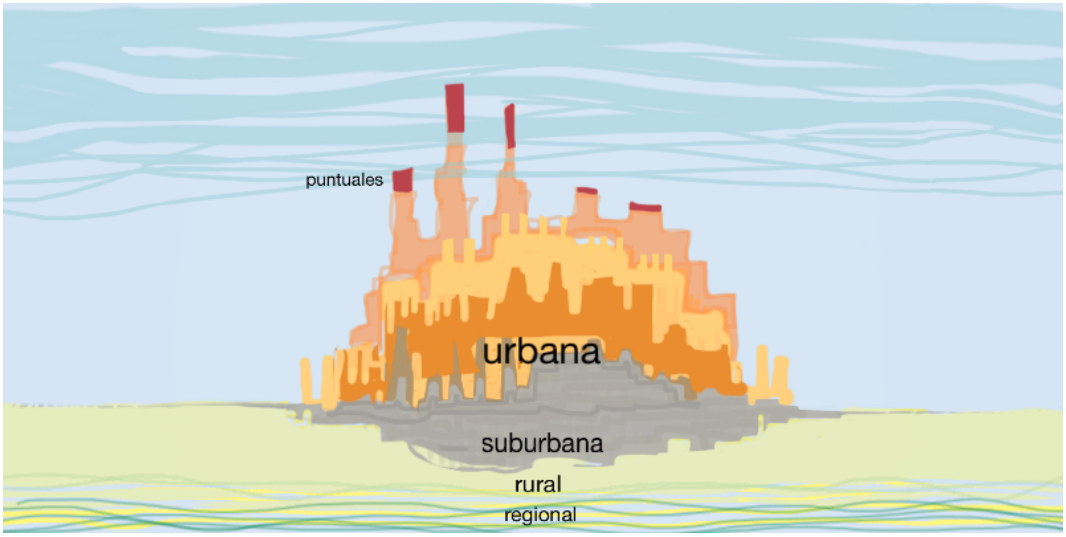
Figura 4.1 Algoritmo Random Forest (bosques aleatorios).

4.3. Análisis de componentes principales

El ACP es un método estadístico conocido desde principios de siglo (Pearson, 1901) y que consiste en describir la variación producida por la observación de p variables aleatorias, en términos de un conjunto

de nuevas variables aleatorias incorreladas entre sí (denominadas *componentes principales*, -CCPP-), cada una de las cuales sea combinación lineal de las variables originales. Estas nuevas variables son obtenidas en orden de importancia, de manera que la primera CP incorpora la mayor variación debida a las variables originales; la segunda CP se elige de forma que explique la mayor cantidad posible de variacion que resta sin explicar por la primera CP, sujeta a la condición de ser incorrelada con la primera CP, y así sucesivamente. De esta manera, se reduce la dimensionalidad de los datos al considerar un número de variables $q\,(q < p)$ y sin perder apenas información relevante.

La aplicación perseguida del ACP en este trabajo, además de reducir la dimensionalidad, ha sido la determinación de grupos homogéneos entre las estaciones estudiadas, a partir de los valores de μ_m y cv_m obtenidos (Tabla 5.2 y Figuras 5.3 y 5.4). Para su implementación mediante R se utilizó la función prcomp (librería stats). La implementación del ACP puede consultarse en el Anexo C.2.3.



Recreación de la aproximación incremental de Lenschow, mostrando las diferentes contribuciones en un área urbana.

5

Caracterización y mejora de las redes de vigilancia de la calidad del aire

Resumen En ocasiones, los planes de monitorización de la calidad del aire no se actualizan convenientemente en concordancia con las cambiantes condiciones locales, repercutiendo en la información atmosférica que proporcionan, bien dejando de detectar nuevas fuentes de contaminación o duplicando cierta información. Además, posibles mantenimientos deficientes del equipamiento de las redes de monitorización suponen a aquel un inconveniente añadido. Para abodar estos aspectos, se ha recurrido a una combinación de métodos estadísticos para la optimización de los recursos empleados en la monitorización, introduciendo nuevos criterios para su mejora. Datos de monitorización de contaminantes clave como el monóxido de carbono (CO), dióxido de nitrógeno (NO_2), ozono (O_3), material particulado (PM_{10}) y dióxido de azufre (SO_2) fueron obtenidos de 12 estaciones de monitorización de la calidad del aire en Sevilla (España). Un total de 49 conjuntos de datos se modelizaron mediante mixturas finitas gaussianas utilizando el algoritmo de esperanza-maximización (EM). Para resumir estos 49 modelos, se calculó la media (μ_m) y coeficiente de variación (cv_m) de cada mixtura, y a partir de ellos, se realizó un análisis clúster jerárquico (ACJ) para estudiar el agrupamiento de las estaciones de acuerdo con estos estadísticos. El valor de los parámetros no monitorizados en las estaciones de medición fue imputado aplicando un algoritmo basado en bosques aleatorios, utilizando los valores de μ_m y cv_m conocidos. Posteriormente, el análisis de componentes principales (ACP) permitió comprender la relación intrínseca entre las estaciones de la red, así como la concordancia en su clasificación. Todas las técnicas se aplicaron utilizando el software estadístico gratuito y de código abierto R. Se ha analizado un ejemplo de atribución y contribución de fuentes utilizando la modelización mediante mixturas finitas, y el potencial de estos modelos se propone para caracterizar tendencias de contaminación. Los estadísticos de la mixturas μ_m y cv_m representan su huella dactilar, y su empleo es nuevo en la caracterización de los modelos mixtos en el área de la gestión de la calidad del aire. La técnica de imputación empleada ha permitido la estimación de valores de concentración de parámetros no monitorizados y el planteamiento de nuevos esquemas de monitorización para esta red. El empleo posterior del ACP ha confirmado una clasificación errónea de una estación detectada inicialmente mediante el ACJ.

El contenido de este capítulo es una adaptación de **Gómez-Losada, Á., Lozano-García, A., Pino-Mejías, R., Contreras-González, J.** 2014. Finite mixture models to characterize and refine air quality monitoring networks. *Science of the Total Environment*, 485-486: 292-9. Factor de impacto de 4.099 (2014) y posicionamiento 18/223 en la categoría "Environmental Sciences".

5.1. Introducción

El diseño de una red de monitorización de la calidad del aire conlleva básicamente determinar el número de estaciones y su emplazamiento, clase y número de contaminantes a monitorizar, sin dejar de considerar sus objetivos, costes y recursos disponibles. En la mayoría de los casos, las redes de monitorización en áreas metropolitanas se diseñan para medir contaminantes de importancia sanitaria,

como el CO, NO_2, O_3, PM_{10} y SO_2 (Chang y Tseng, 1999). Si no se reevalúa la representatividad de estos contaminantes para la detección de nuevas fuentes o niveles de contaminación, pueden llegar a no satisfacer la demanda informativa que la sociedad requiere al respecto. En ocasiones, algunos de estos contaminantes se monitorizan en estaciones vecinas, lo que conlleva una duplicidad de la información obtenida o la detección de similares niveles de contaminación (la redundancia en el equipamiento fue introducida de forma analítica por Pires et al., 2008). Además, un problema inherente en los equipamientos de monitorización es que están sometidos a rigurosos programas de mantenimiento que, en caso de no cumplirse, conducen a que las estaciones no operen a un nivel satisfactorio.

El propósito de este capítulo es realizar una aproximación integrada para la optimización de la información suministrada por las redes de monitorización de calidad del aire y obtener criterios para su mejora. Esta mejora consiste esencialmente en replantear la monitorización de parámetros contaminantes en estaciones de inmisión, detectar posibles duplicidades, reclasificar los tipos de estaciones y, finalmente, ayudar al gestor de redes a efectuar consecuentes y progresivas modificaciones. Incluso, si se considerara que estas mejoras no son necesarias, la información obtenida mediante estos métodos estadísticos sigue constituyendo una valiosa fuente para la caracterización y conocimiento de las redes de monitorización.

Se ha realizado la aproximación objeto de este trabajo mediante funciones específicamente diseñadas para la ocasión (algoritmo EM) o ya implementadas (ACJ, BA, ACP) a través del software libre R (R Core Team, 2015). La aplicación metodológica es secuencial y responde a los siguientes objetivos: 1) obtener mediante cada modelo de mixtura asociado a cada contaminante (Tabla 5.1) información de su nivel de impacto (μ_m) y variabilidad (σ_m) en las áreas de influencia de cada estación, además de evaluar la atribución de fuentes de contaminación; 2) identificar el agrupamiento de las estaciones de inmisión en base a μ_m y cv_m empleando el ACJ; 3) estimar mediante la técnica de imputación basada en BA (Breiman, 2001) los valores de μ_m y cv_m para contaminantes no monitorizados; y 4) estudiar la reclasificación de las estaciones a través del ACP, añadiendo la nueva información obtenida mediante la imputación a la ya conocida por modelización mediante mixturas.

El procedimiento descrito ha sido aplicado en la red de monitorización de calidad del aire de Sevilla, tanto por el mayor número de estaciones que la componen como por su centralización. La provincia de Sevilla se localiza en el sur de España y cubre una superficie de 14 036 km^2, y durante el año de estudio tenía una población total de 1 935 364 habitantes (IECA, 2012). El área metroplitana de Sevilla es la principal aglomeración urbana de la región de Andalucía y tenía, para el mismo periodo, una población de 1 217 811 habitantes (SG, 2012).

5.2. Datos y métodos

5.2.1. Estaciones de monitorización

La red de monitorización de calidad del aire de Andalucía incluye 89 estaciones de inmisión, siendo monitorizados simultánemante los contaminantes CO, NO_2, O_3, PM_{10} y SO_2 en, aproximadamente, unas 43 estaciones fijas de medida. Esta red es gestionada por la Agencia de Medio Ambiente y Agua de Andalucía, agencia pública empresarial adscrita a la administración ambiental de la Junta de Andalucía (en la fecha de redacción de esta memoria, denominada como Consejería de Medio Ambiente y Ordenación del Territorio).

El presente estudio se concentró en el análisis de los contaminantes arriba citados durante el año 2012. Los datos observacionales provienen de 12 estaciones de monitorización, 10 de ellas situadas en el área metropolitana de Sevilla y 2 en la provincia, en un entorno rural, por lo que se dispuso de un diverso rango de concentraciones, localizaciones y ejemplos de atribuciones de fuentes.

Estas estaciones se han clasificado de acuerdo con el tipo de área donde se localizan (R-Rural, S-Suburbana, U-Urbana) y su fuente de emisión predominante (F-Fondo, I-Industrial, T-Tráfico). (Las

principales características de las estaciones y contaminantes analizados en cada una de ellas se indican en la Tabla 5.1).

Dado que los datos monitorizados se obtienen a intervalos de tiempo de longitud diez minutos, dichas concentraciones fueron promediadas para obtener un valor único diario, asegurando de esta forma la independencia estadística de las observaciones. Estos valores medios se han calculado únicamente cuando se ha dispuesto de, al menos, el 80 % de todas las observaciones horarias durante el día (19 de 24 horas). A lo largo de este trabajo, todas las unidades de concentración se representan en $\mu g/m^3$. Los métodos de monitorización de referencia establecidos en la Directiva Europea 2008/50/EC fueron utilizados para los contaminantes CO, NO_2, O_3 y SO_2, el método de monitorización β-atenuación para PM_{10}. En este estudio, el factor de corrección a los datos de PM_{10} no fue aplicado ni se tomó en consideración la contribución natural de partículas; los valores obtenidos mediante las estaciones de monitorización se utilizaron directamente.

La red de monitorización estudiada se encuentra sujeta a un exhaustivo programa de mantenimiento que asegura la obtención de valores correctos. Dichos valores son validados por la administración ambiental autonómica de forma previa a cualquier uso.

Estación y abreviatura	Clasificación	Emplazamiento		Contaminantes estudiados				
Alcalá de Guadaíra (Alc)	U-F	248974	4136631	CO	NO_2	O_3	PM_{10}	SO_2
Aljarafe (Alj)	S-F	230473	4137017	·	NO_2	O_3	PM_{10}	SO_2
Bermejales (Ber)	U-F	236063	4137554	CO	NO_2	O_3	PM_{10}	SO_2
Centro (Cen)	U-F	235156	4142125	CO	NO_2	O_3	·	SO_2
Cobre Las Cruces (Cob)	R-I	231798	4160779	CO	NO_2	O_3	PM_{10}	SO_2
Dos Hermanas (Dos)	U-F	241677	4130413	CO	NO_2	O_3	·	SO_2
Príncipes (Pri)	U-F	233863	4140741	CO	NO_2	·	PM_{10}	SO_2
Ranilla (Ran)	U-T	237965	4141611	CO	NO_2	·	·	SO_2
San Jerónimo (Saj)	S-I	236286	4146731	·	NO_2	O_3	·	·
Santa Clara (Sac)	S-F	238720	4143149	CO	NO_2	O_3	PM_{10}	·
Sierra Norte (Sie)	R-F	265817	4208544	·	NO_2	O_3	PM_{10}	SO_2
Torneo (Tor)	U-T	234151	4142873	CO	NO_2	O_3	PM_{10}	SO_2

Tabla 5.1 Contaminantes analizados y clasificación de las estaciones de monitorización de donde los datos fueron obtenidos (los emplazamientos son expresados en coordenadas X,Y ETRS89-UTM, zona 30). Los puntos representan contaminantes no monitorizados (11 casos). R-Rural, S-Suburbana, U-Urbana; F-Fondo, I-Industrial, T-Tráfico.

5.2.2. Estimación de los modelos

La implementación computacional de los MMF se realizó mediante funciones en R (Anexo A), utilizando como criterio de parada la diferencia relativa (2.30) y estableciendo el valor de $\epsilon = 1 \cdot 10^{-6}$. Se emplearon 11 modelos ($K = 1, \ldots, 11$), seleccionándose el "mejor" de ellos en función del criterio BIC. A efectos de validación de los resultados obtenidos con estas funciones, se empleó la librería mclust, sin encontrarse diferencias significativas entre ambos resultados. El resto de técnicas auxiliares (ACJ, BA y ACP) empleadas se explicaron en el Capítulo 4, acompañándose su implementación en la Sección C.2.

5.3. Resultados y discusión

Los modelos de distribuciones mixtas resultantes de la aplicación del algoritmo EM sobre cada conjunto de datos se muestran en el Anexo C, indicando, además de la estimación de los parámetros y sus errores asociados, el tamaño de cada clúster (componente).

5.3.1. Análisis descriptivo y atribución de fuentes

En este apartado se explica una modelización mediante mixturas finitas entre todas las obtenidas (Anexo C) y su atribución de fuentes asociada. En este caso se ha elegido el modelo de PM_{10} en "Alc"

(Alcalá de Guadaíra, Tabla C.1), representándose gráficamente en la Figura 5.1.

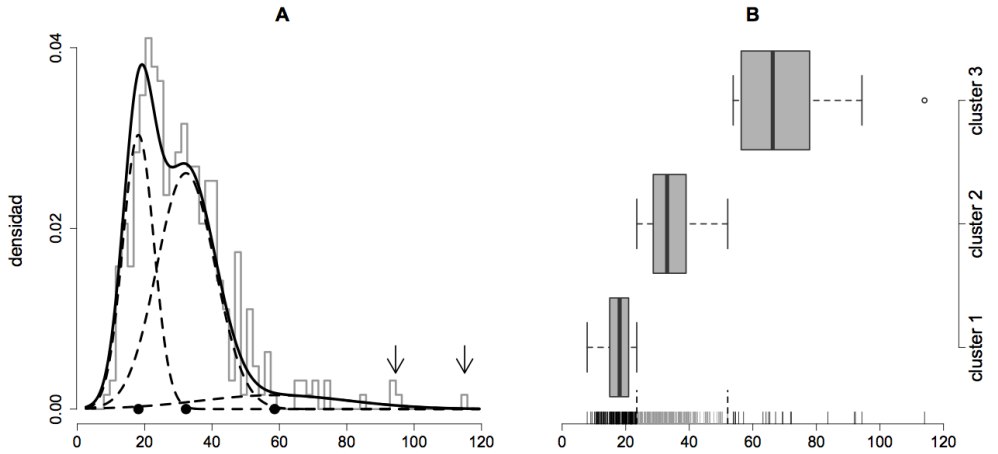


Figura 5.1 A. La densidad estimada de la mixtura (línea continua) y de las componentes (línea discontinua) se muestran sobreimpuestas al histograma de los datos observados (línea gris). Los puntos negros sobre el eje horizontal indican la media de cada componente y las flechas señalan las cuatro observaciones de mayor concentración. **B**. Diagramas de caja y bigote de cada clúster. Las observaciones que pertenecen a cada clúster se muestran en la base del gráfico en escala de grises y se muestran separadas unas de otras mediante una pequeña línea discontinua.

La primera componente (izquierda en la Figura 5.1A) está relacionada con la contaminación de fondo de PM_{10} en el área de "Alc", y la concentración mínima obtenida fue de 7.88 $\mu g/m^3$. La segunda componente se extiende desde el rango de concentraciones de 23.50 a 52.10 $\mu g/m^3$ conteniendo el mayor número de observaciones (n_2 = 193). Esta segunda componente está relacionada con emisiones antropogénicas de PM_{10} de tipo primario, generadas por el tráfico viario y fundamentalmente por procesos industriales, representados por una industria de cemento presente en la localidad, cuya capacidad de producción anual es 775 641 t (E-PRTR, 2011), así como por pequeñas y dispersas canteras dedicadas a la extracción de albero. Esta presencia industrial podría llevar a pensar que esta segunda componente debería haber registrado un mayor valor que contribuyera a un aumento de la media del modelo mixto (μ_m). Sin embargo, ha de considerarse la moderada actividad industrial de la zona debido a la actual crisis económica, lo que también se refleja en un descenso del tráfico rodado. Desafortunadamente, no existen para el año de estudio datos relativos al consumo de energía eléctrica en esta localidad, si bien su consumo eléctrico industrial descendió en el periodo 2007 a 2010, de 920 127 MWh (DP, 2009) a 766 200 MWh (DP, 2012). La tercera componente en la mixtura (derecha) refleja las concentraciones más altas de PM_{10} y está relacionada con fenómenos meteorológicos especiales, como las intrusiones de polvo sahariano, las cuales invaden la región de Andalucía con frecuencia. Las flechas en la Figura 5.1A indican niveles de concentración de partículas de 91.96, 114.02, 92.27 y 94.32 $\mu g/m^3$, que fueron registrados, respectivamente, el 27 y 28 de junio y el 10 y 22 de agosto. Todas estas fechas coinciden con intrusiones de polvo sahariano relevantes durante el verano y, en particular la última concentración señalada, con agresivos incendios en los países mediterráneos, incluyendo España (CIQSO, 2012). La concentración de 114.02 $\mu g/m^3$ se clasifica como un dato anómalo en la Figura 5.1B (outlier), aunque se explica por estos fenómenos.

La comparación de las diferentes componentes de las mixturas a través de diferentes años permite caracterizar la tendencia de la contaminación y de sus componentes en una escala de tiempo. Además, un mismo contaminante puede estudiarse a través del tiempo en una localidad, o bien ese mismo contaminante puede estudiarse para un mismo año en diferentes localidades, añadiendo una componente espacial al estudio.

5.3.2. Obtención de los momentos de las mixturas

Los momentos de las distribuciones de mixturas gaussianas pueden obtenerse fácilmente y representan de forma resumida la parametrización de estos modelos (Anexo C.1).

Estación	Contaminante	n	K	μ_m	σ_m
Alc	CO	361	3	308.57	51.00
	NO_2	360	2	21.06	10.22
	O_3	356	2	61.12	20.92
	PM_{10}	358	3	28.84	13.92
	SO_2	360	2	4.82	1.65
Alj	NO_2	364	3	18.44	9.39
	O_3	366	2	62.86	21.50
	PM_{10}	361	3	30.53	15.22
	SO_2	360	3	6.78	2.92
Ber	CO	365	2	480.55	148.03
	NO_2	342	3	21.64	14.37
	O_3	365	2	51.91	19.71
	PM_{10}	341	2	33.55	16.01
	SO_2	326	3	5.02	1.89
Cen	CO	340	2	677.75	276.24
	NO_2	345	2	21.46	9.03
	O_3	348	2	53.54	21.44
	SO_2	348	2	2.80	1.08
Cob	CO	302	2	206.70	74.56
	NO_2	236	3	6.54	4.34
	O_3	351	2	55.91	16.92
	PM_{10}	311	3	17.04	12.29
	SO_2	311	2	2.76	1.82
Dos	CO	320	2	463.84	123.08
	NO_2	348	2	19.44	7.17
	O_3	349	2	57.39	21.56
	SO_2	344	1	5.79	1.12
Pri	CO	342	2	412.38	156.77
	NO_2	316	2	29.39	12.40
	PM_{10}	355	2	29.82	13.88
	SO_2	351	2	5.36	1.66
Ran	CO	361	3	221.04	141.94
	NO_2	359	2	36.15	13.09
	SO_2	358	3	5.49	1.52
Saj	NO_2	358	2	22.79	7.94
	O_3	362	2	53.57	21.05
Sac	CO	358	2	367.10	86.60
	NO_2	351	2	20.87	11.04
	O_3	351	2	47.94	21.67
	PM_{10}	357	2	24.74	13.19
Sie	NO_2	349	2	3.78	1.99
	O_3	349	2	61.42	17.42
	PM_{10}	338	3	19.75	14.87
	SO_2	347	3	3.37	1.38
Tor	CO	365	2	465.83	181.32
	NO_2	365	1	33.63	15.37
	O_3	363	2	39.25	17.72
	PM_{10}	313	3	29.72	12.40
	SO_2	365	2	3.67	0.92

Tabla 5.2 Momentos de los 49 conjuntos de datos analizados mediante modelos mixtos, indicando el número de observaciones de cada uno (n) y las componentes detectadas (K), según el criterio de información BIC.

Con un propósito descriptivo, el valor de σ_m siempre debe entenderse en relación con el de μ_m, y el coeficiente de variación, cv_m, representa un estadístico adimensional y una medida normalizada de la variabilidad de la mixtura. Como sucede con los momentos, el uso de este último estadístico es desconocido en la literatura ambiental. El conjunto de valores μ_m y σ_m de los 49 modelos obtenidos se muestra en la Tabla 5.2. El valor del coeficiente de variación (cv_m) se refleja en la Tabla 5.4 junto con los resultados del segundo proceso de imputación.

Por tanto, μ_m y cv_m de cada modelo mixto nos permite obtener el nivel cuantitativo y la variabilidad de cada contaminante, respectivamente, y su significado cobra un mayor sentido cuando se analiza conjuntamente con la clasificación de las estaciones de monitorización, según su área y fuente de contaminación principal de donde provienen las observaciones (Tabla 5.1). En este estudio, la obtención de estos valores representa un punto de partida útil para la aplicación de posteriores análisis estadísticos con el fin de caracterizar la red de inmisión en estudio. Estos análisis se desarrollan en los siguientes apartados.

5.3.3. ACJ de las estaciones previo a la imputación

La información hasta ahora obtenida sugiere estudiar la relación de similitud entre las estaciones de monitorización con respecto a los valores μ_m y cv_m. Los resultados del ACJ se muestran en la Figura 5.2.

La Figura 5.2A muestra cómo las estaciones son agrupadas de acuerdo con el nivel cuantitativo de contaminación (μ_m) presente en sus áreas de representatividad. Dos grupos de estaciones son claramente segregadas del resto: las urbanas de tráfico "Ran" y "Tor", y las rurales "Cob" y "Sie". Las estaciones de fondo, como "Dos", "Alc" y "Alj", exhiben cercanía en la clasificación, y el resto de estaciones se sitúan en el dendrograma de acuerdo con sus características de contaminación. Se detecta una gran similitud entre dos estaciones "U-B" ("Ber" y "Cen") y una estación "S-I" ("Saj"). Esta última fue clasificada como tal por su cercanía a una pequeña industria de coches, localizada en el interior de la ciudad, donde no se realizan labores de fabricación propiamente dichas, sino el montaje de piezas inertes no manufacturadas *in situ*. Por tanto, en el dendograma, teniendo en cuenta la verdadera naturaleza

urbana de "Saj", y de foma más lógica, esta se sitúa próxima a estaciones clasificadas como "U-F", por lo que se plantea una posible clasificación errónea de esta estación.

La Figura 5.2B muestra el agrupamiento de las estaciones en función de los valores de cv_m. Inicialmente, "Cob" es situada en una hoja aparte, ya que esta estación se localiza en un área industrial representada por una de las minas de cobre más grandes de Europa. Como en la Figura 5.2A, "Sie" muestra la mayor similitud con "Cob"; y "Alc", con "Alj" (ver los valores similares de μ_m and σ_m en la Tabla 5.2). El resto de las localizaciones de las estaciones en los dendrogramas no se comentan en detalle, ya que responden a los valores que presentan para los estadísticos μ_m y cv_m.

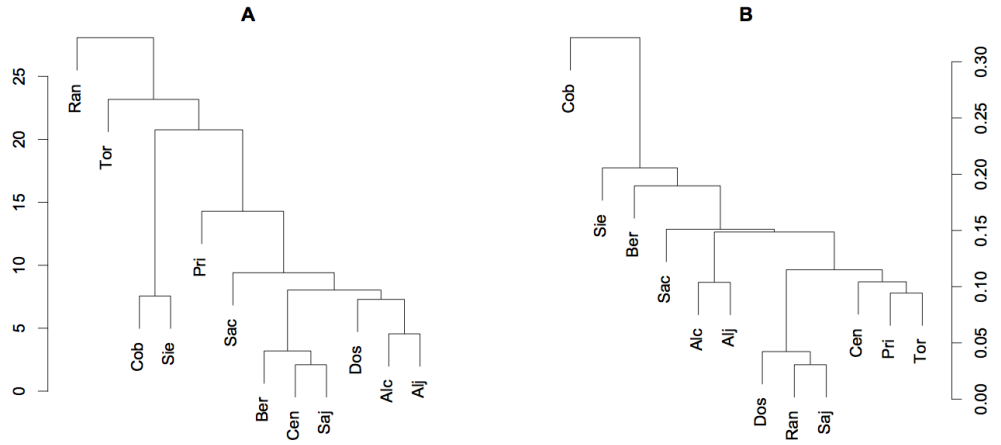


Figura 5.2 Los dendrogramas muestran la jerarquía de las estaciones en función de sus valores de μ_m (**A**) y cv_m (**B**) de acuerdo con la información de la Tabla 5.2. Los ejes verticales indican la distancia entre las estaciones, representando su similitud (baja distancia) o diferencia.

5.3.4. Imputación de los estadísticos μ_m y cv_m

Como se mencionó anteriormente, la falta de monitorización de contaminantes en las estaciones de inmisión representa una ausencia de información al respecto. Pollice (2009) utilizó la imputación para solventar el problema de falta de datos por un funcionamiento deficiente de sensores de PM_{10}, y Vaidyanathan et al. (2013), para estimar concentraciones de $PM_{2.5}$ en estaciones y días sin datos monitorizados. Para ampliar esta estimación más allá del material particulado, se procedió a imputar los valores ausentes de μ_m y cv_m de aquellos contaminantes sin datos observacionales, utilizando la información del resto de contaminantes monitorizados que habían sido modelizados con estos estadísticos (Tabla 5.2). El uso de la imputación en este contexto permite optimizar el diseño de las redes de monitorización, ya que este procedimiento es equivalente al establecimiento de sensores virtuales en estaciones que no disponen de unos reales. Las Tablas 5.3 y 5.4 muestran los resultados del primer y segundo proceso de imputación (μ_m y cv_m).

Al analizar los valores de la Tabla 5.3, puede detectarse alguna duplicidad en la información. Por ejemplo, "Ber" y "Cen" (estaciones U-B) monitorizan ambas NO_2, obteniendo valores casi iguales de este contaminante. Sin embargo, PM_{10} no es medido en Cen. Como una propuesta de mejora, "Cen" podría monitorizar PM_{10}, en lugar de NO_2. Si esta opción no es considerada, monitorizar NO_2 en "Cen"/"Ber" es también posible, ya que una información similar es obtenida de "Ber"/"Cen". "Ran" y "Tor" (estaciones U-T) registran similares niveles de PM_{10} (el valor en "Ran" es imputado). Por tanto, el gestor de red podría intercambiar la monitorización de estos contaminantes entre estaciones si similares niveles de PM_{10} son de nuevo detectados después de un periodo de estudio. Otra duplicidad puede encontrarse en "Dos" y "Pri" (estaciones U-B) con respecto a SO_2. Como se deduce de estos análisis, pueden obtenerse otras posibilidades de configuración de la red de inmisión, además de las

Estaciones	Contaminantes				
	CO	NO_2	O_3	PM_{10}	SO_2
Alc	308.57	21.06	61.12	28.84	4.82
Alj	**394.12**	18.44	62.86	30.55	6.78
Ber	480.55	21.64	51.91	33.53	5.02
Cen	677.75	21.46	53.54	**27.93**	2.80
Cob	206.70	6.54	55.91	17.04	2.76
Dos	463.84	19.44	57.39	**27.69**	5.79
Pri	412.38	29.39	**54.68**	29.82	5.36
Ran	221.04	36.15	**54.83**	**28.16**	5.49
Saj	**420.50**	22.79	53.57	**28.85**	**4.87**
Sac	367.10	20.87	47.94	24.74	**4.56**
Sie	**328.00**	3.78	61.42	19.75	3.37
Tor	465.83	33.63	39.25	29.72	3.67

Tabla 5.3 Valores imputados (en $\mu g/Nm^3$) de μ_m de contaminantes no monitorizados en las estaciones (en negrilla). Los valores no imputados coinciden con aquellos de la Tabla 5.2 y son mostrados de nuevo con propósito comparativo.

Estaciones	Contaminantes				
	CO	NO_2	O_3	PM_{10}	SO_2
Alc	0.17	0.49	0.34	0.48	0.34
Alj	**0.33**	0.51	0.34	0.50	0.43
Ber	0.31	0.66	0.38	0.48	0.38
Cen	0.41	0.42	0.40	**0.52**	0.39
Cob	0.36	0.66	0.30	0.72	0.66
Dos	0.27	0.37	0.38	**0.52**	0.19
Pri	0.38	0.42	**0.39**	0.47	0.31
Ran	0.64	0.36	**0.40**	**0.47**	0.28
Saj	**0.37**	0.35	0.39	**0.48**	**0.32**
Sac	0.24	0.53	0.45	0.53	**0.33**
Sie	**0.31**	0.53	0.28	0.75	0.41
Tor	0.39	0.46	0.45	0.42	0.25

Tabla 5.4 Valores calculados e imputados (en negrilla) de cv_m de los contaminantes en las diferentes estaciones.

reseñadas, basadas en el conocimiento que de esta posea su gestor y de las necesidades de explotación de la misma.

Otra aplicación de la imputación es la de proporcionar una estimación cuando el número mínimo de observaciones requerido durante un año para un contaminante sea inferior al deseado. En el caso de NO_2-Cob, con n = 236, el valor modelizado (6.54, tabla 5.2) podría ser eliminado de la matriz de imputación y realizar de nuevo el proceso de imputación para obtener una estimación de su media anual.

El proceso de imputacion explicado en este apartado considera dos ratios de información. El primero es la proporción de valores observados a no observados (49/11), y el segundo, el número de estaciones estudiadas y su clasificación por tipo (12 estaciones/3 - Urbanas, Suburbanas y Rurales). En líneas generales, no se recomienda reducir ninguna de estas proporciones (aspecto tenido en cuenta en la propuesta de monitorización de Cen-PM_{10}) cuando se estudie una reconfiguración en la red de inmisión por parte de la administración ambiental. Los resultados de la Tabla 5.3 se pueden analizar conjuntamente con los de la Tabla 5.4, aunque la información de esta última no debería ser determinante para adoptar ninguna decisión, sino para complementar la información de la Tabla 5.3.

Una vez obtenida la información completa de los niveles de contaminación tras la imputación, el estudio sugiere aplicar un ACP para mejorar la comprensión de la red en cuanto a la agrupación de las estaciones en función de los niveles y variabilidad de los contaminantes.

5.3.5. ACP

Para proporcionar una diferente aproximación del ACJ, se realizaron dos ACP para estudiar el agrupamiento de las estaciones en función de sus valores μ_m y cv_m, basándonos en la información mostrada en las Tablas 5.3 y 5.4. Debido a las características de este estudio, en ambos análisis fue necesario

introducir una tercera componente para explicar una cantidad aceptable de varianza, ya que inicialmente solo se obtuvo el 77,0 % y 79,0 %, respectivamente. Los valores de μ_m fueron normalizados, dada la diferente magnitud de las concentraciones en los diferentes parámetros estudiados.

Mediante un análisis con tres componentes, la varianza explicada fue del 94,6 % y 94,3 %, considerándose valores óptimos para el propósito perseguido. Las coordenadas de cada punto (estaciones) en un sistema de tres dimensiones se muestran en el Anexo C (Tablas C.13 y C.14), mientras que, en esta sección, el gráfico que relaciona las componentes principales (CCPP) CP2 a CP3 se omite por no resultar decisivo para demostrar el agrupamiento de las estaciones.

La Figura 5.3 muestra los resultados del ACP con respecto a μ_m. Como puede verse en la Tabla C.15, las cargas más altas se encuentran en las variables NO_2 y PM_{10} para CP1 (signo negativo), O_3 y SO_2 para CP2 (signo negativo), y CO para CP3 (signo positivo). El ACP sitúa a "Cob" (RI) , "Sie" (RB) y "Tor" (UT) bien separadas del resto, y en menor alcance "Ran"(UB) y "Cen"(UB). El resto de estaciones situadas dentro del elipsoide permanecen más o menos distantes del centro de esta figura dependiendo de sus características. La interpretación que a continuación se incluye precisa que el gestor de la red posea un conocimiento en detalle tanto de la configuración de la red de inmisión como de las características de sus estaciones.

En la Figura 5.3, los valores más altos de PM_{10} y NO_2 aproximan las estaciones hacia la izquierda del eje CP1, mientras que los valores superiores de SO_2 y O_3 las sitúan en la parte inferior de CP2. Como resultado, las estaciones de tráfico "Tor" y "Ran" caen a la izquierda de CP1 y las estaciones U-B, "Ber" y "Pri", entre ellas. Esta últimas estaciones poseen una gran influencia de tráfico debido a su localización próxima a vías principales y, por tanto, el ACP las sitúa a la izquierda (mayores niveles de PM_{10} y NO_2) de "Ran". Las estaciones rurales ("Sie" y "Cob") se localizan en la parte derecha de CP1, ya que registran los valores más bajos de PM_{10} y NO_2.

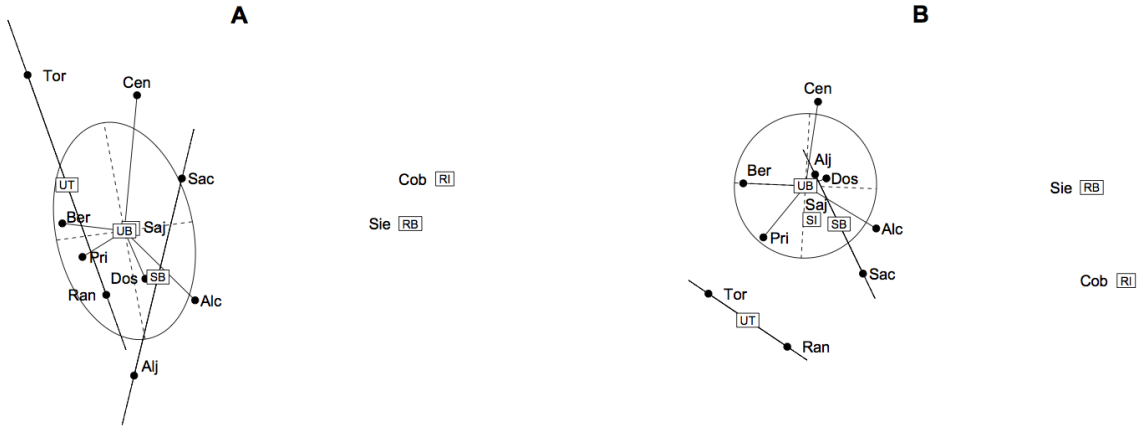


Figura 5.3 PCA basado en valores normalizados de μ_m. CP1 se representa en el eje horizontal. Las etiquetas dentro de un recuadro conectan estaciones con la misma clasificación. **A**. Primera y segunda componente (CP2 en el eje vertical). **B**. Primera y tercera componente (CP3 en el eje vertical).

Las estaciones localizadas en el centro urbano de Sevilla ("Cen" y "Sac") registran un valor más bajo de O_3 y se localizan en la parte superior del CP2. A medida que nos acercamos hacia el área suburbana de la ciudad, la concentración de ese contaminante aumenta, lo que se refleja en las estaciones suburbanas ("Alj", "Alc" y "Dos"), colocándolas más abajo en el eje CP2. Teniendo en cuenta que Sevilla no es una ciudad altamente industrializada, la contribución de SO_2 es inferior en comparación con otros parámetros analizados. En la Figura 5.4, las estaciones se localizan en la parte superior del eje CP2, donde se reflejan las concentraciones superiores de CO, y, por tanto, las estaciones en el centro urbano aparecen en esa posición.

El PCA con respecto a cv_m (Figura 5.4) describe el agrupamiento de las estaciones respecto a la variabilidad de las concentraciones de los contaminantes monitorizados basados en modelos mixtos, aunque la interpretación de estas figuras es menos intuitiva. Las cargas más altas en los CCPP destacan en SO_2 para CP1 (signo negativo), CO para CP2 (signo positivo), y NO_2 y PM_{10} para CP3 (signos positivo y negativo, respectivamente). Por el contrario, el O_3 no se relaciona fácilmente con ningún CP (Tabla C.4). Como puede percibirse, las estaciones rurales ("Sie" y "Cob") exhiben una variabilidad en la concentración de contaminantes diferente de las estaciones urbanas, y, entre estas, pueden distinguirse claramente las estaciones de fondo (dentro del elipsoide) y de tráfico ("Tor" y "Ran"). En la Figura 5.4A, las estaciones se encuentran posicionadas sobre una fracción más pequeña de variación para el SO_2. Las estaciones localizadas en el centro urbano muestran una pequeña variación de este contaminante, ya que, como se ha mencionado, Sevilla no es una ciudad industrial. La mayor variación de CO coloca a las estaciones en la parte superior del eje CP2. La estación de tráfico "Ran" presenta la mayor variación en sus niveles de contaminantes. Esta interpretación cualitativa completa aquella cuantitativa basada en el ACP considerando μ_m.

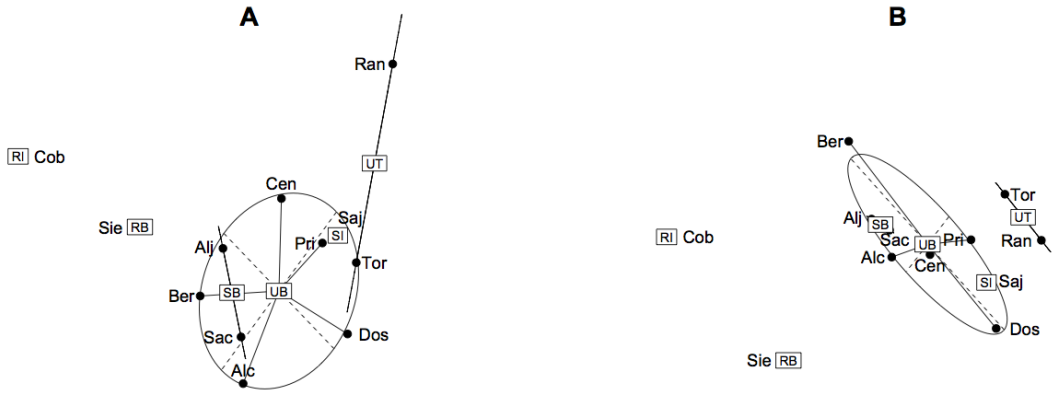


Figura 5.4 ACP basado en valores no normalizados de cv_m con los ejes y las distribución de las CCPP como en la Figura 5.3.

No obstante, los resultados del ACP deben interpretarse analizando conjuntamente los resultados del ACJ realizado previamente a la imputación de los estadísticos para confirmar los agrupamientos de las estaciones. Como era de esperar, en la Figura 5.2A y 5.3, "Ran", "Tor", "Cob" y "Sie" son establecidos aparte antes y después de la imputación, y también "Saj" permanece entre estaciones del tipo U-B y S-B, confirmando los resultados del ACJ. Las Figuras 5.2B y 5.4 también mantienen estos emplazamientos. La información añadida después de la imputación puede recolocar estaciones en posiciones diferentes cuando se utiliza cualquier técnica de agrupamiento, y esto puede variar los resultados del ACJ inicial. El conocimiento de la red y la experiencia del gestor son cruciales para dar sentido a estas recolocaciones. Los resultados de los ACJ y ACP basados en μ_m y cv_m, como se esperaba, revelan que, a pesar de las condiciones cambiantes locales atmosféricas y ambientales en las estaciones durante el periodo de estudio, pueden asumirse unos niveles de contaminación y variabilidad en concordancia con sus emplazamientos y predominantes fuentes de emisión.

5.4. Conclusiones

Se han empleado MMF para ajustar distribuciones de datos observacionales de 5 contaminantes clave monitorizados, según disponibilidad, en 12 estaciones de inmisión durante 2012 en Sevilla. Los 49 modelos resultantes caracterizaron de forma precisa los contaminantes monitorizados y constituyen una

herramienta útil para el gestor, por su capacidad para detectar fuentes de contaminación y comparar las tendencias de estas a lo largo del tiempo y el espacio. Pero los modelos de mixturas permiten ser utilizados no solo con un propósito descriptivo. La media y desviación típica de cada modelo (μ_m y σ_m) se pueden obtener fácilmente para su caracterización, y, derivado de ellos, el coeficiente de variación, representando una huella dactilar de los mismos. Su aplicación en este trabajo supone una aproximación que abre nuevas vías para la caracterización de los modelos mixtos en la disciplina de la contaminación atmosférica. Basado en ellos, el ACJ detectó una clara separación entre las estaciones de acuerdo con su nivel de contaminación (U-T) o emplazamiento (estaciones rurales) del resto.

Cuando el ratio de información perdida es apropiado, la imputación de μ_m y cv_m representa una estrategia valiosa para estimar el nivel y variabilidad de los contaminantes no monitorizados en las estaciones, o cuando la ausencia de datos entre los monitorizados es destacable. Esta imputación permite optimizar el diseño de la red en estudio. Una vez que se obtienen los valores imputados, pueden emplearse análisis estadísticos posteriores para completar la información de la red obtenida hasta el momento, y también para plantear alternativas de configuración en la red, como fue el caso de "Cen" y "Ber" con respecto a la monitorización de NO_2 y PM_{10}. Atendiendo a todo lo anterior, es esencial que el gestor disponga de un conocimiento exhaustivo de la red para poder interpretar correctamente los resultados obtenidos mediante estos análisis estadísticos, con el fin de poder adoptar medidas de mejora si fuera necesario. Este podría ser el caso de la posible clasificación incorrecta de "Saj".

Aunque una de las principales ventajas de la aplicación secuencial de estos procedimientos estadísticos es la de conocer en profundidad la red de monitorización en operación, su aplicación a lo largo de varios años puede resultar incluso más valiosa. Esta orientación puede ayudar a entender cómo la contaminación ha evolucionado durante un periodo de tiempo concreto y cómo la red se ha comportado en consonancia.

6

Análisis de contribución de fuentes mediante modelos ocultos de Markov

Resumen De forma previa a la realización de los análisis de contribución de fuentes de emisión (ACF), se llevan a cabo análisis estadísticos sencillos con los que caracterizar, de forma inicial, las fuentes de emisión en estudio. A su vez, técnicas como el modelado de receptores se basan en la especiación química que precisa de ensayos prolongados y costosos. En este capítulo, los modelos ocultos de Markov (MOM) se proponen como una herramienta exploratoria y rutinaria para su empleo en los ACF de PM_{10}. Estos modelos fueron empleados con series temporales (SSTT) anuales obtenidas de 33 emplazamientos de fondo rural localizados en la Península Ibérica y en los Archipiélagos de las Azores, Balear y Canarias. Los MOM permiten la creación de grupos de observaciones de las SSTT de PM_{10} con similares valores de concentración, definiendo diferentes regímenes de este contaminante. Los resultados de esta modelización sobre las SSTT incluyen la estimación de las contribuciones de estos regímenes, la probabilidad de cambio entre ellos y sus contribuciones a la concentración media anual de PM_{10}. La contribución anual media de PM_{10} procedente de regiones áridas norteafricanas fue estimada y comparada con la estimación de otros estudios. Además, se propone un procedimiento nuevo para cuantificar la contribución natural del desierto a las concentraciones medias diarias de PM_{10} medidas en estaciones de monitorización de fondo. Este nuevo procedimiento parece corregir las estimación de la carga neta de los desiertos obtenida con el método más empleado en la actualidad, y de referencia en la Unión Europea, basado en el percentil 40 medio móvil mensual.

El contenido de este capítulo es una adaptación de **Gómez-Losada, Á., Pires, J.C.M., Pino-Mejías, R.** 2015. Time series clustering for estimating particulate matter contributions and its use in quantifying impacts from deserts. *Atmospheric Environment*, 117: 271-81. Factor de impacto de 3.281 (2014) y posicionamientos 42/223 en la categoría "Environmental Sciences" y 15/77 en "Meteorology & Atmospheric Sciences".

6.1. Introducción

Numerosos estudios relacionados con los aerosoles atmosféricos basan sus conclusiones en el desarrollo de análisis de contribución de fuentes (ACF). Un ACF consiste en estudiar el origen de las diversas fuentes de contaminación y la estimación de sus contribuciones a los niveles de la contaminación atmosférica ambiente. Esta información es crucial para el desarrollo e implementación de políticas para la protección de la salud humana y el medio ambiente, así como para el diseño de estrategias de reducción de la contaminación efectivas, en una escala local o más amplia, donde los umbrales legislados se superen. Existen tres grupos de técnicas de ACF (Viana et al., 2008): 1) métodos que implican la valoración de los datos de monitorización, 2) métodos que dependen de inventarios de emisiones y/o modelos de dispersión atmosférica, y 3) métodos que se basan en la evaluación estadística de datos químicos de material particulado (MP) a partir de la información obtenida de estaciones de monitorización receptoras (modelos receptores o MR). El primer grupo se apoya en tratamientos básicos de datos (Belis et al., 2013). Incluye también el modelado mediante series temporales (SSTT) de datos

que pueden ser utilizados, por ejemplo, para estimar las contribuciones naturales de PM_{10} (Escudero et al., 2007a). El segundo grupo incluye modelos para simular la formación de emisiones de aerosoles, transporte y deposición, aunque se encuentran limitados por la exactitud de los inventarios de emisiones, cuando estos existen. El tercer grupo se emplea especialmente para el transporte por el aire del MP. El principio básico de los MR se basa en el equilibrio de masas entre el emisor y el receptor, que asume que las masas y las especies permanecen constantes entre uno y otro, o experimentan mínimos cambios.

Aparte de esta clasificación, es recomendable, de forma previa al desarrollo de un ACF, llevar a cabo un análisis estadístico básico de los datos en estudio, que puede incluir, si lo permiten, un análisis de tendencias de las observaciones o la estimación de la distribución estadística que mejor describe tal conjunto de datos (Belis et al., 2014). Procedimientos estadísticos sencillos como el análisis de correlaciones o modelización mediante SSTT se utilizan como una aproximación inicial en el ACF, o como una tarea previa a la aplicación de los MR, de mayor coste económico y requerimientos de tiempo. Estos métodos exploratorios son diversos y no pueden considerarse realmente orientados hacia las prácticas del ACF, además de carecer de un soporte estadístico sólido. Es lógico pensar que los resultados del ACF serán más robustos si se combinan diferentes tipos de modelizaciones, ya que, aisladamente, ninguna de ellas resulta completamente satisfactoria debido a sus requisitos teóricos de aplicación.

Los modelos ocultos de Markov (MOM) no se utilizan frecuentemente para la predicción de la calidad del aire debido a su limitada capacidad para predecir las concentraciones de contaminantes (Dong et al., 2009). Esta limitación se debe a la propiedad de Markov, por la cual solo el estado actual del sistema proporciona algún conocimiento respecto al comportamiento futuro del proceso (la información sobre el pasado del proceso no revela nada sobre su futuro). Pero si no se pretende realizar una estadística predictiva con respecto a la concentración de los contaminantes, los MOM se revelan como unos modelos "flexibles" de uso generalista para el análisis de SSTT univariantes (Cappé et al., 2005) y multivariantes (Zucchini y MacDonald, 2009), a la vez que son fácilmente interpretables.

Los MOM pueden constituir un punto de partida para los ACF y en este capítulo se utilizarán para el estudio y la caracterización de las SSTT de PM_{10}, agrupando sus observaciones en el tiempo en grupos homogéneos o *regímenes de concentración*. Así, MOM gaussianos han sido aplicados a SSTT univariantes de PM_{10} obtenidas de estaciones de vigilancia de la calidad del aire de fondo en la Península Ibérica y Archipiélagos de las Azores, Balear y Canario. Se analizó, de forma particular, la ST de PM_{10} obtenida en la estación de Temisas en Las Palmas de Gran Canaria (Islas Canarias, España) durante el año 2013. El ACF en este archipiélago ha sido previamente estudiado por otros autores, confirmándose las altas contribuciones de MP debido al transporte de masas de aire desde los Desiertos del Sahel y Sáhara (África del Norte).

Este estudio que se expone en este capítulo pretende: 1) proponer el uso de HMM homogéneos como una técnica estadística exploratoria y rutinaria para complementar otras técnicas de ACF con las que estimar contribuciones de PM_{10}, y 2) presentar un nuevo método para la estimación de la carga neta de polvo en PM_{10} procedente de las regiones norteafricanas utilizando los MOM.

6.2. Datos y métodos

6.2.1. Estaciones de monitorización

Conjuntos de datos de concentraciones medias diarias de PM_{10} obtenidos en 33 estaciones de fondo (en su mayoría rurales) de la Península Ibérica y los Archipiélagos de las Azores, Balear y Canario fueron estudiados en diferentes intervalos de tiempo (Tabla 6.1). De estas estaciones, 28 pertenecen a la administración ambiental del Estado español (en la fecha de redacción de esta memoria, Ministerio de Agricultura, Alimentación y Medio Ambiente, MAAM) y pertenecen a la red ibérica de fondo para la detección de episodios africanos (Querol et al., 2013a), con 13 de ellas también incluidas en la red EMEP (EMEP, 2014). La *Comissão de Coordenação da Direcção Regional* (CCDR) *do Centro*, CCDR

do Alentejo y *Direcção Regional do Ambiente dos Açores* de Portugal gestiona 5 de estas estaciones de monitorización. Los datos empleados en este estudio fueron proporcionados por estas instituciones portuguesas y MAAM tras su validación.

Las concentraciones de PM_{10} obtenidas de las estaciones de monitorización se determinaron utilizando métodos gravimétricos y automáticos (β-atenuación y TEOM). Por tanto, para proceder a la armonización de los datos de las SSTT, las mediciones fueron corregidas mediante la aplicación de factores de corrección obtenidos mediante una comparación con el método gravimétrico (EN-12341, 1998). La aparición de episodios diarios de intrusiones de MP durante el año 2013 debido al transporte de masas de aire norteafricanas considerados en este estudio fueron determinados por Pérez et al. (2014), empleando una combinación de métodos (Querol et al., 2009), incluyendo la modelización HYSPLIT (Draxler and Rolph, 2003).

Estación	Entorno	Localización (Lat. Long.)	Altitud (m, s.n.m.)	Zona geográfica	Periodo
Niembro (NI)	R-EMEP	43°38'05"N 04°51'00"W	134	Norte (N)	2009-2013
Valderejo (VA)	R	42°52'31"N 03°13'53"W	911	Norte	2009-2013
Pagoeta (PA)	R	43°15'02"N 02°09'18"W	225	Norte	2009-2013
Els Torms (TO)	R-EMEP	41°23'38"N 00°44'05"E	470	Noreste (NE)	2013
Cabo de Creus (CR)	R-EMEP	42°09'19"N 03°18'57"E	23	Noreste	2013
Monagrega (MG)	R	40°56'48"N 00°17'27"W	570	Noreste	2013
Montseny (MS)	R	41°45'46"N 02°21'29"E	693	Noreste	2013
Zarra (ZA)	R-EMEP	39°04'58"N 01°06'54"W	885	Este (E)	2013
Morella (MR)	R-I	40°38'14"N 00°05'28"W	1150	Este	2013
El Pinós (PI)	R	38°27'06"N 01°03'53"W	642	Este	2013
Víznar (VI)	R-EMEP	37°14'14"N 03°32'03"W	1230	Sureste (SE)	2009-2013
Alcornocales (AL)	R	36°14'02"N 05°39'49"W	189	Sureste	2009-2013
Sierra Norte (SN)	R	37°59'40"N 05°40'01"W	573	Suroeste (SW)	2013
Barcarrota (BA)	R-EMEP	38°28'22"N 06°55'25"W	393	Suroeste	2013
Doñana (DO)	R-EMEP	37°03'07"N 06°33'19"W	5	Suroeste	2013
Ervedeira (ER)	R	39°55'26"N 08°53'30"W	32	Oeste (W)	2013
Fornelo do Monte (FM)	R	40°38'28"N 08°06'02"W	741	Oeste	2013
Fundão (FU)	R	40°13'59"N 07°18'07"W	473	Oeste	2013
Montemor-o-Velho (MO)	R	40°10'58"N 08°40'36"W	96	Oeste	2013
O Saviñao (SA)	R-EMEP	42°38'05"N 07°42'17"W	506	Noroeste (NW)	2013
Noia (NO)	R-EMEP	42°43'14"N 08°55'25"W	685	Noroeste	2013
Peñausende (PE)	R-EMEP	41°14'20"N 05°53'51"W	985	Centro (C)	2009-2013
Campisábalos (CA)	R-EMEP	41°16'27"N 03°08'33"W	1360	Centro	2009-2013
S. Pablo (MN)	R-EMEP	39°32'49"N 04°21'02"W	917	Centro	2009-2013
El Atazar (AT)	R	40°54'37"N 03°28'00"W	995	Centro	2009-2013
Monfagrüe (MF)	R	39°50'57"N 05°56'23"W	376	Centro	2009-2013
Faial (FA)	Re	38°36'18"N 28°37'53"W	310	Archipiélago de las Azores (AZ)	2009-2013
C. de Bellver (BE)	SU-B	39°33'48"N 02°37'14"E	117	Archipiélago Balear (BA)	2013
Mahón (MA)	R-EMEP	39°52'31"N 04°18'59"E	78	Archipiélago Balear	2013
El Río (RI)	R-I	28°08'42"N 16°31'25"W	500	Archipiélago Canario (CA)	2013
Temisas (TE)	R	27°53'54"N 15°29'18"W	460	Archipiélago Canario	2013
Echedo (EC)	R	27°49'54"N 17°55'18"W	375	Archipiélago Canario	2013
Tefia (TF)	R	28°31'38"N 14°00'16"W	160	Archipiélago Canario	2013

Tabla 6.1 Estaciones de monitorización de PM_{10} de fondo para la obtención de las SSTT y periodos de estudio. R: regional; I: Industrial; Re-Remota; SU-B: Suburbana de fondo.

6.2.2. Aplicación de los MOM al estudio de SSTT de PM_{10}

En las aplicaciones prácticas, existe, con frecuencia, un significado físico asociado a los estados ocultos de un MOM. Por ejemplo, en economía, estos estados pueden relacionarse con el "estado de la economía" (expansión y receso), centrándose el interés en el estudio de la dinámica entre ellos (Dias et al., 2010); en psicología del desarrollo, los estados de un MOM se emplean para cuantificar el conocimiento que sujetos en estudio manifiestan ante una determinada tarea de aprendizaje (Visser et al., 2002); o en estudios sobre las etapas del sueño, los estados de un MOM se corresponden con diferentes etapas, tales como el sueño REM ("rapid eye movement"), sueño profundo o vigilia (Flexer et al., 2002).

En este capítulo, los estados ocultos de un MOM representarán diferentes rangos de concentración de PM_{10} en cada ST modelizada (77 conjuntos de datos; ver Tabla 6.1 o Tablas D.1-D.5 en el Anexo D). Por simplicidad, y una aproximación más intuitiva, el concepto de estado oculto y el término "rangos de concentración" se unificarán para ser referidos como "regímenes de concentración" o simplemente como "regímenes". Como consecuencia del empleo de una técnica de agrupamiento (técnica *clúster*), estos regímenes agrupan observaciones de la ST con una concentración similar de PM_{10} que, al mismo tiempo, difieren en su concentración de otras observaciones agrupadas en otros regímenes. No obstante, los valores de las observaciones agrupadas en unos regímenes pueden exhibir cierto grado de solapamiento con las agrupadas en otros, lo cual es necesario para capturar convenientemente la forma de la distribución de los datos en la ST.

Los MOM proporcionan los valores media y desviación típica para cada uno de los regímenes de la ST, quedando así definida cada distribución gaussiana asignada a cada régimen, proporcionando una infomación valiosa al caracterizar, de esta manera, las SSTT obtenidas en las distintas estaciones de monitorización y periodos de tiempo. Esta abundante parametrización, sin embargo, puede ser resumida, como sucedía con los MMF, a través de expresiones algebráicas sencillas (2.7). Estas expresiones calculan la media y desviación típica de cada ST empleando semejantes valores de cada régimen en un suma que es ponderada con la representatividad (π) de cada régimen respecto a la distribución general de los datos de la ST. Estos valores, que definen a cada ST y se denotan por μ_m y σ_m (m representa *mixtura*), son cuantitativamente similares a los valores de la media aritmética ($\bar{x}$) y desviación estándar (s), que podrían haberse obtenido a partir de los datos de la ST sin considerar en ellos una dependencia temporal, calculando simplemente estos estadísticos a partir del conjunto de datos de la ST. Por tanto, μ_m puede emplearse como un indicador cuantitativo del nivel medio de exposición anual de la población a PM_{10}, y σ_m, representa el nivel de variabilidad de este contaminante alrededor de μ_m, tras tener en cuenta la dependencia temporal que existe entre los datos en estudio.

Además, los resultados de la modelización de una ST con MOM incluyen la matriz de probabilidades de transición (m.p.t.), la cual representa la parte dinámica de los MOM, al indicar como valores de probabilidad, el paso de un regímen a otro en la ST. Por tanto, es posible determinar cómo de probable es que, tras haber observado una concentración de PM_{10} asignada por el modelo a un régimen en particular, la siguiente observación en la ST sea similar (permanezca en el mismo régimen de concentraciones) o no (cambie a otro régimen). Los resultados asociados a esta m.p.t. conducen al concepto de estabilidad de la ST, la cual se deduce del valor de los elementos de su diagonal principal (ver Figura 3.3, página 42). Esta diagonal se relaciona con la probabilidad de que el proceso permanezca en un régimen determinado a largo plazo. Así, mientras más cercanos al valor 1 se encuentran cada uno de los elementos de la diagonal principal de la m.p.t., mayor es la probabilidad de que las observaciones de PM_{10} que se sucedan en el tiempo permanezcan en el mismo régimen. En la Figura 3.3, los elementos de $\mathbf{A}$ exhiben una marcada inestabilidad en la hipotética ST, ya que los valores de la diagonal principal se encuentran más cercanos al 0 que al 1 ($a_{11} = 0.2 \ll 1$; $a_{22} = 0.3 \ll 1$).

6.2.3. Estimación de los modelos

La implementación computacional de los MOM se realizó mediante la librería de R (R Core Team, 2015) depmixS4. Los detalles de la implementación se encuentran en la Sección D.2 (pág. 144). Los resultados obtenidos mediante esta librería fueron comparados con los obtenidos mediante la implementación realizada en R al efecto para esta tesis (Anexo B, pág. 122), sin encontrarse diferencias importantes (Sección B.5, pág 126).

6.3. Resultados y discusión

6.3.1. Modelización con MOM y estimaciones

Los valores numéricos obtenidos tras modelizar con MOM la ST de concentraciones medias diarias de PM_{10}, obtenidas en la estación de Temisas durante el año 2013, se muestran en la Tabla 6.2. Una descripción gráfica de la modelización mediante MOM se muestra en la Figura 3.3 (pág. 42), donde puede observarse el agrupamiento de las concentraciones medias diarias de PM_{10}. Esta estación y el resto, indicadas en la Tabla 6.1, han sido empleadas por otros autores (Pérez et al., 2014; Pey et al., 2013, Querol et al., 2013a) para cuantificar las contribuciones de los episodios de polvo procedentes del Norte de África en el Archipiélago Canario y en la Península Ibérica.

					$\mu g/m^3$					
Régimen	δ	n	Rango ($\mu g/m^3$)	π	μ	σ	μ_m	$\bar{x}$	σ_m	s
1	1	200	5-17	0.532	10.3	2.4	21.3	21.0	25.3	25.3
2	0	95	10-29	0.265	17.7	4.3				
3	0	64	7-96	0.181	42.8	17.7				
4	0	6	103-237	0.022	153.3	62.6				

$$\mathbf{A} = \begin{array}{c} \\ \text{del 1} \\ \text{del 2} \\ \text{del 3} \\ \text{del 4} \end{array} \begin{array}{c} \begin{array}{cccc} \textit{Régimen} & \text{al 1} & \text{al 2} & \text{al 3} & \text{al 4} \end{array} \\ \begin{pmatrix} \mathbf{0.871} & 0.110 & 0.019 & \approx 0 \\ 0.221 & \mathbf{0.679} & 0.066 & 0.034 \\ 0.025 & 0.170 & \mathbf{0.782} & 0.023 \\ \approx 0 & \approx 0 & 0.666 & \mathbf{0.333} \end{pmatrix} \end{array}$$

Tabla 6.2 Resultados de la modelización con MOM de la ST obtenida en Temisas durante el año 2013 y comparación de μ_m y σ_m con la media aritmética ($\bar{x}$) y desviación estándar (s) del conjunto de datos analizados. n indica el número de observaciones (medias diarias) agrupadas en cada régimen y **A** la m.p.t. (se destacan los elementos de la diagonal principal en **A**). Los elementos de las filas de **A** suman 1.

Las Figuras 6.1A y 6.1B son equivalentes entre sí y muestran cómo la modelización forma 4 grupos (clústeres) en las observaciones diarias obtenidas durante 2013 en la estación de Temisas, representados mediante una ST e histograma, respectivamente. Estos grupos se corresponden con los denominados régimenes (estados ocultos) del MOM mencionados tras la Definicion 3.7, en la página 40 (un MOM de 4 estados). En la Figura 6.1A cada observación (media diaria de PM_{10}) se etiqueta con el número del régimen asignado por el MOM, o bien, se agrupa bajo una curva gaussiana en la Figura 6.1B. Esta Figura 6.1B muestra cómo la densidad resultante de la mixtura (línea negra) captura la forma de la distribución de los datos (histograma, en línea gris) de forma mucho más precisa que cualquier función de densidad individual, independientemente de la familia a la que perteneciese (p. ej., log-normal). En la Tabla 6.2 se observa que el rango de valores de cada régimen muestra un típico solapamiento, y también, la similitud en los valores de media y desviacion típica de la mixtura (μ_m y σ_m) con la media aritmética ($\bar{x}$) y desviación estándar (s) de los datos que forman la ST.

Una vez que los datos han sido modelizados con MOM, a cada uno de los regímenes establecidos se les debe proporcionar un significado. A continuación, se proponen los siguientes, para la concentración media de cada régimen (μ_i, $i = 1, 2, 3, 4$), considerando la estación de vigilancia Temisas y el periodo en estudio (año 2013):

- μ_1 (10.3 $\mu g/m^3$) representa la concentración subyacente o umbral en la que no se esperan grandes cambios a lo largo del tiempo si las condiciones atmosféricas y de contaminación permanecen aproximadamente constantes, siendo una concentración característica del área de estudio.
- μ_2 (17.7 $\mu g/m^3$) es la concentración media de PM_{10} en días afectados por moderadas contribuciones de origen antropogénico, debido a actividades que se realizan en la región. El valor μ_2 está sujeto a mayores variaciones en comparación con el de μ_1 de un año para otro.

- μ_3 (42.8 μg/m^3) es la concentración media de PM_{10} de aquellos días afectados por contribuciones características, y frecuentes, debidas a intrusiones norteafricanas. Estas contribuciones son muy variables en concentración y son las principales responsables de la superaciones del valor límite diario establecido por la Directiva 2008/50/EC (Directiva, 2008) para este contaminante.

- μ_4 (153.3 μg/m^3) representa la concentración media de PM_{10} en aquellos días afectados por episodios severos, y por lo general poco frecuentes, de contribuciones naturales norteafricanas.

Es de destacar que las definiciones dadas para los regímenes después de modelizar esta ST con MOM coinciden, desde un punto de vista conceptual, con la separación de niveles de PM_{10} realizada por Lenschow et al. (2001) en la región de Berlín (Alemania), por la cual se distinguía una fracción local, urbana y regional de fondo. Estos autores dieron, a lo que se ha denominado aquí como regímenes, el intuitivo nombre de *perfiles horizontales de concentraciones de PM_{10} ambiente*, y en particular, lo que aquí se define como primer régimen (régimen 1 en Figura 6.1) fue denominado como concentración de fondo *natural*. El trabajo posterior de Escudero et al. (2007) menciona la utilidad de interpretar la variabilidad de los niveles de fondo regionales de PM_{10}, ya que las contribuciones locales pueden así ser identificadas y, por tanto, los planes y programas para la mejora de la calidad del aire pueden ser implementados de forma apropiada.

Hay que recalcar que las definiciones dadas deben adaptarse al área geográfica en estudio. En general, las concentraciones estudiadas de estos regímenes poseen un efecto acumulativo, ya que diferentes fuentes de contaminación pueden contribuir de forma simultánea a la calidad del aire (p.ej., una observación de la ST que pertenece al régimen 3 incluye contribuciones de las fuentes asociadas con los regímenes 1 y 2). A partir de las definiciones dadas, y empleando este supuesto aditivo, pueden estimarse, de manera aproximada, las siguientes cantidades:

- $\mu_2 - \mu_1$: concentración media debida a contribuciones antropogénicas de la región (7.4 μg/m^3).

- $\mu_2 - \mu_1$: concentración media asociada con las contribuciones características procedentes del Norte de África, cuando estas ocurren (25.1 μg/m^3).

- $\mu_4 - \mu_2$: concentración media asociada con las contribuciones de naturaleza más severas y ocasionales procedentes del Norte de África, cuando estas ocurren (135.6 μg/m^3).

Si bien las anteriores estimaciones proceden del cálculo de dos valores medios, y por tanto deben ser consideradas como aproximadas, pueden obtenerse, no obstante, otras estimaciones más precisas como las contribuciones de cada régimen a la concentración media anual de PM_{10}. Estas estimaciones pueden calcularse fácilmente empleando la información de la Tabla 6.2, multiplicando la representatividad de cada régimen (π_i, $i = 1, 2, 3, 4$) por su valor medio (μ_i, $i = 1, 2, 3, 4$). La Figura 6.2 muestra la contribución de cada régimen.

La m.p.t. (**A**, Tabla 6.2) muestra cómo los regímenes son, de alguna forma, inestables, dado que los elementos de la diagonal principal no son cercanos a 1 (los elementos de cada fila de **A** suman 1). Transiciones poco probables entre regímenes son aquellas cuyo valor de probabilidad es cercano a 0, a saber: a_{41}, del régimen 4 al 1, y a_{42}, ambas debidas al tiempo de residencia atmosférico después de que una contribución norteafricana suceda; también a_{14}, dada practicamente la imposibilidad de un incremento repentino en la concentración de PM_{10} de un día para otro. También de **A**, se deduce que la probabilidad de dos episodios severos consecutivos es baja ($a_{14} = 0.333$).

6.3.2. Comportamiento de los regímenes en la Península Ibérica y archipiélagos

Los MOM, descritos en el Capítulo 3, se aplicaron a las SSTT obtenidas de las 33 estaciones de monitorización durante el año 2013 indicadas en la Tabla 6.1. La Figura 6.3A muestra los resultados obtenidos, cada punto de color representando el valor medio de cada régimen de PM_{10} de las SSTT

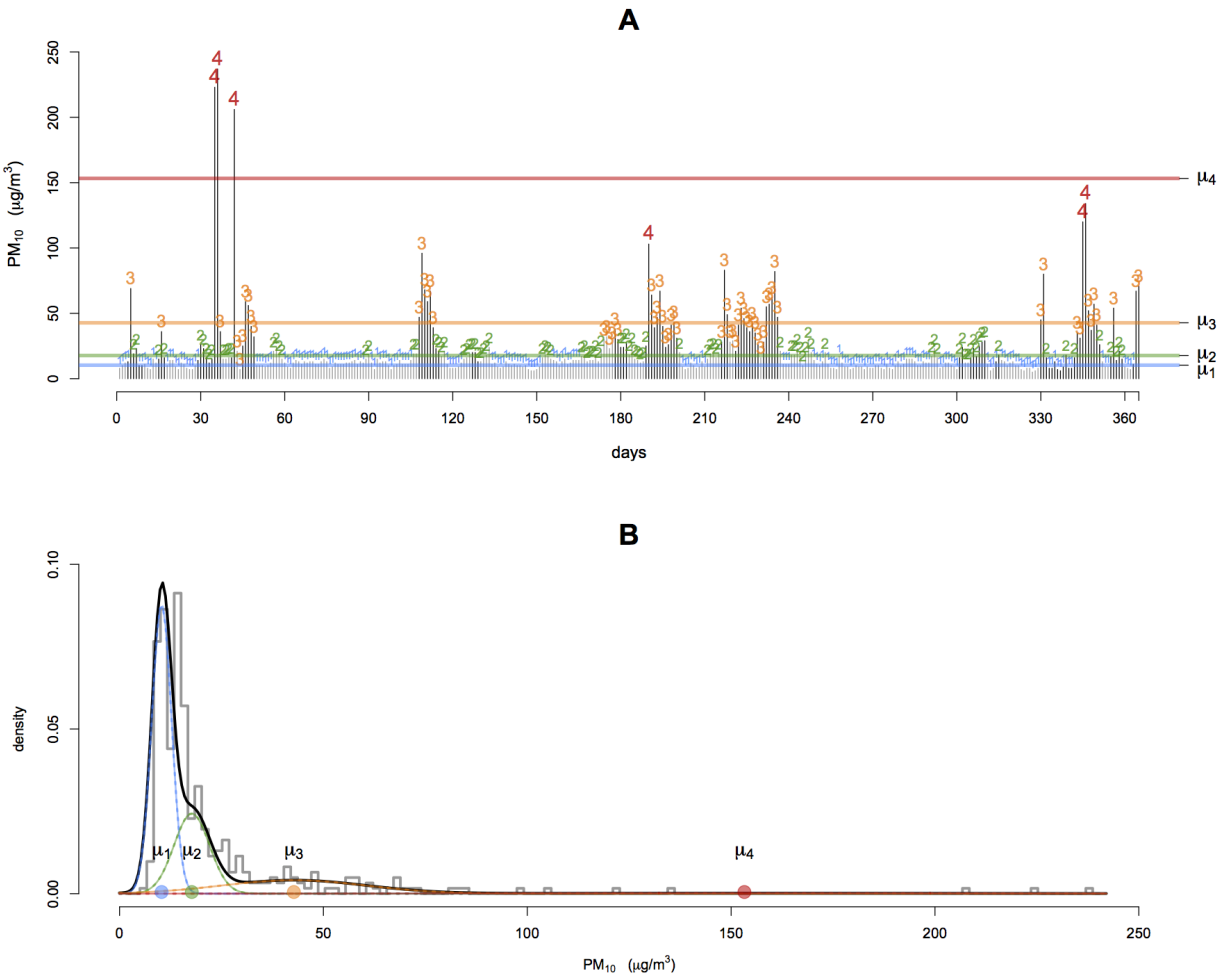


Figura 6.1 A. Efecto del agrupamiento de las observaciones de la ST de PM_{10} obtenida en el emplazamiento de Temisas. Cada observación es numerada después de haber sido asignada a un régimen por el MOM (régimen 1, 2, 3 y 4 en azul, verde, naranja y rojo, respectivamente). Los valores medios de cada régimen (clúster) se indican mediante μ_1, μ_2, μ_3 y μ_4, manteniendo las líneas el mismo código de color. **B**. Mixtura de distribuciones gaussianas capturando la forma de la distribución de los datos de la ST (histograma en gris). **A** y **B** se corresponden en color. Las líneas verticales de color negro en la ST indican días en los que ha sido detectada una intrusión sahariana.

analizadas (el numero de regímenes detectados y otras medidas definiéndolos, junto con sus errores *bootstrap*, se indican en la Tabla D.1), obteniéndose un máximo de 4 (μ_1, μ_2, μ_3 y μ_4). A efectos de simplificación, la Figura 6.3B resume toda esta información mediante diagramas de caja y bigote, dado que el propósito de este análisis no es el de describir cada uno de los regímenes detectados en cada estación de monitorización, sino el de obtener una visión genérica del comportamiento de los regímenes en cada área geográfica.

La utilización de los MOM para la modelización de SSTT de contaminantes, conlleva, deseablemente, la definición de los regímenes en las SSTT de forma previa a la estimación de las contribuciones de fuentes en un área específica. La definición de regímenes es siempre subjetiva y sujeta a diferentes interpretaciones, siendo necesario al respecto el consenso entre expertos. Debido al número de estaciones estudiadas, esta tarea no pudo llevarse a cabo en este capítulo. No obstante, varios casos merecen analizarse, como el de la estación de la Isla de Faial en el Archipiélago de las Azores (AZ). En este emplazamiento no se detecta un cuarto régimen en su ST, y las concentraciones de los regímenes existentes son especialmente bajas (μ_m=5.80 μg/m^3; μ_1=2.30 μg/m^3; μ_2=5.88 μg/m^3; μ_3=11.13 μg/m^3). Si la concentración registrada en este sitio es influenciada por MP procedente de los desiertos norteafricanos es un factor que no puede ser determinado mediante esta modelización, aunque el aerosol marino

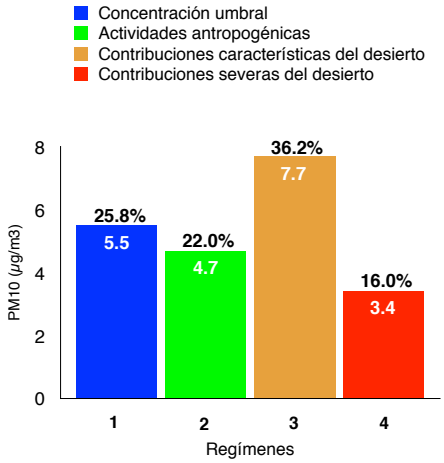


Figura 6.2 Contribuciones de los diferentes regímenes a la concentación media anual de PM_{10} (21.3 $\mu g/m^3$) en el emplazamiento de Temisas. La contribución de cada régimen a la concentración media anual de PM_{10} se indica dentro de cada una de las barras, y su representatividad, en %. Las contribuciones se calculan como $\pi_i \cdot \mu_i$, $i = 1, 2, 3, 4$, a partir de la información de la Tabla 6.2.

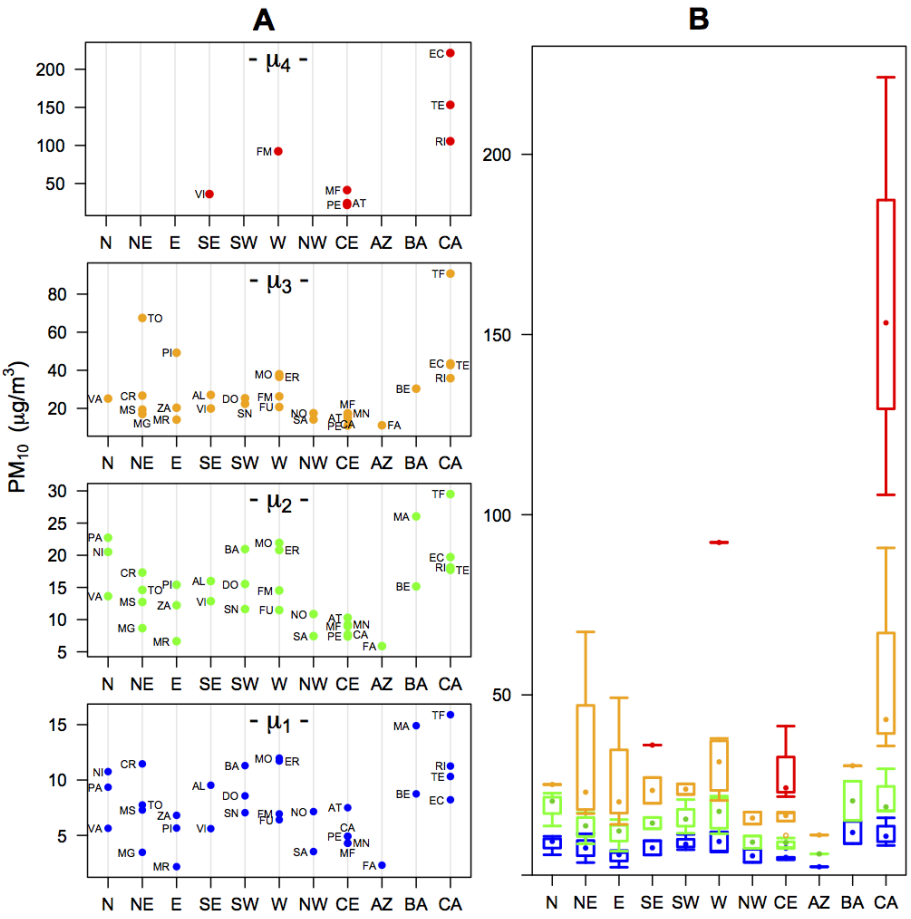


Figura 6.3 A. Estudio de los regímenes en la Península Ibérica y Archipiélagos de las Azores, Balear y Canario, durante el año 2013. Cada punto representa el valor medio de los regímenes de PM_{10} en los emplazamientos de medida, por área geográfica. **B**. Diagramas de caja y bigote resumiendo la información de **A**. El código de color para los regímenes es la misma que en la Figura 6.1. Las áreas geográficas se abrevian como se indica en la Tabla 6.1.

podría estar representado parcialmente por el tercer régimen, considerando el aislamiento de esta isla en el Océano Atlántico. Por otra parte, en dos de los tres emplazamientos analizados en el área Norte no se detecta un tercer régimen. Las estaciones de monitorización en el centro (CE) de la Península Ibérica presentan un cuarto régimen que podría venir representado por resuspensiones de polvo y contribuciones naturales de África del Norte. El resto de regímenes analizados se considera para futuras investigaciones.

Las definiciones de los regímenes dadas para Temisas (pág. 76) pueden ser consideradas de nuevo en esta sección, teniendo en cuenta la naturaleza de fondo de prácticamente la totalidad de los emplazamientos estudiados (Tabla 6.1). Este es el caso del régimen 1, ya que es asumible, en teoría, que en cada emplazamiento se registre una concentración media mínima, subyacente de PM_{10}, que sea característica de dicho emplazamiento y que exhiba escasa variación a lo largo del tiempo mientras las condiciones atmosféricas y de contaminación permanezcan relativamente constantes. Consecuentemente, el régimen 1 (Figura 6.3A, gráfico inferior) puede proporcionar una indicación del nivel de contaminación por PM_{10} en las estaciones de vigilancia.

Los regímenes 3 y 4, definidos para el emplazamiento de Temisas, son también aplicables al resto de emplazamientos del Archipiélago de Canarias, dado que todos ellos están afectados por la misma fuente natural de contribuciones. La Tabla 6.3 compara la contribución estimada de los episodios africanos en el Archipiélago de Canarias dada por Pérez et al. (2014) y aquellas estimadas utilizando los MOM después de considerar, en este último caso, las definiciones de regímenes dadas en la Sección 6.3.1. Como puede apreciarse, los resultados de la estimación son similares. Aquellos autores emplean una metodología basada en la valoración de las SSTT de PM_{10} en estaciones de fondo por medio del cálculo del percentil 40 medio móvil mensual (Escudero et al., 2007a, Querol et al., 2013a, 2013b). Este último método se describe en la Sección 6.3.4, donde se compara con el propuesto en este capítulo.

Estación	MOM	Pérez et al. (2014)
El Río	11.6	10.8
Temisas	11.1	9.0
Echedo	6.1	6.2
Tefia	9.9	10.4
	$\bar{x} = 9.7$	$\bar{x} = 9.1$

Tabla 6.3 Comparación entre las contribuciones estimadas (en $\mu g/m^3$) debidas a los episodios africanos en el Archipiélago de Canarias durante el año 2013 y sus valores medios ($\bar{x}$). Se incluye el emplazamiento de Temisas. Utilizando los MOM, las contribuciones se calculan como se indicó en la Figura 6.2, pero considerando únicamente los regímenes 3 y 4, ya que en ellos se definen las intrusiones saharianas (p. ej., en el caso de Temisas: 11.1 = 7.7 + 3.4).

6.3.3. Estudio de los regímenes a lo largo del tiempo

El mismo estudio sobre el comportamiento de los regímenes realizado en la sección anterior se amplió desde el año 2009 al 2012, por economía de espacio en áreas geográficas seleccionadas (N, SE, CE y AZ), mostrándose los resultados en la Figura 6.4 (ver Tablas D.2-D.5 en el Anexo D). La Figura 6.4A muestra la tendencia general de PM_{10} para ese periodo, empleando de nuevo diagramas de caja y bigote. Cada caja y bigote (en negro) describe los valores μ_m obtenidos mediante el estudio de las SSTT en los emplazamientos de cada área. La Figura 6.4B repite los resultados del año 2013 de la Figura 6.3B para estos emplazamientos seleccionados, siendo incluidos de nuevo por un propósito comparativo con años anteriores. Los emplazamientos del Archipiélago de Canarias (CA) no se han incluido, dado que la red de la calidad del aire en estas islas fue alterado durante los 5 años en los que se desarrolló el estudio.

Como se aprecia en la Figura 6.4A, en general, se detecta una suave tendencia a la baja en las concentraciones de PM_{10}, representada mediante una línea horizontal gris, calculada como la media

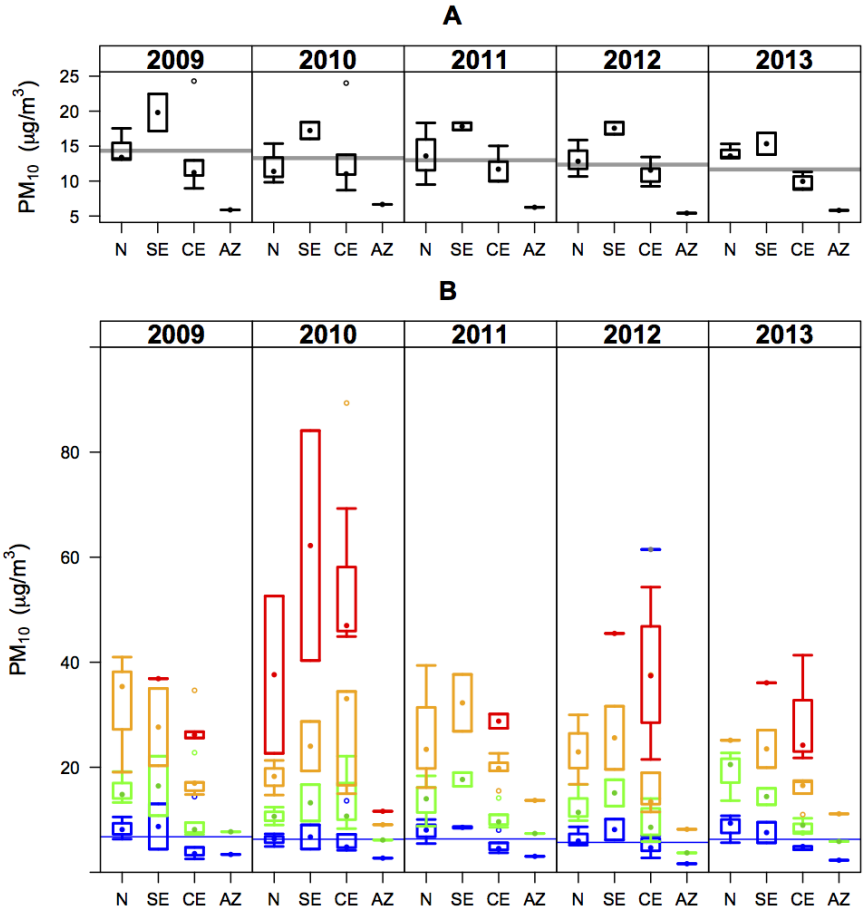


Figura 6.4 A. Representación mediante diagramas de caja y bigote de los valores de μ_m desde el 2009 al 2013 en los emplazamientos de las áreas seleccionadas de la Península Ibérica y Archipiélago de las Azores. **B**. Desagregación de los regímenes de **A** mediante los mismos diagramas. Las abreviaturas y códigos de color son semejantes a los de la Figura 6.3.

aritmética de los valores de μ_m, para cada año. Esta tendencia también se refleja en la desagregación de los valores μ_m por regiones mostrado en la Figura 6.4B. Se detecta, también, un paralelismo cercano entre la tendencia del primer régimen (Figura 6.4B, diagramas de caja y bigote y línea horizontal azules) y la tendencia general (línea gris, Figura 6.4A), respaldando el uso factible de este régimen como un indicador de la contaminación media por PM_{10}. Esto se debe a la alta representatividad del primer régimen en la distribución general de los datos de la ST (p. ej., en el emplazamiento de Temisas, π_1=0.532).

La desagregación de regímenes mostrado en la Figura 6.4B ayuda a entender fenómenos de contribuciones nuevos o alterados, como la aparición de un cuarto régimen en 2010 en el área Norte (caja y bigote roja) que no se presenta en el resto de años estudiados, o un quinto régimen en el área CE, en el año 2012 (detectada en San Pablo). Estas contribuciones no se presumen importantes cuantitativamente, pero podrían describir aportes diferentes o cambios en los aportes esperados habitualmente. En el área Norte (N) durante el año 2010, el cuarto régimen es descrito únicamente en los emplazamientos de Niembro y Pagoeta. Estos emplazamientos se localizan aproximadamente a 1 km y 6 km de la Costa Cantábrica, respectivamente. Además, la modificación en los niveles de concentración de PM_{10} se debe posiblemente a la contribución del aerosol marino, aunque las contribuciones norteafricanas han sido también cuantificadas para los emplazamientos de Niembro y Pagoeta (Pey et al., 2011) empleando el método del percentil 40 medio móvil mensual. Estos autores determinan que los aportes africanos en Niembro y Pagoeta son de 0.9 μg/m^3 y 0.7 μg/m^3, respectivamente. Los MOM establecen que la contribución del cuarto régimen (μ_4) a la concentración media anual de PM_{10} en estos emplazamientos fue de 0.47 μg/m^3 y de 2.33 μg/m^3, respectivamente.

En el caso del área SE en el año 2010, el cuarto régimen está presente en Víznar y Alcornocales. Este régimen contribuye a la concentración media anual de PM_{10} en estos emplazamientos con 4.31 $\mu g/m^3$ y 0.84 $\mu g/m^3$, respectivamente, mientras que Pey et al. (2011) calculan que las contribuciones debidas al polvo Norteafricano son de 3.9 $\mu g/m^3$ y 2.0 $\mu g/m^3$.

No pueden obtenerse conclusiones directas de estos resultados, ya que los regímenes en estas áreas geográficas no han sido definidas. No obstante, resultados similares obtenidos mediante la metodología actual (Escudero et al., 2007a, 2007b) y los MOM sugieren que debería llevarse a cabo una investigación más detallada al respecto. Sería interesante saber cómo ambas metodologías se complementan una a otra, dado que los MOM pueden ofrecer informacion relevante una vez que la definición de regímenes se ha realizado, a saber: 1) contribuciones netas de las diferentes fuentes de emisión; 2) contribuciones respecto la media anual de PM_{10} de cada régimen, y 3) la probabilidad de cambio entre los diferentes regímenes de cada ST. Dada esta valiosa información, los MOM representan una análisis estadístico individual, que destaca por su sustento teórico sólido, con el que caracterizar cualquier ST. Además, deben tenerse en cuenta la facilidad en su interpretación y la reproducibilidad de sus resultados, así como el hecho de que esta modelización puede llevarse a cabo empleando software gratuito disponible para la comunidad científica (R).

6.3.4. Nuevo método para la estimación del aporte de PM_{10} de origen desértico

Las intrusiones de masas de aire africano consisten en procesos de advección de masas de aire con elevada carga particulada desde los desiertos africanos hacia la zona de estudio, con el consiguiente incremento de los niveles de partículas (Querol et al., 2013c). La Directiva 2008/50/EC (Directiva, 2008) establece que podrán descontarse las superaciones de los valores límite de PM_{10} (50 $\mu g/m^3$) del cómputo de superaciones anuales siempre que se demuestre que dichos valores son sobrepasados por la influencia de aportaciones procedentes de fuentes naturales, que, según el artículo 2.15 de dicha Directiva, se definen como"emisiones de agentes contaminantes no causadas directa o indirectamente por actividades humanas, lo que incluye los fenómenos naturales tales como erupciones volcánicas, actividades sísmicas, actividades geotérmicas, o incendios de zonas silvestres, fuertes vientos, aerosoles marinos o resuspensión atmosférica o transporte de partículas naturales procedentes de regiones áridas".

En España y Portugal los episodios naturales con mayor repercusión en los niveles de MP son los episodios de aporte de polvo africano, aunque en episodios y zonas concretos los incendios forestales (zonas forestales en verano), el aerosol marino (cornisa atlántica, islas Madeira y Canarias) e, incluso, la resuspensión (interior peninsular) pueden tener mucha importancia (Querol et al., 2013b). El método que en la actualidad es más ampliamente aceptado (Viana et al., 2014) para cuantificar el impacto de los episodios de intrusión de polvo africano en los niveles de PM_{10} fue presentado por Escudero et al. (2007) y se basa en el cálculo del percentil 40 medio móvil mensual (método P40). Esta metodología fue aceptada por la Unión Europea mediante la publicación del documento"Commission staff working paper establishing guidelines for demonstration and subtraction of exceedances attributable to natural sources under the Directive 2008/50/EC on ambient air quality and cleaner air for Europe" (EC, 2011).

De forma breve, este método obtiene el aporte de PM_{10} africano, de manera individual para cada día en el que se ha detectado una intrusión norteafricana, restando el nivel de fondo regional a la media diaria de PM_{10} determinada en la estación de fondo regional en estudio. Este nivel de fondo regional se obtiene tras aplicar el percentil 40 medio móvil mensual a la ST de PM_{10} en la estación, para cada día de medida, habiendo extraído previamente de la ST aquellos días con influencia de polvo africano.

La motivación para proponer un nuevo método para la estimación del aporte de PM_{10} africano responde a la sugerencia al respecto de Viana et al. (2008, pág. 843), quien reconoce necesario mas investigación al respecto, aun cuando ya existan métodos para estimar este aporte (Collaud Coen et al., 2004; Escudero et al., 2007).

Aunque el método actual (método P40) representa una adecuada y conveniente aproximación para la estimación de los aportes, el que se presenta en esta tesis puede mejorarlo en varios aspectos, ya

que evita: 1) el efecto de suavizado implícito en la aplicación de un procedimiento móvil a una ST, y 2) la aproximación empírica sobre la que se basa la elección del percentil del método (40). Además, el método que aquí se propone permite obtener un intervalo de confianza para las estimaciones de los aportes africanos, a partir del cálculo del error estándar *bootstrap* (estos errores se acompañan en el Anexo D), aspecto que ya se sugería como necesario en las conclusiones del trabajo de Viana et al. (2008) antes mencionado.

El cálculo de la estimación del aporte africano se calculará, como ejemplo, sobre la ST de Temisas (Tabla 6.2), y se utilizará su representación gráfica como ST (Figura 6.1A) para aproximar de forma intuitiva el método propuesto. Este método estima los aportes africanos sobre los días en los que se ha detectado un episodio de intrusión restando a las concentraciones medias diarias de PM_{10} de estos días el valor de μ_2. La elección del valor medio de este régimen como sustraendo se fundamenta en el supuesto aditivo de las concentraciones, anteriormente expuesto, basado en el trabajo de Lenschow et al. (2001). Un ejemplo del método se acompaña en la Tabla 6.4 y se compara con el P40. La Figura 6.5 compara las diferentes estimaciones de las cargas de polvo en PM_{10} obtenidas empleando ambos métodos, para estaciones de vigilancia del Archipiélago Canario (Temisas, Echedo y Tefia), y del Sur (Víznar, Doñana), Este (Zarra) y Centro (San Pablo, Peñausende y Campisábalos) peninsular. La línea diagonal negra representa una hipotética correlación perfecta entre ambos métodos. Como puede apreciarse, el método P40 podría sobreestimar las cargas de polvo atribuible a severos episodios de intrusión que frecuentemente ocurren en el Archipiélago Canario (Figuras 6.5A-6.5C). Esta sobreestimación es menos significativa desde un punto de vista cuantitativo cuando se observan contribuciones menos severas (Figuras 6.5D, F-I; en la Figura 6.5E se observa una ligera subestimación). La Tabla 6.5 muestra la diferencia anual (2013) entre ambos métodos al estimar las contribuciones medias en días en los que se han detectado intrusiones saharianas. Debido la aproximación empírica utilizada por el método P40, no puede derivarse un razonamiento analítico de la discrepancia entre ambos métodos, aunque pudiera estar implicado el efecto de alisado realizado sobre las SSTT por el procedimiento móvil que emplea.

				Contribución del desierto	
Fecha	PM_{10} observado	Régimen	μ_2	Método propuesto	Método P40
5 Enero 2013	69	3		69-18=51	60
4 Febrero 2013	223	4		223-18=205	213
5 Febrero 2013	237	4	$17.7 \approx 18$	237-18=219	227
25 Abril 2013	21	2		21-18=3	10
12 Diciembre 2013	134	4		134-18=116	125

Tabla 6.4 Ejemplo de la estimación del aporte natural africano utilizando la ST de Temisas, empleando el método propuesto. La última columna muestra el resultado de la estimación utilizando el método P40 (en $\mu g/m^3$).

Área	Emplazamiento	Método P40	Método propuesto
Archipiélago Canario	Temisas	29.6	25.4
	Echedo	22.2	18.6
	Tefia	34.7	24.8
Sur de PI	Víznar	8.9	9.9
	Doñana	6.6	9.8
Este de la PI	Zarra	7.3	7.3
Centro de la PI	S. Pablo	9.0	10.2
	Peñausende	8.6	10.5
	Campisábalos	9.9	12.1

Tabla 6.5 Estimaciones de las contribuciones medias de PM_{10} debidas a episodios de intrusión durante el año 2013, empleando ambos métodos, en emplazamientos de la Península Ibérica (PI) y Archipiélago Canario (en $\mu g/m^3$).

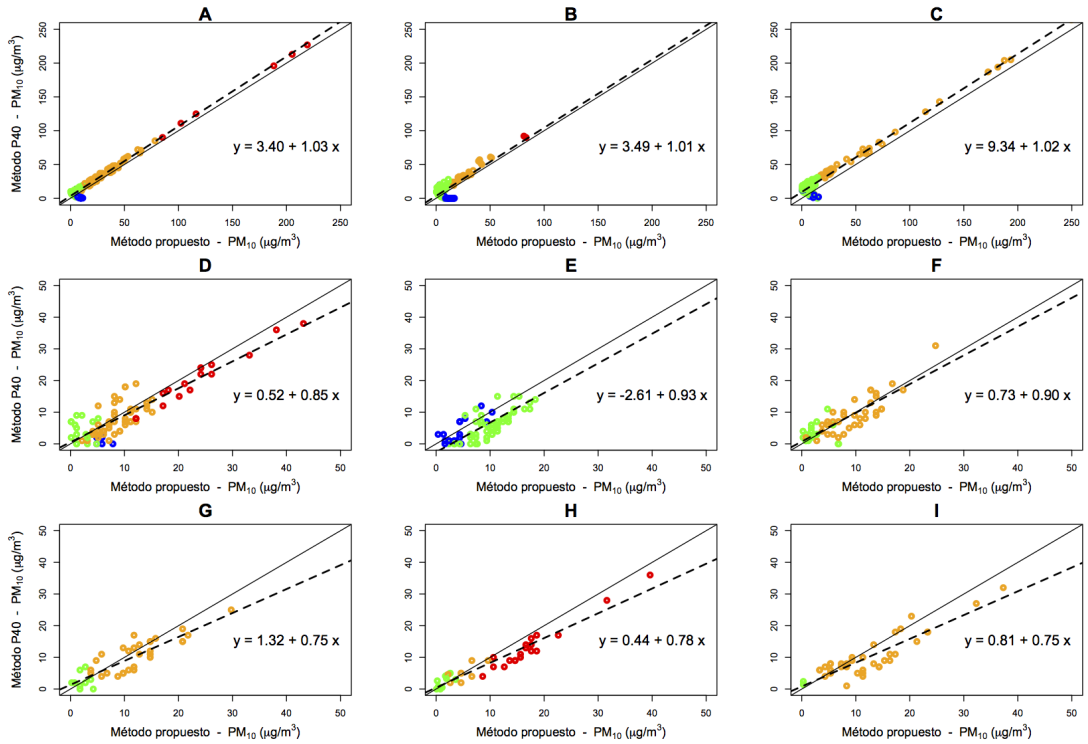


Figura 6.5 Comparación en la estimación de contribuciones de PM_{10} por episodios de intrusiones norteafricanas utilizando ambos métodos. Los puntos de color indican la asignación de cada cantidad a un régimen (código de color como en Figura 6.1). Las líneas de regresión (discontinuas) indican la discrepancia entre ambos métodos, y la línea continua, una hipotética correlación perfecta. La discrepancia se muestra numéricamente mediante una regresión lineal simple, del método P40 (y) sobre el que se propone en este trabajo (x). De **A** a **I**: emplazamientos de Temisas, Echedo, Tefia, Víznar, Doñana, Zarra, San Pablo, Peñausende y Campisábalos, respectivamente.

El método propuesto es intuitivo y simple. Intuitivo, al basarse en un esquema aditivo de concentraciones, y simple, al emplear tan solo una sustracción para obtener la carga neta de PM_{10}. Sin embargo, presenta el inconveniente de la facilidad con la que pueda dotarse de una definición a los regímenes obtenidos mediante la aplicación de MOM sobre las SSTT, y en particular, a su segundo régimen (valor medio μ_2), que es el empleado en este nuevo método. Esta dificultad se extiende más allá de este método, ya que suele presentarse en cualquiera en el que se aplique los MOM. No obstante, parte de este inconveniente puede paliarse mediante la m.p.t. de estos modelos, ayudando sus valores a dar coherencia a las definiciones de los regímenes. En el método aquí propuesto, la m.p.t. **A** (Tabla 6.2) ayuda a verificar las definiciones dadas a los regímenes.

La sustitución de la cantidad μ_2-μ_1 en lugar de μ_2 para obtener la carga neta no se recomienda, ya que procede de la estimación de dos valores medios (μ_2 y μ_1) y, por tanto, resultaría en una estimación más grosera de las contribuciones antropogénicas.

6.4. Conclusiones

En este capítulo se presentan parte de las propiedades y usos de una nueva metodología para el ACF basada en el agrupamiento de las observaciones de las SSTT, así como un método nuevo para las estimaciones de las contribuciones de PM_{10} de los desiertos, no exclusivamente norteafricanos. Se analizaron los resultados de la aplicación de los MOM sobre las SSTT de concentraciones medias diarias de PM_{10} durante diferentes periodos de tiempo. Las contribuciones netas debidas a diferentes fuentes de emisión, contribuciones de cada régimen a la media anual de PM_{10}, así como la probabilidad de cambio entre los diferentes regímenes de la SSTT de Temisas (Archipiélago de Canarias) fueron estimadas, después de definir sus regímenes de concentración. Este emplazamiento se caracteriza por las altas

contribuciones de PM_{10} procedentes de las regiones áridas del Norte de África. El primer régimen de las SSTT analizadas en los diferentes emplazamientos se propone como un indicador de la contaminación de fondo.

La definición de regímenes en cada ST es un paso previo y deseable para la aplicación de los MOM y puede requerir por ello el consenso entre expertos. Estas definiciones proporcionan un sustento teórico a las diferentes fracciones de contaminación presentadas por otros autores (Lenschow et al., 2001), que deben ser consideradas para el diseño de los planes y programas de la calidad del aire. El estudio de los regímenes en una escala espacial ayuda a distinguir y cuantificar las diferentes contribuciones en las distintas áreas geográficas, aunque tales estudios deben completarse con otros tipos de modelizaciones para obtener un ACF más robusto. La contribución anual de los episodios norteafricanos al valor medio de PM_{10} en el Archipiélago Canario coincide, de forma notable, con las mismas estimaciones empleando el método del P40 empleado por Pérez et al. (2014). Mediante la adición de una escala temporal al análisis espacial de las contribuciones, pueden detectarse nuevas fuentes de contribución o la alteración de las habituales en las áreas de estudio.

El método propuesto en este capítulo para la estimación de las contribuciones de los desiertos parece corregir la carga neta de PM_{10} estimada por el método P40, y atribuye menos impacto sobre las áreas que sufren una mayor influencia de los episodios africanos (Archipiélago Canario). La solidez estadística del nuevo método propuesto, basado en los MOM, su sustento conceptual sobre la aproximación de Lenschow et al. (2001), y su capacidad para obtener intervalos de confianza respecto a las estimaciones de las intrusiones, hacen de él una alternativa deseable al método actual, aun considerando el inconveniente ya referido respecto a la necesidad de dotar de definiciones a los regímenes obtenidos por el modelo.

El agrupamiento de las observaciones de las SSTT empleando MOM proporciona una aproximación metodológica importante a los métodos exploratorios empleados en los ACF, pudiendo utilizarse para complementar los resultados de los MR, que requieren ensayos más costosos económicamente (especiaciones químicas) y dilatados en el tiempo. Los resultados de los MOM son sencillos de interpretar y tiene un alto grado de reproducibilidad, y en cuanto a su implementación, puede realizarse mediante software gratuito, sin requerir unos avanzados conocimientos estadísticos o de programación. Por todo lo anterior, se recomienda el uso de MOM en el estudio de la contaminación por PM_{10}.

7

Caracterización de la contaminación atmosférica de fondo en entornos urbanos

Resumen La contaminación atmosférica en las áreas urbanas es el resultado de la presencia de contaminantes emitidos localmente (de diferentes fuentes), de otras contribuciones menores aportadas a escala rural y regional, y del transporte horizontal (contaminación de fondo). Entender los perfiles de contaminación urbanos de fondo (concentraciones más bajas) es clave en estudios epidemiológicos y para valorar la exposición general de la población a la contaminación atmosférica. Por ello, la contaminación del aire en estaciones de monitorización sin la influencia directa de fuentes de emisión locales ha sido estudiada en numerosas ocasiones. Con frecuencia, estos niveles de la contaminación de fondo son caracterizados únicamente a partir del valor medio de sus concentraciones, obtenidas a lo largo de periodos de estudio prolongados, generalmente anuales. Este capítulo muestra cómo una métrica basada en los modelos ocultos de Markov (MOM) puede emplearse para caracterizar más fielmente los perfiles de concentración de fondo de los principales contaminantes primarios del aire. Los MOM se aplicaron a concentraciones medias diarias de CO, NO_2, PM_{10} y SO_2 obtenidas en trece estaciones de monitorización de tres ciudades, desde el año 2010 al 2013. Mediante la métrica que se propone, los valores medios de la concentración ambiente y de fondo en estas estaciones fueron estimados para esos contaminantes, así como su ratio y diferencia. El indicador ratio para estos contaminantes y en estas ciudades, durante el periodo de cuatro años estudiado, establece que la contaminación de fondo representa entre el 48 % y 69 % de la contaminación ambiente del aire, mientras que la diferencia entre ambas concentraciones se sitúa en el rango de 101-193 $\mu g/m^3$, 7-12 $\mu g/m^3$, 11-13 $\mu g/m^3$ y 2-3 $\mu g/m^3$ para el CO, NO_2, PM_{10} y SO_2, respectivamente.

El contenido de este capítulo es una adaptación de **Gómez-Losada, Á., Pires, J.C.M., Pino-Mejías, R.** 2016. Characterization of background air pollution exposure in urban environments using a metric based on Hidden Markov Models. *Atmospheric Environment*, 127: 255-61.

7.1. Introducción

En las aglomeraciones urbanas, resulta de interés el estudio de los perfiles de concentración de fondo (concentraciones más bajas) de los contaminantes del aire, ya que se han observado efectos perjudiciales para la salud en niveles de contaminación considerados seguros para la salud (Lepeule et al., 2014). Esto ha sido descrito para contaminantes como el CO, NO_2, O_3, PM_{10} y SO_2 (Sunyer et al., 2002; Vedal et al., 2003; Latza et al., 2009; REVIHAAP Project, 2013). Los niveles de la contaminación atmosférica de fondo se estiman en áreas donde los efectos directos de las fuentes de contaminación local se consideran prácticamente excluidos. La estimación de estos niveles en las ciudades suscita controversia, ya que se basan en estaciones de monitorización localizadas en áreas que pueden no ser representativas para el propósito de caracterizar esta fracción de la contaminación. En caso de que sí lo sean, los niveles de la contaminación de fondo son, con frecuencia, dados como un valor medio después de promediar sus fluctuaciones sobre largos periodos de tiempo, siendo probable que surjan errores en la estimación de su concentración. Como ya introdujo Stein et al. (2007), la selección inapropiada de los valores de concentración de fondo puede representar una fuente de incertidumbre significativa en los

resultados de las modelizaciones de dispersión, o en técnicas de interpolación empleadas para estimar las características de la exposición ambiente.

El propósito de este capítulo es mostrar cómo una métrica basada en modelos ocultos de Markov (MOM) puede ser utilizado para caracterizar los perfiles de concentración de fondo de contaminantes atmosféricos, exponiéndose el potencial de este tipo de modelización para definir características de exposición.

7.2. Datos y métodos

7.2.1. Estaciones de monitorización

Se obtuvieron datos de la calidad del aire (concentraciones medias horarias de CO, NO_2, PM_{10} y SO_2) de estaciones de monitorización urbanas de las redes de calidad del aire de Jaén, Granada y Sevilla (Andalucía, España), desde el año 2010 al 2013 (Tabla 7.1). Los métodos de monitorización estándares establecidos por la Directiva Europea 2008/50/EC (Directiva, 2008) fueron los utilizados para la determinación de las concentraciones de los contaminantes CO, NO_2 y SO_2, y el método de la β-atenuación para las PM_{10}. Como en el capítulo 5, las concentraciones medias diarias a partir de los datos horarios se han calculado únicamente cuando se ha dispuesto de, al menos, el 80 % de todas las observaciones horarias durante el día (19 de 24 horas). Los datos de calidad del aire fueron proporcionados por la administración ambiental autonómica tras su validación. La población de Jaén, Granada y Sevilla durante 2013 fue de 116 176, 237 818 y 700 169 (IECA, 2015) habitantes, respectivamente. Aunque en la modelización de los datos de la calidad del aire se considera fundamental el estudio de las condiciones meteorológicas que gobiernan el transporte, la transformación y la eliminación de los contaminantes atmosféricos estudiados, estas condiciones no fueron tenidas en cuenta, ya que el interés de este capítulo se centra en los perfiles de concentración de fondo a los que queda expuesta la población, y no en el estudio de los factores que explican, influencian o alteran estos perfiles de contaminación.

Ciudad	Estación y abreviatura	Tipo	Fuente de emisión principal	Localización		Contaminantes estudiados			
Granada	Palacio de Congresos (Pal)	Suburbana	Tráfico	343546	4196517	CO	NO_2	PM_{10}	SO_2
	Granada Norte (Nor)	Urbana	Tráfico	345040	4195364	CO	NO_2	PM_{10}	SO_2
Jaén	Fuentezuelas (Fue)	Suburbana	Fondo	428647	4182208	CO	NO_2	*	SO_2
	Ronda del Valle (Ron)	Urbana	Fondo	431177	4181976	CO	NO_2	PM_{10}	SO_2
Sevilla	Alcalá de Guadaíra (Alc)	Urban	Fondo	248974	4136631	CO	NO_2	PM_{10}	SO_2
	Aljarafe (Alj)	Suburbana	Fondo	230473	4137017	*	NO_2	PM_{10}	SO_2
	Bermejales (Ber)	Urbana	Fondo	236063	4137554	CO	NO_2	PM_{10}	SO_2
	Centro (Cen)	Urbana	Fondo	235156	4142125	CO	NO_2	*	SO_2
	Dos Hermanas (Dos)	Urbana	Fondo	241677	4130413	CO	NO_2	*	SO_2
	Príncipes (Pri)	Urbana	Fondo	233863	4140741	CO	NO_2	PM_{10}	SO_2
	Ranilla (Ran)	Urbana	Tráfico	237965	4141611	CO	NO_2	*	SO_2
	Santa Clara (Sac)	Suburbana	Fondo	238720	4143149	CO	NO_2	PM_{10}	*
	Torneo (Tor)	Urbana	Tráfico	234151	4142873	CO	NO_2	PM_{10}	SO_2

Tabla 7.1 Características de los emplazamientos y contaminantes atmosféricos monitorizados, desde el año 2010 al 2013, en las tres áreas urbanas en estudio (los emplazamientos son expresados en coordenadas X,Y ETRS89-UTM cooreinadas, zona 30). Los asteriscos representan contaminantes no monitorizados.

7.2.2. Primer régimen de las SSTT como estimador de la contaminación de fondo

Los MOM pueden emplearse para el análisis de contribución de fuentes (ACF) al atribuir cada clúster (régimen de observaciones) formado en la ST en estudio a una fuente de emisión contaminante. Así, el

ACF con MMO conlleva la agrupación de medias diarias de concentración aproximadamente similar y se basa en el supuesto de que estas concentraciones son emitidas por una fuente de emisión común. Esta tarea, compleja, requiere un análisis detallado de las fuentes de contaminación atmosférica presentes en un área determinada. Si bien las áreas urbanas comparten características en términos de fuentes de emisión que influencian la calidad del aire, un ACF riguroso requiere la caracterización formal de todas las fuentes de emisión presentes. En este capítulo, tal caracterización no fue realizada debido al número de estaciones de monitorización estudiadas, los cuatro años que abarcó el estudio y los cuatro contaminantes atmosféricos analizados.

En su lugar, la caracterización formal mencionada se sustituyó por la aproximación de Lenschow (Lenschow et al., 2001). Esta aproximación asume que la concentración de un contaminante medida en una estación de monitorización se corresponde con la suma de diferentes contribuciones, a saber, regionales, urbanas de fondo y de naturaleza local. Por ejemplo, en una estación de tráfico, la contaminación ambiente resultante es la suma de las contribuciones del tráfico local en las proximidades de la estación, así como de las contribuciones urbanas de fondo y regionales. En el caso de una estación de fondo, urbana o suburbana, las contribuciones que explican la contaminación ambiente se corresponden con aquellas de los niveles de fondo de la ciudad o área metropolitana, respectivamente, y aquellas del fondo regional. Por tanto, las concentraciones medidas en estaciones de tráfico, así como las urbanas y suburbanas de fondo, son representativas de los efectos del tráfico y niveles de fondo de las áreas urbanas y metropolitanas, respectivamente, sobre la calidad del aire en las ciudades (Salvador et al., 2015). Esta metodología, aditiva, ha sido empleada con frecuencia como una primera aproximación en ACF para contaminantes del aire (Belis et al., 2013), y puede aplicarse a áreas urbanas con un escaso impacto de emisiones industriales, como es el caso de Granada, Jaén y Sevilla.

La aproximación de Lenschow adoptada en este capítulo se empleó únicamente para atribuir fuentes de emisión a los primeros clústeres de las SSTT de concentraciones medias diarias, dado que sobre ellos recae el interés en este capítulo. Por tanto, las concentraciones medias diarias registradas en un emplazamiento de tráfico y agrupadas en el primer clúster de su ST, representan la contribución factible más baja debida al tráfico local, pero también incluye contribuciones de los niveles de fondo regionales y urbanos, igualmente en sus niveles más bajos. Con respecto a las concentraciones medias diarias en un emplazamiento urbano o suburbano, y que se agrupan en el primer clúster de la ST, estos valores de concentración representan contribuciones de fuentes específicas de la aglomeración urbana, igualmente a sus niveles más bajos (p.ej., niveles de emisión de fondo del tráfico, construcción, demolición y calefacciones domésticas), pero también contribuciones mínimas del fondo regional. Mediante esta aproximación, las concentraciones medias diarias del primer grupo en la ST se encuentran influenciadas por las contribuciones regionales, si bien tales contribuciones no serán estudiadas en este capítulo, ya que el interés reside en las áreas urbanas y no en las contribuciones externas a estas áreas.

En este capítulo, el primer clúster de las SSTT agrupa a las observaciones cuyos niveles de concentración es el más bajo registrado durante un periodo anual, en cualquier emplazamiento de monitorización y para cualquier contaminante de los estudiados, recalcándose su utilidad al describir las tendencias más bajas de la contaminación en las SSTT analizadas. Así, la contaminación representada por este primer clúster es sinónimo de contaminación de fondo. Esta contaminación de fondo fue estimada y comparada con la contaminación ambiente de CO, NO_2, PM_{10} y SO_2 en tres ciudades durante un periodo de cuatro años, utilizando datos procedentes de sus redes de monitorización de la calidad del aire. Los valores medios de concentración de la contaminación de fondo y ambiente se utilizaron para comparar la contaminación en las estaciones de monitorización de estas ciudades y para clasificar estas estaciones de acuerdo con estos valores en un gradiente de concentración.

7.2.3. Estimación de los modelos

Como en el capítulo 6, la implementación computacional de los MOM se realizó mediante la librería de R (R Core Team, 2015) depmixS4 (Visser y Speekenbrink, 2014). Los detalles de la implementación se encuentran, igualmente, en la Sección D.2 (pág. 144).

7.3. Resultados y Discusión

7.3.1. Caracterización de la contaminación de fondo por PM_{10} en dos estaciones

La Figura 7.1 muestra los resultados gráficos de la modelización con MOM de la ST de PM_{10} obtenida en la estación de Alcalá de Guadaíra ("Alc", Sevilla) y Palacio de Congresos ("Pal", Granada), durante 2013. La Tabla 7.2 muestra los resultados numéricos correspondientes. En las Figuras 1A ("Alc") y 1B ("Pal"), cada una de las concentraciones medias diarias de PM_{10} es etiquetada en color con el número del clúster o grupo asignado por el MOM. El valor m_1 representa la concentración media de la contaminación de fondo en ambas estaciones y se indica con una línea azul horizontal en ambas SSTT, y coincide, con el valor medio de todas las concentraciones agrupadas en el primer clúster de cada ST. Como se deducirá en la Sección 7.3.2, este valor medio (m_1) es específico de cada emplazamiento de medida y contaminante estudiado. En estas SSTT se detectan dos y tres clústeres, respectivamente, presentando ambas valores medios anuales similares (M), indicados mediante una línea horizontal discontinua negra. Estos valores medios anuales se calculan mediante las expresiones algebráicas (2.7) ya mencionadas anteriormente en los Capítulos 5 y 6, teniendo en cuenta el valor medio de cada clúster y sus representatividades en la ST. Los valores medios de los clústeres restantes se indican mediante líneas en verde (segundo clúster) y naranja (tercer clúster), si bien el interés en este capítulo recae sobre el primer clúster. Las Figuras 7.1C ("Alc") y 7.1D ("Pal") son equivalentes a las representaciones de las Figuras 7.1A y 7.1B, respectivamente, con la única diferencia de que las SSTT se muestran como histogramas. La representación mediante histogramas revela las curvas gaussianas bajo las cuales se agrupan las observaciones que pertenecen a los diferentes clústeres. Así, el primer clúster es caracterizado también por el valor de la desviación estándar, que estima la dispersión de los datos de la ST agrupados en el primer clúster alrededor de m_1, o de forma equivalente, la intensidad de la exposición a PM_{10}.

Estación	K = 2	Porcentaje de días	Tamaño (días)	Rango	Media del clúster	sd	Media ST	SD ST
Alc	Clúster 1	48	167	7.9-32.3	19.6	4.6	27.9	10.0
	Clúster 2	52	185	22.2-61.1	35.5	7.3		

Estación	K = 3	Porcentaje de días	Tamaño (días)	Rango	Media del clúster	sd	Media ST	SD ST
Pal	Clúster 1	26	92	6.8-22.2	13.1	3.8	24.5	9.8
	Clúster 2	40	151	15.7-32.2	23.1	3.8		
	Clúster 3	34	114	25.3-63.3	34.8	6.8		

Tabla 7.2 Resultados de la modelización con MMO para la SSTT de PM_{10} en las estaciones de Alcalá de Guadaíra ("Alc") y Palacio de Congresos ("Pal") durante 2013 (en $\mu g/m^3$). sd: desviación estándar del clúster; SD: desviación estándar (de la ST).

Como se muestra en la Tabla 7.2, en la ST de "Alc" los valores m_1 (19.6 $\mu g/m^3$) y sd_1 (4.6 $\mu g/m^3$) son superiores que los valores de la ST de "Pal" (13.1 $\mu g/m^3$ y 3.8 $\mu g/m^3$, respectivamente). Esto indica que, aunque la exposición a la contaminación de fondo por PM_{10} es alrededor de 7 unidades de concentración superior en "Alc" que en "Pal", la exposición a este contaminante se distribuye de forma más homogénea a lo largo del año en "Alc" (4.6 $\mu g/m^3$) que en "Pal" (3.8 $\mu g/m^3$). No obstante, las diferencias en los valores de sd_1 en estos emplazamientos no son cuantitativamente significativos. Es más, el valor de la representatividad en el primer clúster en "Alc" indica que las concentraciones diarias asociadas a la contaminación de fondo estuvieron presentes durante el 48 % de los días en el periodo anual, de lo que se deduce que no se detectaron influencias destacables de contribuciones externas de PM_{10} (p.ej, intrusiones saharianas influyendo sobre las concentraciones de PM_{10} diarias) durante la mitad del año de estudio. Respecto a la ST de "Pal", el valor de la representatividad del primer régimen

fue menor (26 % de los días) durante el periodo anual. Este análisis comparativo de las SSTT obtenidas en estos emplazamientos muestra que, ante unos similiares valores de la contaminación ambiente de PM_{10} (M), pueden detectarse distintos perfiles de contaminación, y en particular, las características de la contaminación de fondo representada por el primer clúster pueden ser muy diferentes. En la estación de fondo urbano "Alc", la diferencia entre los valores de las concentraciones M y m_1 dan evidencias de la diferencia cuantitativa entre la contaminación ambiente (M) en esta estación urbana de fondo y la concentración de fondo"real" (m_1). Como se verá en la siguiente sección, este valor de la contaminación de fondo puede obtenerse para cualquier contaminante y para cualquier estación de medida.

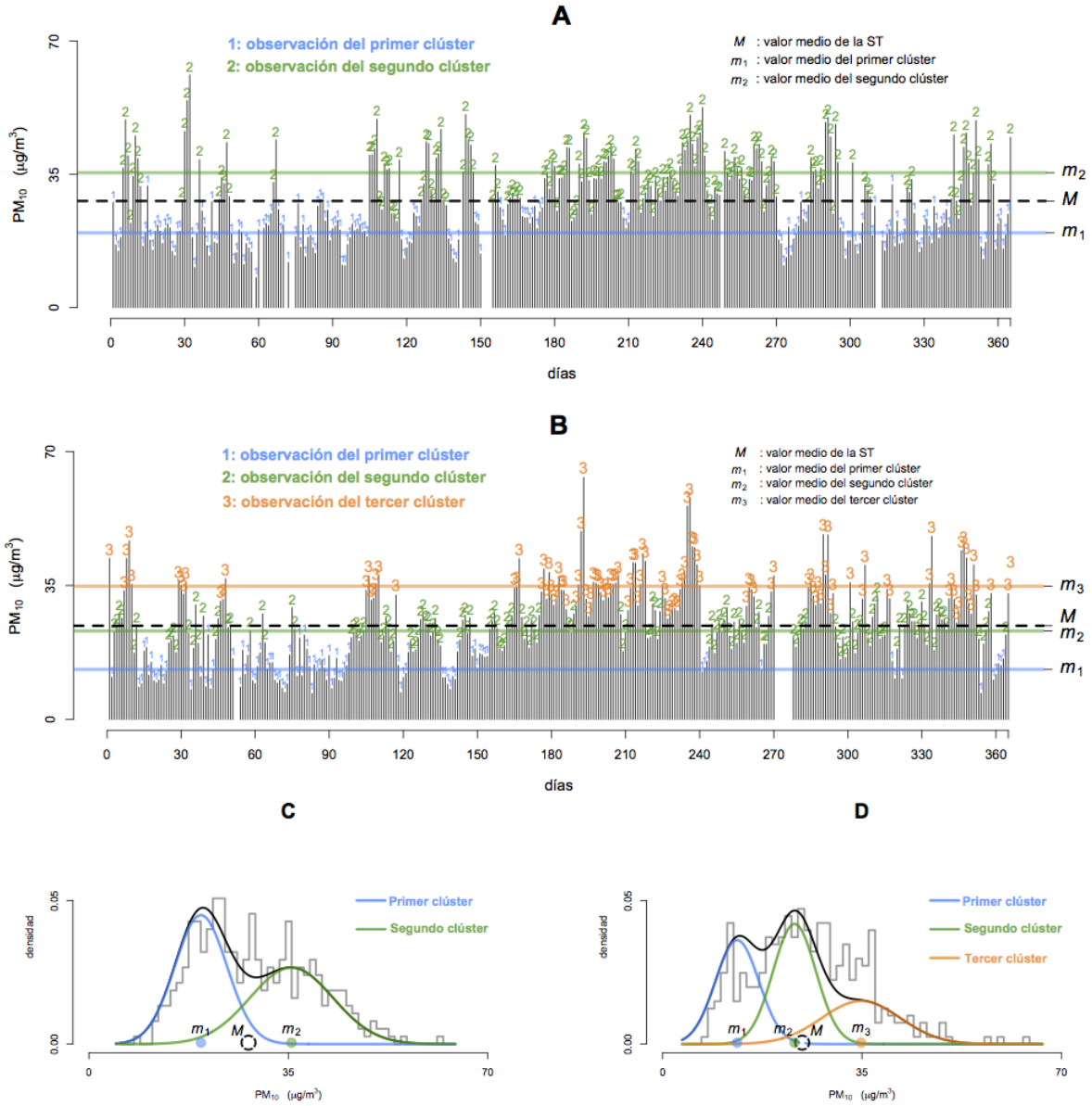


Figura 7.1 Agrupación de las observaciones de las SSTT de concentraciones medias diarias en dos (**A**) y tres (**B**) grupos, en los emplazamientos de "Alc" y "Pal", respectivamente. Representación de las SSTT de Alc (**C**) y Pal (**D**) como un histograma. En **A** y **B**, la contaminación representada por el primer clúster (contaminación de fondo) se representan mediante una línea azul y su valor medio m_1. El valor medio de las SSTT (contaminación ambiente) se representa mediante líneas discontinuas negras (M). En **C** y **D** la contaminación de fondo se caracteriza mediante una curva gaussiana azul. Los valores medios de los clústeres 2 y 3 se representan mediante un punto naranja y verde, y sus valores medios como m_2 y m_3. Las SSTT e histogramas muestran correspondencia en color.

7.3.2. Comparación de las exposiciones de fondo en diferentes estaciones de medida

Siguiendo una semejante aproximación que en la sección anterior, la contaminación de fondo respecto a las concentraciones de CO, NO_2, PM_{10} y SO_2 fue estimada en las estaciones de Granada, Jaén y Sevilla (Tabla 7.1), desde el año 2010 al 2013. Los resultados numéricos tras modelizar con MOM cada ST se acompañan en el Anexo E. Debido al número de emplazamientos estudiados, contaminantes y años de estudio, la representatividad (valores de π) y las desviaciones estándares no fueron analizadas, planificándose su estudio para investigaciones posteriores.

La Figura 7.2 representa de forma intuitiva los resultados obtenidos. Para contextualizar el valor medio de las concentraciones medias más bajas de la ST (m_1), representando el valor medio de la contaminación de fondo, los valores de la contaminación media ambiente (M) también se han incluido. Por continuidad con los símbolos utilizados en la Figura 7.1, los valores m_1 se representan con puntos azules, y los valores M con circunferencias discontinuas. Los resultados del cociente entre M y m_1 (ratios M/m_1) durante los años estudiados se muestran como números, próximos a estos símbolos. Este ratio compara cuántas veces son superiores los valores de la contaminación media ambiente, en términos absolutos, a la de los valores medios de la contaminación de fondo para los contaminantes considerados. Los valores más bajos de los ratios M/m_1 indican que la contaminación en las estaciones se explica principalmente por la contaminación de fondo, mientras que los ratios más elevados indican contribuciones perceptibles de otras fracciones a la contaminación ambiente (M). Como referencia, mientras más cercano a 1 sea el ratio M/m_1, la contaminación en un emplazamiento se explicará mayoritariamente por la contaminación de fondo. Como regla general, un ratio de 2 o 3 significa que la contaminación de fondo es responsable de la mitad (50 %) o un tercio (33 %) de la contaminación ambiente en un emplazamiento, respectivamente.

Para resumir los resultados mostrados en la Figura 7.2, los valores anuales de M y m_1 desde el año 2010 al 2013 fueron promediados por estaciones de monitorización. El resultado gráfico se muestra en la Figura 7.3, en la cual las estaciones de medida de las tres ciudades estudiadas se ordenan en función de un gradiente de concentraciones de M y m_1. Las áreas urbanas a las cuales las estaciones de medida pertenecen se indican entre paréntesis. La diferencia media entre los valores M y m_1 se reconoce más fácilmente, a la vez que puede apreciarse la diferencia entre la curvas superior e inferior de estaciones (más alta en PM_{10} y más baja en NO_2 y SO_2). Esta distancias relativas parecen depender del contaminante estudiado. Excepto para PM_{10}, muchos niveles de contaminación de fondo en algunos emplazamientos son superiores a los niveles de contaminación ambiente en otros, de la misma área urbana o no (p. ej., la contaminación de fondo de CO en “Cen” -Sevilla- es superior a la contaminación ambiente en “Alc” -Sevilla- y “Pal” -Granada-).

Los promedios de los ratios y diferencias obtenidos en el periodo de 4 años en cada una de las ciudades se muestran en la Tabla 7.3. El ratio M/m_1 se mantiene relativamente constante para las tres ciudades y la diferencia M-m_1 exhibe una variación que depende del contaminante. Estos dos indicadores, ratio y diferencia, podrían evidenciar que la contaminación de fondo en las ciudades con un perfil urbano mantienen una relación casi constante respecto a la contaminación ambiente. En el caso de los contaminantes y ciudades estudiadas, el indicador ratio muestra un rango de valores de 1.5 a 2.1, lo que significa que la concentración de fondo representa entre el 48 % y 69 % de la contaminación ambiente. La diferencia entre la contaminación de fondo y ambiente en estas ciudades varía entre 101-193 $\mu g/m^3$, 7-12 $\mu g/m^3$, 11-13 $\mu g/m^3$ y 2-3 $\mu g/m^3$ para el CO, NO_2, PM_{10} y SO_2, respectivamente. Desafortunadamente, solo se dispuso de dos estaciones de monitorización en Granada y Jaén para el estudio de la contaminación ambiente y de fondo, por lo que es recomendable ampliar los resultados obtenidos en ciudades que dispongan de un mayor número y tipología de estaciones de monitorización.

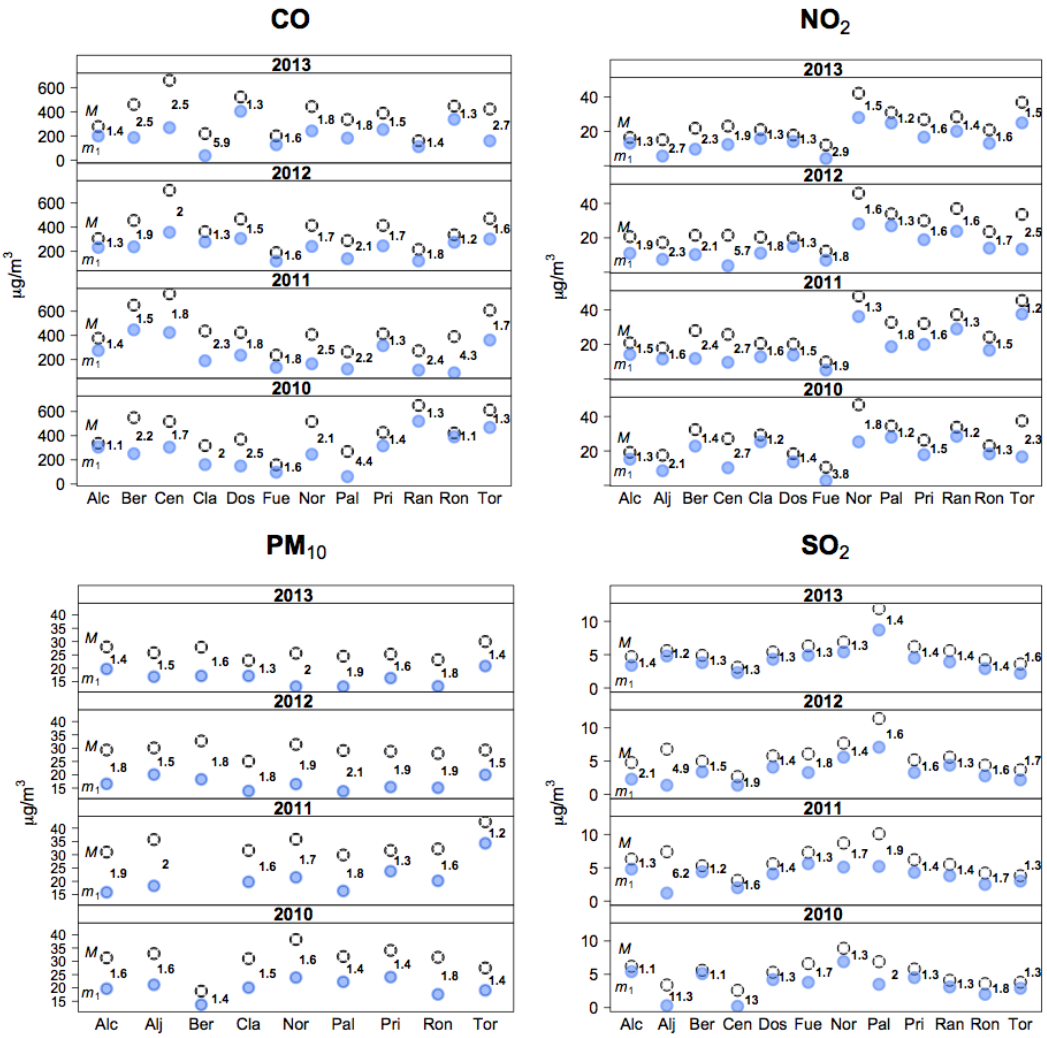


Figura 7.2 Comparación cuantitativa de los valores de contaminación ambiente (M) y de fondo (m_1) desde el 2010 al 2013, para los contaminantes atmosféricos y emplazamientos estudiados. Los valores M y m_1 se indican mediante circunferencias discontinuas y puntos azules, respectivamente. Los números indican el ratio M/m_1 ratio. El significado de los colores es semejante al de la Figura 7.1.

Contaminante	Ciudad	M	m_1	$M-m_1$	M/m_1	%
CO	Granada	367.5	174.8	192.7	2.1	48
	Jaén	298.0	196.6	101.4	1.5	66
	Sevilla	444.3	266.8	177.5	1.7	66
NO_2	Granada	39.4	27.0	12.4	1.5	69
	Jaén	17.0	10.1	6.9	1.7	59
	Sevilla	25.5	15.7	9.8	1.6	62
PM_{10}	Granada	30.8	17.6	13.2	1.8	57
	Jaén	28.7	16.5	12.2	1.7	57
	Sevilla	29.7	19.2	10.5	1.5	64
SO_2	Granada	9.1	5.9	3.2	1.5	65
	Jaén	5.3	3.5	1.8	1.5	66
	Sevilla	5.0	3.3	1.7	1.5	66

Tabla 7.3 Valores medios de la contaminación ambiente (M) y de fondo (m_1) para las ciudades y contaminantes estudiados, desde el año 2010 al 2013 (en $\mu g/m^3$). Se calcula la diferencia (M-m_1) y ratio (M/m_1) entre ellas, así como la contribución de la contaminación de fondo respecto a la contaminación ambiente (en %).

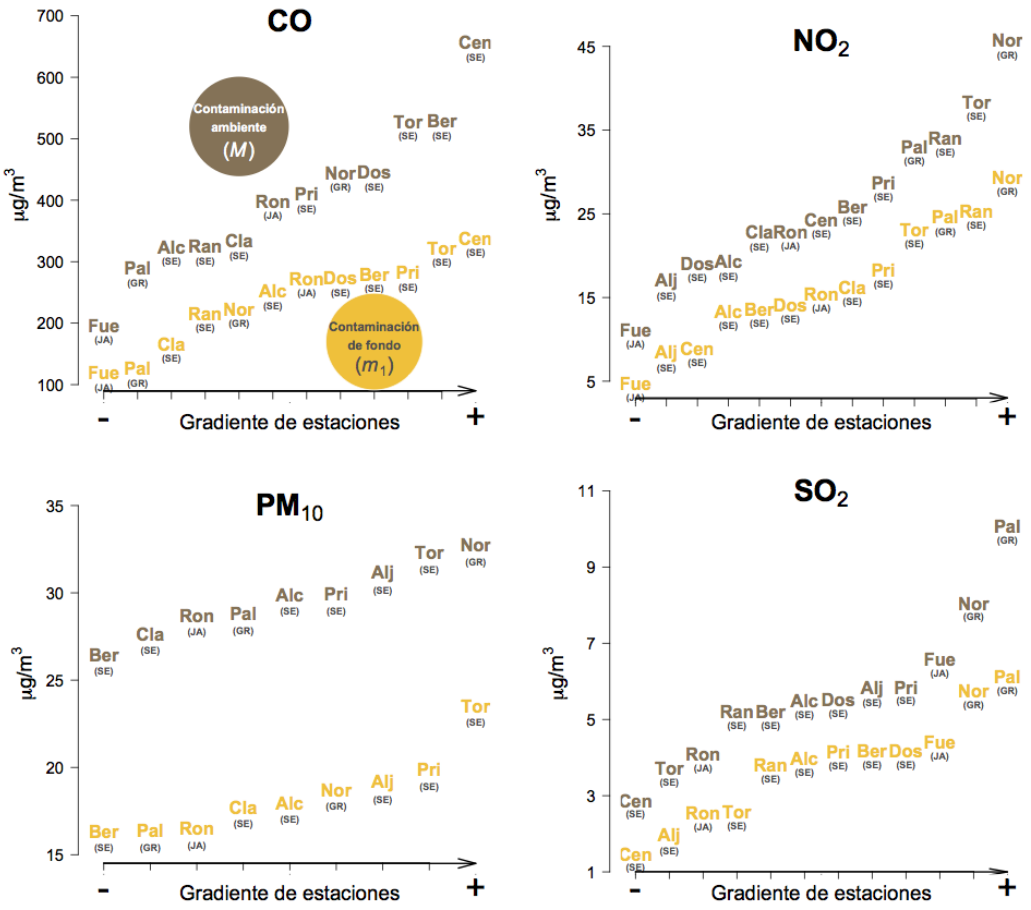


Figura 7.3 Estaciones de monitorización ordenados por un gradiente de concentración (de valores inferiores a superiores); la fila superior representa la contaminación ambiente (M) y la inferior, la contaminación de fondo (m_1). Las ciudades se indican bajo el nombre de las estaciones de monitorización: Granada (GR), Jaén (JA) y Sevilla (SE).

7.4. Conclusiones

Con el propósito de caracterizar los perfiles más bajos de contaminación del aire en áreas urbanas, se modelizaron con MOM SSTT de concentraciones medias diarias de CO, NO_2, PM_{10} y SO_2, obtenidas en trece estaciones de monitorización de tres áreas urbanas sin influencia industrial, desde el año 2010 al 2013. Los resultados de los MOM incluyen la agrupación de las concentraciones medias diarias más bajas de estas SSTT, definiendo así con su valor medio el nivel de la contaminación de fondo. Esta contaminación de fondo, de acuerdo con la métrica utilizada, se cuantificó aproximadamente alrededor del 48 %-69 % de la contaminación ambiente para los contaminantes y áreas urbanas estudiadas. La diferencia entre la contaminación ambiente y la contaminación de fondo en estas áreas también fue estimada para cada uno de los contaminantes CO, NO_2, PM_{10} y SO_2.

8

Conclusiones generales

Si bien las conclusiones de este trabajo de tesis se han ido exponiendo a lo largo de sus capítulos correspondientes, se presentan nuevamente para ofrecer una visión conjunta de ellas.

1. La aplicación combinada de los MMF y las técnicas de minería de datos (ACJ, BA y ACP) ha permitido plantear alternativas en la configuración de la red de calidad del aire de Sevilla que pueden redundar en la mejora de su gestión y aprovechamiento de los recursos económicos empleados en su explotación. Esta metodología es aplicable a cualquier red de vigilancia de la calidad del aire.

2. Mediante esta combinación de MMF y el resto de técnicas, se han clasificado las estaciones de la red de vigilancia de la calidad del aire de Sevilla atendiendo a su carga de contaminación, mediante criterios estadísticos. La metodología empleada permite una caracterización exhaustiva de la calidad del aire ambiente en entornos urbanos y rurales.

3. La aplicación de los MOM a las series temporales (SSTT) de PM_{10} ha permitido caracterizar sus "regímenes" o "perfiles de concentración". La concentración media de estos regímenes, así como su representatividad en las SSTT, permiten obtener la contribución de cada uno de ellos a la concentración media ambiente de PM_{10}, lo que es especialmente importante en los análisis de contribuciones de fuentes.

4. La matriz de probabilidades de transición de los MOM permite verificar la coherencia de las definiciones dadas a los regímenes de PM_{10} de las SSTT. Resulta de gran interés dotar de un significado a cada uno de los regímenes establecidos por esta modelización.

5. La metodología propuesta para la estimación de las intrusiones saharianas de PM_{10} mejora al método actual de referencia empleado en la Unión Europea (percentil 40 medio móvil mensual -método P40-), en tanto que: (i) elimina el caracter empírico sobre el que el método P40 descansa, por la selección del percentil y la orden de la media móvil; (ii) se apoya en una metodología especialmente indicada para el tratamiento de SSTT; (iii) evita el uso de una técnica de alisado de SSTT (media móvil); (iv) justifica el sustraendo empleado (μ_2) en el principio de la aproximación incremental de Lenschow, y (v) permite obtener un intervalo de confianza para la estimación de la carga de polvo aportada por estas intrusiones. La metodología propuesta requiere de un conocimiento de las principales fuentes de emisión de contaminantes de PM_{10} en el área de estudio.

6. La concepción de los regímenes de concentración en las SSTT se ha podido aplicar al resto de contaminantes primarios estudiados, CO, NO_2 y SO_2, permitiendo, mediante el primer régimen, determinar la concentración de fondo de estos contaminantes en áreas urbanas. Esta concentración es especialmente significativa, ya que puede ser asociada a una contaminación crónica presente en las ciudades y que padecen sus habitantes.

7. Las características de los primeros regímenes o perfiles de las SSTT permiten efectuar comparaciones entre la contaminación de fondo y ambiente en distintas áreas urbanas, así como cuantificar la contribución de la contaminación de fondo a la contaminación ambiente.

8. Los modelos empleados en esta tesis pueden ser asequiblemente implementados, bien mediante el código propio proporcionado, bien mediante librerías del entorno R. Destaca, además, su fácil interpretabilidad, razón, junto con la anterior, por la que se propone su empleo de forma rutinaria.

9

Bibliografía

Aitkin, M. y Aitkin, I. 1996. A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, 6: 127-30.

Aitkin, M. y Rubin, D. 1985. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society*, Series B (Methodological), 47(1): 67-75.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions On Automatic Control*, 19: 716-23.

Akaike, H. 1978. On the likelihood of a time series model. *The Statistician*, 27: 217-35.

Amato, F., Pandolfi, M., Escrig, A., et al. 2009. Quantifying road dust resuspension in urban environment by Multilinear Engine: a comparison with PMF2. *Atmospheric Environment*, 43: 2770-80.

Amato, F., Schaap, M., Reche, C., et al. Road Traffic: A Major Source of Particulate Matter in Europe. En: Urban Air Quality in Europe, M. Viana (ed.), *The Handbook of Environmental Chemistry* 26: 165-94, Springer-Verlag Berlin Heidelberg, 2013.

Análisis de la Calidad del Aire en España. Evolución 2001-2012. Subdirección General de la Calidad del Aire y Medio Ambiente Industrial. Ministerio de Agricultura, Alimentación y Medio Ambiente. Madrid. 2014.

Bahreini, R., Middlebrook, A.M., De Gouw, J.A., et al. 2012. Gasoline emissions dominate over diesel in formation of secondary organic aerosol mass. *Geophysical Research Letters*, 39(6): Art. no. L06805.

Baker, S.G. 1992. A simple method for computing the observed information matrix when using the EM algorithm. *Journal of Computational and Graphical Statistics*, 1: 63-76.

Baum, L.E. y Petrie, T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37: 1554-63.

Baum, L.E., Petrie, T., Soules, G., et al. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1): 164-71.

Baum, L.E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In Proc. Third Symposium on Inequalities, O. Shisha (ed.), 1-8. Academic Press, New York.

Belin, T.R., y Rubin, D.B. 1995. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90: 694-707.

Belis, C.A., Karagulian, F., Larsen, B.R., et al., 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. *Atmospheric Environment*, 69: 94-108.

Belis, C.A., Larsen, B.R., Amato, F., et al. European guide on air pollution source apportionment with receptor models. Luxembourg. Joint Research Centre Reference Reports; 2014. Report No.: EUR 26080 EN.

Blanco, M.R., Johnson-Buck, A.E. y Walter, N.G. Hidden Markov Modeling in Single-Molecule Biophysics. En: Encyclopedia of Biophysics, Gordon C. K. Roberts Eds. Springer, 2013. New York. pp. 971-5.

Bickel, P. J., Ritov, Y. y Rydén, T. 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4): 1614-35.

Biernacki, C., Celeux, G. y Govaert, G. 2000. Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22: 719-25.

Blischke, W. R. 1962. Moment estimators for the parameters of a mixture of two binomial distributions. *The Annals of Mathematical Statistics*, 33(2): 444-54.

Boulter, P.G., Borken-Kleefeld, J., y Ntziachristos, L. The Evolution and Control of NO_x Emissions from Road Transport in Europe. En: Urban Air Quality in Europe, M. Viana (ed.), *The Handbook of Environmental Chemistry* 26: 31-54, Springer-Verlag Berlin Heidelberg, 2013.

Bozdogan, H. 1993. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. En: O. Optiz, B. Lausen y R. Klar (Eds.), Information and Classification, pp. 40-54. Springer-Verlag.

Brandt, C., Kunde, R., Dobmeier, B. et al. 2011. Ambient PM_{10} concentrations from wood combustion-emission modeling and dispersion calculation for the city area of Augsburg, Germany. *Atmospheric Environment*, 45: 3466-74

Breiman, L. 2001. Random forests. *Machine Learning*, 45: 5-32.

Breiman, L. y Cutler, A. 2014. R package "randomForest". Breiman and Cutler's random forests for classification and regression. `http://stat-www.berkeley.edu/users/breiman/RandomForests`.

Bulla, J. y Berzel, A. 2008. Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, 23: 1-18.

Bulla J., Mergner S., Bulla I., et al. 2011. Markov-switching Asset Allocation: Do Profitable Strategies Exist? *Journal of Asset Management*, 12 (1): 310-21.

Cappé, O., Moulines, E. y Rydén, T. 2005. Inference in hidden Markov models. Springer, New York.

Chang, N.B. y Tseng, C.C. 1999. Optimal design of multi-pollutant air quality monitoring network in a metropolitan region using Kaohsiung, Taiwan as an example. *Environmental Monitoring and Assessment*, 57(2):121-48.

Charlier, C. 1906. Researches into the theory of probability. Hakon Ohlsson: Lund.

Charlier, C. y Wicksell, S. 1924. On the dissection of frequency functions. *Arkiv för Matematik, Astronomi och Fysik*, Bd. 18, No. 6.

CIQSO, 2012 [Internet]. Huelva: Associate Unit CSIC University of Huelva "Atmospheric Pollution", Center for Research in Sustainable Chemistry June, August 2012. [consulta: 27 de agosto de 2015]
Disponible en `http://uhuaerosol.blogspot.com.es/2012/06/`.

Cohen, A. 1967. Estimation in mixtures of two normal distributions. *Technometrics*, 9(1): 15-28.

Collaud Coen, M., Weingartner, E., Schaub, D., et al. 2004. Saharan dust events at the Jungfraujoch: Detection by wavelength dependence of the single scattering albedo and first climatology analysis. *Atmospheric Chemistry and Physics*, 4: 2465-80.

Csiszár, I. y Shields, P.C. 2000. The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28(6): 1601-19.

Day, N. 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3): 463-74.

Delmar, P., Robin, S., Tronik-Le, R., et al. 2005. Mixture model on the variance for the differential analysis of gene expression data. *Journal of the Royal Statistical Society: Series C* (Applied Statistics), 54: 31-50.

Dempster, A., Laird, N. y Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B (Methodological), 39(1): 1-38.

Derwent, R.G. 2008. New directions: prospects for regional ozone in north-west Europe. *Atmospheric Environment*, 42: 1958-60.

Derwent, R.G. y Hjellbrekke, A.-G. Air pollution by ozone across Europe. En: Urban Air Quality in Europe, M. Viana (ed.), *The Handbook of Environmental Chemistry*, 26: 55-74, Springer-Verlag Berlin Heidelberg, 2013.

Dias, J.G., Vermunt, J.K. y Ramos, S. 2010. Mixture hidden Markov models in finance research, En: Fink, A., et al. (Eds.), Advances in Data Analysis, Data Handling and Business Intelligence. Studies in Classification, Data Analysis and Knowledge Organization. Springer-Verlag, Heidelberg, pp. 451-9.

Directiva 2008/50/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008, relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa.

Draxler, R.R. y Rolph, G.D. 2003. HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory) Model Access via NOAA ARL READY Website. NOAA Air Resources Laboratory, Silver Spring, MD.

Dong, M., Yang, D., Kuang, Y., et al. 2009. $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Systems with Applications*, 36: 9046-55.

Douc, R. y Matias, C. 2001. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3): 381-420.

Du, J. 2002. Combined algorithms for fitting finite mixture distributions. Master's thesis, Mc-Master University Hamilton, Ontario.

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.

DP (Diputación de Sevilla), 2009. Anuario estadístico de la provincia de Sevilla 2010. Sevilla (España): Presidencia; 2009.

DP (Diputación de Sevilla), 2012. Anuario estadístico de la provincia de Sevilla 2012. Sevilla (España): Presidencia; 2012.

EC, 2011. Commission staff working paper establishing guidelines for demonstration and subtraction of exceedances attributable to natural sources under the Directive 2008/50/EC on ambient air quality and cleaner air for Europe. Brussels, 15.02.2011.

EEA, 2010. European Union emission inventory report 1990-2008 under the UNECE Convention on Long-range Transboundary Air Pollution (LRTAP). EEA Technical Report No 7/2010. European Environment Agency, Copenhagen.

EMEP, 2014. Transboundary particulate matter, photo-oxidants, acidifying and eutrophying components. Norwegian Meteorological Institute; EMEP Status Report 2014.

EN 12341, 1998. Air Quality - Determination of the PM_{10} fraction of suspended particulate matter - Reference method and field test procedure to demonstrate equivalence of measurement methods, 1998.

Ephraim, Y. y Merhav, N. 2002. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48: 1518-69.

E-PRTR, 2011. PRTR-España; Registro Estatal de Emisiones y Fuentes Contaminantes. `http://www.prtr-es.es`.

Escudero, M., Querol, X., Pey, A., et al. 2007a. A methodology for the quantification of the net African dust load in air quality monitoring networks. *Atmospheric Environment*, 41: 5516-24.

Escudero, M., Querol, X., Ávila, A., et al. 2007b. Origin of the exceedances of the European daily PM limit value in regional background areas of Spain. *Atmospheric Environment*, 41: 730-44.

España. Real Decreto 812/2007, de 22 de junio, sobre evaluación y gestión de la calidad del aire ambiente en relación con el arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos. Boletín Oficial del Estado, 23 de junio de 2007, núm. 150, pp. 27171-77.

España. Ley 34/2007, de 15 de noviembre, de calidad del aire y protección de la atmósfera. Boletín Oficial del Estado, 16 de noviembre de 2007, núm. 275, pp. 46962-87.

España. Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire. Boletín Oficial del Estado, 29 de enero de 2011, núm. 25, pp. 9574-626.

Etheridge, A.M., et al. 2008. Handbook of hidden Markov models in bioinformatics. Chapman & Hall/CRC. Mathematical and Computational Biology Series. Londres. 161 pp.

Everitt, B. 1996. An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, 5(2): 107-27.

Everitt, B. y Hand, D. 1981. Finite Mixture Distributions. Chapman and Hall, London.

Falls, L. 1970. Estimation of parameters in compound Weibull distributions. *Technometrics*, 12(2): 399-407.

Feller, W. 1968. An Introduction to Probability Theory, with Applications. Wiley, New York.

Ferguson, J.D. 1980. Hidden Markov analysis: an introduction. En: Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech, Ferguson, J.D., (Eds.) Princeton, NJ: IDA, Communications Research Division.

Finch, S.J., Mendel, N.R. y Thode, H.C.Jr. 1989. Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistical Association*, 84(408): 1020-3.

Flexer, A., Sykacek, P., Rezek, I., et al. 2002. An automatic, continuous and probabilistic sleep stager based on a hidden Markov model. *Applied Artificial Intelligence*, 16: 199-207.

Forney, G.D. 1973. The Viterbi algorithm. *Proceedings of the IEEE*, 61 (3): 268-78.

Fraley, C., Raftery, A.E., Murphy, T.B. et al. 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597. Department of Statistics University of Washington, Seattle, Estados Unidos.

Freedman, D. 1975. Markov Chains. Holden-Day, San Francisco.

Frühwirth-Schnatter, S. 2010. Finite Mixture and Markov Switching Models. Springer, New York.

Gómez-Losada, Á., Lozano-García, A., Pino-Mejías, R., Contreras-González, J. 2014. Finite mixture models to characterize and refine air quality monitoring networks. *Science of the Total Environment*, 485-486: 292-9.

Gómez-Losada, Á., Pires, J.C.M., Pino-Mejías, R. 2015. Time series clustering for estimating particulate matter contributions and its use in quantifying impacts from deserts. *Atmospheric Environment*, 117: 271-81.

Gómez-Losada, Á., Pires, J.C.M., Pino-Mejías, R. 2016. Characterization of background air pollution exposure in urban environments using a metric based on Hidden Markov Models. *Atmospheric Environment*, 127: 255-61.

Guedon, Y. 2007. Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computational Statistics & Data Analysis*, 51: 2379-409.

Ghahramani, Z., y Jordan, M. I. 1997. Factorial hidden Markov models. *Machine Learning*, 29: 245-73.

Grimmett, G.R. y Stirzaker, D.R. 2001. Probability and Random Processes, third edition. Oxford University Press, Oxford.

Hainsch, A. 2003. Ursachenanalyse der PM_{10}-Immission in urbanen Gebieten am Beispiel der Stadt Berlin. Dissertation. Technische Universität, Berlin.

Hasselblad, V. 1966. Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8(3): 431-44.

Hasselblad, V. 1969. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64(328): 1459-71.

Healy, M. y Westmacott, M. 1956. Missing values in experiments analysed on automatic computers. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 5(3): 203-6.

Hendriks, C., Kranenburg, R., Kuenen, J., et al. 2013. The origin of ambient particulate matter concentrations in the Netherlands. *Atmospheric Environment*, 69: 289-303.

Holgersson, M. y Jorner, U. 1978. Decomposition of a mixture into normal components: A review. *International Journal of Bio-Medical Computing*, 9(5): 367-92.

Hurvich, C.F. y Tasi, C.L. 1989. Regression and time series model selection in small samples. *Biometrika*, 76(2): 297-307.

IECA, 2012, 2015. Instituto de Estadística y Cartografía de Andalucía (IECA). Consejería de Economía, Innovación, Ciencia y Empleo. Sevilla (Spain).

IPCC, 2001. Climate Change 2001: The Scientific Basis, Cambridge University Press.

Isaacson, D. y Madsen, R. 1976. Markov Chains, Theory and Applications. Wiley, New York.

Jamshidian, M. y Jennrich, R.I. 2000. Standard errors for EM estimation. *Journal of the Royal Statistical Society*, series B, 62(2): 257-70.

Jelinek, F. 1998. Statistical Methods for Speech Recognition. MIT Press. 305 pp.

Karanasiou, A., Mihalopulos, N. Road Traffic: A Major Source of Particulate Matter in Europe. En: Urban Air Quality in Europe, M. Viana (ed.), *The Handbook of Environmental Chemistry* 26: 219-38, Springer-Verlag Berlin Heidelberg, 2013.

Karlis, D. y Xekalaki, E. 2003. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4): 577-90.

Kim, C.-J. 1994. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60: 1-22.

Krogh, A. 1998. An introduction to hidden Markov models for biological sequences. En: S. L. Salzberg, D. B. Searls y S. Kasif (Eds.), Computational methods in molecular biology (pp. 45-63). Amsterdam: Elsevier.

Langeheine, R. y Van de Pol, F. 1990. A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, 18(4): 416-41.

Latza U., Gerdes S. y Baur X. 2009. Effects of nitrogen dioxide on human health: systematic review of experimental and epidemiological studies conducted between 2002 and 2006. *International Journal of Hygiene and Environmental Health* 212, 271-87.

Lenschow, P., Abraham, H.-J., Kutzner, et al. 2001. Some ideas about the sources of PM_{10}. *Atmospheric Environment* 35, Supplement No.1: S23-S33.

Lepeule J., Bind M.A.C., Baccarelli A.A., et al. 2014. Epigenetic influences on associations between air pollutants and lung function in elderly men: the normative aging study. *Environmental Health Perspectives*, 122: 566-72.

Leroux, B. G. 1992. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications*, 40: 127-43.

Li, S., Batterman, S., Su, F. y Mukherjee, B. 2013. Addressing extrema and censoring in pollutant and exposure data using mixture of normal distributions. *Atmospheric Environment*, 77: 464-73.

Lim, S.S., Vos, T., Flaxman, A.D., et al. 2013. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380: 2224-60.

Little, R.J.A. y Rubin, D.B. 2002. Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics, Hoboken.

Mamon, R.S. y Elliot, R.J. Hidden markov models in finance. International series in operations research and management science advancing the state-of-the-art. Springer, 2007. New York. 203 pp.

Miller, G. A. 1952. Finite Markov processes in psychology. *Psychometrika*, 17: 149-67.

McCulloch, C.E. 1998. Review of "EM Algorithm and Extensions". *Journal of the American Statistical Association* 93: 403-4.

McKendrick, A. 1926. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44: 98-130.

McLachlan, G. y Basford, K. 1988. Mixture models: inference and applications to clustering. Dekker, New York.

McLachlan, G. J. y Jones, P. N. 1988. Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 44(2): 571-78.

McLachlan, G. y Krishnan, T. 1997. The EM Algorithm and Extensions. Wiley Series in Probability and Statistics, New York.

McLachlan, G. y Peel, D. 2000. Finite Mixture Models. Wiley Series in Probability and Statistics, New York.

Medgyessy, P. 1961. Decomposition of superpositions of distribution functions. Publishing House of the Hungarian Academy of Sciences, Budapest.

Meng, X.L. y Pedlow, S. 1992. EM: A bibliographic review with missing articles. *Statistical Computing Section, Prococeedings of the American Statistical Association*, 24-7.

Meng, X.L. y Rubin, D.B. 1991. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86(416): 899-909.

Mengerse, K., Robert, C. y Titterington, D. 2011. Mixtures: Estimation and Applications. Wiley Series in Probability and Mathematical Statistics. New York.

Moharir, P. 1992. Estimation of the compounding distribution in the compound Poisson process model for earthquakes. *Journal of Earth System Science*, 101(4): 347-59.

Murray, G.D. 1977. Contribution to discussion of paper by A.P. Dempster, N.M. Laird and D.B. Rubin. *Journal of the Royal Statistical Society*, Series B, 39: 23-4.

Nagl, C., Ansorge, C., Moosmann, L., et al. Critical Areas for Compliance with PM_{10} and NO_2 Limit Values in Europe. En: Urban Air Quality in Europe, M. Viana (ed.). *The Handbook of Environmental Chemistry*, 26: 3-30, Springer-Verlag Berlin Heidelberg, 2013.

Newcomb, S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4): 343-66.

Orchard, T. y Woodbury, M. 1972. A missing information principle: theory and applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistic and Probability*, 1: 697-715.

OMS, 2002. Air quality guidelines for Europe, Second edition. World Health Organization Regional Office for Europe. WHO Regional Publications, European Series, No. 91.

OMS, 2006. Guías de calidad del aire de la OMS relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre. Actualización mundial 2005. Resumen de evaluación de los riesgos. Organización Mundial de la Salud.

Ozone Position Paper. 1999. Office for Official Publications of the European Communities, Luxemburgo.

Pérez, N., Querol, X., Alastuey, A., et al. 2014. Episodios Naturales de Partículas 2013. CSIC, AEMet, Ministerio de Agricultura, Alimentación y Medio Ambiente-Subdirección General de Calidad del Aire y Medio Ambiente Industrial; Abril 2014.

Pey, N., Querol, X., Alastuey, A., et al. Episodios Naturales de Partículas 2010. CSIC, AEMet, Ministerio de Medio Ambiente y Medio Rural y Marino-Subdirección General de Calidad del Aire y Medio Ambiente Industrial; Marzo 2011. `http://bit.ly/1CWTHRW`.

Pey, N., Pérez, N., Querol, X., et al. 2013. Episodios Naturales de Partículas 2012. CSIC, AEMet, Ministerio de Medio Ambiente y Medio Rural y Marino-Subdirección General de Calidad del Aire y Medio Ambiente Industrial; Mayo 2013.

Pires, J.C.M., Sousa, S.I.V., Pereira, M.C., et al. 2008. Management of air quality monitoring using principal component and cluster analysis-Part I: SO_2 and PM_{10}. *Atmospheric Environment*, 42: 1249-60.

Pollice, A. y Lasinio, G. J. 2009. Two approaches to imputation and adjustment of air quality data from a composite monitoring network. *Journal of Data Science*, 7:43-59.

Poritz, A.B. 1988. Hidden Markov models: a guided tour. Institute for Defense Analyses, Communications Research Division. Princenton, Estados Unidos.

Querol, X., Pey, J., Pandolfi, M., et al. 2009. African dust contributions to mean ambient PM_{10} mass-levels across the Mediterranean Basin. *Atmospheric Environment* 43, 4266-77.

Querol, X., Alastuey, A., Pey, J., et al. 2013a. Procedimiento para la identificación de episodios naturales de PM_{10} y $PM_{2.5}$, y la demostración de causa en lo referente a las superaciones del valor límite diario de PM_{10}. Instituto de Diágnostico Ambiental y Estudios del Agua (IDAEA), CSIC, Universidad Nova de Lisboa, AEMet-Izaña, CIEMAT, Universidad de Huelva. Ministerio de Medio Ambiente, Medio Rural y Marino, Ministério Do Ambiente, Ordenamiento Do Territorio e Desenvolvimiento Regional (Portugal), Agência Portuguesa do Ambiente (Portugal); Abril, 2013.

Querol, X., Alastuey, A., Pey, J., et al. 2013b. Procedimiento para la identificación de episodios naturales de PM_{10} y $PM_{2.5}$, y la demostración de causa en lo referente a las superaciones del valor límite diario de PM_{10}. Instituto de Diágnostico Ambiental y Estudios del Agua (IDAEA), CSIC, Universidad Nova de Lisboa, AEMet-Izaña, CIEMAT, Universidad de Huelva. Ministerio de Medio Ambiente, Medio Rural y Marino, Ministério Do Ambiente, Ordenamiento Do Territorio e Desenvolvimiento Regional (Portugal), Agência Portuguesa do Ambiente (Portugal); Abril, 2013.

Querol, X., Viana, M.M., Alastuey, A., et al. 2013c. Niveles de PM_{10} y $PM_{2.5}$ en España: Aragón, Asturias, Castilla La Mancha, y Madrid. Instituto de Diágnostico Ambiental y Estudios del Agua (IDAEA-CSIC), CIEMAT, Instituto de Salud Carlos III, Ministerio de Agricultura, Alimentación y Medio Ambiente, S.D.G. de Calidad del Aire y Medio Ambiente Industrial. Abril, 2013.

Quass, U., John, A.C., y Kuhlbusch, T.A.J. Souce apportionment of airborne dust in Germany: methods ans results. En: Urban Air Quality in Europe, M. Viana (ed.), *The Handbook of Environmental Chemistry*, 26: 195-218, Springer-Verlag Berlin Heidelberg, 2013.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the Institute of Electrical and Electronics Engineers*, 77(2): 267-95.

Rao, C. 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society*, Series B (Methodological), 10(2): 159-203.

Redner, R. y Homer, W. F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2): 195-239.

REVIHAAP Project, 2013. Review of Evidence on Health Aspects of Air Pollution, REVIHAAP Project: Technical Report, 2013. World Health Organization. Regional Office for Europe, p. 309.

Rydén, T., Teräsvirta, T. y Åsbrink, S. 1998. Stylized facts of daily returns series and the hidden Markov model. *Journal of Applied Econometrics*, 13: 217-244.

Rydén, T. 1995a. Estimating the order of hidden Markov models. Statistics. *A Journal of Theoretical and Applied Statistics*, 26(4): 345-54.

Rydén, T. 1995b. Consistent and asymptotically normal parameter estimates for Markov modulated Poisson processes. *Scandinavian Journal of Statistics*, 22(3): 295-303.

Salvador P., Artíñano B., Viana M.M., et al. 2015. Multicriteria approach to interpret the variability of the levels of particulate matter and gaseous pollutants in the Madrid metropolitan area, during 1999-2012 period. *Atmospheric Environment*, 109: 205-16.

Schlattmann, P. 2009. Medical Applications of Finite Mixture Models. Springer, Berlin Heidelberg.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461-4.

Seidel, W., Mosler, K. y Alker, M. 2000. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52(3): 481-7.

Segal, M.R., Bacchetti, P. y Jewell, N.P. 1994. Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society*, Series B: 56: 345-52.

SG (Sevilla Global), 2012. Barómetro de economía urbana 2012. Agencia Urbana de Desarrollo Integral. Sevilla (España): Ayuntamiento de Sevilla; 2012.

Stein A.F., Isakov V., Godowitch J., et al. 2007. A hybrid modelling approach to resolve pollutant concentrations in an urban area. *Atmospheric Environment*, 41: 9410-26.

Steyn, D.G. 2014. Air Pollution Modeling and its Application XXII. NATO Science for Peace and Security Series C: Environmental Security. Steyn, D.G., Builtjes, P.J.H., Timmermans, R.M.A. (eds.)

Simar, L. 1976. Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, 4(6): 1200-9.

Sunyer J., Basagana X., Belmonte J., et al. 2002. Effect of nitrogen dioxide and ozone on the risk of dying in patients with severe asthma. *Thorax*, 57: 687-93.

Tan, W. Y. y Chang, W. C. 1972. Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association*, 67(339): 702-8.

Tanner, M.A. 1996. Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions, 3ª edición. Springer Series in Statistics, New York.

Teicher, H. 1961. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1): 244-8.

Teicher, H. 1963. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4): 1265-9.

Thorpe, A.J. y Harrison, R.M. 2008. Sources and properties of non-exhaust particulate matter from road traffic: a review. *Science of the Total Environment*, 400: 270-82.

Timmermann, A. 2000. Moments of Markov switching models. *Journal of Econometrics*, 96: 75-111.

Titterington, D., Smith, A. y Makov, U. 1985. Statistical Analysis of finite mixture distributions. John Wiley & Sons, Chichester.

Unión Europea. Directiva 2004/107/CE del Parlamento Europeo y del Consejo de 15 de diciembre de 2004 relativa al arsénico, el cadmio, el mercurio, el níquel y los hidrocarburos aromáticos policíclicos en el aire ambiente. Diario Oficial de la Unión Europea L 23/3, 26 de enero de 2005, pp. 3-16.

Unión Europea. Directiva 2008/50/CE del Parlamento Europeo y del Consejo de 21 de mayo de 2008 relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa. Diario Oficial de la Unión Europea L 152/1, 11 de junio de 2008, pp. 1-44.

Vaidyanathan, A., Dimmick, W.F., Kegler, S.R, et al. 2013. Statistical air quality predictions for public health surveillance: evaluation and generation of county level metrics of $PM_{2.5}$ for the environmental public health tracking network. *International Journal of Health Geographics*, 2: 1-12.

Vallero, D.A. 2014. Fundamentals of Air Pollution, 5ª ed. Academic Press.

Vedal S., Brauer M., White R., et al. 2003. Air pollution and daily mortality in a city with low levels of pollution. *Environmental Health Perspectives* 111, 45-51.

Viana, M., Kuhlbusch, T.A.J., Querol, X., et al. 2008. Source apportionment of particulate matter in Europe: A review of method and results. *Journal of Aerosol Science* 39: 827-49.

Viana, M. 2013. Urban Air Quality in Europe, M. Viana (ed.). *The Handbook of Environmental Chemistry 26*. Springer-Verlag Berlin Heidelberg.

Viana, M., Pey, J., Querol, X., et al. 2014. Natural sources of atmospheric aerosols influencing air quality accross Europe. *Science of the Total Environment*, 472: 825-33.

Visser, I., Raijmakers, M.E.J., Molenaar, P.C.M. 2002. Fitting hidden Markov models to psychological data. *Scientific Programming*, 10: 185-99.

Visser, I., Raijmakers, M.E.J., Maas, H.L.J. 2009. Hidden Markov models for individual time series. En: Valsiner, J., Molenaar, P.C.M., Lyra, M.C.D.P., Chaudhary, N. (Eds.), Dynamic process methodology in the Social and Developmental Sciences. Springer, Heidelberg, pp. 269-89.

Visser, I. 2011. Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology*, 55: 403-15.

Visser, I. y Speekenbrink, M. 2010. depmixS4: an R-package for hidden Markov models. *Journal of Statistical Software*, 36 (7): 1-21.

Viterbi, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 (2): 260-9.

Warneck, P. 1999. Chemistry of the natural atmosphere. Academic, San Diego.

Welch, L.R. 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* 53 (4): 10-3.

Wickens, T. D. 1982. Models for behavior: Stochastic processes in psychology. San Francisco: W. H. Freeman. 384 pp.

Wilks DS. 2006. Statistical methods in the atmospheric sciences. 2ª ed. Academic Press. 996 pp.

Wu, C. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1): 95-103.

Yakowitz, S.J. y Spragins, J.D. 1968. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1): 209-14.

Zucchini, W. y MacDonald, I.L. 2009. Hidden Markov Models for Time Series. An Introduction Using R. Monographs on Statistics and Applied Probability 110. CRC Press, 2009. Florida, 278 pp.

Anexos

A

Implementación computacional de los modelos de mixturas finitas

A lo largo de esta sección, se explican las funciones diseñadas en R con las que se ha implementado la modelización mediante mixturas finitas utilizando el algoritmo EM. Se acompaña un ejemplo de uso en cada una de ellas.

A.1. Obtención de los valores iniciales para el algoritmo EM

Dado un conjunto de datos inicial (`muestra`), la función obtiene la media y desviación típica del número de particiones (`nc`) equidistantes realizado en el conjunto de datos. Su resultado, los valores iniciales del algoritmo EM (`vi.Q`), son almacenados como una variable global, de tal forma que quedan disponibles en el entorno para ser utilizados por una siguiente función.

```
v.i<-function(muestra,nc) {

     partes<-seq(0,1,1/nc)
     intervalos<-quantile(muestra, prob=partes)
     elem<-findInterval(muestra, intervalos, all.inside=TRUE)
     medias<-desviaciones<-numeric(nc)

     for (i in 1:nc) {
     cluster<-muestra[elem==i]
     medias[i]<-mean(cluster)
     desviaciones[i]<-sd(cluster)
}

     pi<-rep(1/nc,nc)
     vi.Q<<-c(pi, medias,desviaciones)
     vi.Q
}
```

Uso:

```
set.seed(123)
cluster.1<-rnorm(100,10,3) # mu1:  10.271218, sd1: 2.738448
cluster.2<-rnorm(100,30,4) # mu2:  29.569813, sd2: 3.867946
cluster.3<-rnorm(100,50,5) # mu3:  50.602326, sd3: 4.749395
```

```
datos<-c(cluster.1, cluster.2, cluster.3)

> v.i(datos,3)
[1]  0.3333333  0.3333333  0.3333333 10.2712177 29.5523448 50.6197935  2.7384476  3.8103611  4.7176350
```

La secuencia de valores obtenidos se corresponden, para la mixtura de 3 componentes creada, con los de $\hat{\Psi}^{(0)} = \{\pi_1^{(0)}, \pi_2^{(0)}, \pi_3^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}, \sigma_3^{(0)}\}$.

A.2. Función de log-verosimilitud

`logLik` calcula el valor de la función de log-verosimilitud, dados un conjunto de datos (`muestra`) y los valores de los estimadores de los parámetros de la mixtura (`r.EM`). Los objetos `ind.1` y `ind.2` representan las posiciones en $\hat{\Psi}^{(t)}$ de las medias (`ind.1`) y las desviaciones típicas (`ind.2`). Reproduce el cálculo de la expresión (2.10).

```
logLik<-function(muestra, r.EM) {

        nc<-length(r.EM)/3
        ind.1<-seq(nc+1,3*nc)
        ind.2<-ind.1+nc
        f<-length(muestra)
        pesos<-r.EM[1:nc]
        m<-matrix(NA,f,nc)

        for (i in 1:nc){
        m[,i]<-dnorm(muestra,r.EM[ind.1[i]],r.EM[ind.2[i]])
        }

        resultado<-sum(apply(m%*%pesos,2,log))
        resultado
}
```

Uso:

```
> logLik(datos, vi.Q)
[1] -1139.322
```

A.3. Criterios de información

- Bayesiano (BIC)

  Calcula el valor del estadístico BIC a partir de los argumentos explicados en las funciones `v.i` y `logLik`. Utiliza el número de parámetros independientes de la mixtura (`n.par.indep`).

  ```
  BIC<-function(muestra, r.EM){

       n.par.indep<-length(r.EM)-1
       2*logLik(muestra, r.EM) - (n.par.indep*log(length(muestra)))
  }
  ```

  Uso:

  ```
  > BIC(datos, vi.Q)
  [1] -2324.273
  ```

- Akaike (AIC)

  Calcula el valor del estadístico AIC.

```
AIC<-function(muestra, r.EM){

    n.par.indep<-length(r.EM)-1
    -2*logLik(muestra, r.EM) + (2*n.par.indep)
}
```

Uso:

```
> AIC(datos, vi.Q)
[1] 2294.643
```

- Akaike corregido (AIC_c)
  Calcula el valor del estadístico AICc.

```
AIC.c<-function(muestra, r.EM){

       n.par.indep<-length(r.EM)-1
       n<-length(muestra)
       AIC<-(-2*logLik(muestra, r.EM) + (2*n.par.indep))
       num_penalizacion<-2*n.par.indep+1*(n.par.indep+1)
       dem_penalizacion<-(n-n.par.indep-1)
       penalizacion<-num_penalizacion/dem_penalizacion
       AIC+penalizacion
}
```

Uso:

```
> AIC.c(datos, vi.Q)
[1] 2294.729
```

- Integrated classification likelihood (ICL)
  Calcula el valor del estadístico ICL, utilizando un objeto array (`m`) en donde se almacena toda la información necesaria. En la primera capa, las densidades (`m[,i,1]`); en la segunda, las *responsabilites* (`m[,,2]`), y en la tercera, el clúster al que pertenece cada observación (`m[i,1,3]`) y sus probabilidades de asignación (`m[i,2,3]`). El objeto `penalty` almacena la penalización del criterio BIC.

```
ICL<-function(data, param) {

     nc<-length(param)/3
     ind.1<-seq(nc+1, 2*nc)
     ind.2<-ind.1+nc
     pesos<-param[-c(ind.1,ind.2)]
     f<-length(data)
     m<-array(NA,dim=c(f,nc,3))

     for (i in 1:nc){
     m[,i,1]<-pesos[i]*dnorm(data,param[ind.1[i]],param[ind.2[i]])
     }

     m[,,2]<-m[,,1]/apply(m[,,1],1,sum)

     for (i in 1:f){
     m[i,1,3]<-which.max(m[i,,2])
     m[i,2,3]<-m[i,,2][which.max(m[i,,2])]
     }

     # m<<-m

     penalty<-0

     for (i in 1:nc){
     termino<-subset(m[,,3], m[,1,3]==i, select=2)
     termino<-as.vector(termino)
     penalty<-penalty+sum(log(termino))
     }

     penalty<<-penalty
     resultado<-BIC(data,param)+2*penalty
     return(resultado)
}
```

Uso:

```
> ICL(datos,vi.Q)
[1] -2327.473
```

A.4. Desarrollo de una iteración del algoritmo EM

El propósito de la siguiente función es su anidamiento con la función principal del algoritmo EM (siguiente función `EM`). En el array `m` se almacenan todos los cálculos intermedios. `iter` calcula una iteración del algoritmo, dados los datos de entrada (`data`) y $\hat{\Psi}^{(t)}$ (`param`). En este caso, el ejemplo que se muestra corresponde a la obtención de $\hat{\Psi}^{(1)}$, dado que `param`= $\hat{\Psi}^{(0)}$. En la línea de código 12, se calcula la expresión (2.17); en la 13, la (2.25); (2.28) en la 14, y (2.29) en la 18. Esta función está diseñada para efectuar el cálculo de la primera iteración sobre una mixtura de cualquier g.

```
iter<-function(data, param) {                                                    1

      nc<-length(param)/3                                                        2
      ind.1<-seq(nc+1, 2*nc)                                                     3
      ind.2<-ind.1+nc                                                            4
      pesos<-param[-c(ind.1,ind.2)]                                              5
      f<-length(data)                                                            6
      m<-array(NA,dim=c(f,nc,3))                                                 7

      for (i in 1:nc){                                                           8
      m[,i,1]<-pesos[i]*dnorm(data,param[ind.1[i]],param[ind.2[i]])              9
      }                                                                          10

     den<-apply(m[,,1],1,sum)                                                    11

     m[,,2]<-m[,,1]/den # Bayes                                                  12

     param[-c(ind.1,ind.2)]<-apply(m[,,2],2,mean)                                13
     param[ind.1]<-apply(m[,,2]*data,2,sum)/apply(m[,,2],2,sum)                  14

     m[,,3]<-outer(data,param[ind.1],"-")                                        15
     m[,,3]<-m[,,3]^2                                                            16

     param[ind.2]<-apply(m[,,2]*m[,,3],2,sum)/apply(m[,,2],2,sum)                17
     param[ind.2]<-sqrt(param[ind.2]) # sd                                       18

     param                                                                       19
}                                                                                20
```

Uso:

```
> iter(datos,vi.Q)
[1]  0.3330608  0.3270753  0.3398639 10.2664917 29.3481776 50.4006326  2.7208532  3.5833521  4.9106148
```

La secuencia de valores obtenidos se corresponde, para la mixtura de 3 componentes creada y en esta primera iteración, con los de $\hat{\Psi}^{(1)} = \{\pi_1^{(1)}, \pi_2^{(1)}, \pi_3^{(1)}, \mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}, \sigma_3^{(1)}\}$.

A.5. Algoritmo EM

`EM` efectúa el desarrollo iterativo propiamente dicho del algoritmo a partir de un conjunto de datos (`datos`), $\Psi^{(t)}$ (`vi`), y un valor de ϵ prestablecido (`eps`). En su implementación incorpora a las funciones `logLik` y `iter` anteriores. `err` almacena el cálculo de la expresión (2.30), por lo que se le asigna inicialmente el valor 1.

Los objetos `n.param` se corresponden con el número de parámetros de la mixtura; `c`, con el número de componentes, e `iteraciones` almacena todas las iteraciones del algoritmo hasta su convergencia final, necesarias, por una parte, para el cálculo posterior de los errores SEM, y por otra, para evaluar la convergencia del algoritmo si así se requiriese. El resultado final de las estimaciones de los parámetros de la mixtura y el número de iteraciones + 1 ocurridas se devuelve en un formato de lista.

```
EM<-function(datos,vi, eps=1/1000000){                                                                    1

    n.param<-length(vi)                                                                                   2
    iteraciones<<-matrix(rep(NA,n.param), nrow=1)                                                         3
    c<-n.param/3                                                                                          4

    nuevos.p<-vi                                                                                          5
    err<-1                                                                                                6
    iter<-1                                                                                               7

    while(err>eps) {                                                                                      8

    vi<-iter(data=datos,param=vi)                                                                         9
    antig.p<-nuevos.p                                                                                     10
    nuevos.p<-vi                                                                                          11

    iteraciones<<-rbind(iteraciones, nuevos.p)                                                            12

    err<-abs((logLik(datos, antig.p)-logLik(datos, nuevos.p))/logLik(datos, nuevos.p))                    13

    iter<-iter+1                                                                                          14

    }                                                                                                     15

    r.EM<<-vi                                                                                             16

    return (list(componentes=c,p.EM=c(vi[1:c]),                                                           17
            mu.EM=c(vi[(c+1):(2*c)]),                                                                     18
            sd.EM=c(vi[(2*c+1):(3*c)]), iter=iter))                                                       19
}                                                                                                         20
```

El blucle while se interrumpe (líneas 8 a 15) cuando, a lo largo de las iteraciones, la diferencia absoluta relativa (`err`) entre la función de log-verosimilitud parametrizada según `antig.p` y `nuevos.p` es mayor que la tolerancia asumida (ϵ), devolviendo en ese caso el resultado de $\hat{\Psi}^{(6)}$, indicando mediante "†" en el siguiente ejemplo.

Uso:

```
> EM(datos,vi.Q)
$componentes
[1] 3

$p.EM
[1] 0.3332530 0.3196722 0.3470748

$mu.EM
[1] 10.26989 29.16179 50.14222

$sd.EM
[1] 2.723816 3.354494 5.179383

$iter
[1] 7
```

```
> iteraciones
              [,1]      [,2]      [,3]     [,4]     [,5]     [,6]     [,7]     [,8]     [,9]
                NA        NA        NA       NA       NA       NA       NA       NA       NA
nuevos.p 0.3330608 0.3270753 0.3398639 10.26649 29.34818 50.40063 2.720853 3.583352 4.910615
nuevos.p 0.3331725 0.3237826 0.3430449 10.26844 29.26456 50.28866 2.722516 3.475382 5.026074
nuevos.p 0.3332177 0.3218843 0.3448981 10.26925 29.21704 50.22175 2.723239 3.418273 5.095954
```

```
nuevos.p 0.3332373 0.3207569 0.3460058 10.26960 29.18882 50.18140 2.723558 3.385458 5.138234
nuevos.p 0.3332474 0.3200800 0.3466726 10.26978 29.17193 50.15699 2.723724 3.366066 5.163858
nuevos.p 0.3332530 0.3196722 0.3470748 10.26989 29.16179 50.14222 2.723816 3.354494 5.179383  †
```

Para la validación de resultados obtenidos de esta implementación, se efectuó su comparación con los de la función Mclust, sin obtener diferencias significativas al respecto:

Comparación de los resultados de la implementación con Mclust, con g especificado (`G=3`):

```
require(mclust)
modelo<-Mclust(datos, modelNames="V", G=3)
summary(modelo, parameters=TRUE)

----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust V (univariate, unequal variance) model with 3 components:

 log.likelihood   n df       BIC       ICL
      -1137.624 300  8 -2320.878 -2326.321

Clustering table:
  1   2   3
100  96 104

Mixing probabilities:
        1         2         3
0.3332530 0.3196725 0.3470745

Means:
       1        2        3
10.26989 29.16180 50.14223

Variances:
        1         2         3
 7.419175 11.252671 26.825923
```

A.6. Obtención de los errores de $\hat{\Psi}$ mediante *bootstrap*

Con esta función se obtienen los valores $\hat{\Psi}$ a través del procedimiento descrito en la sección 2.5.2., con $B = 1000$, así como sus errores. Utiliza la función de R, sample, para la generación de las muestras bootstrap, y las funciones `v.i` y `EM` descritas con anterioridad. Todos los resultados de estimación de $\hat{\Psi}$ sobre cada muestra bootstrap se almacenan en la matriz `m`.

```
errores.boots<-function(sample, nc, rep){

               num.param<-3*nc
               m<-matrix(NA,nrow=rep, ncol=num.param, byrow=TRUE)

               for (i in 1:rep){
               muestra.boot<-sample(sample,replace=TRUE)
               valores.ini<-v.i(muestra.boot,nc)
               EM(muestra.boot, valores.ini)
               m[i,]<-r.EM
               }

               param.medios<-apply(m,2,mean)
               errores.medios<-apply(m,2,sd)

               return (list(componentes=nc,pis.boot=c(param.medios[1:nc]),
                       mus.boot=c(param.medios[(nc+1):(2*nc)]),
                       sigmas.boot=c(param.medios[(2*nc+1):(3*nc)]),
```

```
                    errores.pis.boot=c(errores.medios[1:nc]),
                    errores.mus.boot=c(errores.medios[(nc+1):(2*nc)]),
                    errores.sigmas.boot=c(errores.medios[(2*nc+1):(3*nc)])))
}
```

Uso:

```
> errores.boots(sample=datos,nc=3,rep=1000)
$componentes
[1] 3

$pis.boot
[1] 0.3344449 0.3193336 0.3462215

$mus.boot
[1] 10.26426 29.16339 50.16348

$sigmas.boot
[1] 2.701940 3.336825 5.114673

$errores.pis.boot
[1] 0.02737678 0.02762136 0.02815098

$errores.mus.boot
[1] 0.2717807 0.3765336 0.5633902

$errores.sigmas.boot
[1] 0.1888861 0.2829260 0.4245305
```

Los nombres de los componentes de la lista que muestran los errores bootstrap son `errores.pis.boot`, `errores.mus.boot` y `errores.sigmas.boot`, en referencia a $\hat{\pi}_i$, $\hat{\mu}_i$ y $\hat{\sigma}_i$, con $i = 1, \ldots, 3$. Compárense los resultados de $\hat{\Psi}$ (`pis.boot`, `mus.boot`, `sigmas.boot`) con los obtenidos mediante la implementación del algoritmo EM (función `EM`).

A.7. Obtención de los errores de $\hat{\Psi}$ mediante el método SEM

La función final que los calcula, `SEM`, precisa de la utilización de tres funciones previas: `param_ik`, `r` y `stop`. `r` anida a `param_ik` y `stop` a `r`, por lo que el tiempo computacional empleado en la obtención de estos errores es superior a todas las funciones empleadas hasta ahora.

- `param_ik`

  A partir del historial de iteraciones del algoritmo, obtiene la expresion (2.32). Como se aprecia en el ejemplo de uso, únicamente se considera la evolución de uno solo de los estimadores de $\hat{\Psi}$, en este caso $\hat{\mu}_1$, permaneciendo el resto fijos. La línea 4 del código obtiene el valor de $\hat{\Psi}$ a partir del historial de iteraciones `iteraciones`.

```
param_ik<-function(i) {                                                    1

          n.filas<-nrow(iteraciones)                                       2
          n.colum<-ncol(iteraciones)                                       3
          resultados.iteraciones<-iteraciones[n.filas,]                    4

          matriz<-iteraciones                                              5

          matriz.ficticia<-matrix(0, n.filas, n.colum)                     6

          for(j in 1:n.filas) {                                            7
          matriz.ficticia[j,i]<-matriz[j,i]                                8
          matriz.ficticia[j,-i]<-resultados.iteraciones[-i]                9
          }                                                                10

          return(matriz.ficticia)                                          11
}                                                                          12
```

Uso:

```
> param_ik(4)
         [,1]      [,2]      [,3]     [,4]     [,5]     [,6]     [,7]     [,8]     [,9]
[1,] 0.333253 0.3196722 0.3470748 10.26649 29.16179 50.14222 2.723816 3.354494 5.179383
[2,] 0.333253 0.3196722 0.3470748 10.26844 29.16179 50.14222 2.723816 3.354494 5.179383
[3,] 0.333253 0.3196722 0.3470748 10.26925 29.16179 50.14222 2.723816 3.354494 5.179383
[4,] 0.333253 0.3196722 0.3470748 10.26960 29.16179 50.14222 2.723816 3.354494 5.179383
[5,] 0.333253 0.3196722 0.3470748 10.26978 29.16179 50.14222 2.723816 3.354494 5.179383
[6,] 0.333253 0.3196722 0.3470748 10.26989 29.16179 50.14222 2.723816 3.354494 5.179383
```

- `r`

Obtiene el valor de la expresión (2.33). En el ejemplo de uso, se obtiene el valor de $r_{3,2}$.

```
r<-function(i,j, muestra) {

   matriz.ficticia<-param_ik(i)
   n.filas<-nrow(matriz.ficticia)-1
   convergencias<-numeric(n.filas)

   for (k in 1:n.filas){
   num<-iter(data=muestra,matriz.ficticia[k,])[j]-(r.EM)[j]
   den<-matriz.ficticia[k,i]-(r.EM[i])
   convergencias[k]<-num/den
   }

   return(convergencias)
}
```

Uso:

```
> r(3,2, datos)
[1] 0.02005329 0.04702889 0.09900547 0.21606495 0.59737938
```

- `stop`

Representa una modificación del método SEM, ya que obtiene el valor de r_{ij} en lugar de como se propone en (2.34), como la menor diferencia entre los valores de la secuencia $r_{ij}^{(k)}, r_{ij}^{(k+1)}, r_{ij}^{(k+2)}, \ldots$. Así, se consigue no tener que aplicar diferentes ϵ según los pares de valores i, j de $r_{i,j}$ para obtener la convergencia de $r_{ij}^{(k)}$. En historiales de iteración reducidos del algoritmo EM, como el que se plantea en el ejemplo, con tan solo 6 iteraciones, si el ϵ propuesto es muy laxo, puede no obtenerse ningún $r_{i,j}^{(k)}$ por ausencia de convergencia.

```
stop<-function(i,j,muestra) {

      conv<-r(i,j, muestra)
      elegido<-which(abs(diff(conv))==min(abs(diff(conv))))
      conv[elegido+1]
}
```

Uso:

```
> stop(3,2, datos)
[1] 0.04702889
```

- `SEM`

Esta función conjuga las anteriormente descritas a través de la función `stop` (línea 30 del código) para la obtención de la matriz DM, implementando el resto de cálculos necesarios de la expresión (2.31). Los elementos de la matriz I_{oc} cuadrada (línea 15) se insertan a través de sus índices mediante el bucle de las líneas 16 a 23.

```
SEM<-function(muestra, param) {                                    1

     nc<-length(param)/3                                           2
     nparam<-length(param)                                         3
     ind.1<-seq(nc+1, 2*nc)                                        4
     ind.2<-ind.1+nc                                               5
```

```
    pesos<-param[-c(ind.1,ind.2)]                                                                    6
    f<-length(muestra)                                                                               7
    m<-array(NA,dim=c(f,nc,2))                                                                       8

    for (i in 1:nc){                                                                                 9
    m[,i,1]<-pesos[i]*dnorm(muestra,param[ind.1[i]],param[ind.2[i]])                                 10
    }                                                                                                11

    denominador<-apply(m[,,1],1,sum)                                                                 12
    m[,,2]<-m[,,1]/denominador                                                                       13
    suma.Ez<-apply(m[,,2],2,sum)                                                                     14
    Ioc<-matrix(0,nparam,nparam)                                                                     15

    for (i in 1:nc){                                                                                 16
       Ioc[i,i]<-(1/param[i]^2)*suma.Ez[i]                                                           17
       Ioc[i+nc,i+nc]<-(1/param[i+2*nc]^2)*suma.Ez[i]                                                18
       Ioc[i+2*nc,i+2*nc]<-((-1/(param[i+2*nc]^2))*suma.Ez[i]) +                                     19
       ((1/param[i+2*nc]^4)*(sum(m[,i,2]*3*(muestra-param[i+nc])^2)))                                20

       Ioc[i+nc,i+2*nc]<-(2/param[i+2*nc]^3)*(sum(m[,i,2]*(muestra-param[i+nc])))                    21
       Ioc[i+2*nc,i+nc]<-(2/param[i+2*nc]^3)*(sum(m[,i,2]*(muestra-param[i+nc])))                    22
       }                                                                                             23

       v.i(muestra,nc)                                                                               24
       EM(muestra,vi.Q)                                                                              25
       iteraciones<<-iteraciones[-1,]                                                                26

       dm<-matrix(0,nparam,nparam)                                                                   27

       for(i in 1:nparam) {                                                                          28
               for (j in 1:nparam) {                                                                 29
               dm[i,j]<-stop(i,j, muestra)                                                           30
       }                                                                                             31
       }                                                                                             32

       identidad<-matrix(0,nparam,nparam)                                                            33
       diag(identidad)<-1                                                                            34

       Ioc.inv<-solve(Ioc)                                                                           35
       inv.dm<-solve(identidad-dm)                                                                   36
       AV<-Ioc.inv%*%dm%*%inv.dm                                                                     37
       V.COV<-Ioc.inv+AV                                                                             38
       errores.SEM<-sqrt(diag(V.COV))                                                                39

       return(errores.SEM)                                                                           40
}                                                                                                    41
```

Uso:

```
> SEM(datos, r.EM)
[1] 0.03254683 0.03180168 0.03313496 0.26613849 0.35582356 0.51770878 0.18846662 0.25985883 0.39235493
```

Estos resultados se corresponden con los errores de $\hat{\pi}_i$, $\hat{\mu}_i$ y $\hat{\sigma}_i$, con $i = 1, \ldots, 3$. Compárense con los obtenidos mediante el método bootstrap (función `errores.boot`) y sus resultados `errores.pis.boot`, `errores.mus.boot` y `errores.sigmas.boot`.

A.8. Obtención de la incertidumbre de asignación de y_j

La función `cuantiles` determina las incertidumbres de pertenencia de la observación muestral $(y_j, i = 1, \ldots, n)$ a su componente asignada de la mixtura, según las expresiones (2.17) y (2.18). Como resultado, devuelve en un conjunto de g listas (línea 2 del código), el número de observaciones asignadas a cada componente (`$n_clusters`) y los cuantiles de las incertidumbres (`$cuantiles_de_incertidumbres`) en cada una de ellas. Como en otras funciones anteriores, toda la información intermedia de los cálculos se almacena en un objeto array (`m`, línea 8).

```
cuantiles<-function(data, param) {

           probs.en.clusters<-vector("list")

           nc<-length(param)/3
           ind.1<-seq(nc+1, 2*nc)
           ind.2<-ind.1+nc
           pesos<-param[-c(ind.1,ind.2)]
           f<-length(data)
           m<-array(NA,dim=c(f,nc,3))

           for (i in 1:nc){
           m[,i,1]<-pesos[i]*dnorm(data,param[ind.1[i]],param[ind.2[i]])
           }

           m[,,2]<-m[,,1]/apply(m[,,1],1,sum)

           for (i in 1:f){
           m[i,1,3]<-which.max(m[i,,2])
           m[i,2,3]<-m[i,,2][which.max(m[i,,2])]
            }

          #m<<-m

          for (i in 1:nc){
          incertidumbres<-1-subset(m[,,3], m[,1,3]==i, select=2)
          incertidumbres<-as.vector(incertidumbres)
          probs.en.clusters[[i]]<-list(n_clusters=length(incertidumbres),
                                cuantiles_de_incertidumbres=quantile(incertidumbres))
          }

          return(probs.en.clusters)
}
```

Uso:

```
> cuantiles(datos,r.EM)
[[1]]
[[1]]$n_clusters
[1] 100

[[1]]$cuantiles_de_incertidumbres
          0%          25%          50%          75%         100%
1.287859e-14 5.167107e-10 1.299451e-08 4.968867e-07 2.697498e-03


[[2]]
[[2]]$n_clusters
[1] 96

[[2]]$cuantiles_de_incertidumbres
          0%          25%          50%          75%         100%
4.392749e-06 1.315026e-05 6.921459e-05 5.909122e-04 4.710164e-01


[[3]]
[[3]]$n_clusters
[1] 104

[[3]]$cuantiles_de_incertidumbres
          0%          25%          50%          75%         100%
0.000000e+00 5.354051e-14 2.349214e-10 7.465486e-08 3.998059e-01
```

A.9. Cálculo de los momentos de la mixtura

La función `momentos` calcula los momentos de la mixtura, segun las expresiones (2.7).

```
momentos<-function(parametros){

        nc<-length(parametros)/3
        ind.1<-seq(nc+1,3*nc)
        ind.2<-ind.1+nc
        pesos<-r.EM[1:nc]

        mu<-0
        sigma.0<-0

        for (i in 1:nc){
        mu<-mu+pesos[i]*r.EM[ind.1[i]]
        sigma.0<- sigma.0 + (pesos[i]*(r.EM[ind.1[i]]^2+r.EM[ind.2[i]]^2))
        }

        sigma<-sigma.0-mu^2

        lista<-list("mu mixtura"=mu, "sigma mixtura"=sqrt(sigma))

        return(lista)
}
```

Uso:

```
> momentos(r.EM)
$`mu mixtura`
[1] 29.89482

$`sigma mixtura`
[1] 16.81013
```

El argumento `r.EM` representa a $\hat{\Psi}$.

A.10. Selección del mejor modelo

Una vez que todas las funciones anteriores han sido descritas, la última etapa del proceso de modelización consiste en seleccionar aquella mixtura cuyo número de componentes mejor describa los datos experimentales en estudio, sobre un número de componentes prefijados ($K = 1, \ldots, 11$). La siguiente función `modelos` calcula el modelo más apropiado para ese conjunto de datos, pero también genera un resumen de dicha modelización resultante. En este resumen, se incluye, entre otros resultados, la comparación de la implementación del algoritmo EM con los resultados de la función Mclust (componente del objeto lista `$resultados.Mclust`), esta vez sin prefijar en esa función el número de componentes de la mixtura (argumento G ausente).

Se sintetiza a continuación el contenido de la función:

- Líneas 2-3: resultados de la función Mclust a efectos de comparación.
- Líneas 4-7: vectores numéricos donde almacenar los resultados de los criterios de información según las $K = 11$ modelos propuestos.
- Líneas 10-16: caso particular de mixtura $K = 1$.
- Líneas 17-24: obtención del valor de los criterios de información para $K = 2, \ldots, 11$ según todas las modelizaciones realizadas, que se almacenan en una lista (`resultados.EM[[i]]`).
- Líneas 26-29: obtención del mejor modelo según el valor de los criterios de información para $K = 2, \ldots, 11$.
- Línea 30: obtención del mejor K según el valor del criterio de información BIC para $K = 2, \ldots, 11$.
- Línea 33: obtención del mejor modelo según K, obtenido mediante el valor del criterio de información BIC.
- Líneas 34-37: operación sobre el mejor modelo utilizando funciones descritas en esta sección.
- Líneas 38-42: devolución de resultados en forma de lista.

```
modelos<-function(muestra){

        modelo.mclust<-Mclust(muestra, modelNames="V")
        res.mclust<-summary(modelo.mclust, parameters=TRUE)

        bic<-numeric(11)
        icl<-numeric(11)
        aic<-numeric(11)
        aic.c<-numeric(11)

        resultados.EM<-vector("list")

        # K=1

        media<-mean(muestra) ; desv.tip<-sd(muestra)
        res.1N<-c(1,media,desv.tip) # se especifica peso unico 1

        resultados.EM[[1]]<-res.1N
        bic[1]<-BIC(muestra,res.1N)
        aic[1]<-AIC(muestra,res.1N)
        aic.c[1]<-AIC.c(muestra,res.1N)
        icl[1]<-bic[1]

        for (i in seq(2,11)) {

                        v.i(muestra,i)
                        resultados.EM[[i]]<-EM(muestra,vi.Q)
                        bic[i]<-BIC(muestra,r.EM)
                        aic[i]<-AIC(muestra,r.EM)
                        aic.c[i]<-AIC.c(muestra,r.EM)
                        icl[i]<-ICL(muestra,r.EM)
        }

       # seleccion mediante BIC

       modelo.bic<-which.max(bic)
       modelo.icl<-which.max(icl)
       modelo.aic<-which.min(aic)
       modelo.aic.c<-which.min(aic)

       optimo.bic<-bic[modelo.bic]
       icl.deducido<-icl[modelo.bic]

       eleccion.EM<-resultados.EM[[modelo.bic]]

       recuperamos.EM<-c(eleccion.EM$p.EM,eleccion.EM$mu.EM,eleccion.EM$sd.EM)

       incertidumbres<-cuantiles(muestra,recuperamos.EM)

       errores.sem<-SEM(muestra,recuperamos.EM)

       info.bootstrap<-errores.boots(sample=muestra,nc=modelo.bic,rep=1000)

       param.mixtura<-momentos(recuperamos.EM)

       lista<-list(n=length(muestra),bic=bic,icl=icl, aic=aic, aic.c=aic.c, componentes.bic=modelo.bic,
       componentes.icl=modelo.icl, componentes.aic=modelo.aic, componentes.aic.c=modelo.aic.c,
       optimo.bic=optimo.bic, icl.deducido=icl.deducido,resultados.em=eleccion.EM,
       resultados.Mclust=res.mclust,bootstrap=info.bootstrap,Errores.SEM=errores.sem,
       quantiles_incertidumbres=incertidumbres, parametros.mixtura=param.mixtura)

       return(lista)
}
```

Uso:

```
> modelos(datos)
$n
[1] 300

$bic
 [1] -2559.698 -2523.896 -2320.878 -2337.298 -2355.186 -2364.849 -2381.561 -2396.403 -2412.017 -2425.248 -2439.138
```

```
$icl
 [1] -2559.698 -2601.566 -2326.321 -2408.516 -2466.665 -2511.345 -2555.924 -2590.889 -2634.452 -2634.509 -2649.808

$aic
 [1] 2552.291 2505.377 2291.248 2296.557 2303.333 2301.885 2307.485 2311.216 2315.719 2317.839 2320.616

$aic.c
 [1] 2552.314 2505.431 2291.333 2296.675 2303.484 2302.069 2307.704 2311.469 2316.008 2318.165 2320.980

$componentes.bic
[1] 3

$componentes.icl
[1] 3

$componentes.aic
[1] 3

$componentes.aic.c
[1] 3

$optimo.bic
[1] -2320.878

$icl.deducido
[1] -2326.321

$resultados.em
$resultados.em$componentes
[1] 3

$resultados.em$p.EM
[1] 0.3332530 0.3196722 0.3470748

$resultados.em$mu.EM
[1] 10.26989 29.16179 50.14222

$resultados.em$sd.EM
[1] 2.723816 3.354494 5.179383

$resultados.em$iter
[1] 7


$resultados.Mclust
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust V (univariate, unequal variance) model with 3 components:

 log.likelihood   n df       BIC       ICL
      -1137.624 300  8 -2320.878 -2326.321

Clustering table:
  1   2   3
100  96 104

Mixing probabilities:
        1         2         3
0.3332530 0.3196725 0.3470745

Means:
       1        2        3
10.26989 29.16180 50.14223

Variances:
        1         2         3
 7.419175 11.252671 26.825923

$bootstrap
$bootstrap$componentes
[1] 3
```

```
$bootstrap$pis.boot
[1] 0.3331474 0.3197393 0.3471133

$bootstrap$mus.boot
[1] 10.28150 29.16102 50.14907

$bootstrap$sigmas.boot
[1] 2.702722 3.307517 5.146982

$bootstrap$errores.pis.boot
[1] 0.02634813 0.02812192 0.02717928

$bootstrap$errores.mus.boot
[1] 0.2711535 0.3919006 0.5621406

$bootstrap$errores.sigmas.boot
[1] 0.1873681 0.2792534 0.4352822


$Errores.SEM
[1] 0.03254683 0.03180168 0.03313496 0.26613849 0.35582356 0.51770878 0.18846662 0.25985883 0.39235493

$quantiles_incertidumbres
$quantiles_incertidumbres[[1]]
$quantiles_incertidumbres[[1]]$n_clusters
[1] 100

$quantiles_incertidumbres[[1]]$cuantiles_de_incertidumbres
          0%          25%          50%          75%         100%
1.874167e-12 5.725421e-09 8.768767e-08 2.255958e-06 9.604646e-03


$quantiles_incertidumbres[[2]]
$quantiles_incertidumbres[[2]]$n_clusters
[1] 96

$quantiles_incertidumbres[[2]]$cuantiles_de_incertidumbres
          0%          25%          50%          75%         100%
1.230909e-05 4.246857e-05 1.740833e-04 1.440555e-03 4.810307e-01


$quantiles_incertidumbres[[3]]
$quantiles_incertidumbres[[3]]$n_clusters
[1] 104

$quantiles_incertidumbres[[3]]$cuantiles_de_incertidumbres
          0%          25%          50%          75%         100%
0.000000e+00 5.637463e-12 7.335078e-09 1.000904e-06 4.116807e-01



$parametros.mixtura
$parametros.mixtura$‘mu mixtura‘
[1] 29.89482

$parametros.mixtura$‘sigma mixtura‘
[1] 16.81013
```

B

Implementación computacional de los modelos ocultos de Markov

Este anexo presenta cinco secciones. La primera de ellas presenta la implementación de la función `gauss.HMM.lalpha`, la cual permite obtener el valor de l_T a partir de los valores escalados de $\boldsymbol{\alpha}_t$. En la segunda, se explica la función `gauss.HMM.lalphabeta`, para el escalado de los vectores de probabilidades $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$. En la tercera sección, la función `viterbi.Gauss` implementa el algoritmo de Viterbi para solventar el problema de la *decodificación global* en los MMO; y en la cuarta, se desarrolla el algoritmo EM mediante la función `gauss.HMM.EM`, con la que se estiman los parámetros de los MMO. Esta última función anida a `gauss.HMM.lalphabeta`. Finalmente, en la sección B.5 se comparan los resultados de la implementación realizada en R y la obtenida mediante la librería depmixS4.

B.1. Obtención de l_T a partir de los valores escalados de $\boldsymbol{\alpha}_t$

Un ejemplo del uso de la función `gauss.HMM.lalpha`, que implementa el algoritmo 3.17, se recoge en el Ejemplo 3.5 (página 47). Esta función sirve de introducción a la que se presenta en la Sección B.2, ya que comparte el procedimiento para el escalado de $\boldsymbol{\alpha}_t$.

```
gauss.HMM.lalpha<-function(x,m,mu,sigma,gamma, delta=NULL){

        if (is.null(delta)) delta<-solve(t(diag(m)-gamma+1), rep (1,m))
        n<-length(x)
        lalpha<-matrix(NA,m,n)
        allprobs<-matrix(NA,n,m)
        l<-numeric(3)

        for (i in 1:m) { allprobs[,i]<-dnorm(x,mu[i],sigma[i]) }

        foo<-delta*allprobs[1,]
        sumfoo<-sum(foo)
        lscale<-log(sumfoo)
        l[1]<-lscale
        foo<-foo/sumfoo
        lalpha[,1]<-log(foo)+lscale

        for (i in 2:n) {

                foo<-foo%*%gamma*allprobs[i,]
                sumfoo<-sum(foo)
                lscale<-lscale+log(sumfoo)
                l[i]<-lscale
                foo<-foo/sumfoo
                lalpha[,i]<-log(foo)+lscale
                }
```

```
        list(lalpha=lalpha, l=l)
}
```

B.2. Obtención de los valores escalados de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$

Si bien la función `gauss.HMM.lalphabeta` fue introducida en el Ejemplo 3.7 (página 49), a continuación se explica su implementación en R de forma más detallada.

El argumento `x` representa la secuencia de observaciones y `m` el número de estados en la CM. Las distribuciones dependientes de estados, mixtura de distribuciones gaussianas, vienen representadas cada una por un vector de medias, `mu`, y de desviaciones típicas, `sigma`. El resto de argumentos son `gamma`, la m.t.p. $\boldsymbol{\Gamma}$, y `delta`, $\boldsymbol{\delta}$, la distribución inicial. En caso de que $\boldsymbol{\delta}$ no sea proporcionada como argumento, esta se calcula como la distribución estacionaria, segun (3.6), y utilizada como distribución inicial. Esta función comparte los valores iniciales proporcionados a la función `gauss.HMM.EM`.

En las líneas 4 y 5 se crean matrices vacías en donde se almacenarán los valores de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$, y los valores $p_i(x)$, respectivamente. El algoritmo 3.17, desarrollado en la página 47, se corresponde con las líneas de código 6 a 18, y el algoritmo 3.26, en la página 49, en las líneas 19 a 28. Finalmente, el resultado de los valores escalados de $\boldsymbol{\alpha}_t$ y $\boldsymbol{\beta}_t$ es devuelto en formato de lista en la línea 30.

Los valores α_1 y β_T se asignan en las líneas 7 y 19, respectivamente.

```
gauss.HMM.lalphabeta<-function(x,m,mu,sigma,gamma, delta=NULL){                          1

        if (is.null(delta)) delta<-solve(t(diag(m)-gamma+1), rep (1,m))                  2
        n<-length(x)                                                                     3
        lalpha<-lbeta<-matrix(NA,m,n)                                                    4
        allprobs<-matrix(NA,n,m)                                                         5

        for (i in 1:m) { allprobs[,i]<-dnorm(x,mu[i],sigma[i]) }                         6

        foo<-delta*allprobs[1,]                                                          7
        sumfoo<-sum(foo)                                                                 8
        lscale<-log(sumfoo)                                                              9
        foo<-foo/sumfoo                                                                  10
        lalpha[,1]<-log(foo)+lscale                                                      11

        for (i in 2:n) {                                                                 12

                foo<-foo%*%gamma*allprobs[i,]                                            13
                sumfoo<-sum(foo)                                                         14
                lscale<-lscale+log(sumfoo)                                               15
                foo<-foo/sumfoo                                                          16
                lalpha[,i]<-log(foo)+lscale                                              17

        }                                                                                18

        lbeta[,n]<-rep(0,m)                                                              19
        foo<-rep(1/m,m)                                                                  20
        lscale<-log(m)                                                                   21

        for(i in (n-1):1) {                                                              22

                foo<-gamma%*%(allprobs[i+1,]*foo) # escalado de las backward.            23
                lbeta[,i]<-log(foo)+lscale                                               24
                sumfoo<-sum(foo)                                                         25
                foo<-foo/sumfoo                                                          26
                lscale<-lscale+log(sumfoo)                                               27

        }                                                                                28

        list(lalpha=lalpha, lbeta=lbeta)                                                 29
}                                                                                        30
```

B.3. Algoritmo de Viterbi

Un ejemplo de uso y resultados de la implementación de la función `viterbi.Gauss` fueron introducidos en la sección 3.2.2 (página 54). El escalado de las filas de la matriz de probabilidades $\{\xi_{tj}\}$ para que estas sumen 1 se realiza en las líneas de código 9 y 12, dividiendo cada una de las probabilidades $\{\xi_{t\cdot}\}$ por la suma de cada fila (`sum(foo)`).

```
viterbi.Gauss<-function(x,mu,sd,gamma,delta=NULL ,...) {                                    1

        n<-length(x)                                                                        2
        m<-length(mu)                                                                       3

        if(is.null(delta)) delta<-solve(t(diag(m)-gamma+1),rep(1,m))                        4

        probs<-matrix(NA,nrow=n,ncol=m)                                                     5

        for(i in 1:m) { probs[,i] <- outer(x,mu[i], dnorm, sd[i])  }                        6

        xi<-matrix(0,n,m)                                                                   7

        foo<-delta*probs[1,]                                                                8

        xi[1,]<-foo/sum(foo)                                                                9

        for(i in 2:n) {                                                                     10
                foo<-apply(xi[i-1,]*gamma,2,max)*probs[i,]                                  11
                xi[i,]<-foo/sum(foo)                                                        12
        }                                                                                   13

        iv<-numeric(n)                                                                      14
        iv[n]<-which.max(xi[n,])                                                            15

        for(i in (n-1):1) {                                                                 16

        iv[i]<-which.max(gamma[,iv[i+1]]*xi[i,])                                            17
        }                                                                                   18

        print(list(mu=mu,sd=sd,delta=delta, xi=round(xi,6), x=x, iv=iv))                    19

}                                                                                           20
```

B.4. Algoritmo EM

La función `gauss.HMM.EM` implementa el algoritmo EM utilizando mixturas gaussianas. Requiere semejantes argumentos que `gauss.HMM.lalphabeta`, y además, el máximo número de iteraciones del algoritmo (`maxiter`) y el error o tolerancia entre sucesivas iteraciones (`tol`). Como se mencionó anteriormente, `gauss.HMM.EM` en la línea 11, incluye a la función `gauss.HMM.lalphabeta`.

La concepción de este algoritmo EM es diferente del empleado en las mixturas finitas (Anexo A). Entre las principales diferencias se encuentra que es una única función la que implementa el algoritmo EM, el cálculo del criterio de parada (líneas 27-28), y que el cálculo de la función de log-verosimilitud (línea 15), que lleva implícito el cálculo de $\boldsymbol{\alpha}_t$, debe tener en cuenta el problema del desbordamiento. Por esto último, el valor de $l\,(\log L_T)$, asignado al objeto `llk`, que se calcula tanto en el paso E como M, se obtiene como sigue (Zucchini and MacDonald, 2009):

$$l = \log\left(\sum_i^m \alpha_n(i)\right) = c + \log\left(\sum_i^m \exp\left(\alpha_n(i)\right) - c\right),$$

y el valor de c, un escalar, se selecciona de tal forma que reduzca las probabilidades de desbordamiento en la exponenciación (línea 14).

```
gauss.HMM.EM <- function(x,m,pi,mu,sigma,gamma,delta, maxiter=1000,tol=1e-6,...){   1
```

```
 n<-length(x)                                                                                                    2
 pi.sig<-pi                                                                                                      3
 mu.sig<- mu                                                                                                     4
 sigma.sig<- sigma                                                                                               5
 gamma.sig<- gamma                                                                                               6
 delta.sig<- delta                                                                                               7

 lallprobs<-matrix(NA,n,m)                                                                                       8

 for (iter in 1:maxiter){                                                                                        9

   for (i in 1:m) { lallprobs[,i]<-dnorm(x,mu[i],sigma[i], log=TRUE) }                                           10

   fb  <-  gauss.HMM.lalphabeta(x,m,mu,sigma,gamma,delta=delta)                                                  11
   la  <-  fb$lalpha                                                                                            12
   lb  <-  fb$lbeta                                                                                             13
   c   <-  max(la[,n])                                                                                          14
   llk <- c+log(sum(exp(la[,n]-c)))                                                                             15

   for (j in 1:m) {                                                                                             16

     for (k in 1:m)   {                                                                                         17

       gamma.sig[j,k] <- gamma[j,k]*sum(exp(la[j,1:(n-1)] + lallprobs[2:n,k]+lb[k,2:n]-llk))                    18

     }                                                                                                          19

   pi.sig[j] <- sum(exp(la[j,]+lb[j,]-llk))/n                                                                   20

   mu.sig[j] <- sum(exp(la[j,]+lb[j,]-llk)*x) / sum(exp(la[j,]+lb[j,]-llk))                                     21

   sigma.sig[j] <- sqrt(sum(exp(la[j,]+lb[j,]-llk)*(x-mu.sig[j])^2) / sum(exp(la[j,]+lb[j,]-llk)))              22

   }                                                                                                            23

   gamma.sig <- gamma.sig/apply(gamma.sig,1,sum)                                                                24
   delta.sig <- exp(la[,1]+lb[,1]-llk)                                                                          25
   delta.sig <- delta.sig/sum(delta.sig)                                                                        26
   crit       <- sum(abs(mu-mu.sig)) +  sum(abs(sigma-sigma.sig)) + sum(abs(pi-pi.sig)) +                       27
                 sum(abs(gamma-gamma.sig)) + sum(abs(delta-delta.sig))                                          28

   if(crit<tol) {                                                                                               29

     np     <- m*m+m-1                                                                                          30
     AIC    <- -2*(llk-np)                                                                                      31
     BIC    <- -2*llk+np*log(n)                                                                                 32
     return(list(pi=pi,mu=mu,sigma=sigma, gamma=round(gamma,4),delta=delta,                                     33
                 mllk=-llk,AIC=AIC,BIC=BIC, iter=iter))                                                         34
   }                                                                                                            35

   pi <- pi.sig                                                                                                 36
   mu <- mu.sig                                                                                                 37
   sigma <- sigma.sig                                                                                           38
   gamma <- gamma.sig                                                                                           39
   delta <- delta.sig                                                                                           40
}                                                                                                               41
 print(paste("No existe convergencia tras",maxiter,"iteraciones"))                                              42

}                                                                                                               43
```

Uso:

Se crea a continuación una muestra artificial, semejante a la del apartado A.1. Se calcularán sus valores iniciales mediante la función `v.i`, que serán suministrados a la función `gauss.HMM.EM`. El resto de parámetros iniciales son la m.t.p. $\mathbf{\Gamma}$ (`gamma.3`), con todos sus elementos con valor 1/3, y $\boldsymbol{\delta} = (0, 1, 0)$ (`delta`). Se acompañan los resultados de la función `gauss.HMM.EM` en formato de lista.

```
set.seed(123)
cluster.1<-rnorm(100,10,3) # mu1:  10.271218, sd1: 2.738448
cluster.2<-rnorm(100,30,4) # mu2:  29.569813, sd2: 3.867946
cluster.3<-rnorm(100,50,5) # mu3:  50.602326, sd3: 4.749395
datos<-c(cluster.1, cluster.2, cluster.3)
```

```
gamma.3<-matrix(rep(1/3,9), ncol=3 ,byrow=TRUE)

> v.i(datos,3)
[1]  0.3333333  0.3333333  0.3333333 10.2712177 29.5523448 50.6197935  2.7384476  3.8103611  4.7176350

vi<-v.i(datos,3)

> gauss.HMM.EM(datos,m=3,pi=c(vi[1],vi[2],vi[3]), mu=c(vi[4],vi[5],vi[6]),
sigma=c(vi[7],vi[8],vi[9]),delta=c(0,1,0),gamma=gamma.3)

$pi
[1] 0.3299998 0.3366669 0.3333333

$mu
[1] 10.29094 29.35940 50.60233

$sigma
[1] 2.731334 4.369446 4.725589

$gamma
       [,1]   [,2]   [,3]
[1,] 0.9899 0.0101 0.0000
[2,] 0.0099 0.9802 0.0099
[3,] 0.0000 0.0000 1.0000

$delta
[1] 0 1 0

$mllk
[1] 846.1932

$AIC
[1] 1718.386

$BIC
[1] 1766.536

$iter
[1] 5
```

B.5. Comparación entre la implementación propia de MOM y depmixS4

A continuación se modeliza una ST sintética mediante la implementación propia y la librería depmixS4, sin encontrarse diferencias significativas. Para la implementación propia se suministran a la función `gauss.HMM.EM` los valores iniciales mediante `v.i.` y `gamma.3` (la m.p.t de 3 estados); los resultados son devueltos en formato de lista.

```
set.seed(123)
cluster.1<-rnorm(100,10,3) # mu1:  10.27122, sd1: 2.738448
cluster.2<-rnorm(200,30,4) # mu2:  30.02584, sd2: 3.851468
cluster.3<-rnorm(100,50,5) # mu3:  49.81889, sd3: 5.193906

# ----------------------
# Implementación propia:
# ----------------------

> vi<-v.i(datos,3)

> vi
[1]  0.3333333  0.3333333  0.3333333 13.8463514 29.7455081 46.3914950  6.7024468  1.9244248  7.4746984

> gamma.3<-matrix(rep(1/3,9), ncol=3 ,byrow=TRUE); gamma.3
          [,1]      [,2]      [,3]
[1,] 0.3333333 0.3333333 0.3333333
[2,] 0.3333333 0.3333333 0.3333333
[3,] 0.3333333 0.3333333 0.3333333
```

```
> gauss.HMM.EM(datos,m=3,pi=c(vi[1],vi[2],vi[3]), mu=c(vi[4],vi[5],vi[6]),
sigma=c(vi[7],vi[8],vi[9]),delta=c(0,1,0),gamma=gamma.3)

$pi
[1] 0.247500 0.502432 0.250068

$mu
[1] 10.29094 29.91718 49.81480

$sigma
[1] 2.731335 4.125322 5.173048

$gamma
       [,1]   [,2]  [,3]
[1,] 0.9899 0.0101 0.000
[2,] 0.0050 0.9900 0.005
[3,] 0.0000 0.0000 1.000

$delta
[1] 0 1 0

$mllk
[1] 1134.303

$AIC
[1] 2290.607

$BIC
[1] 2334.513

$iter
[1] 7

# ----------------------------
# Implementación con depmixS4:
# ----------------------------

> muestra<-data.frame(y=datos)
> m3<-depmix(y~1, data=muestra, ns=3, ntimes=nrow(muestra))
> fm3<-fit(m3, em=em.control(maxit=2000, tol=1e-08, crit="relative"))

> summary(fm3)
Initial state probabilties model
pr1 pr2 pr3
  1   0   0


Transition matrix
              toS1         toS2         toS3
fromS1 9.90000e-01 1.000000e-02 3.713822e-43
fromS2 2.17230e-61 9.949993e-01 5.000691e-03
fromS3 4.65613e-91 3.308672e-11 1.000000e+00

Response parameters
Resp 1 : gaussian
    Re1.(Intercept)   Re1.sd
St1        10.27122 2.724721
St2        30.02516 3.841676
St3        49.81477 5.173060

> probs<-posterior(fm3)
> colMeans(probs[,2:4])
       S1        S2        S3
0.2499984 0.4992151 0.2507864

> logLik(fm3)
'log Lik.' -1113.115 (df=14)
> AIC(fm3)
[1] 2254.231
> BIC(fm3)
[1] 2310.111
```

C

Material suplementario del Capítulo 5

El contenido de este anexo se corresponde con el material suplementario de la publicación **Gómez-Losada, A., Lozano-García, A., Pino-Mejías, R., Contreras-González, J.** 2014. Finite mixture models to characterize and refine air quality monitoring networks. *Science of the Total Environment*, 485-486: 292-9.

C.1. Parametrizaciones de las mixturas

En este apartado se recogen los valores de $\hat{\Psi}$ de los 49 modelos resultantes tras aplicar el algoritmo EM a los 49 conjuntos de datos según los parámetros analizados en las estaciones de monitorización de Sevilla (Tabla 5.1).

Cada tabla recoge el número de componentes resultantes según el criterio de información elegido (K_{BIC}, K_{AIC}, K_{AIC_c} y K_{ICL}), el número de observaciones en cada componente (n_k) según K_{BIC}, el parámetro en cuestión (θ_k) y los errores de estos utilizando el método SEM (e.e. SEM) o *bootstrap* ($\widehat{se_B}$).

Cuando una de las componentes ha presentado una representatividad cercana al cero ($\pi_i \simeq 0$, $i = 1, 2, 3$), el valor de π se ha incluido en la tabla correspondiente y resaltado en negrita. Este ha sido el motivo, en la mayoría de los casos, por el que no se han podido obtener los errores SEM en la parametrización de algún modelo por interrupción del algoritmo.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Alc	CO	3	83 67 211	π_1	0.23	0.026	0.024	6	6	2
				π_2	0.15	0.026	0.098			
				μ_1	230.78	1.36	1.68			
				μ_2	289.13	1.70	4.38			
				μ_3	337.47	3.28	9.71			
				σ_1	10.40	1.068	1.27			
				σ_2	5.84	1.61	4.81			
				σ_3	38.41	2.38	6.66			
	NO_2	2	176 184	π_1	0.43	0.040	0.049	3	3	1
				μ_1	12.84	0.32	0.66			
				μ_2	26.80	0.85	0.94			
				σ_1	3.60	0.24	0.43			
				σ_2	9.39	0.56	0.50			
	O_3	2	169 187	π_1	0.50	0.042	0.11	3	3	1
				μ_1	44.44	2.37	5.30			
				μ_2	75.37	0.94	2.77			
				σ_1	15.79	1.56	2.33			
				σ_2	10.83	0.66	1.69			
	PM_{10}	3	145 193 20	π_1	0.35	0.051	0.049	3	3	2
				π_2	0.56	0.069	0.057			
				μ_1	18.16	0.88	0.86			
				μ_2	32.28	1.25	1.61			
				μ_3	58.59	5.34	7.68			
				σ_1	4.61	0.60	0.41			
				σ_2	8.62	0.65	0.90			
				σ_3	21.95	3.60	3.33			
	SO_2	2	128 232	π_1	0.34	0.032	0.027	5	5	2
				μ_1	2.73	0.050	0.050			
				μ_2	5.80	0.079	0.073			
				σ_1	0.49	0.036	0.027			
				σ_2	1.07	0.061	0.067			

Tabla C.1 Alcalá de Guadaíra.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Alj	NO_2	3	91 171 102	π_1	0.22	0.033	0.050	3	3	1
				π_2	0.45	0.042	0.063			
				μ_1	7.46	0.30	0.51			
				μ_2	14.99	0.67	0.90			
				μ_3	27.39	1.18	1.58			
				σ_1	1.60	0.22	0.34			
				σ_2	4.14	0.41	0.39			
				σ_3	7.05	0.63	0.76			
	O_3	2	167 199	π_1	0.49	0.037	0.090	3	3	1
				μ_1	46.67	1.21	4.49			
				μ_2	81.09	0.98	2.93			
				σ_1	15.12	1.18	2.26			
				σ_2	11.76	0.74	1.40			
	PM_{10}	3	126 216 19	π_1	0.30	–	0.067	3	3	2
				π_2	0.62	–	0.068			
				$\boldsymbol{\pi_3}$	0.076	0.020	0.037			
				μ_1	18.06	–	1.25			
				μ_2	32.18	–	1.40			
				μ_3	63.14	6.49	7.50			
				σ_1	4.34	0.15	0.71			
				σ_2	8.75	0.47	0.59			
				σ_3	22.09	4.18	3.62			
	SO_2	3	69 158 133	π_1	0.18	0.023	0.022	3	3	1
				π_2	0.35	0.057	0.084			
				μ_1	1.38	0.097	0.10			
				μ_2	7.80	0.12	0.16			
				μ_3	8.00	0.22	0.32			
				σ_1	0.75	0.072	0.062			
				σ_2	0.83	0.11	0.22			
				σ_3	2.42	0.19	0.24			

Tabla C.2 Aljarafe.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Ber	CO	2	305 60	π_1	0.72	0.073	0.12	2	2	1
				μ_1	410.82	9.73	16.76			
				μ_2	577.80	32.17	44.10			
				σ_1	101.78	6.12	8.26			
				σ_2	167.90	18.55	17.35			
	NO_2	3	46 183 113	$\boldsymbol{\pi_1}$	0.11	0.020	0.080	3	3	1
				π_2	0.51	–	0.088			
				π_3	0.37	0.026	0.069			
				μ_1	2.97	0.11	2.13			
				μ_2	15.59	0.90	2.12			
				μ_3	35.13	1.00	2.95			
				σ_1	1.84	0.24	1.28			
				σ_2	6.64	–	0.75			
				σ_3	11.57	0.88	1.31			
	O_3	2	73 292	π_1	0.19	0.031	0.087	2	2	1
				μ_1	23.57	1.92	4.15			
				μ_2	58.04	1.14	2.60			
				σ_1	6.08	0.96	2.55			
				σ_2	15.33	0.80	1.21			
	PM_{10}	2	315 26	π_1	0.87	0.062	0.10	3	3	2
				μ_1	29.14	0.73	1.80			
				μ_2	56.67	5.71	8.51			
				σ_1	10.83	0.54	0.95			
				σ_2	20.14	3.44	3.13			
	SO_2	3	116 129 81	π_1	0.31	0.070	0.052	5	5	2
				π_2	0.39	0.14	0.069			
				π_3	0.30	0.056	0.076			
				μ_1	3.45	0.052	0.080			
				μ_2	4.74	–	0.26			
				μ_3	7.01	0.31	0.54			
				σ_1	0.29	0.036	0.047			
				σ_2	0.76	0.20	0.15			
				σ_3	1.93	0.16	0.25			

Tabla C.3 Bermejales.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Cen	CO	2	266 74	π_1	0.79	0.051	0.16	8	8	2
				μ_1	572.25	9.95	35.58			
				μ_2	1188.43	33.52	154.41			
				σ_1	141.34	7.81	35.37			
				σ_2	158.95	23.24	74.02			
	NO_2	2	129 216	π_1	0.35	0.045	0.072	2	2	1
				μ_1	12.17	0.62	1.063			
				μ_2	26.77	0.75	1.14			
				σ_1	4.17	0.42	0.57			
				σ_2	7.24	0.50	0.59			
	O_3	2	92 256	π_1	0.25	0.036	0.12	3	3	1
				μ_1	24.41	1.77	5.88			
				μ_2	62.26	1.35	4.038			
				σ_1	8.09	1.09	3.36			
				σ_2	15.78	0.96	1.93			
	SO_2	2	253 95	π_1	0.64	0.066	0.064	2	2	1
				μ_1	2.23	0.059	0.074			
				μ_2	3.66	0.17	0.17			
				σ_1	0.58	0.041	0.043			
				σ_2	1.099	0.087	0.072			

Tabla C.4 Centro.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Cob	CO	2	152	π_1	0.50	0.038	0.029	2	2	2
			150	μ_1	140.06	2.53	3.03			
				μ_2	282.64	2.26	2.64			
				σ_1	33.19	1.85	2.050			
				σ_2	29.39	1.67	2.26			
	NO_2	3	23	$\boldsymbol{\pi_1}$	0.084	0.018	0.084	5	5	1
			126	π_2	0.48	–	0.16			
			87	π_3	0.44	–	0.18			
				μ_1	0.33	0.13	0.84			
				μ_2	5.37	0.77	1.34			
				μ_3	9.94	0.66	2.42			
				σ_1	0.21	0.042	0.56			
				σ_2	2.62	0.32	0.58			
				σ_3	4.00	0.67	0.86			
	O_3	2	126	π_1	0.38	0.046	0.092	3	3	1
			225	μ_1	40.37	1.67	3.67			
				μ_2	67.83	1.057	2.19			
				σ_1	10.34	1.062	2.00			
				σ_2	10.70	0.69	0.97			
	PM_{10}	3	143	π_1	0.39	–	–	4	4	2
			161	π_2	0.56	–	–			
			7	$\boldsymbol{\pi_3}$	0.045	0.015	–			
				μ_1	8.41	0.52	–			
				μ_2	20.40	–	–			
				μ_3	50.073	9.26	–			
				σ_1	3.61	0.19	–			
				σ_2	8.51	–	–			
				σ_3	21.52	5.22	–			
	SO_2	2	30	$\boldsymbol{\pi_1}$	0.083	0.018	0.13	5	5	2
			281	π_2	0.92	0.055	0.13			
				μ_1	0.083	–	0.55			
				μ_2	3.13	0.10	0.34			
				σ_1	0.071	–	0.35			
				σ_2	1.67	0.072	0.13			

Tabla C.5 Cobre Las Cruces.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Dos	CO	2	157	π_1	0.57	0.059	0.069	6	6	1
			163	μ_1	430.70	15.74	22.86			
				μ_2	519.01	9.033	11.44			
				σ_1	139.17	11.53	13.69			
				σ_2	65.83	8.50	11.57			
	NO_2	2	203	π_1	0.51	0.037	0.081	4	4	1
			145	μ_1	15.11	0.57	0.77			
				μ_2	25.13	0.51	1.22			
				σ_1	3.73	0.35	0.43			
				σ_2	7.31	0.51	0.42			
	O_3	2	161	π_1	0.46	0.041	0.12	3	3	1
			188	μ_1	40.72	1.52	5.98			
				μ_2	74.41	1.11	3.65			
				σ_1	14.67	1.082	3.01			
				σ_2	12.13	0.81	1.81			
	SO_2	1	344	μ	5.79	–	0.058	5	5	1
				σ	1.12	–	0.039			

Tabla C.6 Dos Hermanas.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Pri	CO	2	339 3	π_1	0.99	0.054	–	2	2	2
				$\boldsymbol{\pi_2}$	0.0088	0.0051	–			
				μ_1	404.42	7.19	–			
				μ_2	1311.85	–	–			
				σ_1	132.34	5.083	–			
				σ_2	80.94	33.05	–			
	NO_2	2	56 260	π_1	0.15	0.0083	0.14	2	2	1
				μ_1	12.78	2.01	3.86			
				μ_2	33.09	0.50	2.64			
				σ_1	3.61	0.88	2.22			
				σ_2	10.95	0.29	1.03			
	PM_{10}	2	338 17	π_1	0.91	0.061	0.17	3	3	2
				$\boldsymbol{\pi_2}$	0.091	0.030	0.17			
				μ_1	26.59	0.72	2.38			
				μ_2	52.79	6.29	14.17			
				σ_1	10.53	0.54	1.45			
				σ_2	21.14	3.66	4.75			
	SO_2	2	108 243	π_1	0.27	0.040	0.043	6	6	1
				μ_1	3.44	0.095	0.14			
				μ_2	5.83	0.13	0.11			
				σ_1	0.56	0.063	0.083			
				σ_2	1.42	0.082	0.076			

Tabla C.7 Príncipes.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Ran	CO	3	165 130 66	π_1	0.40	0.043	0.050	6	6	3
				π_2	0.39	0.046	0.043			
				μ_1	126.81	1.61	2.75			
				μ_2	188.45	5.41	9.077			
				μ_3	424.78	24.54	26.10			
				σ_1	14.05	1.19	1.81			
				σ_2	41.51	3.93	4.46			
				σ_3	176.028	14.016	19.37			
	NO_2	2	115 244	π_1	0.29	0.047	0.12	3	3	2
				μ_1	24.080	1.47	2.97			
				μ_2	42.53	1.16	2.15			
				σ_1	6.27	0.86	1.77			
				σ_2	12.17	0.80	0.73			
	SO_2	3	166 135 57	π_1	0.42	0.057	0.095	3	3	1
				π_2	0.37	0.052	0.12			
				μ_1	4.33	0.057	0.11			
				μ_2	5.71	0.13	0.23			
				μ_3	7.78	0.45	0.47			
				σ_1	0.47	0.040	0.088			
				σ_2	0.70	0.089	0.23			
				σ_3	1.76	0.22	0.21			

Tabla C.8 Ranilla.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Saj	NO_2	2	225 133	π_1	0.52	0.075	0.076	3	3	1
				μ_1	18.41	0.64	1.00			
				μ_2	27.55	1.61	1.15			
				σ_1	5.54	0.40	0.43			
				σ_2	8.42	0.79	0.68			
	O_3	2	131 231	π_1	0.36	0.043	0.094	3	3	1
				μ_1	28.83	1.75	4.42			
				μ_2	64.84	1.32	3.05			
				σ_1	10.91	1.13	2.44			
				σ_2	13.01	0.88	1.62			

Tabla C.9 San Jerónimo.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Sac	CO	2	160 198	π_1	0.39	0.039	0.039	3	3	2
				μ_1	297.38	2.16	2.77			
				μ_2	406.64	6.45	6.33			
				σ_1	19.95	1.83	2.55			
				σ_2	76.034	3.99	3.93			
	NO_2	2	145 206	π_1	0.38	0.050	0.12	3	3	1
				μ_1	11.47	0.92	2.38			
				μ_2	26.066	1.015	1.77			
				σ_1	5.82	0.59	1.40			
				σ_2	9.41	0.69	0.59			
	O_3	2	78 273	π_1	0.20	0.025	0.10	3	3	1
				μ_1	19.55	1.22	4.68			
				μ_2	56.70	1.13	3.57			
				σ_1	5.14	0.68	3.045			
				σ_2	17.40	0.76	1.56			
	PM_{10}	2	318 39	π_1	0.82	0.070	0.084	3	3	1
				μ_1	21.92	0.74	1.14			
				μ_2	41.50	4.052	5.49			
				σ_1	9.26	0.54	0.64			
				σ_2	16.87	2.26	1.78			

Tabla C.10 Santa Clara.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Sie	NO_2	2	267 82	π_1	0.73	0.063	0.065	5	5	2
				μ_1	2.88	0.085	0.14			
				μ_2	6.59	0.43	0.65			
				σ_1	0.97	0.062	0.089			
				σ_2	1.85	0.23	0.32			
	O_3	2	160 189	π_1	0.49	0.038	0.090	4	4	1
				μ_1	48.74	1.00	4.19			
				μ_2	75.69	0.84	2.24			
				σ_1	13.40	1.045	1.91			
				σ_2	9.72	0.65	1.10			
	PM_{10}	3	156 173 9	π_1	0.39	0.083	–	8	8	2
				π_2	0.56	0.089	–			
				$\boldsymbol{\pi_3}$	0.049	0.018	–			
				μ_1	10.54	0.55	–			
				μ_2	22.81	1.39	–			
				μ_3	58.32	10.73	–			
				σ_1	3.64	0.40	–			
				σ_2	8.95	0.86	–			
				σ_3	35.08	7.035	–			
	SO_2	3	121 73 153	π_1	0.36	0.033	0.032	3	3	2
				π_2	0.20	0.029	0.073			
				μ_1	1.88	0.061	0.060			
				μ_2	3.42	0.12	0.17			
				μ_3	4.67	0.064	0.11			
				σ_1	0.54	0.049	0.064			
				σ_2	0.37	0.058	0.15			
				σ_3	0.48	0.042	0.056			

Tabla C.11 Sierra Norte.

Sitio	Contaminante	K_{BIC}	n_k	θ_k	valor $\hat{\theta}_k$	e.e. SEM	$\widehat{se_B}$	K_{AIC}	K_{AIC_c}	K_{ICL}
Tor	CO	2	218 147	π_1	0.53	0.040	0.054	2	2	1
				μ_1	339.62	7.80	10.037			
				μ_2	615.76	14.43	31.96			
				σ_1	87.37	5.72	7.041			
				σ_2	172.71	13.78	17.29			
	NO_2	1	365	μ	33.63	–	0.78	3	3	1
				σ	15.37	–	0.51			
	O_3	2	101 262	π_1	0.27	0.034	0.065	3	3	2
				μ_1	16.83	0.99	2.16			
				μ_2	47.24	1.065	2.034			
				σ_1	5.41	0.66	1.39			
				σ_2	13.078	0.77	1.096			
	PM_{10}	3	81 205 27	π_1	0.22	0.0099	0.064	3	3	1
				π_2	0.63	0.041	0.14			
				π_3	0.15	0.045	0.12			
				μ_1	17.16	0.78	1.24			
				μ_2	30.35	0.38	1.71			
				μ_3	45.20	3.75	9.55			
				σ_1	3.71	0.52	0.76			
				σ_2	7.93	0.84	1.53			
				σ_3	15.21	2.60	3.25			
	SO_2	2	319 46	π_1	0.77	0.090	0.15	4	4	1
				μ_1	3.40	0.059	0.088			
				μ_2	4.56	0.26	0.47			
				σ_1	0.66	0.047	0.12			
				σ_2	1.20	0.13	0.17			

Tabla C.12 Torneo.

C.2. Implementación del ACJ, BA y ACP

A lo largo de esta sección, se explican las funciones utilizadas en R de librerías existentes en ese entorno, con las que se han llevado a cabo los análisis complementarios a la modelización mediante mixturas. Estas funciones o librerías han permitido realizar el ACJ, imputaciones de μ_m y cv_m, y el ACP.

C.2.1. ACJ

▷ Dendrograma según los valores de μ_m (Figura 5.2A).

```
alc.m<-c(308.57,21.06,61.12,28.84,4.82)
alj.m<-c(NA,18.44,62.86,30.53,6.78)
ber.m<-c(480.55,21.64,51.91,33.55,5.02)
cen.m<-c(677.75,21.46,53.54,NA,2.80)
cob.m<-c(206.7,6.54,55.91,17.04,2.76)
dos.m<-c(463.84,19.44,57.39,NA,5.79)
pri.m<-c(412.38,29.39,NA,29.82,5.36)
ran.m<-c(221.04,36.15,NA,NA,5.49)
saj.m<-c(NA,22.79,53.57,NA,NA)
sac.m<-c(367.1,20.87,47.94,24.74,NA)
sie.m<-c(NA,3.78,61.42,19.75,3.37)
tor.m<-c(465.83,33.63,39.25,29.72,3.67)

m<-matrix(c(alc.m,alj.m,ber.m,cen.m,cob.m,
            dos.m,pri.m,ran.m,saj.m,sac.m,
            sie.m,tor.m),ncol=5, byrow=T)

dimnames(m)<-list(estaciones=c("Alc","Alj","Ber",
             "Cen","Cob","Dos","Pri","Ran","Saj",
             "Sac","Sie","Tor"), par=c())

d<-dist(m,method="euclidean")
figura<-hclust(d,method="single")
par(mar=c(0.15,4,0.85,0.15), cex=1.5)
plot(figura, main="A", ylab="", xlab="", axes=FALSE)
axis(side=2, line=0.8)
```

▷ Dendrograma según los valores de cv_m (Figura 5.2B).

```
alc.sd<-c(51,10.22,20.92,13.92,1.65)
alj.sd<-c(NA,9.39,21.50,15.22,2.92)
ber.sd<-c(148.03,14.37,19.71,16.01,1.89)
cen.sd<-c(276.24,9.03,21.44,NA,1.08)
cob.sd<-c(74.56,4.34,16.92,12.29,1.82)
dos.sd<-c(123.08,7.17,21.56,NA,1.12)
pri.sd<-c(156.77,12.40,NA,13.88,1.66)
ran.sd<-c(141.94,13.09,NA,NA,1.52)
saj.sd<-c(NA,7.94,21.05,NA,NA)
sac.sd<-c(86.60,11.04,21.67,13.19,NA)
sie.sd<-c(NA,1.99,17.42,14.87,1.38)
tor.sd<-c(181.32,15.37,17.72,12.40,0.92)

alc.cv<-alc.sd/alc.m
alj.cv<-alj.sd/alj.m
ber.cv<-ber.sd/ber.m
cen.cv<-cen.sd/cen.m
cob.cv<-cob.sd/cob.m
dos.cv<-dos.sd/dos.m
pri.cv<-pri.sd/pri.m
ran.cv<-ran.sd/ran.m
saj.cv<-saj.sd/saj.m
sac.cv<-sac.sd/sac.m
sie.cv<-sie.sd/sie.m
tor.cv<-tor.sd/tor.m

cv<-matrix(c(alc.cv,alj.cv,ber.cv,cen.cv,
             cob.cv,dos.cv,pri.cv,ran.cv,
             saj.cv,sac.cv,sie.cv,tor.cv),
             ncol=5, byrow=T)

dimnames(cv)<-list(estaciones=c("Alc","Alj","Ber",
              "Cen","Cob","Dos","Pri","Ran","Saj","Sac","Sie","Tor"), par=c())

d.cv<-dist(cv,method="euclidean")
figura.cv<-hclust(d.cv,method="single")
par(mar=c(0.15,0.5,0.85,3.15), cex=1.5)
plot(figura.cv, main="B", ylab="", xlab="", axes=FALSE)
axis(side=4, line=0.8)
```

C.2.2. BA

▷ Imputación de los valores de μ_m (Tabla 5.3).

```
Alc<-c(308.57,21.06,61.12,28.84,4.82)
Alj<-c(NA,18.44,62.86,30.53,6.78)
Ber<-c(480.55,21.64,51.91,33.55,5.02)
Cen<-c(677.75,21.46,53.54,NA,2.80)
Cob<-c(206.7,6.54,55.91,17.04,2.76)
Dos<-c(463.84,19.44,57.39,NA,5.79)
Pri<-c(412.38,29.39,NA,29.82,5.36)
Ran<-c(221.04,36.15,NA,NA,5.49)
Saj<-c(NA,22.79,53.57,NA,NA)
Sac<-c(367.1,20.87,47.94,24.74,NA)
Sie<-c(NA,3.78,61.42,19.75,3.37)
Tor<-c(465.83,33.63,39.25,29.72,3.67)

m<-matrix(c(Alc,Alj,Ber,Cen,Cob,Dos,Pri,Ran,Saj,Sac,Sie,Tor),ncol=5, byrow=T)

library(randomForest)
set.seed(222)

tipo<-c("U","S","U","U","R","U","U","U","S","S","R","U")
tipo<-factor(tipo, labels=c("R","S","U"))

marco<-data.frame(Alc,Alj,Ber,Cen,Cob,Dos,Pri,Ran,Saj,Sac,Sie,Tor)
marco<-t(marco); marco
marco<-as.data.frame(marco); marco
marco<-data.frame(marco,tipo); marco
```

```
colnames(marco)<-c("CO","NO2","O3","PM10","SO2","tipo")
marco.imp <- rfImpute(tipo ~ ., marco, iter=250, ntree=2500)

> marco.imp

    tipo       CO   NO2       O3     PM10      SO2
Alc    U 308.5700 21.06 61.12000 28.84000 4.820000
Alj    S 394.1254 18.44 62.86000 30.53000 6.780000
Ber    U 480.5500 21.64 51.91000 33.55000 5.020000
Cen    U 677.7500 21.46 53.54000 27.93820 2.800000
Cob    R 206.7000  6.54 55.91000 17.04000 2.760000
Dos    U 463.8400 19.44 57.39000 27.69083 5.790000
Pri    U 412.3800 29.39 54.68649 29.82000 5.360000
Ran    U 221.0400 36.15 54.83417 28.16964 5.490000
Saj    S 420.4997 22.79 53.57000 28.85198 4.876696
Sac    S 367.1000 20.87 47.94000 24.74000 4.563775
Sie    R 328.0027  3.78 61.42000 19.75000 3.370000
Tor    U 465.8300 33.63 39.25000 29.72000 3.670000
```

▷ Imputación de los valores de cv_m (Tabla 5.4).

```
alc.sd<-c(51,10.22,20.92,13.92,1.65)
alj.sd<-c(NA,9.39,21.50,15.22,2.92)
ber.sd<-c(148.03,14.37,19.71,16.01,1.89)
cen.sd<-c(276.24,9.03,21.44,NA,1.08)
cob.sd<-c(74.56,4.34,16.92,12.29,1.82)
dos.sd<-c(123.08,7.17,21.56,NA,1.12)
pri.sd<-c(156.77,12.40,NA,13.88,1.66)
ran.sd<-c(141.94,13.09,NA,NA,1.52)
saj.sd<-c(NA,7.94,21.05,NA,NA)
sac.sd<-c(86.60,11.04,21.67,13.19,NA)
sie.sd<-c(NA,1.99,17.42,14.87,1.38)
tor.sd<-c(181.32,15.37,17.72,12.40,0.92)

Alc<-alc.sd/Alc
Alj<-alj.sd/Alj
Ber<-ber.sd/Ber
Cen<-cen.sd/Cen
Cob<-cob.sd/Cob
Dos<-dos.sd/Dos
Pri<-pri.sd/Pri
Ran<-ran.sd/Ran
Saj<-saj.sd/Saj
Sac<-sac.sd/Sac
Sie<-sie.sd/Sie
Tor<-tor.sd/Tor

cv<-matrix(c(Alc,Alj,Ber,Cen,Cob,Dos,Pri,Ran,Saj,Sac,Sie,Tor),ncol=5, byrow=T)

marco<-data.frame(Alc,Alj,Ber,Cen,Cob,Dos,Pri,Ran,Saj,Sac,Sie,Tor); marco
marco<-t(marco); marco
marco<-as.data.frame(marco); marco
marco<-data.frame(marco,tipo); marco

colnames(marco)<-c("CO","NO2","O3","PM10","SO2","tipo")

set.seed(222)

tipo<-c("U","S","U","U","R","U","U","U","S","S","R","U")
tipo<-factor(tipo, labels=c("R","S","U"))

marco.imp.cv <- rfImpute(tipo ~ ., cv, iter=250, ntree=2500)

colnames(marco.imp.cv)<-c("CO","NO2","O3","PM10","SO2")
rownames(marco.imp.cv)<-c("Alc","Alj","Ber","Cen","Cob",
                          "Dos","Pri","Ran","Saj","Sac","Sie","Tor")

> marco.imp.cv
           CO       NO2        O3      PM10       SO2
Alc 0.1652785 0.4852802 0.3422775 0.4826630 0.3423237
Alj 0.3331619 0.5092191 0.3420299 0.4985260 0.4306785
Ber 0.3080429 0.6640481 0.3796956 0.4771982 0.3764940
Cen 0.4075839 0.4207829 0.4004483 0.5230816 0.3857143
Cob 0.3607160 0.6636086 0.3026292 0.7212441 0.6594203
```

```
Dos 0.2653501 0.3688272 0.3756752 0.5202061 0.1934370
Pri 0.3801591 0.4219122 0.3969689 0.4654594 0.3097015
Ran 0.6421462 0.3621024 0.3993177 0.4765102 0.2768670
Saj 0.3757580 0.3483984 0.3929438 0.4830967 0.3156961
Sac 0.2359030 0.5289890 0.4520234 0.5331447 0.3382175
Sie 0.3138394 0.5264550 0.2836210 0.7529114 0.4094955
Tor 0.3892407 0.4570324 0.4514650 0.4172275 0.2506812
```

C.2.3. ACP

▷ ACP. Gráficos 5.3A y 5.3B.

```
marco.imp<-marco.imp[,-1]

prcomp(marco.imp, scale=TRUE)

tipo<-c("UB","SB","UB","UB","RI","UB","UB","UT","SI","SB","RB","UT")
tipo<-factor(tipo, labels=c("RB","RI","SB","SI","UB","UT"))

row.names(marco.imp)<-c("Alc","Alj","Ber", "Cen","Cob","Dos","Pri","Ran","Saj","Sac","Sie","Tor")

# gráfico: componente 1 y 2 (eje X CP1, eje Y CP2)
plot(c(-1.5,3.2), c(-2.5,2.5), type="n", axes=FALSE, xlab=NA, ylab=NA, cex.main=2, main="A")
s.class(dfxy=acp$li, fac=tipo, xax=1, yax=2, grid=FALSE, cgrid=FALSE, cpoint=2, addaxes=FALSE,
        origin=c(0,0), add.plot=TRUE)

text(0.4951411,-1.10395170,"Alc")
text(-0.2675646,-2.1002008902,"Alj")
text(-1.1648300,0.13564444,"Ber")
text(-0.4309082,2.13402535,"Cen")
text(3.0160943,0.69518489,"Cob")
text(-0.5869739,-0.77521440,"Dos")
text(-0.91134462,-0.46211689,"Pri")
text(-1.1153932,-1.02288270,"Ran")
text(-0.2093059,-0.03934922,"Saj")
text(0.3702242,0.70132950,"Sac")
text(2.6062083,0.03300452,"Sie")
text(-1.462459,2.23441524,"Tor")

# gráfico: componente 1 y 3 (eje X CP1, eje Y CP3)
plot(c(-1.5,3.2), c(-2.5,2.5), type="n", axes=FALSE, xlab=NA, ylab=NA, cex.main=2, main="B")
s.class(dfxy=acp$li, fac=tipo, xax=1, yax=3, grid=FALSE, cgrid=FALSE, cpoint=2, addaxes=FALSE,
        origin=c(0,0), add.plot=TRUE)

text(0.4951411,-0.04475918,"Alc")
text(-0.3675646,0.95426525,"Alj")
text(-1.1648300,0.82312057,"Ber")
text(-0.4309082,2.03067982,"Cen")
text(3.0160943,-0.81560159,"Cob")
text(-0.0969739,0.69652987,"Dos")
text(-0.91134462,-0.17651229,"Pri")
text(-0.453932,-1.79500809,"Ran")
text(-0.45093059,0.28768354,"Saj")
text(0.3702242,-0.71284352,"Sac")
text(2.6062083,0.56236212,"Sie")
text(-1.462459,-1.00991649,"Tor")
```

▷ ACP. Gráficos 5.4A y 5.4B.

```
acp.cv<-dudi.pca(df=marco.imp.cv, scannf=F,nf=3, scale=FALSE)
prcomp(marco.imp.cv, scale=FALSE)

plot(c(-0.35,0.32), c(-0.25,0.25), type="n", axes=FALSE, xlab=NA, ylab=NA, cex.main=2, main="A")
s.class(dfxy=acp.cv$li, fac=tipo, xax=1, yax=2, grid=FALSE, cgrid=FALSE, cpoint=2, addaxes=FALSE,
        origin=c(0,0), add.plot=TRUE)

text(-0.02502040,-0.163603600,"Alc")
text(-0.08760213,0.0005875967,"Alj")
text(-0.1243724,-0.0634414867,"Ber")
text(0.03680707,0.0876650573,"Cen")
```

```
text(-0.337088,0.1243761226,"Cob")
text(0.18614,-0.1145911354,"Dos")
text(0.075,0.0078347129,"Pri")
text(0.180,0.2487364153,"Ran")
text(0.14793443,0.04,"Saj")
text(-0.065,-0.1185534264,"Sac")
text(-0.238774,0.0293482729,"Sie")
text(0.1867166,-0.0185632943,"Tor")

# ejes 1 y 3

acp.cv$li

par(cex=1)

plot(c(-0.35,0.32), c(-0.25,0.25), type="n", axes=FALSE, xlab=NA, ylab=NA, cex.main=2, main="B")
s.class(dfxy=acp.cv$li, fac=tipo, xax=1, yax=3, grid=FALSE, cgrid=FALSE, cpoint=2, addaxes=FALSE,
        origin=c(0,0), add.plot=TRUE)

text(-0.055,-0.011283883,"Alc")
text(-0.08760213,0.039893456,"Alj")
text(-0.1243724,0.144555121,"Ber")
text(0.03680707,-0.021980026,"Cen")
text(-0.337088,0.015915171,"Cob")
text(0.19,-0.107255301,"Dos")
text(0.075,0.012104212,"Pri")
text(0.17652367,0.011361296,"Ran")
text(0.16793443,-0.045423463,"Saj")
text(-0.02,0.011999056,"Sac")
text(-0.238774,-0.150225873,"Sie")
text(0.1867166,0.073340232,"Tor")
```

C.3. Resultados numéricos del ACP

Est.	X	Y	Z
Alc	0.29	-1.10	-0.044
Alj	-0.46	-2.22	0.75
Ber	-1.36	0.03	0.62
Cen	-0.43	1.93	1.83
Cob	3.41	0.69	-0.81
Dos	-0.32	-0.78	0.69
Pri	-1.11	-0.46	-0.17
Ran	-0.81	-1.02	-1.79
Saj	-0.50	-0.039	0.087
Sac	0.13	0.70	-0.71
Sie	2.98	0.033	0.56
Tor	-1.79	2.23	-1.01

Tabla C.13 Coordenadas de los puntos representando las estaciones en la Figura 5.3.

Est.	X	Y	Z
Alc	-0.025	-0.18	-0.011
Alj	-0.057	0.00	0.039
Ber	-0.094	-0.063	0.14
Cen	-0.036	0.067	0.0079
Cob	0.38	0.12	0.016
Dos	0.14	-0.11	-0.10
Pri	0.10	0.0078	0.012
Ran	0.21	0.24	0.011
Saj	0.12	0.017	-0.045
Sac	-0.028	-0.11	0.024
Sie	-0.19	0.029	-0.15
Tor	-0.15	-0.018	0.073

Tabla C.14 Coordenadas de los puntos representando las estaciones en la Figura 5.4.

Variable	CP1	CP2	CP3
CO	-0.333	0.356	**0.778**
NO_2	**-0.566**	0.0363	-0.428
O_3	0.302	**-0.621**	0.410
PM_{10}	**-0.605**	-0.172	0.199
SO_2	-0.330	**-0.674**	0.0402

Tabla C.15 Cargas factoriales del ACP correspondiente a la Figura 5.3 basado en valores normalizados de μ_m. Los valores en negrilla se corresponden con las contribuciones más altas a las CCPP.

Variable	CP1	CP2	CP3
CO	-0.252	**0.922**	0.185
NO_2	-0.530	-0.116	**0.604**
O_3	0.213	-0.0571	0.295
PM_{10}	-0.474	0.203	**-0.690**
SO_2	**-0.619**	0.301	0.188

Tabla C.16 Cargas factoriales del ACP correspondiente a la Figura 5.4 basado en valores no normalizados de cv_m. Los valores en negrilla se corresponden con las contribuciones más altas a las CCPP.

D

Material suplementario del Capítulo 6

El contenido de este anexo es una adaptación del material suplementario de la publicación **Gómez-Losada, A., Pires, J.C.M., Pino-Mejías, R.** 2015. Time series clustering for estimating particulate matter contributions and its use in quantifying impacts from deserts. *Atmospheric Environment*, 117: 271-81.

D.1. Parametrización de las SSTT

La Tabla D.1 muestra los valores de las medias (μ_m) y desviaciones estándares (σ_m) de las SSTT durante 2013. En cada una de estas series, N indica el número de regímenes detectados, y μ_i y σ_i ($i = 1, \ldots, 4$), la media y desviación estándar de cada uno de ellos (en $\mu g/m^3$). Los errores estándares calculados mediante el método *bootstrap* se indican entre paréntesis. Las Tablas D.2-D.5 muestran semejante información de las SSTT, desde los años 2009 a 2012.

Área	Emplazamiento	N	μ_m	σ_m	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4
	ER	3	20.0	9.7	11.7	20.8	36.6	-	3.1	4.8	8.3	-
			(0.7)	(0.7)	(0.7)	(0.8)	(0.7)	-	(0.5)	(0.6)	(0.8)	-
	FM	4	13.8	17.1	6.9	14.5	26.3	92.3	2.9	3.4	7.2	53.9
			(0.8)	(0.6)	(0.8)	(0.7)	(0.7)	(1.9)	(0.4)	(0.4)	(0.4)	(2.8)
W	FU	3	11.4	6.7	6.4	11.5	20.7	-	2.1	2.4	7.3	-
			(0.8)	(0.8)	(0.7)	(0.8)	(0.7)	-	(0.6)	(0.7)	(0.7)	-
	MO	3	18.8	9.5	12.0	21.9	37.9	-	3.5	4.4	10.2	-
			(0.7)	(0.7)	(0.8)	(0.8)	(0.7)	-	(0.6)	(0.7)	(0.7)	-
AZ	FA	3	5.8	3.9	2.3	5.9	11.1	-	1.1	2.0	3.7	-
			(0.4)	(0.4)	(0.4)	(0.6)	(0.7)	-	(0.2)	(0.2)	(0.5)	-
	BE	3	12.9	6.3	8.8	15.2	30.3	-	2.4	3.4	9.6	-
			(0.8)	(0.5)	(0.6)	(0.6)	(0.7)	-	(0.4)	(0.4)	(0.5)	-
BA	MA	2	17.3	6.7	14.9	26.0	-	-	4.3	7.6	-	-
			(0.7)	(0.5)	(0.7)	(0.7)	-	-	(0.8)	(0.8)	-	-
	RI	4	27.2	39.8	11.3	18.1	35.8	105.5	2.9	3.4	11.4	98.0
			(0.7)	(0.7)	(0.8)	(0.7)	(1.0)	(3.7)	(0.7)	(0.7)	(0.4)	(3.3)
	TE	4	21.3	25.3	10.3	17.7	42.8	153.2	2.4	4.3	17.7	62.6
			(0.7)	(0.7)	(0.7)	(0.7)	(1.0)	(4.2)	(0.7)	(0.8)	(0.4)	(3.3)
CA	EC	4	17.8	31.1	8.2	19.7	43.6	221.4	3.3	4.9	12.8	200.2
			(0.7)	(0.8)	(0.7)	(0.7)	(0.7)	(6.5)	(0.7)	(0.8)	(0.4)	(5.4)
	TF	3	29.8	30.6	15.9	29.5	90.8	-	4.6	5.5	61.1	-
			(1.0)	(0.7)	(0.6)	(0.8)	(1.9)	-	(0.5)	(0.7)	(1.1)	-
	PE	4	8.8	5.3	4.9	7.4	11.0	21.8	1.2	1.4	2.3	7.2
			(0.4)	(0.3)	(0.4)	(0.5)	(0.8)	(0.4)	(0.4)	(0.3)	(0.6)	(0.7)
	CA	3	8.9	5.8	4.9	7.7	16.5	-	0.9	1.7	6.4	-
			(0.6)	(0.7)	(0.5)	(0.6)	(0.7)	-	(0.4)	(0.3)	(0.4)	-
CE	MN	3	10.0	6.2	4.3	9.2	17.4	-	1.6	2.5	5.3	-
			(0.8)	(0.7)	(0.7)	(0.7)	(0.8)	-	(0.4)	(0.4)	(0.7)	-
	AT	4	11.3	4.0	7.5	10.3	15.0	24.2	1.0	1.3	1.8	4.0
			(0.8)	(0.7)	(0.6)	(0.8)	(0.7)	(0.4)	(0.5)	(0.5)	(0.6)	(0.6)
	MF	4	10.7	7.1	4.3	9.0	17.2	41.3	1.6	2.2	4.9	4.1
			(0.7)	(0.7)	(0.8)	(0.8)	(0.8)	(0.3)	(0.4)	(0.4)	(0.5)	(0.7)
	ZA	3	11.6	5.8	6.8	12.2	20.3	-	1.7	2.7	4.9	-
			(0.7)	(0.7)	(0.7)	(0.8)	(0.7)	-	(0.2)	(0.2)	(0.6)	-
E	MR	3	7.9	4.9	2.2	6.6	14.1	-	1.2	2.3	2.8	-
			(0.8)	(0.7)	(0.7)	(0.6)	(0.7)	-	(0.3)	(0.4)	(0.3)	-
	PI	3	12.0	7.7	5.7	15.4	49.2	-	2.7	4.4	18.9	-
			(1.0)	(0.8)	(0.7)	(0.7)	(0.7)	-	(0.4)	(0.5)	(0.8)	-
	TO	3	11.3	6.5	7.8	14.6	67.5	-	2.6	3.9	13.5	-
			(0.7)	(0.7)	(0.8)	(0.4)	(0.4)	-	(0.1)	(0.4)	(0.5)	-
	CR	3	16.8	7.5	11.5	17.3	26.7	-	2.3	2.4	10.9	-
			(1.1)	(0.8)	(1.0)	(0.2)	(0.2)	-	(0.3)	(0.3)	(0.5)	-
NE	MG	3	10.4	6.5	3.5	8.7	17.1	-	1.0	2.7	6.3	-
			(0.7)	(0.7)	(0.7)	(0.4)	(0.3)	-	(0.2)	(0.5)	(0.5)	-
	MS	3	12.7	5.8	7.3	12.7	19.3	-	2.4	2.7	4.8	-
			(1.0)	(1.0)	(0.7)	(0.4)	(0.4)	-	(0.2)	(0.4)	(0.7)	-
	SA	3	6.7	4.2	3.5	7.4	14.1	-	1.3	2.0	2.9	-
			(0.5)	(0.3)	(0.4)	(0.7)	(0.2)	-	(0.2)	(0.3)	(0.4)	-
NW	NO	3	9.7	3.8	7.2	10.9	17.5	-	1.4	2.3	3.7	-
			(0.8)	(0.7)	(0.5)	(0.7)	(0.4)	-	(0.4)	(0.2)	(0.5)	-
	NI	2	15.3	7.0	10.8	20.5	-	-	3.5	6.4	-	-
			(0.7)	(0.7)	(0.6)	(0.7)	-	-	(0.6)	(0.7)	-	-
N	VA	3	13.3	8.1	5.6	13.6	25.1	-	2.2	3.8	7.0	-
			(0.8)	(0.8)	(0.5)	(0.8)	(0.4)	-	(0.3)	(0.7)	(0.8)	-
	PA	2	13.6	8.7	9.4	22.7	-	-	4.1	8.9	-	-
			(0.7)	(0.8)	(0.6)	(0.8)	-	-	(0.4)	(0.7)	-	-
	VI	4	13.8	8.6	5.6	12.9	19.9	36.1	2.2	2.2	3.5	8.6
			(0.7)	(0.7)	(0.5)	(0.6)	(0.3)	(0.3)	(0.3)	(0.6)	(0.4)	(0.6)
SE	AL	3	16.9	7.4	9.5	16.0	27.1	-	2.4	3.5	5.6	-
			(0.8)	(0.7)	(0.7)	(0.8)	(0.3)	-	(0.2)	(0.3)	(0.4)	-
	SN	3	12.8	7.8	7.1	11.6	22.4	-	1.9	2.4	7.4	-
			(0.8)	(0.7)	(0.7)	(0.8)	(0.4)	-	(0.1)	(0.2)	(0.2)	-
SW	BA	2	14.8	7.8	11.3	21.0	-	-	4.2	9.4	-	-
			(0.7)	(0.6)	(0.7)	(0.8)	-	-	(0.8)	(0.7)	-	-
	DO	3	14.6	7.7	8.6	15.5	25.3	-	2.3	3.3	6.1	-
			(0.5)	(0.4)	(0.7)	(0.7)	(0.4)	-	(0.2)	(0.2)	(0.3)	-

Tabla D.1 Parametrización de las SSTT en la Península Ibérica y Archipiélagos de las Azores, Balear y Canario durante el año 2013. Abreviaturas como en la Tabla 6.1 (pág. 74).

Área	Emplazamiento	Año	N	μ_m	σ_m	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4
AZ	FA	2012	3	5.4	3.3	1.6	3.7	8.2	-	0.5	1.1	2.7	-
				(0.4)	(0.2)	(0.4)	(0.5)	(0.6)	-	(0.3)	(0.4)	(0.3)	-
		2011	3	6.3	4.4	3.0	7.4	13.7	-	1.1	2.4	5.8	-
				(0.4)	(0.3)	(0.2)	(0.3)	(0.6)	-	(0.4)	(0.3)	(0.4)	-
		2010	4	6.6	3.9	2.7	6.2	9.1	11.6	1.1	1.5	1.5	4.4
				(0.5)	(0.4)	(0.2)	(0.2)	(0.5)	(0.4)	(0.1)	(0.2)	(0.4)	(0.4)
		2009	2	5.9	3.4	3.4	7.7	-	-	1.2	3.3	-	-
				(0.6)	(0.2)	(0.2)	(0.5)	-	-	(0.3)	(0.5)	-	-

Tabla D.2 Archipiélago de las Azores.

Área	Emplazamiento	Año	N	μ_m	σ_m	μ_1	μ_2	μ_3	μ_4	μ_5	σ_1	σ_2	σ_3	σ_4	σ_5
CE	PE	2012	4	9.3	8.5	4.1	7.1	13.4	35.5	-	1.1	1.5	3.4	18.8	-
				(0.7)	(0.7)	(0.4)	(0.6)	(0.3)	(0.4)	-	(0.3)	(0.4)	(0.6)	(0.8)	-
		2011	3	10.0	6.9	4.5	9.1	19.3	-	-	1.3	2.3	7.9	-	-
				(0.7)	(0.7)	(0.3)	(0.4)	(0.3)	-	-	(0.1)	(0.4)	(0.8)	-	-
		2010	3	8.7	8.3	4.2	10.2	33.1	-	-	1.4	3.2	23.6	-	-
				(0.8)	(0.8)	(0.4)	(0.4)	(0.2)	-	-	(0.2)	(0.5)	(1.0)	-	-
		2009	3	9.0	5.2	3.6	7.7	14.8	-	-	1.2	2.1	4.7	-	-
				(1.0)	(0.8)	(0.2)	(0.4)	(0.3)	-	-	(0.2)	(0.7)	(0.7)	-	-
	CA	2012	4	9.9	11.9	2.7	5.9	11.5	39.4	-	0.9	1.5	3.2	21.3	-
				(0.7)	(0.4)	(0.7)	(0.8)	(0.8)	(0.8)	-	(0.3)	(0.5)	(0.5)	(0.5)	-
		2011	3	10.0	8.3	3.7	8.6	15.5	27.4	-	1.5	2.5	2.8	14.4	-
				(0.7)	(0.6)	(0.5)	(0.5)	(0.8)	(0.7)	-	(0.4)	(0.3)	(0.2)	(0.6)	-
		2010	4	11.0	10.8	4.7	8.3	15.0	47.0	-	1.2	2.1	4.0	26.5	-
				(0.8)	(0.7)	(0.3)	(0.4)	(0.4)	(0.6)	-	(0.2)	(0.2)	(0.4)	(0.7)	-
		2009	3	10.8	7.5	2.6	7.2	17.1	-	-	1.3	1.8	6.5	-	-
				(0.8)	(0.4)	(0.6)	(0.8)	(0.7)	-	-	(0.2)	(0.3)	(0.3)	-	-
	MN	2012	5	11.8	12.4	4.7	8.6	13.0	21.5	61.4	1.4	1.7	1.9	5.7	26.0
				(0.7)	(0.5)	(0.4)	(0.6)	(0.8)	(0.7)	(2.1)	(0.3)	(0.2)	(0.4)	(0.4)	(0.3)
		2011	3	11.7	8.2	4.3	9.6	20.9	-	-	1.6	2.7	8.9	-	-
				(0.7)	(0.4)	(0.4)	(0.3)	(0.4)	-	-	(0.3)	(0.3)	(0.3)	-	-
		2010	4	10.9	11.0	4.8	10.7	16.6	44.9	-	1.8	1.9	3.0	25.0	-
				(0.8)	(0.5)	(0.4)	(0.3)	(0.4)	(0.3)	-	(0.4)	(0.4)	(0.3)	(0.4)	-
		2009	4	11.2	7.3	3.2	8.1	15.5	25.6	-	1.2	2.2	3.5	10.6	-
				(0.8)	(0.6)	(0.4)	(0.7)	(0.7)	(0.7)	-	(0.3)	(0.6)	(0.8)	(0.7)	-
	AT	2012	4	13.4	9.4	6.7	12.2	19.0	54.3	-	1.4	2.3	4.5	14.2	-
				(0.8)	(0.8)	(0.3)	(0.4)	(0.3)	(0.4)	-	(0.3)	(0.3)	(0.4)	(0.5)	-
		2011	4	15.0	7.2	8.0	14.1	19.8	30.1	-	1.9	2.4	3.3	10.0	-
				(0.8)	(0.8)	(0.4)	(0.4)	(0.5)	(0.5)	-	(0.3)	(0.3)	(0.4)	(0.3)	-
		2010	4	24.0	14.5	13.6	22.1	34.4	69.3	-	3.4	4.0	4.2	30.2	-
				(1.0)	(0.8)	(0.6)	(0.7)	(0.8)	(1.2)	-	(0.2)	(0.4)	(0.8)	(0.7)	-
		2009	3	24.3	9.4	14.4	22.8	34.6	-	-	2.0	4.2	7.3	-	-
				(1.2)	(0.7)	(0.5)	(0.6)	(1.1)	-	-	(0.3)	(0.5)	(0.7)	-	-
	MF	2012	3	11.6	9.4	6.5	14.0	37.6	-	-	2.6	4.0	19.2	-	-
				(0.7)	(0.4)	(0.5)	(0.7)	(1.0)	-	-	(0.4)	(0.5)	(0.8)	-	-
		2011	3	12.8	8.1	5.6	11.0	22.7	-	-	2.0	3.0	8.2	-	-
				(0.8)	(0.7)	(0.4)	(0.5)	(0.7)	-	-	(0.4)	(0.5)	(0.5)	-	-
		2010	3	13.8	16.6	7.2	17.0	89.4	-	-	2.5	4.7	45.7	-	-
				(0.7)	(0.2)	(0.5)	(0.7)	(1.0)	-	-	(0.3)	(0.6)	(0.7)	-	-
		2009	4	13.0	7.1	4.8	9.5	17.0	26.8	-	1.5	2.4	3.5	7.8	-
				(0.7)	(0.7)	(0.6)	(0.7)	(0.8)	(0.8)	-	(0.5)	(0.6)	(0.7)	(0.7)	-

Tabla D.3 Centro.

Área	Emplazamiento	Año	N	μ_m	σ_m	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4
N	NI	2012	3	15.9	8.9	8.7	16.7	30.0	-	3.1	4.4	10.9	-
				(1.2)	(0.7)	(0.7)	(0.8)	(1.0)	-	(0.8)	(0.7)	(0.8)	-
		2011	3	18.3	9.3	10.1	18.4	39.4	-	2.8	4.6	9.1	-
				(0.7)	(0.8)	(0.7)	(0.7)	(0.8)	-	(0.8)	(0.8)	(0.7)	-
		2010	4	15.4	7.9	7.3	12.4	21.3	52.6	1.8	2.7	5.5	25.4
				(0.7)	(0.7)	(0.7)	(0.8)	(0.7)	(0.8)	(0.3)	(0.4)	(0.7)	(0.8)
		2009	3	17.5	9.8	10.5	19.2	41.0	-	3.1	5.2	14.2	-
				(0.8)	(0.7)	(0.8)	(0.7)	(0.7)	-	(0.7)	(0.8)	(0.8)	-
	VA	2012	3	10.7	4.7	5.2	9.8	16.7	-	1.7	2.3	3.7	-
				(0.8)	(0.6)	(0.3)	(0.4)	(0.3)	-	(0.4)	(0.4)	(1.0)	-
		2011	3	9.5	4.9	5.5	8.8	16.1	-	1.2	2.3	5.2	-
				(0.7)	(0.7)	(0.5)	(0.4)	(0.6)	-	(0.4)	(0.4)	(0.8)	-
		2010	3	9.8	4.4	6.4	10.6	18.3	-	1.2	2.6	6.4	-
				(0.7)	(0.6)	(0.3)	(0.6)	(0.7)	-	(0.4)	(0.3)	(0.8)	-
		2009	3	13.4	5.8	8.1	13.3	19.1	-	1.9	1.9	5.3	-
				(0.7)	(0.2)	(0.7)	(0.7)	(0.8)	-	(0.4)	(0.3)	(0.7)	-
	PA	2012	3	12.8	8.0	6.0	11.4	22.9	-	2.0	3.3	7.9	-
				(0.8)	(0.6)	(0.4)	(0.7)	(0.8)	-	- (0.3)	(0.4)	(0.4)	-
		2011	3	13.6	8.7	8.1	14.0	23.4	-	2.6	3.5	8.2	-
				(0.8)	(0.7)	(0.3)	(0.4)	(0.7)	-	(0.3)	(0.4)	(0.5)	-
		2010	4	11.4	5.9	4.9	9.0	14.7	22.6	0.9	2.0	2.8	6.9
				(0.8)	(0.7)	(0.8)	(0.7)	(1.2)	(0.3)	(0.7)	(0.7)	(1.3)	(1.1)
		2009	3	13.1	8.2	6.3	14.8	35.4	-	2.1	4.7	4.2	-
				(0.7)	(0.7)	(0.7)	(2.1)	-	-	(0.4)	(0.7)	(1.0)	-

Tabla D.4 Norte.

Área	Emplazamiento	Año	N	μ_m	σ_m	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4
SE	VI	2012	4	16.7	13.9	6.2	12.6	19.6	45.5	2.1	2.4	4.0	17.7
				(0.7)	(0.7)	(0.3)	(0.6)	(0.8)	(1.0)	(0.4)	(0.6)	(0.7)	(0.8)
		2011	3	17.3	11.7	8.7	19.0	37.7	-	3.6	5.1	12.6	-
				(0.7)	(0.3)	(0.3)	(0.7)	(0.4)	-	(0.3)	(0.5)	(0.6)	-
		2010	4	16.0	12.4	4.4	9.8	19.3	40.3	1.3	2.9	3.5	20.0
				(0.8)	(0.6)	(0.6)	(0.8)	(0.7)	(0.8)	(0.7)	(0.3)	(0.3)	(0.4)
		2009	4	17.2	11.0	4.4	10.8	20.3	36.9	1.5	3.2	4.5	11.4
				(1.0)	(0.6)	(0.4)	(0.8)	(0.2)	(0.5)	(0.8)	(0.7)	(0.8)	(0.8)
	AL	2012	3	18.4	11.3	10.1	17.7	31.6	-	2.6	3.5	12.8	-
				(0.7)	(0.5)	(0.4)	(0.8)	(0.4)	-	(0.3)	(0.4)	(0.3)	-
		2011	3	18.3	8.4	8.4	16.4	26.9	-	2.12	3.8	7.0	-
				(1.0)	(0.6)	(0.6)	(0.7)	(0.4)	-	(0.4)	(0.4)	(0.4)	-
		2010	4	18.4	10.8	9.0	16.7	28.7	84.1	2.5	3.5	5.8	22.3
				(1.1)	(0.6)	(0.7)	(0.8)	(0.4)	(0.4)	(0.4)	(0.8)	(0.8)	(0.7)
		2009	3	22.4	9.6	13.0	22.1	35.0	-	3.3	4.3	8.0	-
				(1.3)	(0.4)	(0.5)	(0.7)	(0.8)	-	(0.4)	(0.7)	(0.8)	-

Tabla D.5 Sureste.

D.2. Implementación computacional de los MOM con **depmixS4**

El siguiente código en R se presenta como ejemplo de la modelización con MOM de las SSTT analizadas, el cual obtiene los resultados de la Tabla 6.2 (pág. 76). No obstante, se desarrollaron otra rutinas en R, basadas en la que sigue, para obtener las parametrizaciones de todas las series (77).

```
library(depmixS4)

# El objeto "data" es un vector numérico conteniendo la serie temporal (ST).
# Se crea un marco de datos:

sample<-data.frame(y=data)

# Mediante la función "depmix", se generan 7 modelos diferentes para ajustar la ST,
# desde un estado oculto (ns=1) a siete (ns=7).

m1<-depmix(y~1, data=sample, ns=1, ntimes=nrow(sample))
m2<-depmix(y~1, data=sample, ns=2, ntimes=nrow(sample))
m3<-depmix(y~1, data=sample, ns=3, ntimes=nrow(sample))
m4<-depmix(y~1, data=sample, ns=4, ntimes=nrow(sample))
m5<-depmix(y~1, data=sample, ns=5, ntimes=nrow(sample))
m6<-depmix(y~1, data=sample, ns=6, ntimes=nrow(sample))
m7<-depmix(y~1, data=sample, ns=7, ntimes=nrow(sample))

# Se obtienen los parámetros de cada modelo mediante la función "fit".
# Se especifican los valores de los parámetros del algoritmo EM ("maxit", "tol"),
# y el criterio de convergencia seleccionado ("crit").

fm1<-fit(m1, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm2<-fit(m2, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm3<-fit(m3, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm4<-fit(m4, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm5<-fit(m5, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm6<-fit(m6, em=em.control(maxit=2000, tol=1e-08, crit="relative"))
fm7<-fit(m7, em=em.control(maxit=2000, tol=1e-08, crit="relative"))

# El estadístico BIC se calcula para cada modelo mediante la función "BIC".
# Los valores se almacenan en un vector denominado "bic".

bic<-c(BIC(fm1),BIC(fm2),BIC(fm3),BIC(fm4),BIC(fm5),BIC(fm6),BIC(fm7))

# El "mejor" modelo es aquel con menor BIC:

> which.min(bic)
[1] 4

# El modelo de 4 estados, "fm4", describe la ST mejor.
# La parametrización completa del modelo elegido se obtendrá
# mediante la función "summary": incluye los parámetros
# de las cuatro curvas gaussianas (mu y sigma) y la matriz de probabilidades
# de transición (m.p.t.).

> summary(fm4)
Initial state probabilties model
pr1 pr2 pr3 pr4
  1   0   0   0

Transition matrix
                toS1         toS2       toS3         toS4
fromS1  8.706526e-01 0.1099947038 0.01935269 6.945374e-56
fromS2  2.213438e-01 0.6793470074 0.06635697 3.295224e-02
fromS3  2.534007e-02 0.1697684571 0.78155134 2.334013e-02
fromS4 1.312388e-110 0.0001419835 0.66654146 3.333166e-01

Response parameters
Resp 1 : gaussian
    Re1.(Intercept)    Re1.sd
St1        10.32563  2.418200 #mu1 y sigma1
St2        17.72293  4.346021 #mu2 y sigma2
St3        42.75554 17.701547 #mu3 y sigma3
St4       153.25253 62.565625 #mu4 y sigma4
```

```
# Los valores "St" significan Estado.
# Generalmente, estas parejas de valores habrán de ser ordenadas,
# de menor a mayor valores de mu.

# A continuación se calculan los valores "pi" de cada curva gaussiana:

probs<-posterior(fm4)
colMeans(probs[,2:5])

# Se obtiene a continuación los valores pi1, pi2, pi3 y pi4,
# cuya suma es 1.

> colMeans(probs[,2:5])
        S1         S2         S3         S4
0.53226225 0.26543964 0.18137000 0.01992811

# Para solventar el problema del número de dígitos significativos,
# el valor "S4" puede obtenerse como 1-(S1+S2+S3),
# debiendo asumirse cierta pérdida de precisión, aunque es poco signifcativa:
# 0.01992811 vs 0.022 (ver Tabla 6.2).
# El mismo criterio puede ser aplicado a la última columna de la m.p.t.
```

E

Material suplementario del Capítulo 7

El contenido de este anexo es una adaptación del material suplementario de la publicación **Gómez-Losada, A., Pires, J.C.M., Pino-Mejías, R.** 2016. Characterization of background air pollution exposure in urban environments using a metric based on Hidden Markov Models. *Atmospheric Environment*, 127: 255-61.

E.1. Parametrización de las SSTT

Las Tablas E.1 a E.13 contienen la media (M) y desviación estándar (SD) de las SSTT de los contaminantes monitorizados en estaciones de Jaén, Granada y Sevilla, desde el año 2010 al 2013, así como los valores de la media ($\bar{x}$) y desviación estándar (s) de los conjuntos de datos analizados (en μg/m^3). En cada ST, T indica la longitud de la ST, K el número de regímenes (clústers), y m_1 y sd_1, la media y la desviación estándar del régimen asociado a la contaminación de fondo (primer clúster), respectivamente. La notación M, SD, m_1, sd_1 se corresponde, una a una, con μ_m, σ_m, μ_1, σ_1 del Capítulo 6.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Alc	CO	2013	359	4	277.6	278.0	64.4	63.8	199.4	27.6
			2012	356	4	306.3	305.5	54.1	53.8	233.0	12.6
			2011	327	4	376.6	376.8	57.5	57.8	275.2	42.0
			2010	352	3	337.0	337.6	42.1	42.3	305.5	9.5
		NO_2	2013	360	2	16.4	16.5	8.0	8.1	13.1	3.9
			2012	355	3	20.7	20.6	10.2	10.2	11.0	2.6
			2011	342	3	20.9	20.9	10.0	10.0	14.1	3.8
			2010	345	2	19.2	19.2	8.6	8.6	15.2	4.8
		PM_{10}	2013	352	2	28.0	27.9	10.1	10.0	19.7	4.5
			2012	352	4	29.3	29.5	14.4	14.5	16.6	4.1
			2011	339	4	31.0	31.1	13.3	13.4	15.9	3.6
			2010	306	3	31.3	31.3	15.3	15.3	19.7	4.6
		SO_2	2013	352	5	4.7	4.7	1.3	1.4	3.4	0.3
			2012	355	5	4.8	4.8	1.7	1.7	2.3	0.3
			2011	340	4	6.3	6.3	1.1	1.1	4.8	0.3
			2010	347	4	6.2	6.2	0.7	0.7	5.4	0.3

Tabla E.1 Alcalá de Guadaíra.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Alj	NO_2	2013	343	3	15.1	15.0	8.2	8.1	5.6	1.3
			2012	361	3	17.3	17.2	9.0	9.0	7.5	1.5
			2011	361	3	17.9	17.9	10.2	10.2	11.5	4.5
			2010	362	3	17.5	17.5	10.5	10.4	8.5	2.5
		PM_{10}	2013	346	3	25.8	26.0	9.3	9.4	16.8	3.6
			2012	359	3	30.1	30.3	14.7	14.9	20.1	5.4
			2011	361	3	35.7	35.8	16.4	16.4	18.3	4.7
			2010	358	3	32.9	33.1	16.4	16.6	21.2	5.7
		SO_2	2013	353	2	5.6	5.6	1.3	1.3	4.8	0.8
			2012	358	4	6.8	6.8	3.1	3.1	1.4	0.
			2011	360	4	7.4	7.4	3.3	3.3	1.2	0.7
			2010	365	4	3.4	3.4	2.7	2.7	0.3	0.2

Tabla E.2 Aljarafe.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Ber	CO	2013	333	4	459.3	458.8	185.0	186.0	186.4	83.3
			2012	363	5	456.4	458.0	143.7	145.4	237.8	27.3
			2011	363	4	650.6	654.1	168.0	168.0	446.5	65.6
			2010	349	4	549.0	550.0	146.6	146.7	250.6	83.2
		NO_2	2013	327	3	21.8	21.6	12.8	12.8	9.6	4.8
			2012	340	3	21.5	21.5	14.1	14.2	10.3	6.3
			2011	361	3	28.0	28.0	15.0	15.1	11.8	6.0
			2010	334	3	32.4	32.9	18.4	18.7	22.8	8.8
		PM_{10}	2013	332	3	27.9	28.0	11.8	11.8	17.1	5.1
			2012	336	3	32.7	32.9	15.3	15.5	18.3	5.5
			2011	45	-	-	-	-	-	-	-
			2010	342	3	18.8	18.9	8.4	8.5	13.7	3.0
		SO_2	2013	331	2	4.9	4.9	1.5	1.5	3.8	0.5
			2012	324	3	5.0	5.0	1.8	1.8	3.4	0.2
			2011	360	2	5.3	5.3	1.7	1.7	4.4	0.6
			2010	332	2	5.6	5.6	1.5	1.5	5.1	0.7

Tabla E.3 Bermejales.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Cen	CO	2013	325	6	659.5	657.3	276.8	275.1	268.0	41.4
			2012	340	7	705.8	704.6	291.9	292.2	357.7	73.1
			2011	361	5	744.1	746.0	326.3	325.3	424.0	57.2
			2010	358	5	517.3	517.1	160.0	160.3	304.5	36.1
		NO_2	2013	343	3	23.0	22.9	8.0	8.0	12.3	2.3
			2012	344	3	21.5	21.6	9.4	9.4	3.8	3.8
			2011	359	3	25.8	25.8	10.3	10.3	9.6	5.4
			2010	337	2	27.1	27.2	11.0	11.0	10.2	7.1
		SO_2	2013	348	3	3.1	3.1	0.7	0.7	2.3	0.4
			2012	347	3	2.7	2.7	1.0	1.1	1.4	0.2
			2011	363	4	3.1	3.1	1.0	1.0	2.0	0.2
			2010	364	4	2.6	2.6	0.9	0.9	0.2	0.3

Tabla E.4 Centro.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Cla	CO	2013	357	6	218.9	217.5	160.2	159.3	36.9	25.9
			2012	356	5	363.5	363.3	80.7	80.6	278.5	14.3
			2011	339	6	436.7	436.7	163.3	162.0	190.2	53.5
			2010	361	4	317.9	317.8	144.2	144.3	160.8	36.6
		NO_2	2013	328	2	21.0	21.0	10.2	10.2	15.8	6.7
			2012	349	3	20.4	20.5	10.9	10.9	11.1	5.4
			2011	341	3	20.6	20.6	11.0	11.1	12.8	5.7
			2010	362	2	29.3	29.5	9.8	9.9	25.3	7.6
		PM_{10}	2013	365	3	22.9	23.1	9.4	9.6	17.1	5.2
			2012	355	3	25.1	25.3	13.1	13.3	13.9	5.6
			2011	335	3	31.6	31.7	15.3	15.5	19.8	6.0
			2010	362	3	31.0	31.2	14.4	14.4	20.1	5.4

Tabla E.5 Santa Clara.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Dos	CO	2013	324	4	521.8	523.8	93.3	93.3	403.4	36.1
			2012	315	4	467.3	469.0	122.7	122.0	306.7	78.7
			2011	345	4	423.7	423.9	146.3	145.6	236.3	47.2
			2010	353	4	370.3	368.9	131.3	131.6	149.4	21.0
		NO_2	2013	341	2	17.8	17.8	7.6	7.6	14.0	3.7
			2012	344	2	19.9	20.0	7.6	7.6	15.1	3.7
			2011	355	3	20.2	20.2	8.0	8.1	13.8	2.9
			2010	355	3	18.4	18.4	7.1	7.1	13.6	2.8
		SO_2	2013	320	5	5.4	5.4	1.0	1.0	4.3	0.2
			2012	339	4	5.8	5.8	1.1	1.1	4.1	0.4
			2011	352	4	5.6	5.6	0.8	0.8	4.1	0.4
			2010	358	4	5.3	5.3	0.7	0.7	4.2	0.8

Tabla E.6 Dos Hermanas.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Pri	CO	2013	346	4	388.4	389.8	118.5	118.5	252.3	64.0
			2012	334	5	414.0	413.1	158.7	158.0	245.9	72.6
			2011	350	4	413.4	417.7	187.1	193.0	315.9	55.0
			2010	360	4	427.5	428.8	111.0	111.2	316.1	125.9
		NO_2	2013	307	3	26.8	26.9	11.4	11.4	16.6	6.2
			2012	310	3	30.0	30.1	12.5	12.6	18.9	7.3
			2011	281	3	32.1	32.2	13.1	13.2	20.0	6.7
			2010	334	2	26.3	26.4	12.4	12.4	17.8	6.7
		PM_{10}	2013	330	3	25.3	25.3	9.9	9.9	16.3	4.3
			2012	348	4	28.8	28.9	13.9	14.1	15.4	4.8
			2011	351	2	31.6	31.8	12.8	12.9	23.8	7.2
			2010	219	3	34.1	34.3	14.3	14.6	24.1	8.7
		SO_2	2013	354	4	6.2	6.2	1.7	1.7	4.5	0.4
			2012	343	4	5.2	5.2	1.6	1.6	3.3	0.6
			2011	351	4	6.2	6.2	1.4	1.4	4.3	0.4
			2010	359	3	5.8	5.8	1.2	1.2	4.5	0.6

Tabla E.7 Príncipes.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Ran	CO	2013	351	4	159.5	159.3	100.1	100.8	110.8	7.1
			2012	358	4	216.0	215.7	143.4	143.6	121.1	10.8
			2011	352	5	273.8	272.7	250.3	249.1	113.1	7.7
			2010	352	3	650.4	651.7	138.4	140.2	519.6	41.1
		NO_2	2013	347	3	28.4	28.2	11.8	11.6	20.0	6.3
			2012	356	3	37.0	37.1	13.6	13.6	23.8	6.1
			2011	346	2	37.3	37.5	15.1	15.2	29.0	8.9
			2010	353	2	33.8	33.8	13.5	13.6	28.5	9.6
		SO_2	2013	349	4	5.6	5.6	1.6	1.6	3.9	0.4
			2012	356	3	5.6	5.6	1.6	1.6	4.4	0.5
			2011	347	4	5.5	5.5	2.0	2.0	3.8	0.5
			2010	309	4	4.1	4.1	1.2	1.2	3.1	0.1

Tabla E.8 Ranilla.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Sevilla	Tor	CO	2013	323	4	422.0	421.6	172.4	174.3	158.9	53.9
			2012	361	4	471.3	471.2	193.3	193.6	302.1	72.8
			2011	362	3	608.7	609.6	168.1	169.0	361.5	39.4
			2010	349	3	612.7	612.9	107.9	108.3	468.2	81.1
		NO_2	2013	315	3	36.8	36.8	11.7	11.6	24.9	7.3
			2012	360	3	33.6	33.6	15.4	15.4	13.4	6.8
			2011	355	2	45.5	45.4	12.1	12.1	37.6	8.2
			2010	339	4	37.5	37.4	21.5	21.3	16.6	6.7
		PM_{10}	2013	346	3	30.0	30.2	10.3	10.4	20.8	5.0
			2012	306	3	29.3	29.7	12.1	12.2	20.0	5.4
			2011	330	3	42.4	42.6	11.3	11.5	34.3	5.3
			2010	339	3	27.5	27.4	12.2	12.2	19.1	2.4
		SO_2	2013	321	4	3.6	3.5	1.0	1.0	2.2	0.4
			2012	361	4	3.7	3.7	0.9	1.0	2.2	0.3
			2011	362	4	3.8	3.8	0.9	0.9	3.0	0.2
			2010	349	4	3.8	3.8	0.9	0.9	2.9	0.5

Tabla E.9 Torneo.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Jaén	Fue	CO	2013	351	3	200.5	200.6	83.6	83.6	126.9	12.8
			2012	361	4	190.3	190.2	69.3	69.3	119.8	14.1
			2011	356	4	237.7	237.2	80.1	80.4	135.6	42.4
			2010	180	3	158.8	158.7	83.2	83.2	97.7	31.4
		NO_2	2013	357	4	12.0	12.0	6.5	6.5	4.2	0.6
			2012	301	3	12.3	12.3	6.2	6.3	7.0	1.6
			2011	338	4	9.8	9.7	5.1	5.1	5.1	2.1
			2010	316	3	10.5	10.5	7.2	7.2	2.8	0.6
		SO_2	2013	357	6	6.3	6.3	1.3	1.3	4.9	0.2
			2012	363	5	6.1	6.1	1.7	1.7	3.3	0.8
			2011	344	4	7.3	7.3	1.9	1.9	5.6	0.3
			2010	354	5	6.6	6.6	1.4	1.4	3.8	0.3

Tabla E.10 Fuentezuelas.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Jaén	Ron	CO	2013	354	3	446.3	446.5	118.9	120.5	336.8	55.3
			2012	360	3	337.1	337.6	131.3	131.3	275.0	43.5
			2011	351	4	390.9	391.2	151.2	152.9	90.9	47.9
			2010	327	2	422.1	423.1	134.6	134.6	390.1	95.5
		NO_2	2013	353	3	20.8	20.8	10.4	10.4	13.0	13.0
			2012	358	3	23.6	23.6	11.1	11.1	14.0	3.7
			2011	350	3	24.2	24.3	11.5	11.5	16.6	4.5
			2010	353	2	23.0	23.0	9.7	9.7	18.3	5.3
		PM_{10}	2013	364	3	23.2	23.2	11.2	11.2	13.2	3.5
			2012	358	4	28.0	28.0	16.5	16.5	15.1	3.8
			2011	355	3	32.2	32.2	14.1	14.1	20.2	5.3
			2010	351	3	31.5	31.5	15.9	16.1	17.6	5.1
		SO_2	2013	331	3	4.2	4.2	1.6	1.6	2.9	0.7
			2012	361	3	4.4	4.4	1.6	1.6	2.8	0.7
			2011	349	3	4.2	4.2	1.5	1.6	2.5	0.7
			2010	299	2	3.6	3.6	1.7	1.7	2.0	0.6

Tabla E.11 Ronda del Valle.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Granada	Pal	CO	2013	365	4	335.4	338.4	125.2	126.2	182.0	27.8
			2012	365	4	288.9	288.7	135.4	135.4	139.2	29.0
			2011	365	4	264.5	263.7	123.7	122.8	122.4	31.2
			2010	365	4	269.4	269.7	270.9	270.6	61.6	27.4
		NO_2	2013	354	2	30.9	31.2	11.8	11.8	24.8	7.0
			2012	365	2	34.0	34.0	12.1	12.1	27.0	7.4
			2011	363	3	32.8	33.0	12.1	12.1	18.7	4.5
			2010	361	2	34.5	34.6	11.9	12.0	28.0	7.6
		PM_{10}	2013	357	3	24.5	24.5	9.8	9.8	13.1	3.8
			2012	365	4	29.1	29.0	15.4	15.4	13.8	4.3
			2011	364	4	29.9	29.9	11.8	11.7	16.4	3.9
			2010	218	3	31.8	31.9	16.1	16.1	22.3	6.5
		SO_2	2013	365	3	11.9	11.8	3.2	3.2	8.7	1.7
			2012	364	3	11.4	11.4	3.7	3.7	7.1	1.8
			2011	365	4	10.1	10.2	4.5	4.5	5.2	1.4
			2010	365	4	6.9	6.9	3.0	3.1	3.5	0.3

Tabla E.12 Palacio de Congresos.

.

Ciudad	Estación	Contaminante	Año	T	K	M	$\bar{x}$	SD	s	m_1	sd_1
Granada	Nor	CO	2013	362	4	443.0	443.1	179.7	178.5	241.2	55.9
			2012	363	4	413.9	415.4	136.4	135.8	241.1	50.9
			2011	365	4	407.6	408.4	176.3	176.8	165.7	36.0
			2010	355	4	517.1	514.4	196.9	197.2	245.7	75.1
		NO_2	2013	363	3	42.1	42.1	17.6	17.5	28.0	7.2
			2012	364	3	45.9	46.1	17.8	17.8	28.1	7.5
			2011	360	3	48.1	48.2	17.2	17.2	36.2	8.7
			2010	359	3	46.8	46.9	19.6	19.7	25.3	6.2
		PM_{10}	2013	364	3	25.6	25.7	11.5	11.5	13.1	3.1
			2012	362	3	31.4	31.5	15.3	15.6	16.5	5.2
			2011	259	3	35.8	35.9	16.1	16.0	21.5	5.6
			2010	308	3	38.2	38.2	20.7	20.8	23.9	7.8
		SO_2	2013	365	3	6.9	7.0	1.6	1.6	5.4	0.7
			2012	365	3	7.7	7.7	1.9	2.0	5.6	0.6
			2011	365	4	8.7	8.7	2.2	2.2	5.1	0.6
			2010	361	3	8.9	8.9	2.5	2.5	6.9	0.8

Tabla E.13 Granada Norte.