Control Ambiental en el Puerto de Tarragona Análisis Bayesiano

José-Miguel Bernardo
Universitat de València, España

Memoria Final

Junio 2005

Documento producido para el Ente Público *Puertos del Estado* en virtud del convenio de colaboración suscrito en Junio de 2004 entre *Puertos del Estado* y la *Universitat de València*.

Valencia, Junio de 2005.

Índice

Capítulo 1.

Introducción

1.1. EL PROYECTO HADA

En el informe metodológico que precede a este documento ya se describe el esfuerzo realizado en España por Puertos del Estado para un propiciar una gestión medioambiental adecuada de las actividades portuarias:

> La notable importancia que los ciudadanos actualmente conceden a los problemas medioambientales aparece reflejada en las directivas de calidad medioambiental, crecientemente exigentes, que son emitidas por la Unión Europea y transcritas por las Administraciones de los Estados miembros a sus respectivas legislaciones. Por otra parte, más allá de las exigencias legales, los actores económicos son ya conscientes de que una parte importante de su imagen depende del interés que demuestren en minimizar los posibles impactos negativos de sus actividades sobre el medio ambiente.
>
> En esta línea, el Ente Público *Puertos del Estado* tomó la iniciativa de potenciar, en el contexto de los proyectos europeos LIFE-Medio Ambiente, el diseño de una *Herramienta Automática de Diagnóstico Ambiental* (HADA) que deberá permitir una gestión automatizada, en tiempo real, de los problemas medioambientales que puedan derivarse de las operaciones que se desarrollan en los puertos marítimos españoles.
>
> El ámbito de intervención especificado en el proyecto HADA es muy amplio, incluyendo, entre otros aspectos, el desarrollo y ordenación del uso del territorio, la gestión del agua, el control de los impactos de las actividades económicas y la gestión de residuos, todo ello coordinado en el marco una política de producto integrado.

En una primera fase, desde principios de 2001 *Puertos del Estado* coordinó los esfuerzos de las autoridades portuarias de ocho puertos españoles, específicamente los de *la Coruña, Barcelona, Bilbao, Cartagena, Huelva, Santander, Tarragona y Valencia* para identificar los problemas problemas potenciales que las actividades portuarias pudieran generar, y para estudiar la forma de prevenirlos.

Entre las tareas que se han puesto en marcha como consecuencia del desarrollo del Proyecto HADA se incluyen:

1. La implantación en varios puertos de un sistema de monitorización en tiempo real de las variables que describen calidad ambiental.
2. El desarrollo de un sistema de modelización atmosférica que deberá permitir predicciones microclimáticas a corto plazo en tiempo real.
3. El diseño de estrategias que faciliten la toma racional de decisiones frente a episodios de contaminación portuaria.
4. El diseño de un sistema informático de control de la contaminación atmosférica y de ayuda en la toma de decisiones.
5. La puesta en marcha de una campaña de medidas de pequeñas partículas en suspensión (PM_{10}) .
6. La modelización de los factores de que contribuyen emisión de partículas en suspensión en áreas portuarias.
7. El estudio comparativo de posibles medidas atenuantes y correctoras de impactos ambientales no deseados.
8. La elaboración de un protocolo de seguimiento, evaluación y control de ruidos generados por las actividades portuarias.

A medida que son obtenidos, los resultados de estos trabajos están siendo divulgados a nivel internacional, con espacial énfasis en los países miembros de la Unión Europea.

En esta línea, *Puertos del Estado* y la *Universitat de València* firmaron en Junio de 2001 un primer convenio de colaboración para estudiar las características que debería tener un sistema de gestión automatizada de control ambiental de actividades portuarias, y para determinar la información que debería ser recogida, analizada y modelizada para poder hacer operativo tal sistema de gestión. Los resultados de ese trabajo fueron sintetizados en una Memoria presentada en Julio de 2002.

Con objeto de implementar la metodología propuesta en una situación real, *Puertos del Estado* llegó a un acuerdo con la *Autoridad Portuaria de Tarragona* para que el Puerto de Tarragona hiciese de centro piloto en la gestión automatizada del control ambiental de actividades portuarias. Con este objeto, fué instalada una cabina de medición automática de las variables que definen la calidad del aire, y se solicitaron los recurso humanos necesarios para crear una base de datos en la que, de

manera continua, quedaran reflejadas tanto la calidad del aire como las actividades portuarias y las covariables climáticas que pudieran alterarla.

En este contexto *Puertos del Estado* y la *Universitat de València* firmaron en Junio de 2004 un nuevo convenio de colaboración para, partiendo de esa base de datos, realizar un estudio estadístico que permita implementar en el Puerto de Tarragona una estrategia de control ambiental basada en la teoría Bayesiana de la decisión. En este documento se resumen los resultados obtenidos.

1.2. LOS PROBLEMAS PLANTEADOS

De acuerdo con la línea de trabajo adoptada en el conjunto de todo el proyecto HADA, la atención se ha centrado en los problemas de contaminación atmosférica que pueden derivarse de la manipulación de graneles sólidos pulverulentos, tanto en operaciones de carga y descarga, como en el transporte y en el almacenamiento temporal de tales graneles en el interior las áreas portuarias, con especial atención a los niveles de partículas PM_{10}, cuyos valores límite para la protección de la salud humana fueron regulados por una directiva de la Unión Europea publicada en 1999, implementada en la legislación española en 2002, y que entró en vigor el 1 de Enero de 2005.

Aunque en este trabajo se procede al análisis específico de los datos correspondientes al Puerto de Tarragona (relativos al periodo comprendido entre Mayo de 2004 y Marzo de 2005), es importante subrayar que la metodología descrita en este documento para diseñar las estrategias apropiadas frente a este tipo de contaminación es de *aplicación universal* y, consecuentemente, puede ser adaptada con facilidad a las condiciones de cualquier otro puerto y, con algunas modificaciones, a los problemas planteados por otros episodios de contaminación ambiental.

La gestión medioambiental de un sistema tan complejo como el de un puerto marítimo exige una consideración integrada de los problemas enfrentados, de sus probables interrelaciones, de los factores de los que dependen los resultados de sus posibles soluciones, de la información disponible y de la valoración política de sus posibles consecuencias. Esto es posible en el marco de la *teoría de la decisión*, una construcción normativa que prescribe la *única* forma de tomar decisiones en ambiente de incertidumbre compatible con un comportamiento racional.

Naturalmente, las decisiones deben ser tomadas haciendo uso de toda la información relevante que resulte accesible, de forma que resulta crucial disponer de un mecanismo que permita la actualización inmediata de la información disponible en función de los datos que se van obteniendo. Los *métodos estadísticos bayesianos* constituyen la *única* forma consistente de incorporar información experimental adicional a la información inicialmente disponible.

En este trabajo se describe la forma en la que la teoría de la decisión y la metodología estadística bayesiana pueden ser utilizadas para resolver los problemas

de decisión en ambiente de incertidumbre a los que se enfrenta la Autoridad Portuaria ante el riesgo de contaminación por valores excesivos de PM_{10}.

Desde el punto de vista de los objetivos perseguidos deben distinguirse dos tipos de problemas que requieren un tratamiento diferenciado: los derivados de garantizar el cumplimiento de las normas medio ambientales vigentes, en los que este trabajo está fundamentalmente centrado, y los que son consecuencia de las propias iniciativas del puerto para controlar su impacto ambiental y para mejorar su imagen pública.

Según el horizonte temporal contemplado aparecen asimismo dos tipos de problemas: los que permiten considerar la construcción o la mejora de infraestructuras (que generalmente se plantean a medio o largo plazo) y los que requieren decisiones inmediatas, cualesquiera que sean los medios disponibles.

Finalmente, desde la perspectiva de los datos necesarios para una gestión responsable, se identifican aquellas series de datos temporales relevantes, cuyo control permite una predicción probabilística de los sucesos inciertos relevantes, con especial atención a aquellos que pueden implicar un incumplimiento de la normativa medioambiental legalmente vigente.

1.3. CONTENIDO DE LA MEMORIA

Los fundamentos metodológicos sobre los que descansan las soluciones propuestas ya fueron descritos en el informe metodológico presentado en Noviembre de 2004. En particular, en ese informe se especifica la estructura formal de los problemas de decisión en ambiente de incertidumbre y se describen brevemente los resultados más importantes de la teoría de la decisión. Se fefine y se describe una medida general de asociacin estóadística, basada en la teoría de la información, la medida de asociación intrínseca, que se utiliza de forma intensiva en este trabajo. Se analizan las características básicas de los métodos estadísticos bayesianos. Finalmente, el informe metodológico contiene una bibliografía comentada para quienes deseen profundizar en los fundamentos y aplicaciones de los métodos descritos.

Esta memoria contiene cinco capítulos, de los que el primero es esta introducción, seguidos por dos apéndices.

En en capítulo segundo se detallan los elementos básicos del problema de decisión al que se enfrenta la Autoridad Portuaria comprometida en una gestión racional del riesgo de contaminación atmosférica que puede derivarse de la manipulación de graneles sólidos pulverulentos, con especial atención a los niveles de PM_{10}. En particular, se discuten las alternativas posibles, se identifican los sucesos inciertos relevantes, se describen las posibles consecuencias de las alternativas de que se dispone y se propone una forma general de precisar la estructura de preferencias de la Autoridad portuaria.

En el capítulo tercero se analiza con detalle el modelo probabilístico que, a partir de los datos disponibles, va a permitir construir una distribución de probabilidad sobre el conjunto de sucesos inciertos relevantes. El capítulo empieza con un análisis descriptivo de la serie temporal multivariante proporcionada por la Autoridad Portuaria de Tarragona que contiene información horaria sobre calidad del aire, condiciones climatológicas y actividades portuarias a lo largo de 11 meses consecutivos, de Mayo de 2004 a Marzo de 2005. Se describen además los datos de calibrado de las medidas automáticas de PM_{10} en términos de medidas gravimétricas de referencia proporcionados por el Consejo Superior de Investigación Científicas. Los datos observados de calidad del aire no pueden ser considerados como una muestra de ningún modelo probabilístico estándar. En el apéndica A.1 se describe una metodología bayesiana objetiva para el análisis no-paramétrico de las observaciones disponibles específicamente adaptada para este proyecto; los detalles técnicos han sido redactados, en inglés, en forma de trabajo académico para este material que pueda ser reutilizado en otros contextos. La metodología introducida es utilizada en distintos momentos de este trabajo, pero especialmente para determinar las distribuciones predictivas, marginales y condicionales de la característica medioambiental que se pretende controlar, la concentración media diaria de partículas PM_{10} en el aire. Los resultados obtenidos son validados mediante funciones de evaluación logarítmicas sobre valores no utilizados en la construcción de la distribución predictiva.

En el cuarto capítulo de esta memoria se utilizan modelos bayesianos de regresión no-lineal y no-paramétrica para determinar las covariables que intervienen en los episodios de contaminación atmosférica. En particular, se determinan las variables que pueden utilizarse para predecir la aparición niveles excesivos de PM_{10}, y se utilizan para construir una función de alerta ante peligro de contaminación atmosférica que proporcione a la Autoridad Portuaria una margen de maniobra de 24 horas. El comportamiento predictivo de la función de alerta ha sido validado mediante funciones de evaluación logarítmicas sobre episodios no utilizados en su construcción.

El quinto y último capítulo de esta memoria resume las conclusiones alcanzadas. Se detalla el sistema de alerta de forma que pueda ser profesionalmente programado de forma integrada con los sistemas de medida en tiempo real de que dispone la Autoridad Portuaria. Se sugieren posibles mejoras tanto en la integración de tales sistemas de medida como en la modificación de infraestructuras y en la introducción de buenas prácticas que disminuyan la probabilidad de que salten las alertas. Finalmente, de describen algunos problemas identificados cuya solución debería ser objeto de un estudio futuro.

La memoria concluye con dos apéndices. En el Apéndice 1 se proporciona la base matemática sobre la que descansa la metodología inferencial utilizada en este trabajo. En el Apéndice 2 se detalla la información contenida en el CD que

acompaña a este documento y que, entre otras cosas, contiene los programas (escritos en *Mathematica*), que implementan los algoritmos desarrollados, y una versión electrónica de esta memoria.

Capítulo 2.

Formulación del Problema

2.1. DEFINICIÓN DE ALTERNATIVAS

En condiciones meteorológicas adversas, la polución atmosférica sobre ciudades portuarias como consecuencia de partículas en suspensión levantadas desde parvas situadas al aire libre, o como consecuencia de operaciones de carga o descarga de graneles sólidos, constituye un problema medioambiental al que frecuentemente deben enfrentarse las autoridades portuarias.

En el caso particular del puerto de Tarragona, las parvas de carbón situadas sobre el muelle de Cataluña, así como las operaciones de carga y descarga de graneles pulverulentos en los muelles mas cercanos a la ciudad, pueden dar lugar a episodios de contaminación atmosférica sobre el barrio de la ciudad situado junto al puerto, el barrio de los pescadores (junto a la dársena interior, en la parte superior derecha de la Figura 1).

En su versión más sencilla, el problema planteado por la contaminación atmosférica que las operaciones previstas de carga y descarga de graneles sólidos pueden llegar a producir admite tres acciones alternativas, $\mathcal{A} = \{a_0, a_1, a_2\}$:

- a_0 Autorizar las operaciones portuarias en la forma habitual,
- a_1 Autorizar las operaciones en forma restringida (operando solamente con tolvas especiales, disminuyendo la altura de volcado,...), o
- a_2 Posponer algunas operaciones hasta que disminuya la probabilidad de alcanzar niveles excesivos contaminación.

Figura 1. *Puerto de Tarragona.*

En la Figura 2 se representa de el árbol de decisión correspondiente. Las consecuencias de cualquiera de estas acciones dependen de *forma probabilística* de las condiciones presentes: de las actividades portuarias activas y de la situación meteorológica. En una primera aproximación, tales condiciones pueden ser clasificadas en tres grandes grupos:

θ_0 Buenas condiciones,

θ_1 Condiciones desfavorables y

θ_2 Condiciones muy desfavorables.

Para cada combinación de acción posible a_i y condiciones presentes θ_j es necesario evaluar, en una escala adecuada, las pérdidas $l_{ij} = l(a_i, \theta_j)$, que se generarían, dando lugar a una tabla de la forma:

	θ_0	θ_1	θ_2
a_0	$l_{00} = l(a_0, \theta_0)$	$l_{01} = l(a_0, \theta_1)$	$l_{02} = l(a_0, \theta_2)$
a_1	$l_{10} = l(a_1, \theta_0)$	$l_{11} = l(a_1, \theta_1)$	$l_{12} = l(a_1, \theta_2)$
a_2	$l_{20} = l(a_2, \theta_0)$	$l_{21} = l(a_2, \theta_1)$	$l_{22} = l(a_2, \theta_2)$

Además, desde la perspectiva bayesiana que nos ocupa, la consecuencia básica de la acción a_i si se presentan las condiciones θ_j es una *variable aleatoria*, la cantidad ω_{ij} de partículas PM_{10} recogidas por m^3, cuya distribución, obviamente, depende de a_i

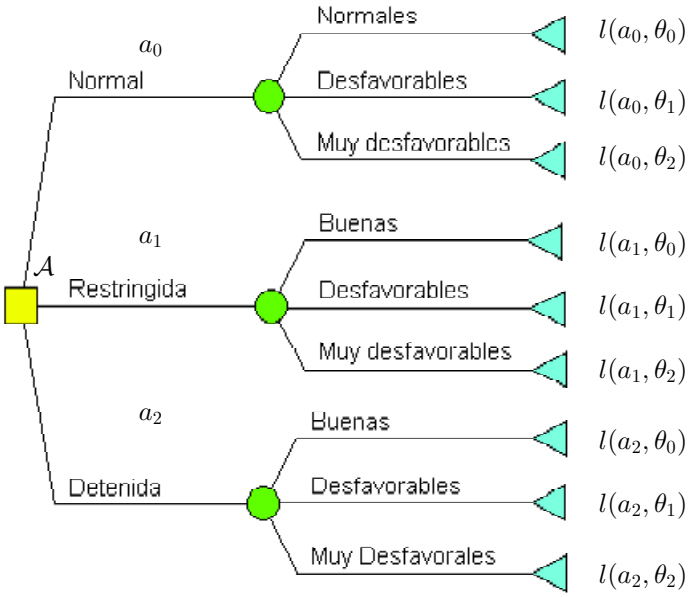


Figura 2. *Árbol de decisión básico en operaciones de carga y descarga*

y de θ_j. Formalmente,

$$l_{ij} = l(a_i, \theta_j) = \int_0^\infty l(c_i, \omega_{ij})\, p(\omega_{ij} \mid a_i, \theta_j, D)\, \mathrm{d}\omega_{ij},$$

de forma que la pérdida sufrida si se toma la acción a_i en las condiciones θ_j es una función del coste c_i de implementar a_i y de la distribución de probabilidad $p(\omega_{ij} \mid a_i, \theta_j, D)$ del nivel ω_{ij} de polución que puede esperarse en esas circunstancias en base a la información proporcionada por la base de datos histórica D. Obsérvese que si se detienen las operaciones, las actividades del puerto no pueden contribuir a la polución de la ciudad; consecuentemente, w_{20}, w_{21} y w_{22} describen, respectivamente, los niveles de polución del sector de la ciudad posiblemente afectado (en las distintas condiciones meteorológicas estudiadas) cuando *no* existen emisiones procedentes del puerto.

La función $l(c, \omega)$ debe describir en una escala conveniente (por ejemplo en una escala [0-1]) la pérdida asociada a una situación en la que con un gasto c se detecta una nivel de polución ω. Determinar esa función es una tarea difícil (pero inevitable) que debe ser el resultado de algún tipo de acuerdo entre todas las partes implicadas. Alternativamente, la función de pérdida puede ser expresada en

términos de la probabilidad de violar la legislación ambiental vigente. Este será el mecanismo adoptado en este trabajo.

Finalmente, una vez determinada la función de pérdida es necesario analizar la base de datos histórica D para calcular las probabilidades $\Pr(\theta_j \mid C, M, D)$ asociadas a cada uno de los tipos de condiciones consideradas, $\{\theta_0, \theta_1, \theta_2\}$, dadas las condiciones de actividad del puerto C y los datos meteorológicos M en el momento de tomar la decisión.

La pérdida esperada de cada una de las alternativas considerada será entonces

$$\overline{l}(a_i \mid C, M, D) = \sum_{j=0}^{2} l(a_i, \theta_j)\ \Pr(\theta_j \mid C, M, D))$$

donde los $l(a_i, \theta_j)$ son los valores esperados de las pérdidas $l(c_i, \omega_{ij})$ asociadas a los niveles de polución ω_{ij}. La decisión óptima a^* en las condiciones definidas por $\{C, M\}$ es entonces aquella que minimiza la pérdida esperada,

$$a^* = a^*(C, M, D) = \arg\min \overline{l}(a_i \mid C, M, D).$$

2.2. DETERMINACIÓN DE LOS SUCESOS RELEVANTES

2.2.1. *Normativa para partículas PM_{10}*

Los valores límites aceptables para la protección de la salud humana con respecto a la cantidad de partículas PM_{10} en condiciones ambientales aplicables a un puerto marítimo español vienen determinados por la Directiva 1999/30 de la Unión Europea y por el Real Decreto 1073/2002, que la introduce en la legislación española. Los valores límite se expresan en μg/m^3 medidos mediante métodos homologados, de tipo gravimétrico y se refieren a dos tipos de control:

(a) *Valor medio diario* (promediado de las 0 a las 24h de cada día)

(b) *Valor medio anual* (promediado en cada año civil)

La normativa consideró dos fases. La primera fase concluyó el 1 de Enero de 2005, mientras que la segunda fase se prolongará hasta el 1 de Enero de 2010. Los valores límites que deberían ser cumplidos a partir del término de la primera fase fase son:

Límites Fase 1 (En vigor desde 1 de Enero de 2005)

(a) Valor medio diario **50**μg/m^3 de PM_{10}, que no se superará más de **35** días al año.

(b) Valor medio anual **40**μg/m^3 de PM_{10}, que no deberá superarse nunca

Los valores límites para la segunda fase son indicativos y están actualmente en periodo de revisión a la luz de la información disponible sobre los efectos de las PM_{10} sobre la salud y el medio ambiente, la viabilidad técnica y la experiencia en la aplicación de los valores límites de la Fase 1 en los estados miembros de la Unión Europea. Los valores límites recomendados en la normativa publicada para la segunda fase son:

Límites Fase 2 (Del 1 de Enero de 2005 al 1 de Enero de 2010)

(a) Valor medio diario **50**μg/m^3 de PM_{10} que no debería superarse más de **7** días al año; se partirá del número máximo de 35 promedios diarios superiores a los **50**μg/m^3 de primera fase y se debería reducir su número progresivamente hasta alcanzar el límite propuesto de 7 promedios diarios mayores de 50μg/m^3 el 1 de Enero de 2010.

(b) Valor medio anual **20**μg/m^3 de PM_{10}, que no debería superarse nunca; se partirá del valor límite de 40μg/m^3 de la primera fase y se debería reducir 4μg/m^3 cada 12 meses hasta alcanzar el límite propuesto de 20μg/m^3 el 1 de Enero de 2010.

Con los datos D disponibles es necesario determinar las probabilidades de incumplimiento, bajo cada uno de los posibles escenarios considerados, de los límites exigidos, es decir de:

(i) La probabilidad condicional $\Pr[A \mid D]$ del suceso A dada la información contenida en el bando de datos D, donde el suceso A consiste en que las medias *diarias*, adecuadamente calibradas, superen los 50μg/m^3 en más de 35 ocasiones al año, y

(ii) La probabilidad condicional $\Pr[B \mid D]$ del suceso B dada la información contenida en D, donde el suceso B consiste en que el promedio de las medias *diarias*, adecuadamente calibradas, a lo largo de un año natural supere los 40μg/m^3.

Capítulo 3.

El Modelo Probabilístico

3.1. EL BANCO DE DATOS DISPONIBLE

La base experimental de este trabajo está constituida por una enorme matriz de datos D que recoge, hora a hora, las condiciones en las que se encontraba en Puerto de Tarragona durante 11 meses consecutivos, de Mayo de 2004 a Marzo de 2005, ambos inclusive. Lamentablemente, problemas laborales en el Puerto de Tarragona nos impidieron completar el año natural con la inclusión de Abril de 2005, como inicialmente estaba previsto. Esta matriz D tiene 8040 filas (24 horas por 335 días), una fila por cada una de las horas recogidas, y 176 columnas, una columna por cada una de las variables utilizadas. Los datos han sido suministrados por la Autoridad Portuaria, por las cabinas de control de la calidad del aire que la Generalitat catalana tiene instaladas en el puerto y en la ciudad de Tarragona, y por la cabina de control del programa HADA, especialmente orientada a la detección de partículas, y situada dentro de las instalaciones portuarias (al oeste del Muelle de Pescadores, ver Figura 3). Andrea Llorente coordinó durante varios meses, desde el propio puerto de Tarragona, el ingente trabajo de recogida, homogeneización y grabación del casi millón y medio (1,415,040) de datos resultantes.

Las primeras 4 columnas de la matriz de datos identifican el momento en el tiempo (día, mes, año y hora). Para un tratamiento cronológico adecuado de los datos, estas cuatro variables fueron reconvertidas en una única variable temporal t que precisa, para cada una de las 8040 horas analizadas, el número exacto de horas transcurridas desde las 0 horas del 1 de Enero de 2004. En nuestro caso, el rango de t se sitúa en el intervalo $[2905, 10944]$.

Figura 3. *Cabina HADA, en el Puerto de Tarragona.*

Las columnas 5 a 115 son variables binarias que describen la presencia o ausencia de distintas actividades portuarias susceptibles de contribuir a la emisión de partículas. Por ejemplo, la columna 57 describe, para cada una de las 8040 horas analizadas, la presencia (1) o ausencia (0) de la actividad "carga de fosfatos en camión". Las columnas 116 a 140 recogen, siempre hora a hora, las condiciones climatológicas del puerto (distintas funciones del viento, temperatura, humedad, presión atmosférica, radiación solar y lluvia acumulada), proporcionadas por las estaciones meteorológicas de Port Control y por la cabina de la Generalitat en el puerto. Las columnas 141 a 169 contienen los datos relativos a la calidad del aire (SO_2, NO,NO_2, NOX, CO y concentraciones de distintos tamaños de partículas), tanto las recogidas por la cabina HADA como las registradas por la cabina de la Generalitat en la ciudad. Finalmente, las columnas 170 a 176 contienen los valores de las variables climatológicas recogidos por la cabina HADA. Los valores no recogidos (por avería de los sistemas o por no disponer de la información relevante) aparecen codificados con un asterisco (*). La relación completa de las 176 variables grabadas, junto son su rango o con las unidades utilizadas para medirlas, está recogida en la Tabla 1 (dividida en cuatro paneles para una presentación adecuada). La matriz completa original, con el nombre <Datos.xls> es aportada en el CD-ROM que acompaña a esta Memoria.

Tabla 1.a *Relación de variables utilizadas (1)*

Variable	Descripción	Rango
01	Día	[1-31]
02	Mes	[1-12]
03	Año	[04-05]
04	Hora	[0-23]
05	Descarga Alfalfa a muelle	Binaria, $\{0, 1\}$
06	Carga Alfalfa a camión	Binaria, $\{0, 1\}$
07	Descarga Alfalfa pellets a muelle	Binaria, $\{0, 1\}$
08	Carga Alfalfa pellets a camión	Binaria, $\{0, 1\}$
09	Descarga Andalucita a muelle	Binaria, $\{0, 1\}$
10	Carga Andalucita a camión	Binaria, $\{0, 1\}$
11	Descarga de Antracita a muelle	Binaria, $\{0, 1\}$
12	Carga de Antracita a camión	Binaria, $\{0, 1\}$
13	Descarga de Avena a silo	Binaria, $\{0, 1\}$
14	Carga de Avena a camión	Binaria, $\{0, 1\}$
15	Descarga de Avena por tolva	Binaria, $\{0, 1\}$
16	Descarga de Bauxita a muelle	Binaria, $\{0, 1\}$
17	Carga de Bauxita a camión	Binaria, $\{0, 1\}$
18	Descarga de Caolín a muelle	Binaria, $\{0, 1\}$
19	Carga de Caolín a camión	Binaria, $\{0, 1\}$
20	Descarga de Carbón de coque a muelle	Binaria, $\{0, 1\}$
21	Carga de Carbón de coque a camión	Binaria, $\{0, 1\}$
22	Descarga de Carbón de hulla a muelle	Binaria, $\{0, 1\}$
23	Carga de Carbón de hulla a camión	Binaria, $\{0, 1\}$
24	Descarga de Bicarbonato por tolva	Binaria, $\{0, 1\}$
25	Carga de Bicarbonato a camión	Binaria, $\{0, 1\}$
26	Descarga de Carbonato sódico por tolva	Binaria, $\{0, 1\}$
27	Carga de Carbonato sódico a camión	Binaria, $\{0, 1\}$
28	Descarga de Cebada a silo	Binaria, $\{0, 1\}$
29	Carga de Cebada a camión	Binaria, $\{0, 1\}$
30	Descarga de Cebada por tolva	Binaria, $\{0, 1\}$
31	Descarga de Cebada a muelle	Binaria, $\{0, 1\}$
32	Descarga de Centeno por tolva	Binaria, $\{0, 1\}$
33	Carga de Centeno a camión	Binaria, $\{0, 1\}$
34	Descarga de Chamota a muelle	Binaria, $\{0, 1\}$
35	Carga de Chamota a camión	Binaria, $\{0, 1\}$
36	Descarga de Clinker a muelle	Binaria, $\{0, 1\}$
37	Carga de Clinker a camión	Binaria, $\{0, 1\}$
38	Descarga de Harina de colza por tolva	Binaria, $\{0, 1\}$
39	Carga de Harina de Colza a camión	Binaria, $\{0, 1\}$
40	Descarga de Pellets de Colza por tolva	Binaria, $\{0, 1\}$
41	Carga de Pellets de colza a camión	Binaria, $\{0, 1\}$
42	Descarga de Copra y palmaste por tolva	Binaria, $\{0, 1\}$
43	Carga de Copra y palmaste a camión	Binaria, $\{0, 1\}$
44	Descarga de Coque petróleo a muelle	Binaria, $\{0, 1\}$

Tabla 1.b *Relación de variables utilizadas (2)*

Variable	Descripción	Rango
45	Carga de Coque petróleo a camión	Binaria, $\{0, 1\}$
46	Descarga de DAP a muelle	Binaria, $\{0, 1\}$
47	Carga de DAP a camión	Binaria, $\{0, 1\}$
48	Descarga de Ferromanganeso a muelle	Binaria, $\{0, 1\}$
49	Carga de Ferromanganeso a camión	Binaria, $\{0, 1\}$
50	Descarga de Fosfato de cal a muelle	Binaria, $\{0, 1\}$
51	Carga de Fosfato de cal a camión	Binaria, $\{0, 1\}$
52	Descarga de Fosfato diamónico	Binaria, $\{0, 1\}$
53	Carga de Fosfato a camión	Binaria, $\{0, 1\}$
54	Descarga de Fosfato monocálcico por tolva	Binaria, $\{0, 1\}$
55	Carga de Fosfato monocálcico a camión	Binaria, $\{0, 1\}$
56	Descarga de Fosfatos por tolva	Binaria, $\{0, 1\}$
57	Carga de Fosfatos a camión	Binaria, $\{0, 1\}$
58	Descarga de Pellets de Girasol por tolva	Binaria, $\{0, 1\}$
59	Carga de Pellets de girasol a camión	Binaria, $\{0, 1\}$
60	Descarga de Torta de girasol por tolva	Binaria, $\{0, 1\}$
61	Carga de Torta de girasol a camión	Binaria, $\{0, 1\}$
62	Descarga de Gluten por tolva	Binaria, $\{0, 1\}$
63	Carga de Gluten a camión	Binaria, $\{0, 1\}$
64	Descarga de Cascarilla de Gluten por tolva	Binaria, $\{0, 1\}$
65	Carga de Cascarilla de Gluten a camión	Binaria, $\{0, 1\}$
66	Descarga de Gluten y destilados por tolva	Binaria, $\{0, 1\}$
67	Carga de Gluten y destilados a camión	Binaria, $\{0, 1\}$
68	Descarga de gravilla a muelle	Binaria, $\{0, 1\}$
69	Carga de Gravilla a camión	Binaria, $\{0, 1\}$
70	Descarga de Guisantes por tolva	Binaria, $\{0, 1\}$
71	Carga de Guisantes a camión	Binaria, $\{0, 1\}$
72	Descarga de Hulla a muelle	Binaria, $\{0, 1\}$
73	Carga de Hulla a camión	Binaria, $\{0, 1\}$
74	Descarga de Maíz a silo	Binaria, $\{0, 1\}$
75	Carga de Maíz a camión	Binaria, $\{0, 1\}$
76	Descarga de Maíz por tolva	Binaria, $\{0, 1\}$
77	Descarga de Mandioca/Tapioca por tolva	Binaria, $\{0, 1\}$
78	Carga de Mandioca/Tapioca a camión	Binaria, $\{0, 1\}$
79	Descarga de Nitrophoska por tolva	Binaria, $\{0, 1\}$
80	Carga de Nitrophoska a camión	Binaria, $\{0, 1\}$
81	Descarga de cenizas de Pirita a muelle	Binaria, $\{0, 1\}$
82	Carga de Cenizas de Pirita a camión	Binaria, $\{0, 1\}$
83	Descarga de Potasa a muelle	Binaria, $\{0, 1\}$
84	Carga de Potasa a camión	Binaria, $\{0, 1\}$
85	Descarga de Pulpa de remolacha a muelle	Binaria, $\{0, 1\}$
86	Carga de Pulpa de remolacha a camión	Binaria, $\{0, 1\}$
87	Descarga de sal común a muelle	Binaria, $\{0, 1\}$
88	Carga de Sal común a camión	Binaria, $\{0, 1\}$

Tabla 1.c *Relación de variables utilizadas (3)*

Variable	Descripción	Rango
89	Descarga de Semilla de algodón a muelle	Binaria, $\{0, 1\}$
90	Carga de Semilla de algodón a camión	Binaria, $\{0, 1\}$
91	Descarga de Silicomanganeso a muelle	Binaria, $\{0, 1\}$
92	Carga de Silicomanganeso a camión	Binaria, $\{0, 1\}$
93	Descarga de Cascarilla de Soja por tolva	Binaria, $\{0, 1\}$
94	Carga de Cascarilla de soja a camión	Binaria, $\{0, 1\}$
95	Descarga de Habas de Soja por tolva	Binaria, $\{0, 1\}$
96	Carga de Habas de Soja a camión	Binaria, $\{0, 1\}$
97	Descarga de Harina de Soja por tolva	Binaria, $\{0, 1\}$
98	Carga de Harina de soja a camión	Binaria, $\{0, 1\}$
99	Descarga de Torta de Soja por tolva	Binaria, $\{0, 1\}$
100	Carga de Harina de Soja a camión	Binaria, $\{0, 1\}$
101	Descarga de Sorgo a silo	Binaria, $\{0, 1\}$
102	Carga de Sorgo a camión	Binaria, $\{0, 1\}$
103	Descarga de Sorgo por tolva	Binaria, $\{0, 1\}$
104	Descarga de Sulfato Amónico por tolva	Binaria, $\{0, 1\}$
105	Carga de Sulfato Amónico a camión	Binaria, $\{0, 1\}$
106	Descarga de Trigo a silo	Binaria, $\{0, 1\}$
107	Carga de Trigo a camión	Binaria, $\{0, 1\}$
108	Descarga de Trigo por tolva	Binaria, $\{0, 1\}$
109	Descarga de Trigo panificable a silo	Binaria, $\{0, 1\}$
110	Carga de Trigo panificable a camión	Binaria, $\{0, 1\}$
111	Descarga de Trigo panificable por tolva	Binaria, $\{0, 1\}$
112	Descarga de Urea por tolva	Binaria, $\{0, 1\}$
113	Carga de Urea a camión	Binaria, $\{0, 1\}$
114	Descarga de Vidrio a muelle	Binaria, $\{0, 1\}$
115	Carga de Vidrio a camión	Binaria, $\{0, 1\}$
116	Velocidad del viento media (Port Control)	m/s
117	Velocidad del viento máxima (Port Control)	m/s
118	Velocidad del viento sig (Port Control)	m/s
119	Procedencia del viento media (Port Control)	[0-360[
120	Procedencia del viento máxima (Port Control)	[0-360[
121	Procedencia del viento media máxima (Port Control)	[0-360[
122	Procedencia del viento sig (Port Control)	[0-360[
123	Temperatura del aire media (Port Control)	^{o}C
124	Temperatura del aire media máxima (Port Control)	^{o}C
125	Temperatura del aire máxima (Port Control)	^{o}C
126	Temperatura del aire mínima (Port Control)	^{o}C
127	Humedad relativa (Port Control)	%
128	Humedad relativa máxima (Port Control)	%
129	Presión atmosférica media (Port Control)	mb
130	Radiación solar media (Port Control)	W/m^2
131	Radiación solar máxima (Port Control)	W/m^2
132	Lluvia acumulada (Port Control)	mm

Tabla 1.d *Relación de variables utilizadas (y 4)*

Variable	Descripción	Rango
133	Lluvia acumulada máxima (Port Control)	mm
134	Velocidad del viento media (Generalitat Puerto)	m/s
135	Procedencia del viento media (Generalitat Puerto)	[0,360[
136	Temperatura media del aire (Generalitat Puerto)	^{o}C
137	Humedad Relativa (Generalitat Puerto)	%
138	Presión atmosférica media (Generalitat Puerto)	mb
139	Radiación Solar media (Generalitat Puerto)	W/m^2
140	Lluvia acumulada (Generalitat Puerto)	mm
141	Partículas Sólidas Totales (Generalitat Puerto)	μg/m^3
142	Concentración media SO2 (cabina HADA)	μg/m^3
143	Concentración media NO (cabina HADA)	μg/m^3
144	Concentración media NO2 (cabina HADA)	μg/m^3
145	Concentración media NOX (cabina HADA)	μg/m^3
146	Concentración media CO (cabina HADA)	μg/m^3
147	Partículas de $PM_{2.5}$ (cabina HADA)	μg/m^3
148	Partículas de PM_{10} (cabina HADA)	μg/m^3
149	Temperatura media interior (cabina HADA)	^{o}C
150	Partículas sólidas totales (cabina HADA)	μg/m^3
151	Partículas PM > 0.3 (cabina HADA)	μg/m^3
152	Partículas PM > 0.4 (cabina HADA)	μg/m^3
153	Partículas PM > 0.5 (cabina HADA)	μg/m^3
154	Partículas PM > 0.65 (cabina HADA)	μg/m^3
155	Partículas PM > 0.8 (cabina HADA)	μg/m^3
156	Partículas PM > 1 (cabina HADA)	μg/m^3
157	Partículas PM > 1.6 (cabina HADA)	μg/m^3
158	Partículas PM > 2 (cabina HADA)	μg/m^3
159	Partículas PM > 3 (cabina HADA)	μg/m^3
160	Partículas PM > 4 (cabina HADA)	μg/m^3
161	Partículas PM > 5 (cabina HADA)	μg/m^3
162	Partículas PM > 7.5 (cabina HADA)	μg/m^3
163	Partículas PM > 10 (cabina HADA)	μg/m^3
164	Partículas PM > 15 (cabina HADA)	μg/m^3
165	Partículas PM > 20 (cabina HADA)	μg/m^3
166	Concentración media CO2 (Generalitat ciudad)	μg/m^3
167	Concentración media NO_2 (Generalitat ciudad)	μg/m^3
168	Concentración media H_2S (Generalitat ciudad)	μg/m^3
169	Concentración media de O_3 (Generalitat ciudad)	μg/m^3
170	Temperatura del aire media máxima (cabina HADA)	^{o}C
171	Temperatura del aire media (cabina HADA)	^{o}C
172	Humedad relativa máxima (cabina HADA)	%
173	Humedad relativa media (cabina HADA)	%
174	Radiación solar máxima (cabina HADA)	W/m^2
175	Radiación solar media (cabina HADA)	W/m^2
176	Lluvia acumulada (cabina HADA)	mm

3.2. ANÁLISIS PREDICTIVO INCONDICIONAL

Para una apreciación adecuada de las condiciones generales en las que se sitúa el problema estudiado, es conveniente describir la distribución predictiva incondicional y el comportamiento temporal de las variables que han resultado ser mas relevantes.

Puesto que tan sólo se trata de un primer análisis descriptivo, estas distribuciones predictivas han sido calculadas con métodos no-paramétricos convencionales de estimación de densidades, utilizando núcleos normales con ventana de referencia, cuyo cálculo es muy rápido. Específicamente, dada una muestra aleatoria de tamaño n, $\boldsymbol{z} = \{x_1, x_2, \ldots, x_n\}$, de la población objeto de estudio, la distribución predictiva correspondiente a una nueva observación x se estima mediante

$$p(x \mid \boldsymbol{z}) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{N}\big(x \mid x_{(j)},\ 3.5\, s_m\, m^{-1/3}\big) \tag{1}$$

donde $\boldsymbol{z}_m = \{x_{(1)}, x_{(2)}, \ldots, x_{(m)}\}$ es una submuestra aleatoria de $\boldsymbol{z}$ de tamaño m, y donde s_m es la desviación típica de $\boldsymbol{z}_m$. La elección del factor de submuestreo $\rho = m/n$ se hace en función del tamaño n original de la muestra, tanto menor cuanto mayor sea n. Para muestras pequeñas se toma $\rho = 1$ y $\boldsymbol{z}_m = \boldsymbol{z}$. Los resultados resultan estables para una amplia selección de valores de ρ. El proceso puede repetirse con k submuestras distintas $\{\boldsymbol{z}_{m_1}^{(1)}, \boldsymbol{z}_{m_2}^{(2)}, \ldots, \boldsymbol{z}_{m_k}^{(k)}\}$, del mismo o de distintos tamaños, y utilizar como resultado final la media correspondiente,

$$p(x \mid \boldsymbol{z}) = \frac{1}{k} \sum_{i=1}^{k} p(x \mid \boldsymbol{z}_{m_i}^{(i)}) \tag{2}$$

Puesto que se trata de densidades de probabilidad, el área debajo de la curva es la unidad, y la probabilidad incondicional de que la variable se sitúe en un determinado intervalo es el área situaba bajo la curva en ese intervalo.

Para cada una de las variables descritas a continuación, todas ellas variables continuas, se presenta una gráfica con dos paneles.

En el panel superior se representa la densidad de probabilidad de su distribución predictiva marginal, basada en todos los datos disponibles (8040 registros posibles, que en la práctica son algunos menos debido a los periodos en los que el sistema estuvo inoperativo).

En el panel inferior de describe la evolución temporal de la variable a lo largo de los 11 meses estudiados. Los puntos azules representan observaciones

individuales (medias horarias), mientras que los puntos negros representan medias diarias (medias móviles de 24 horas). El valor medio anual, correspondiente al centro de gravedad de la distribución representada en el panel superior, se representa mediante una recta roja.

3.2.1 *Vientos*

Tanto la estación meteorológica de Port Control como la cabina de la Generalitat en el puerto proporcionan el valor medio horario del valor absoluto de la velocidad del viento (en m/s). En la Figura 4 se analizan los resultados correspondientes a los datos de Port Control. En el panel superior se reproduce la distribución predictiva incondicional del valor absoluto de la velocidad del viento, basada en las 7416 observaciones realizadas en el periodo estudiado. Como puede observarse, se trata de una distribución claramente asimétrica, centrada alrededor de los 4 m/s y con valores típicamente menores que 15 m/s. La probabilidad incondicional de que la media horaria del valor absoluto de la velocidad del viento sea mayor de 10 m/s es 0.051 (área sombreada).

La evolución temporal de la variable estudiada muestra que las rachas de viento mas fuerte se situaron entre Noviembre y Febrero. A finales de Febrero se registraron medias horarias por encima de los 20 m/s (72 km/h) y una media diaria de 13 m/s (47 km/h).

Los datos procedentes de la cabina de la Generalitat en el puerto muestran una distribución muy diferente, con la moda en cero y una media de solo 2.9 m/s. La distinta ubicación de los medidores puede explicar parcialmente la diferencia, pero en este trabajo se ha optado por utilizar sistemáticamente los datos meteorológicos de Port Control, lo que al menos garantiza un tratamiento unificado del conjunto de datos meteorológicos. Consecuentemente, las ecuaciones predictivas que en esta Memoria involucren datos meteorológicos se entenderán ajustadas con los datos de Port Control.

La distribución predictiva de la dirección de procedencia del viento tiene dos componentes bien definidas (Figura 5).

Existe una primera componente, muy acusada, correspondiente a vientos procedentes del norte (mas precisamente del sector $[000, 030]$), con una probabilidad de 0.147 de que el viento proceda de ese sector (zona sombreada de la Figura 5.).

El resto de las observaciones resultan distribuidas de forma prácticamente uniforme (con la probabilidad complementaria, 0.853). Los datos procedentes de la cabina de la Generalitat en el puerto tienen una estructura parecida, por lo que no parece que la componente norte observada se deba a un sesgo del sistema de medida.

La enorme dispersión de las observaciones en la representación gráfica de la evolución temporal es consistente con la amplia componente uniforme. En su parte

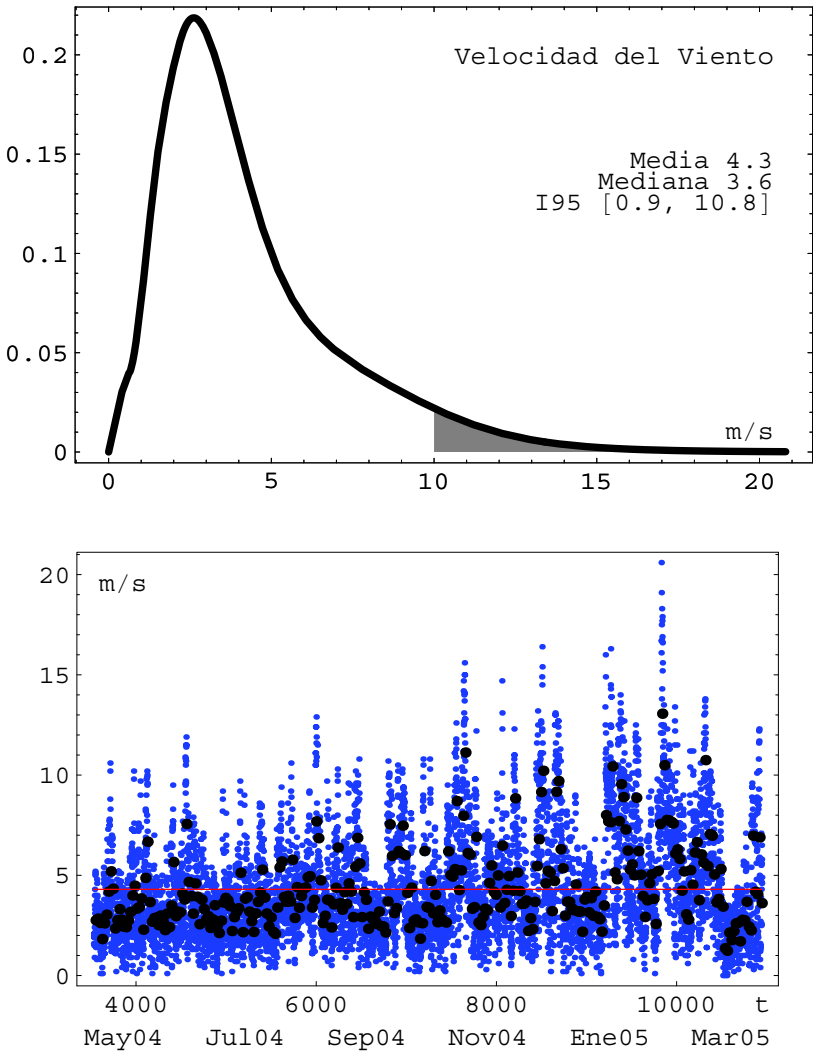


Figura 4. *Velocidad absoluta del viento*

inferior puede observarse la concentración de observaciones correspondientes a la componente norte.

La velocidad y la procedencia del viento pueden ser combinados para predecir vectores direccionales de la velocidad del viento, lo que frecuentemente proporciona

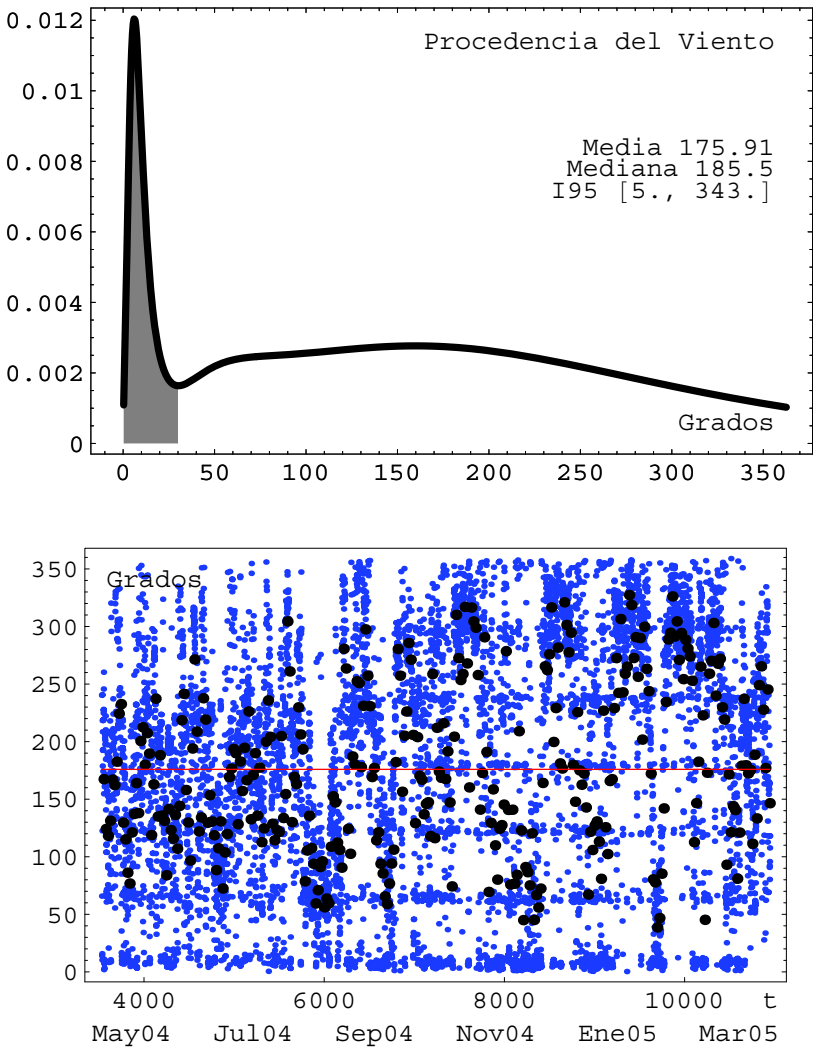


Figura 5. *Procedencia del viento*

variables más interesantes en problemas medioambientales. Por ejemplo,

$$v_{sur}(v_0, \alpha) = v_0 \cos(180 - \alpha)$$

describe la componente sur de la velocidad del viento, donde v_0 el el módulo de la velocidad del viento y α el radial (en grados) del que proviene. Si $\alpha = 180$,

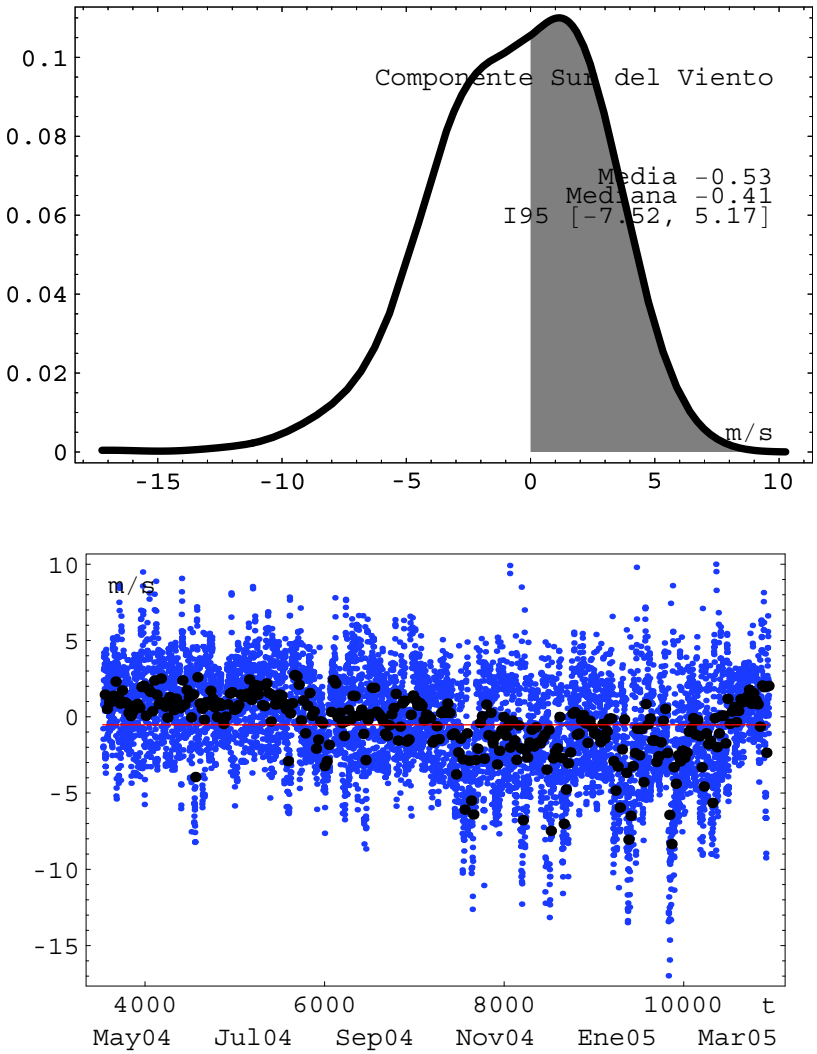


Figura 6. *Componente sur de la velocidad del viento*

el viento proviene del sur y $v_{sur} = v_0$; si $\alpha = 0$, el viento proviene del norte y $v_{sur} = -v_0$.

La Figura 6 describe describe las predicciones incondicionales de la componente sur de la velocidad del viento en el Puerto de Tarragona. Como puede

observarse, el viento proviene de la mitad sur del hoorizonte aproximadamente la mitad del tiempo (la probabilidad de que v_{sur} sea positiva (zona sombreada en la Figura 6) es 0.465, de forma que, a lo largo del año pueden esperarse alrededor de un 46% de medias horarias del viento con componente sur positiva.

El diagrama de la evolución cronológica muestra que, en verano los vientos proceden fundamentalmente del sur (valores medios diarios, en negro, positivos), mientras que en invierno la hacen desde el norte. la media anual (representada por la recta roja es -0.53 m/s, mostrando una ligera predominancia global de los vientos del norte. Confirmando las observaciones anteriores, se las rachas fuertes de viento norte (los puntos azules mas alejados de la línea roja, aparecen entre Noviembre y Febrero, con las rachas de mayor intensidad en Febrero.

3.2.2. *Temperatura del aire*

Todas las variables climatológicas están relacionadas entre si de forma muy compleja, lo que requiere un análisis multivariante no paramétrico. Por ejemplo, la Figura 7 muestra la dirección de procedencia del viento en función de la temperatura del aire. De su observación resulta aparente que, como podía esperarse, las temperaturas relativamente altas, entre 20 y 25^{o}Cse asocian a la componente uniforme de las direcciones del viento, mientras que las temperaturas por debajo de los 20^{o}C se asocian fundamentalmente con vientos del norte.

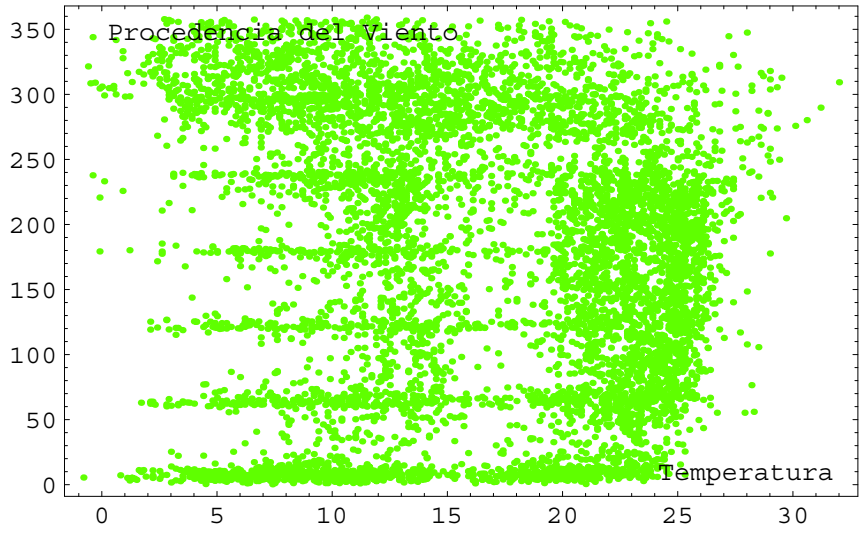


Figura 7. *Procedencia del viento en función de la temperatura del aire*

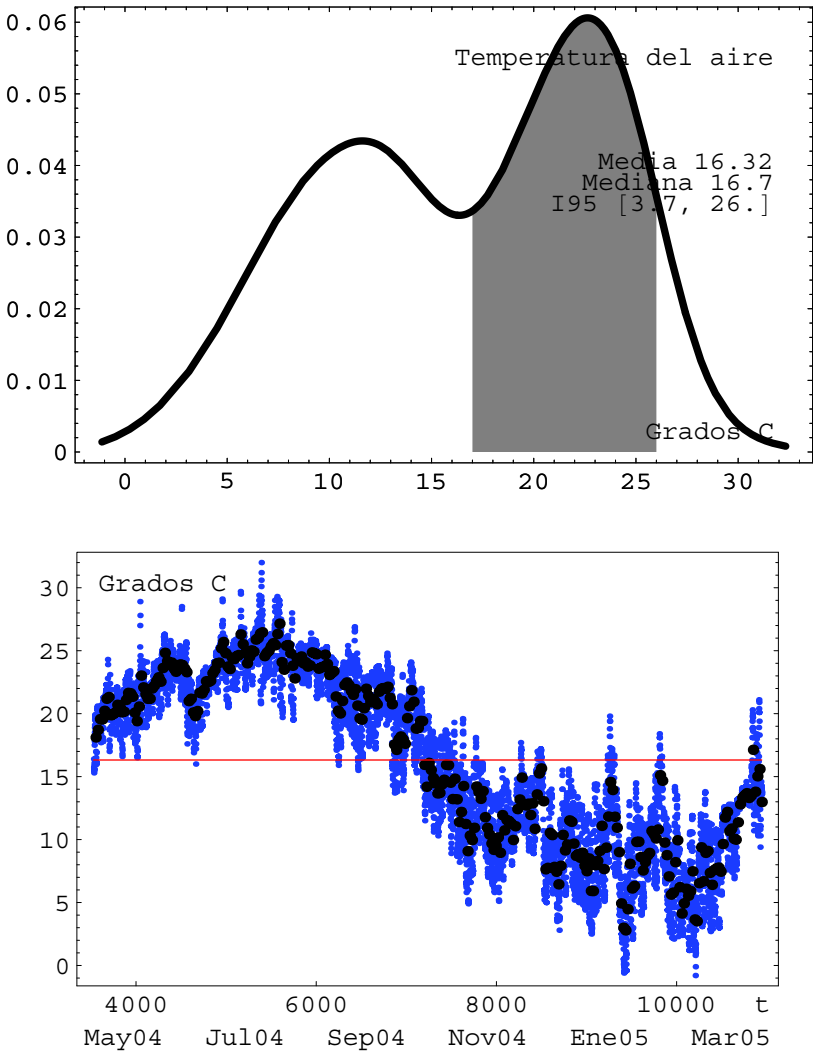


Figura 8. *Temperatura del aire*

La distribución de la temperatura del aire es claramente bimodal, una mixtura de dos componentes, una centrada en 12^{o}C y la otra en 23^{o}C. Esta situación da lugar a un interesante mínimo relativo de densidad de probabilidad alrededor de los 17^{o}C (Figura 8). La probabilidad asociada al intervalo de alta densidad relativa

$[17, 26]$ (área sombreada) es 0.444. Como podía esperarse, y como se claramente observa en el diagrama de evolución temporal, la componente centrada en los 23^{o}C corresponde al periodo [Mayo-Septiembre] y la componente centrada en los 11^{o}C al periodo [Octubre-Marzo]. El mes de Abril (del que lamentablemente no disponemos datos) se situará alrededor de los 17^{o}C. La existencia de dos componentes relativamente bien diferenciadas de la temperatura del aire permitirán utilizar una función de la temperatura (la distancia, con signo, a los 17^{o}C) como un indicador eficiente del periodo anual en la construcción de un índice climático compuesto que permita predicciones de cálculo sencillo mediante una función lineal (ver Capítulo 4)

3.2.3. *Presión atmosférica*

Estudiadas dos a dos, algunas de las variables climatológicas resultan básicamente independientes entre si. Este es el caso del par constituido por la presión atmosférica y la temperatura del aire, como la La Figura 9 pone inmediatamente de manifiesto.

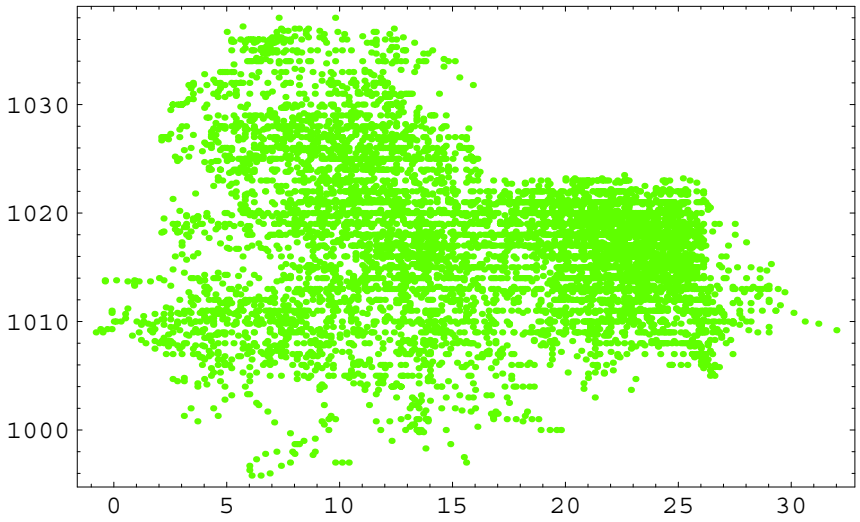


Figura 9. *Presión atmosférica en función de la temperatura del aire*

La distribución de la presión atmosférica en el Puerto de Tarragona a lo largo del año es claramente unimodal, centrada alrededor de una presión relativamente alta, de 1017 mb, con una desviación típica de unos 7 mb, y valores siempre situados entre los 996 y los 1040 mb. La probabilidad incondicional de que la presión se sitúe por encima de los 1013 mb que definen la presión estándar a nivel del mar es 0.75 (zona sombreada de la Figura 10).

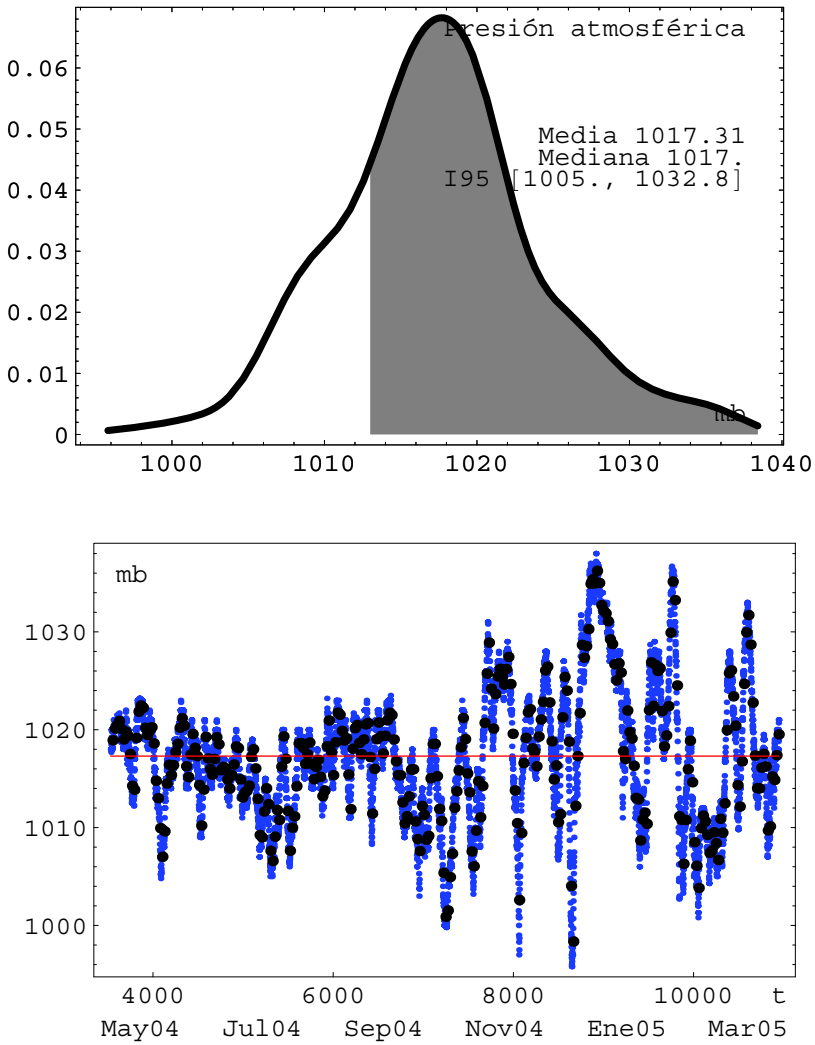


Figura 10. *Presión atmosférica*

Los datos agregados, que dan lugar a una distribución predictiva incondicional muy regular, esconden sin embargo una importante variabilidad local. En efecto, la evolución cronológica muestra una situación relativamente estable en primavera y verano, seguida de grandes oscilaciones en otoño y en invierno, cuando se alcanzan tanto las presiones más altas como las más bajas.

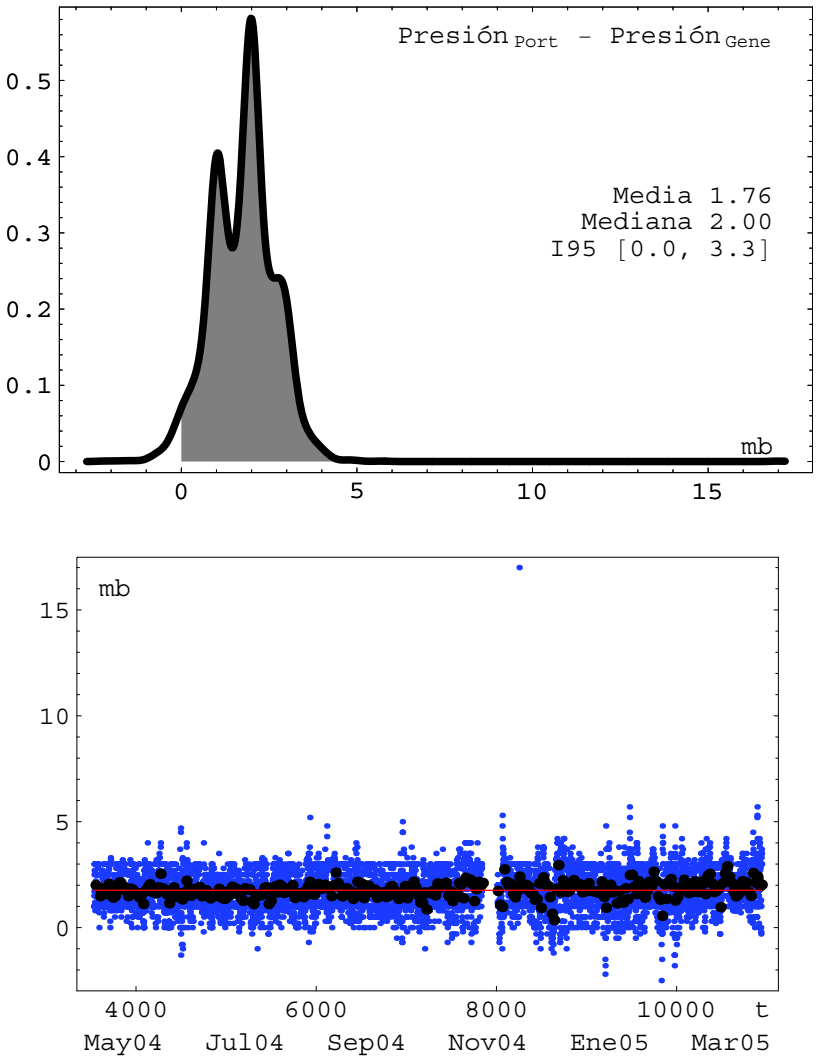


Figura 11. *Diferencias entre las presiones atmosféricas medidas por Port Control y las medidas por la cabina de la Generalitat*

Con objeto de comparar con algún detalle las dos estaciones climatológicas disponibles, se ha estudiado la serie cronológica de las diferencias de las medias horarias de las presiones atmosféricas medidas por ambos sistemas en los 7211

registros (de los 8040 posibles) en los que ambos sistemas estaban funcionando. El resultado (Figura 11) indica diferencias estadísticamente significativas, mostrando un sesgo de alrededor de 2 mb, la mediana de la distribución predictiva de la diferencia). La probabilidad de que la media horaria de las presiones atmosféricas en Port Control sea mayor que la de la cabina de la Generalitat es 0.974 (zona sombreada en la Figura 10), muy lejos del deseable 0.5.

Por otra parte, la distribución predictiva *no* resulta ser una distribución normal, lo que sugiere que se trata de un sesgo producido por el diseño de los aparatos, no el resultado de causas aleatorias independientes. El análisis de la serie cronológica correspondiente pone de manifiesto que se trata de un sesgo constante en el tiempo, muestra la existencia de un fallo de varios días en uno de los sistemas (en Noviembre de 2004) y señala la existencia de una discrepancia inaceptable, de 17 mb (probablemente debida a un fallo serio en uno de los dos sistemas), en Diciembre de 2004.

3.2.4. *Lluvia acumulada*

Aunque tiene aparentemente una estructura de variable continua, la cantidad de lluvia acumulada es, de hecho una variable mixta que, como se comprobará más adelante, es importante tratar como tal. El realidad, se trata de dos variables superpuestas, una variable binaria que indica la presencia o ausencia de precipitación, y una variable continua que indica la cantidad de precipitación recogida, en caso de producirse. En el caso del Puerto de Tarragona, se han registrado $r = 7188$ horas sin ninguna precipitación entre las $n = 7416$ registros válidos a lo largo del periodo estudiado, lo que produce una probabilidad predictiva incondicional de que se produzca algún tipo de precipitación de

$$\Pr[\text{lluvia} \mid D] = \frac{n - r + \frac{1}{2}}{n + 1} = 0.031.$$

La Figura 12 recoge el comportamiento incondicional del nivel de precipitación cuando esta se produce. La distribución predictiva es de tipo exponencial, decreciendo monótonamente desde cero. La probabilidad de que la lluvia acumulada en una hora se sitúe por debajo de los 2 mm (área sombreada) es 0.859.

La serie cronológica correspondiente pone de manifiesto que las precipitaciones, cuando se producen, tienden a agruparse en cortos periodos de unos días de duración. Durante en año estudiado los periodos de lluvia más importantes se situaron a finales de Agosto (cuando se produjeron, con mucho, las precipitaciones más abundantes), y a finales de Noviembre.

La relación entre la presión atmosférica y la lluvia acumulada no es monótona; la mayor parte de las precipitaciones se producen con presiones medias, situadas entre los 1010 y los 1025 mb. Sin embargo, la mayores precipitación registrada con el barómetro también en funcionamiento (de 10 mm) corresponde a una de

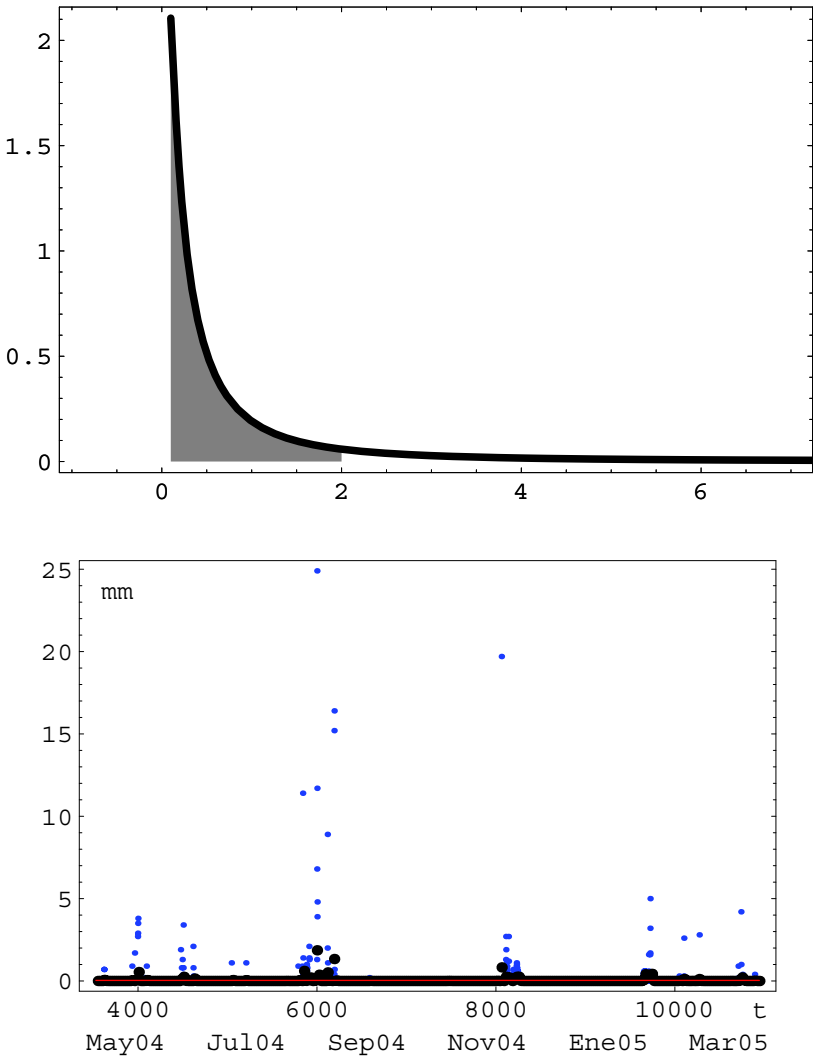


Figura 12. *Lluvia acumulada*

las presiones más bajas, mostrando un punto aparentemente singular, situado en la esquina superior izquierda de la Figura 13. Como en muchos otros casos de observaciones anormales, sería conveniente que la Autoridad Portuaria explorares si estos puntos estadísticamente singulares se deben a circunstancias especiales, o si se trata simplemente de errrores de algún tipo.

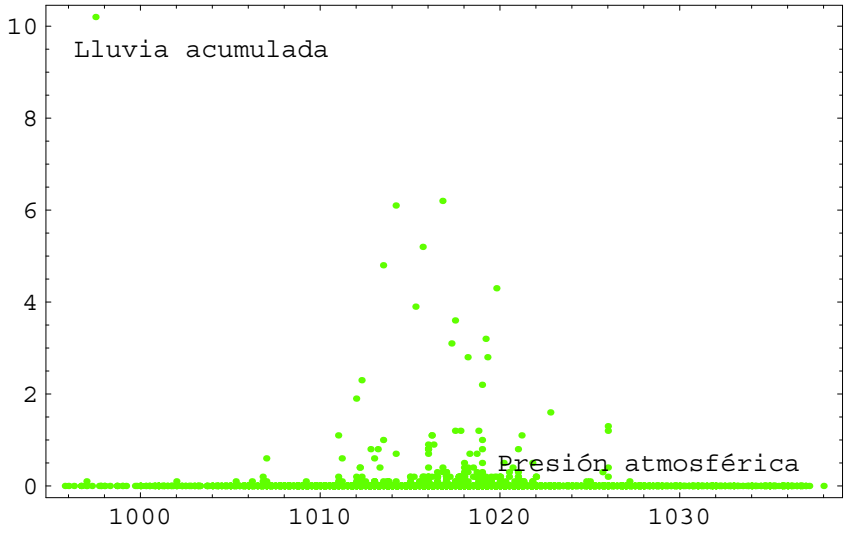


Figura 13. *Lluvia acumulada en función de la presión atmosférica*

3.2.5. *Humedad relativa*

Como podía esperarse, la humedad relativa está correlacionada con la lluvia acumulada. Los periodos lluviosos corresponden a una humedad relativa situada entre el 70% y el 95% (Figura 15).

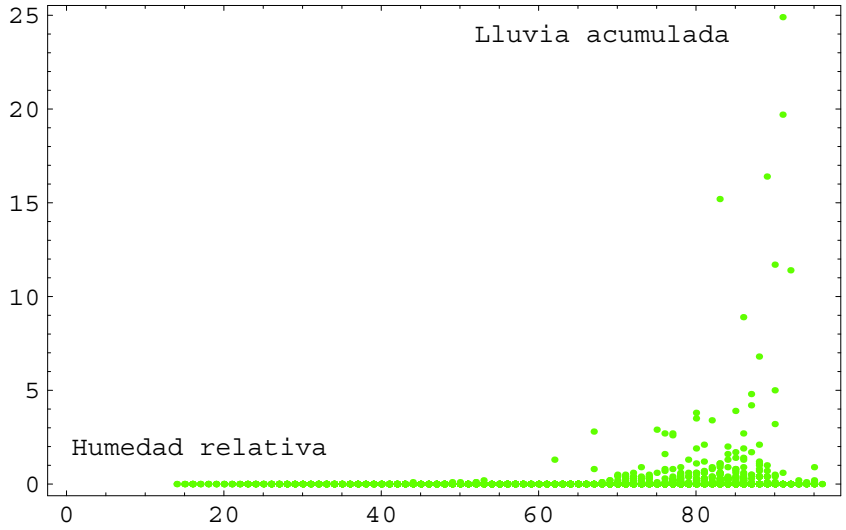


Figura 15. *Lluvia acumulada en función de la humedad relativa*

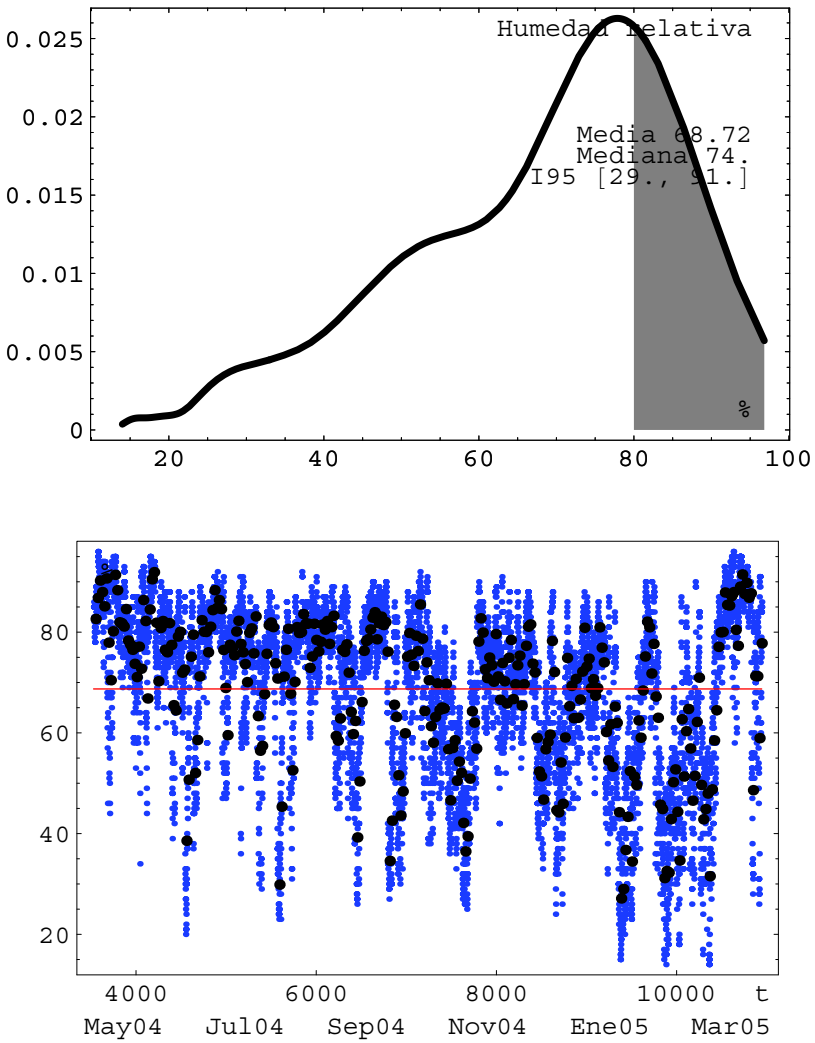


Figura 14. *Humedad relativa*

La distribución predictiva de la media horaria de humedad relativa en el puerto de Tarragona es muy asimétrica, con una moda única en el 78%, situada muy a la derecha de la media (69%). La mediana se sitúa en el 74%. La humedad relativa se sitúa entre el 29% y el 91% con probabilidad 0.95. La probabilidad de que supere el 80% (zona sombreaba el la Figura 14) es 0.287.

Una vez más, la forma unimodal de la distribución incondicional, en la que se agregan datos procedentes de todo el año, esconde importantes variaciones estacionales, que quedan explícitas al analizar su serie cronológica. En efecto, el comportamiento de la humedad relativa en función del tiempo muestra numerosos episodios de baja humedad relativa. Dentro del periodo estudiado aparecen valores bajos a finales de Octubre y entre Diciembre y primeros de Marzo. En particular, Febrero de 2005 contuvo cortos periodos de humedad relativa muy baja, por debajo del 20% (panel inferior de la Figura 14).

Los máximos registrados de lluvia acumulada en las Figuras 13 y 15 no coinciden debido a que la máxima lluvia acumulada (con un registro de 20mm) coincidió (si no se trata de una observación errónea) con un fallo en el sistema de medida de la presión atmosférica.

3.2.6. *Radiación solar*

Como podía esperarse, la radiación solar y la temperatura del aire están relacionadas, pero la estructura de su distribución conjunta (Figura 15) es muy compleja, con un soporte de tipo triangular limitado por una relación aproximadamente lineal ente la temperatura y el *máximo* de la radiación solar registrada.

La media horaria de la radiación solar es otra variable mixta, que vale cero durante las horas nocturnas y en con cielos totalmente cubiertos, y toma valores positivos en otro caso.

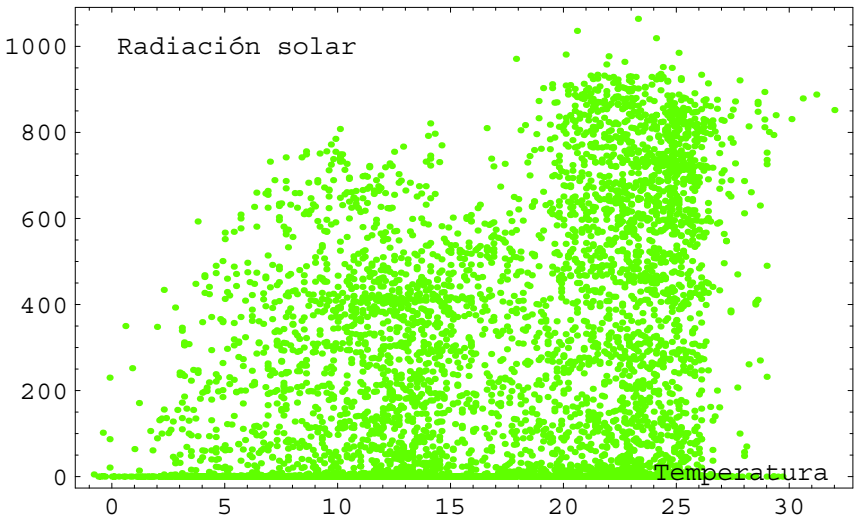


Figura 15. *Radiación solar positiva*

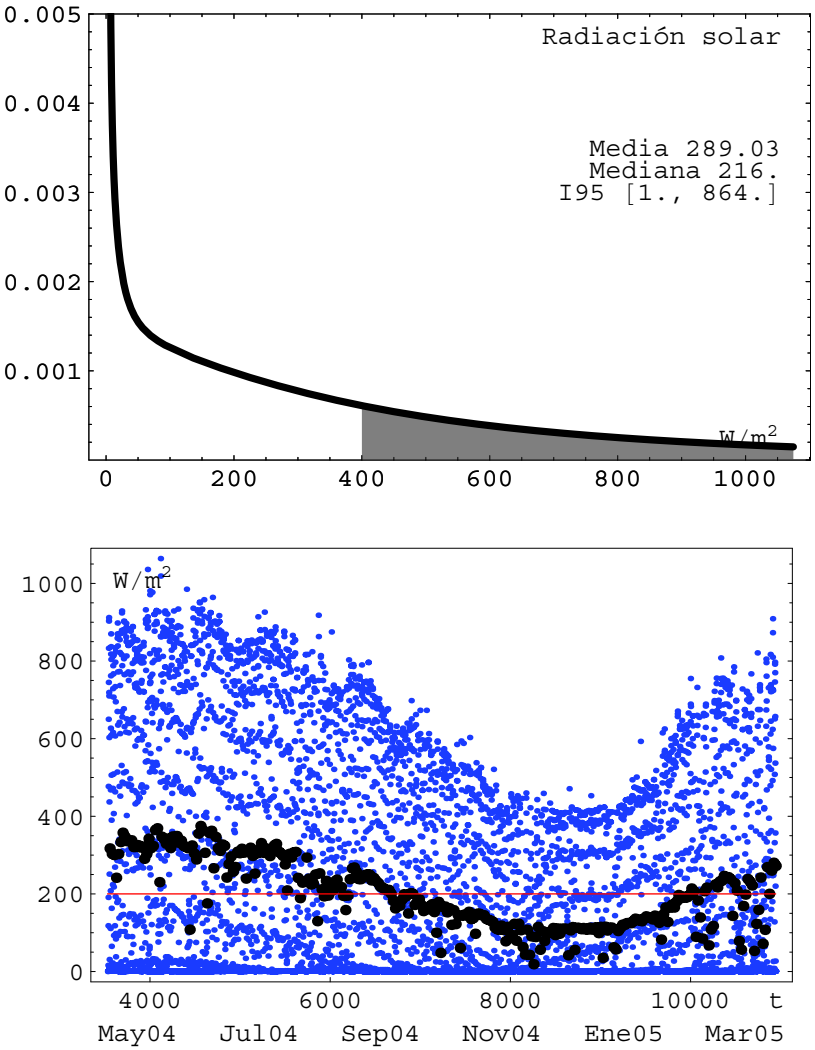


Figura 15. *Radiación solar positiva*

Durante el periodo estudiado, se registraron 5133 horas con radiación media no nula entre las 7416 en las que los sensores funcionaron. La probabilidad predictiva correspondiente, que constituye un estimador de la proporción de horas al año con radiación solar no nula, es $\Pr[\text{Radiación} > 0 \mid D] = 0.692$.

La densidad de probabilidad de la distribución predictiva de la media horaria de radiación solar positiva el puerto de Tarragona decrece monótonamente desde cero, y tiene una mediana de 216 W/m^2 situada muy a la izquierda de la media ($289W/m^2$). La probabilidad de que la radiación horaria media positiva supere los 400 W/m^2 (zona sombreaba el la Figura 16) es 0.313.

Como podía esperarse, la serie cronológica correspondiente es máxima entre Mayo y Julio (con valores medios horarios por encima de los $1000W/m^2$y medias de las 24 horas diarias cercanas a los $400W/m^2$, y mínimos situados en Diciembre, con máximos horarios por debajo de los $500W/m^2$ y medias diarias alrededor de los $100W/m^2$.

3.2.7. *Actividades portuarias*

Las 111 variables binarias que describen las actividades portuarias en el periodo estudiado constituyen, junto a las variables climatológicas, la base predictiva para modelizar el comportamiento de la función objetivo, la concentración en el aire de partículas PM_{10}. Sólamente 28 de estas 111 actividades tuvieron lugar en al menos el 5% de las 8040 horas analizadas.

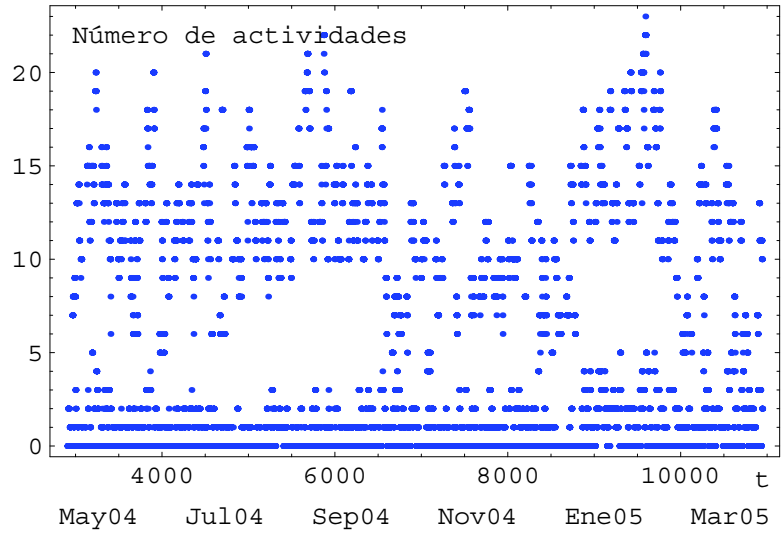


Figura 17. *Número de Actividades por hora*

La Figura 17 representa en número total de las 111 actividades portuarias consideradas que tenían lugar en cada una de las 8040 horas analizadas. Como puede observarse, el número de actividades realizadas simultáneamente oscila entre 0 y 25, con una pauta relativamente uniforme a lo largo del año.

Tabla 2. *Porcentajes de actividad horaria para las 28 actividades mas frecuentes. El código que las identifica es el descrito la Tabla 1.*

Código	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic	Ene	Feb	Mar
22	0	36	26	31	42	37	21	12	52	51	55
44	17	31	36	33	36	35	6	19	52	49	48
37	7	30	27	26	33	32	25	8	24	32	34
98	29	28	19	23	32	8	22	0	18	11	18
86	8	17	24	24	30	13	27	24	23	7	8
72	33	26	27	16	23	14	6	12	13	16	14
78	26	28	27	26	18	11	8	0	10	16	8
61	21	12	29	24	32	0	2	0	23	16	18
97	20	20	24	2	18	15	14	8	14	23	16
113	0	0	5	11	2	16	25	24	32	24	18
90	10	13	24	21	32	0	8	0	8	7	10
43	0	0	0	2	0	18	8	26	32	23	19
106	11	6	3	16	16	13	5	8	14	12	18
63	19	8	11	11	30	0	8	0	7	9	8
51	16	2	2	2	0	5	2	10	21	20	20
94	24	7	11	11	20	3	8	0	5	2	5
62	4	16	4	17	18	3	3	4	10	4	7
39	0	0	21	11	0	0	0	5	19	14	15
77	24	26	0	0	0	8	0	0	13	6	4
36	5	3	3	14	12	7	8	5	5	12	8
108	6	1	0	10	15	14	17	2	4	1	8
45	0	0	2	5	0	0	35	34	0	0	0
23	0	0	0	2	0	0	32	35	2	0	0
59	0	0	2	0	2	0	0	0	29	21	18
70	18	12	0	2	4	3	10	8	6	1	0
67	0	17	3	18	7	2	2	0	3	7	5
107	0	0	0	0	0	7	17	26	11	0	0
26	6	4	5	9	7	4	9	2	2	13	1
32	0	0	0	3	13	10	12	4	2	4	6

La actividades más frecuentes en el periodo estudiado fueron, por este orden, la descarga de carbón de hulla al muelle (32.9% de horas activas), la descarga de coque petróleo al muelle (32.7%) y la carga de clinker a camión (25.2%). Sin embargo, como veremos en la Sección 3.4, estas no son las actividades más problemáticas desde el punto de vista de emisión de partículas PM_{10}.

La Tabla 2 proporciona, para cada una de las 28 actividades más frecuentes, ordenadas de mayor a menor frecuencia, y para cada uno de los 11 meses analizados, el porcentaje de horas de cada mes en el que cada actividad tuvo lugar. Por ejemplo, con base a los datos observados, la variable 44 (“descarga de coque petróleo a muelle” según se detalla en la Tabla 1), fué globalmente la segunda actividad más

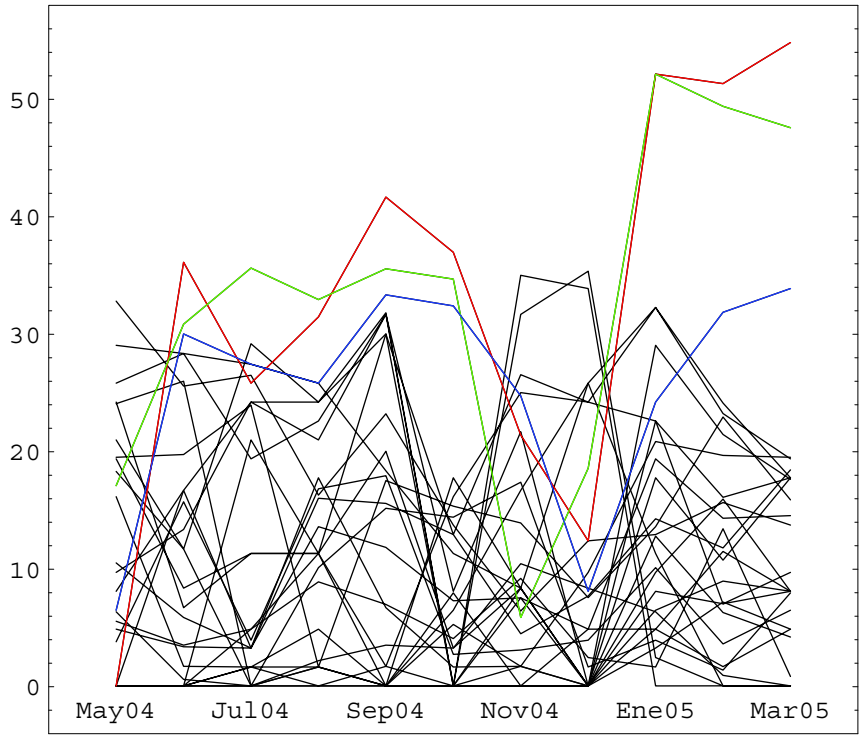


Figura 18. *Porcentaje de horas utilizadas en función de tiempo para cada una de las 28 actividades portuarias más frecuentes*

frecuente en el periodo estudiado, y permaneció activa el 52% de las horas de Enero, pero solo el 6% de las de Noviembre.

La Figura 18 proporciona una representación gráfica de los valores de la Tabla 2. Las líneas roja, verde y azul corresponden, en ese orden, a las tres actividades más frecuentes, ya mencionadas.

Puede observarse que la frecuencia de las distintas actividades varía fuertemente con el periodo del año analizado. Puesto que sólo disponemos de datos de un año, no podemos precisar si se trata o no de cadencias estables.

3.3. LA CONCENTRACIÓN DE PARTÍCULAS PM_{10}

La función objetivo en este trabajo es el control de la concentración media horaria en el aire de partículas PM_{10}, variable que será denotada y_{pm10}. Para ello será necesario determinar las distribuciones predictivas condicionales $p(y_{pm10} \mid D, C)$ de los niveles de PM_{10}, dado el banco de datos disponible D, en función del conjunto de condiciones C que puedan resultar relevantes. En general, las condiciones C serán una función de la situación meteorológica y de las actividades portuarias.

3.3.1. *Calibrado*

Las medidas de la variable y_{pm10} contenidas en el banco de datos D fueron realizadas de forma automatizada mediante un método indirecto (Grimm) basado en las interacciones con un rayo láser de las partículas contenidas en el aire aspirado. En total, se ha dispuesto tan sólo de 6507 medias horarias de las 8040 previstas, debido a los fallos producidos en el sistema de medición automático, cuyo nivel de operatividad se situó alrededor del 80%.

Tabla 3. *Resultados del experimento de calibrado del CSIC*

Medida Automática (*Grimm*)	Medida Gravimétrica (*Andersen*)
48.2	41.2
41.4	39.4
44.1	38.1
50.2	33.4
71.2	48.3
49.0	35.2
14.3	17.5
66.2	44.9
30.0	24.3
36.8	27.8
58.1	50.7
83.1	68.0
110.3	77.3
82.8	61.1
31.8	34.9
15.9	24.0
27.6	29.7
71.7	60.2
26.6	23.5
35.2	27.9

Tratándose de medidas automáticas, no gravimétricas como requieren las normas, es necesario calibrarlas adecuadamente antes de proceder a su análisis.

Para ello hemos dispuesto de los resultados de un experimento de calibración realizado entre Octubre y Diciembre de 2004, a instancias de Puertos del Estado, por el Institut de Ciències de la Terra de Barcelona, dependiente del Consejo Superior de Investigaciones Científicas. Los 20 pares valores que se obtuvieron, $\boldsymbol{z} = \{(x_1, y_1), \ldots, (x_{20}, y_{20})\}$ están reproducidos en la Tabla 3 y representados gráficamente (en rojo) en la Figura 19.

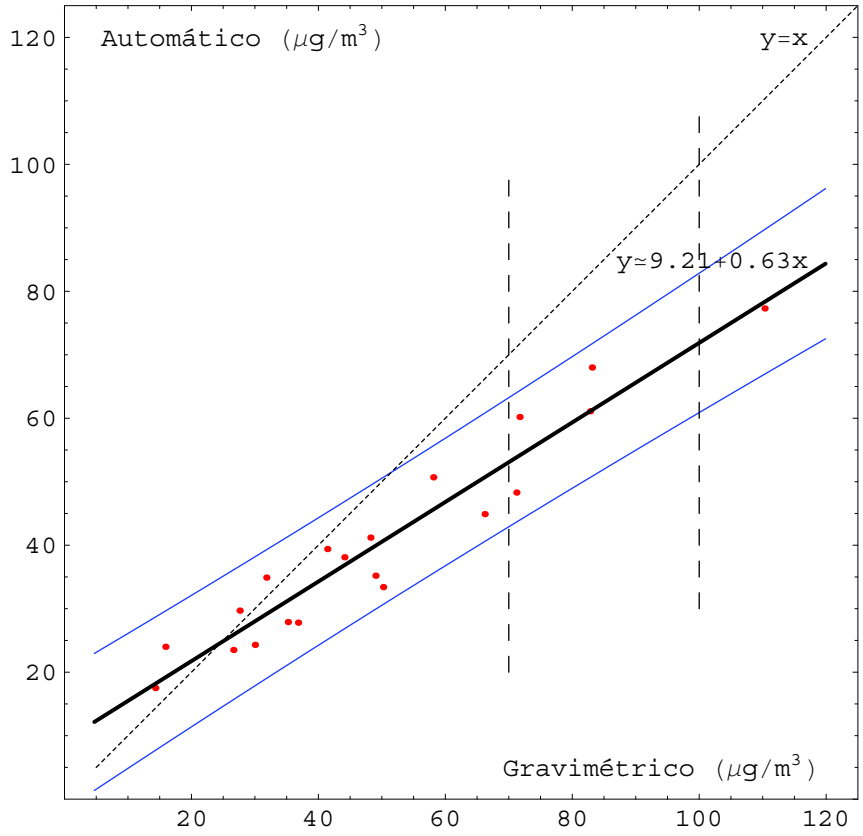


Figura 19. *Calibrado gravimétrico de medidas automáticas de PM_{10}*

Como puede observarse en esa figura, se trata de una dependencia esencialmente lineal. Para valores pequeños (por debajo de 24.7μg/m^3), el método automático de medida (en abscisas en la Figura 19) subestima los valores gravimétricos de referencia (en ordenadas), pero para valores grandes, que son los que resultan más relevantes para cumplir la normativa, el método automático *sobreestima* los valores gravimétricos de referencia en un factor que tiende a $(1/0.627) = 1.595$ cuando crece el valor observado.

Los datos de que se dispone son compatibles con la hipótesis de un modelo lineal homocedástico, de la forma

$$y_{pm10} = \alpha + \beta x_{pm10} + \epsilon, \quad \text{con} \quad p(\epsilon) = \mathrm{N}(\epsilon \,|\, 0, \sigma)$$

En ausencia de información inicial sobre los valores de los parámetros $\{\alpha, \beta, \sigma\}$, procede utilizar una función inicial de referencia (esto es una distribución inicial "no-informativa") que en este caso resulta ser $\pi(\alpha, \beta, \sigma) = \sigma^{-1}$. Consecuentemente, la distribución final conjunta de los parámetros es de la forma

$$\pi(\alpha, \beta, \sigma \,|\, \boldsymbol{z}) \propto \prod_{j=1}^{n} \mathrm{N}(y_j \,|\, \alpha + \beta x_j, \sigma)\ \sigma^{-1},$$

y la distribución predictiva $p(y \,|\, x, \boldsymbol{z})$, que permite calibrar un valor automático x en función de los datos disponibles $\boldsymbol{z}$ es

$$\begin{aligned} p(y \,|\, x, \boldsymbol{z}) &= \int_{\Re^2 \times \Re^+} N(y \,|\, \alpha + \beta\ x,\ \sigma)\, \pi(\alpha, \beta, \sigma \,|\, \boldsymbol{z})\, \mathrm{d}\alpha\, \mathrm{d}\beta\, \mathrm{d}\sigma \\ &= \mathrm{St}(y \,|\, \hat{\alpha} + \hat{\beta}\ x,\ s\sqrt{\frac{n\, f(x)}{n-2}},\ n-2), \end{aligned}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\ \overline{x}, \quad \hat{\beta} = \frac{s_{xy}}{s_{xx}}, \quad f(x) = 1 + \frac{1}{n}\frac{(x - \overline{x})^2 + s_{xx}}{s_{xx}}\ ,$$

donde los estadísticos suficientes son $ns^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$, $n\,\overline{x} = \sum_i x_j$, $n\,\overline{y} = \sum_i y_i$, $n\,s_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$, $n\,s_{xx} = \sum_i (x_i - \overline{x})^2$. Los detalles fueron descritos en el informe metodológico que precede a esta memoria. En el caso que nos ocupa, los estadísticos suficientes resultan ser

$$\begin{aligned} \{n, \overline{x}, \overline{y}, s_{xx}, s_{xy}\} &= \{20, 49.73, 40.37, 588.93, 369.04\} \\ \{\hat{\alpha}, \hat{\beta}, s\} &= \{9.211, 0.627, 4.408\} \end{aligned}$$

de forma que la distribución predictiva del valor gravimétrico y_{pm10} correspondiente a una medida automática x_{pm10} es la distribución Student

$$\begin{aligned} p(y_{pm10} \,|\, x_{pm10}, \boldsymbol{z}) &= \mathrm{St}(y_{pm10} \,|\, 9.211 + 0.627\, x_{pm10},\ s(x_{pm10}),\ 18) \\ s(x_{pm10}) &= 0.4675\sqrt{1 + \frac{1}{20}\frac{(x_{pm10}\ - 49.73)^2 + 588.93}{588.93}}\ . \end{aligned} \tag{3}$$

Obsérvese que el parámetro de escala $s(x_{pm10})$ de la distribución predictiva de y_{pm10} crece con la distancia $\,|\, x_{pm10} - \overline{x}\,|\,$; cuanto más alejado esté el valor automático observado x_{pm10} de la media $\overline{x}$ de los valores automáticos utilizados para calibrar, mayor incertidumbre existirá sobre el valor gravimétrico correspondiente y_{pm10}.

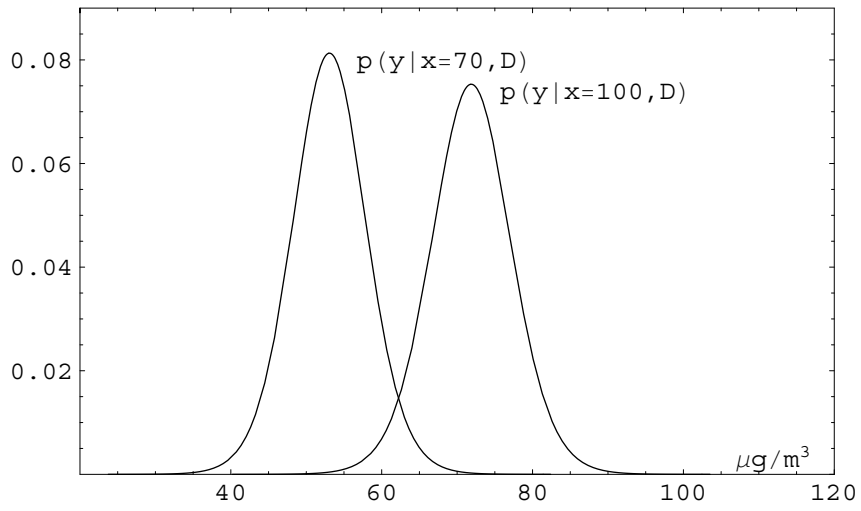
En la Figura 19, sobreimpuestos a los datos de calibrado, se han representado en función de x_{pm10}, tanto el valor medio de la distribución predictiva, esto es la recta de regresión

$$E[y_{pm10} \,|\, x_{pm10}, D] = 9.211 + 0.627\, x_{pm10}$$

(en trazo sólido), como las curvas que limitan los intervalos con contenido probabilístico 0.95 (en azul, ligeramente más próximas entre sí en el centro, cerca de $\overline{x} = 49.73$, que en los extremos).

En la Figura 20 se reproducen las distribuciones predictivas del valor gravimétrico y_{pm10} correspondientes a los valores automáticos $x_{pm10} = 70$ y $x_{pm10} = 100$, que son, respectivamente, $\mathrm{St}(y_{pm10} \,|\, 53.07, 4.84, 18)$ y $\mathrm{St}(y_{pm10} \,|\, 71.87, 5.22, 18)$. Sus regiones con contenido probabilístico 0.95 son, respectivamente, $[42.91, 63.24]$ y $[60.90, 82.85]$, que corresponden a la intersección de las líneas de puntos con las bandas azules en la Figura 19.

Figura 20. *Distribuciones predictivas del valor gravimétrico de PM_{10} correspondientes a valores automáticos de 70 y de 100$\mu g/m^3$.*



El correcto calibrado de las medias horarias de la concentración de PM_{10} tiene importantes implicaciones con respecto al cumplimiento de la normativa vigente. En efecto, de la inversión de la recta de regresión, es decir de la ecuación

$$x_{pm10} = (E[y_{pm10} \,|\, x_{pm10}, D] - 9.211)/0.627,$$

se deduce inmediatamente que los valores automáticos que corresponden en valor medio a los límites medios diarios citadas en la normativa en vigor (40μg/m^3 y 50μg/m^3 en valores gravimétricos de referencia) resultan ser, respectivamente, 49.1μg/m^3 y 65.1μg/m^3 en la escala automática proporcionada por la cabina HADA.

En el resto de este trabajo, toda referencia a valores de concentración de PM_{10} en el aire se entenderá referida a valores gravimétricos, ya calibrados. Todos los algoritmos empleados utilizan un módulo de calibración que transforma los valores automáticos contenidos en el banco de dados original en valores calibrados, teniendo en cuenta (cuando es necesario hacerlo) la incertidumbre adicional que este proceso conlleva, puesto que se substituye el valor constante observable x_{pm10}, por la distribución de probabilidad $p(y_{pm10} \mid x_{pm10}, \boldsymbol{z})$ descrita en la ecuación (1) (pág. 44).

3.3.2. *Distribución incondicional de la media diaria de PM_{10}*

Es esta sección se estudia el comportamiento global de las medias horarias de la concentración en el aire de partículas PM_{10}, recalibradas a escala gravimétrica.

La columna 148 de la matriz de datos D contiene la concentración horaria media de PM_{10} medida por la cabina HADA, que constituye el principal objeto de estudio en este trabajo.

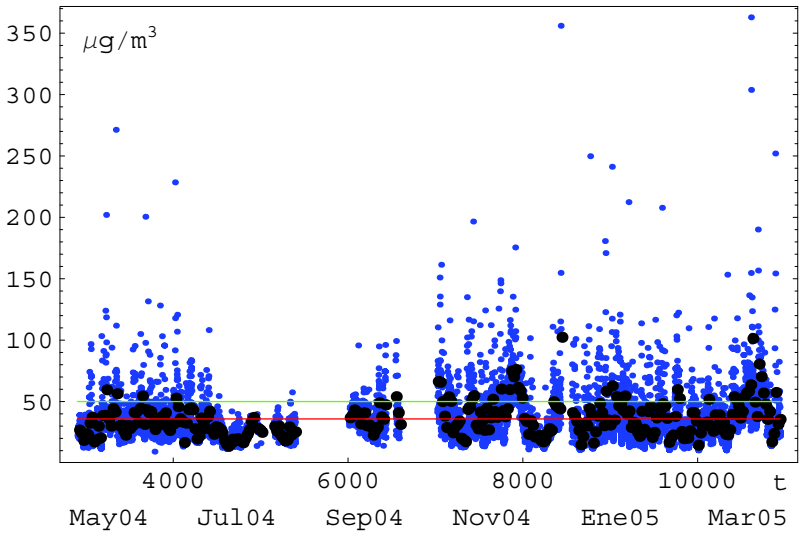


Figura 21. *Evolución temporal de la concentración de PM_{10} (medias horarias en azul, medias diarias en negro, calibradas a escala gravimétrica) de PM_{10}.*

En la Figura 21 se representa la evolución temporal de la medias horarias (en azul) y de las medias diarias (en negro) de las cantidades de PM_{10}, ya calibradas en escala gravimétrica, en función de las horas transcurridas desde el 1 de Enero de 2004. Se indica asimismo con una línea roja la media anual y con una línea verde el límite para las medias diarias ($50\mu g/m^3$) impuesto por la normativa. La falta de datos por fallos del sistema de recogida de datos en Agosto y Octubre del 2004 resulta aparente.

Las medias horarias se sitúan entre 9.2 y 362.9 $\mu g/m^3$, con una media de 35.6 y una desviación típica de 22.0; la mediana es 30.2, y el cuantil 0.99 es 117.17; un 16.75% de las medias horarias superan los $50\mu g/m^3$.

El objeto de atención en este caso no son las medias horarias, de las que disponemos de 6587 valores registrados, sino las medias diarias, de las que solamente disponemos de 279 registros, debido a que la normativa vigente sólo hace referencia a las medias diarias. Debe subrayarse que no se dispone de un modelo paramétrico que permita el calcula analítico de la distribución de las medias diarias a partir de las medias horarias. Tampoco sería razonable obtener la distribución por simulación de medias de 24 valores tomados al azar, porque este procedimiento ocultaría la estructura temporal de la serie. Finalmente, tampoco es aconsejable utilizar medias móviles para determinar la distribución incondicional de las medias diarias porque tales valores estarían fuertemente correlacionados.

La Figura 22 es un histograma normalizado de los 279 registros disponibles. Puede inmediatamente observase su estructura compleja, que al menos sugiere dos componentes principales muy superpuestas, entre los 10 y los $60\mu g/m^3$ y una tercera componente, con valores anormalmente altos alrededor de los $100\mu g/m^3$ cuya importancia en este problema puede ser capital.

La observación de la Figura 22 pone de manifiesto que las medias diarias de PM_{10} tienen una estructura demasiado compleja para que los métodos convencionales de estimación de densidades puedan captarla. En efecto, los métodos tradicionales proporcionan distribuciones artificialmente suavizadas, que ignoran precisamente las observaciones anormales cuya identificación puede ser crucial en un problema como el estudiado.

La Figura 23 muestra el resultado de aplicar la metodología convencional, mediante kernels normales de ventana estándar, construidos sobre los 279 datos calibrados disponibles, lo que da lugar la una distribución muy regular, en la que la probabilidad de superar la cota de $50\mu g/m^3$ (zona sombreada) es 0.174. Comparando este resultado con el histograma de los datos calibrados de la Figura 22 se observa inmediatamente que la regularidad sugerida por el análisis convencional es ilusoria.

Utilizando la metodología bayesiana objetiva no-paramétrica descrita con detalle en el Apéndice 1, se ha obtenido la distribución final predictiva $p(y \mid D)$ donde y es la variable de interés central, esto es la media diaria de la concentración

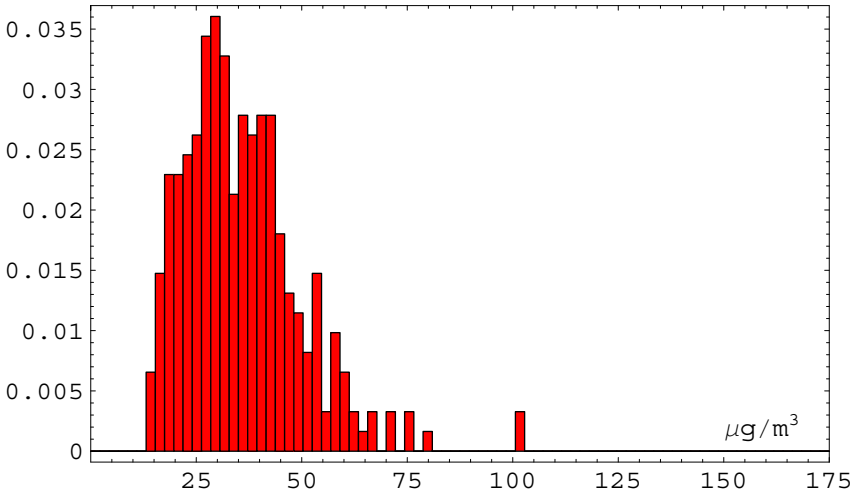


Figura 22. *Histograma de las concentraciones medias diarias de de PM_{10}, calibradas a escala gravimétrica.*

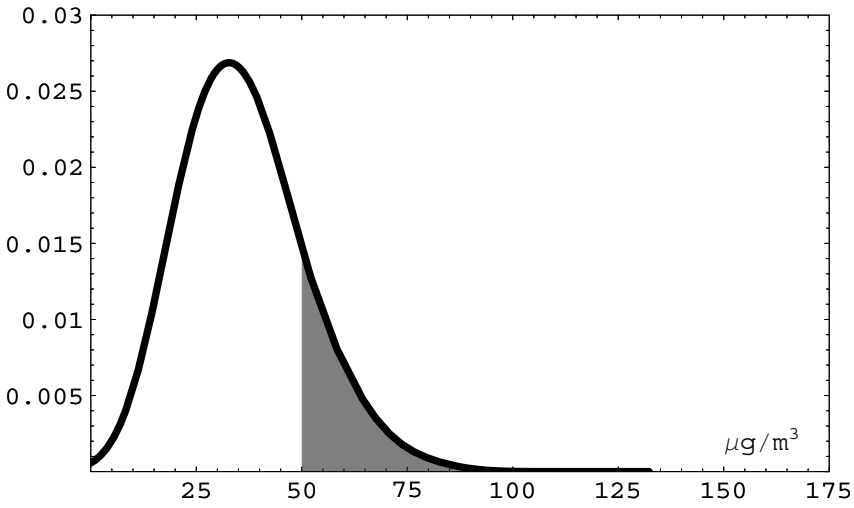


Figura 23. *Distribución predictiva incondicional de las medias diarias de de PM_{10} estimada mediante métodos convencionales.*

de PM_{10} (adecuadamente calibrada en términos gravimétricos en $\mu g/m^3$) y D es el conjunto de todos los datos disponibles.

Se trata de una distribución complicada, una poli-t con 160 núcleos, cuya densidad de probabilidad es la representada en la Figura 24.

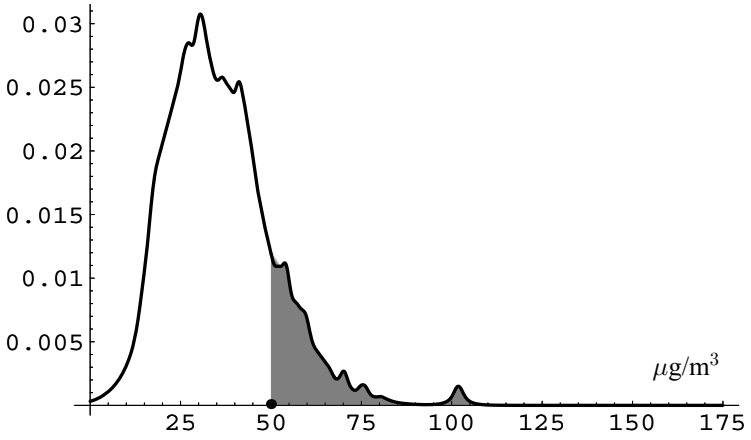


Figura 24. *Distribución predictiva incondicional de las medias diarias de de PM_{10} estimada mediante métodos bayesianos objetivos.*

La comparación visual con el histograma de la Figura 22 ya sugiere que la solución bayesiana si que es capaz de adaptarse a la compleja estructura de los datos. En cualquier caso, el correcto comportamiento predictivo de la distribución obtenida ha sido adecuadamente comprobado mediante técnicas de validación cruzada, utilizando una combinación lineal del logaritmo de la densidad de probabilidad asociada al verdadero valor como función de evaluación propia.

Como puede observarse se trata de una distribución de probabilidad esencialmente concentrada en el intervalo [0–85] $\mu g/m^3$, con una pequeña componente alrededor de $100\mu g/m^3$ que corresponde a los días con claros episodios de polución atmosférica por concentración excesiva de partículas PM_{10}.

3.3.3. *Consecuencias de la distribución incondicional de la media diaria de PM_{10}*

En vista de la normativa vigente, los niveles subyacentes de concentraciones medias de PM_{10} son preocupantes. Analizamos a continuación las consecuencias de la distribución predictiva encontrada en relación a las dos normales vigentes, descritas en la Sección 2.2.

Primera norma.
La primera norma europea exige no superar el límite de 50μg/m^3 más de 35 días en cada año natural.

De acuerdo con los resultados obtenidos, la probabilidad de exceder el límite de 50μg/m^3 fijado por la normativa de la Unión Europea en un día cualquiera (la zona sombreada en la figura), obtenida por integración numérica en la poli-t que describe su comportamiento probabilístico es

$$p_1 = \Pr[x > 50 \,|\, D] = \int_{50}^{\infty} p(x \,|\, D)\, dx = 0.159.$$

Análogamente, la probabilidad de supera los 40μg/m^3 un día cualquiera resulta ser

$$p_2 = \Pr[x > 40 \,|\, D] = \int_{40}^{\infty} p(x \,|\, D)\, dx = 0.342.$$

Bajo la hipótesis simplificadora de independencia entre días sucesivos, la probabilidad de que se supere una cota cualquiera en más 35 días al año, en función de la probabilidad p de que se supere un día cualquiera es

$$\Pr(A \,|\, D) = \Pr(A \,|\, p) = 1 - \sum_{k=0}^{35} \binom{365}{k} p^k (1-p)^{365-k}, \quad p = \Pr[x > 50 \,|\, D],$$

En la Figura 25 se representa $\Pr(A \,|\, p)$ en función de la probabilad p de uqe se supere la cota de 50 μg/m^{3}en un dia cualquiera escogido al azar.

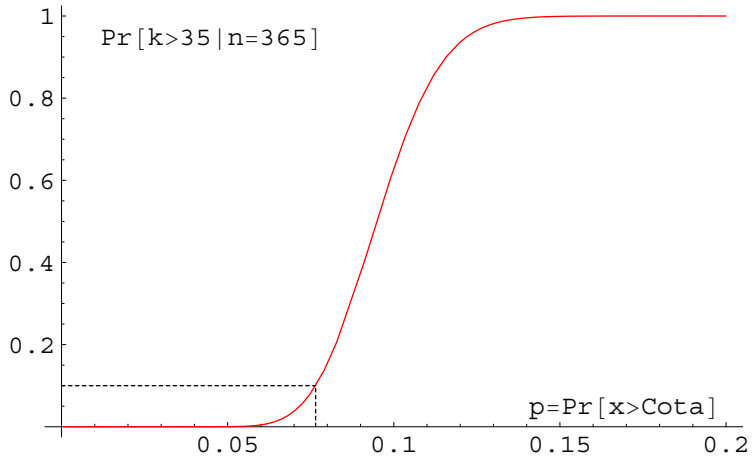


Figura 25. *Probabilidad de superar al menos 35 días al año una cota, en función de la probabilidad p de superarla un día cualquiera, bajo hipótesis de independencia condicional.*

Utilizando una aproximación normal a la función de distribución binomial implícita en la ecuación anterior, resulta

$$\Pr(A \mid D) \approx 1 - \Phi\left(\frac{35 - 365p}{\sqrt{365p(1-p)}}\right),$$

donde $\Phi(t)$ es la función de distribución de una normal tipificada. Puede comprobarse numéricamente que se trata de una aproximación notablemente precisa.

Para la probabilidad obtenida de superar la cota un día cualquiera, $p = 0.159$, la probabilidad de superar la cota más de 35 veces al año, $\Pr(A \mid D)$, es esencialmente la unidad; exactamente, $\Pr(A \mid D) = 0.99983$. Consecuentemente, si no se toman medidas correctoras, la norma comunitaria (a) descrita en la Sección 2.2 será violada con toda seguridad. Si las probabilidades de superar la cota de 50μg/m^3 en días sucesivos no fuesen independientes—como posiblemente sucede–, la probabilidad $\Pr(A \mid D)$ de violar la norma vigente sería todavía mayor.

La Figura 25 puede utilizarse para deducir los valores de p que inducirían una probabilidad $\Pr(A \mid D)$ cualquiera. Por ejemplo, para conseguir $\Pr(A \mid D) \leq 0.10$, es necesario que $p \leq 0.076$, y para que $\Pr(A \mid D) \leq 0.05$, es necesario que $p \leq 0.072$.

Consecuentemente, si se pretende que exista una probabilidad razonablemente razonablemente pequeña de violar la normativa vigente, es necesario reducir hasta valore próximos a 0.07 (desde el valor actual, 0.16) la probabilidad p de que ma media diaria de la concentración de PM_{10} supere los 50μg/m^3.

Durante los 11 meses observados, el valor límite de 50μg/m^3 de media diaria fué superado en 39 ocasiones (un 14% de los días), claramente por encima del $100(35/365) \approx 9.6\%$ exigido por la normativa.

Segunda norma.

La segunda norma europea exige que la media anual no supere los 40μg/m^3.

La probabilidad de que la media aritmética $\overline{x}$ de n observaciones independientes $\{x_1, \ldots, x_n\}$ supere una determinada cota es, salvo casos especiales, una expresión complicada que depende de la densidad de probabilidad $p(x)$ de la variable aleatoria x. En el caso que nos ocupa, el en que $p(x)$ no tiene una expresión analítica sencilla, es necesario recurrir a métodos de simulación.

A partir de la densidad de probabilidad de $p(x)$, una poli-t, es fácil definir su función de distribución (y los lo tanto, su función inversa, la función de cuantiles) mediante una integral numérica unidimensional. La función de cuantiles permite transformar una muestra aleatoria de observaciones uniforme en [0,1] en una muestra aleatoria de la distribución objeto de estudio. A partir de aquí es sencillo obtener por simulación (método de Monte Carlo) una estimación de la probabilidad de

que la media de de n observaciones supere una cota determinada, c. La precisión de la estimación puede ser indefinidamente mejorada aumentando el número de simulaciones.

Partiendo de la distribución predictiva de la media diaria representada en la Figura 23, se generaron 10,000 medias aleatorias de 365 valores. En la Figura 26 se representa el histograma de los $10,000$ valores obtenidos. Como podía esperarse del teorema central del límite, el resultado es una distribución aproximadamente normal. De hecho, la densidad normal con la media $m_{365} = 36.07\mu$g/m^3 y la desviación típica $s_{365} = 079\mu$g/m^3 de los datos obtenidos se ajusta al histograma extraordinariamente bien. La media de los valores diarios obtenidos en el año estudiado fué de 35.84μg/m^3, perfectamente compatible con el resultado obtenido.

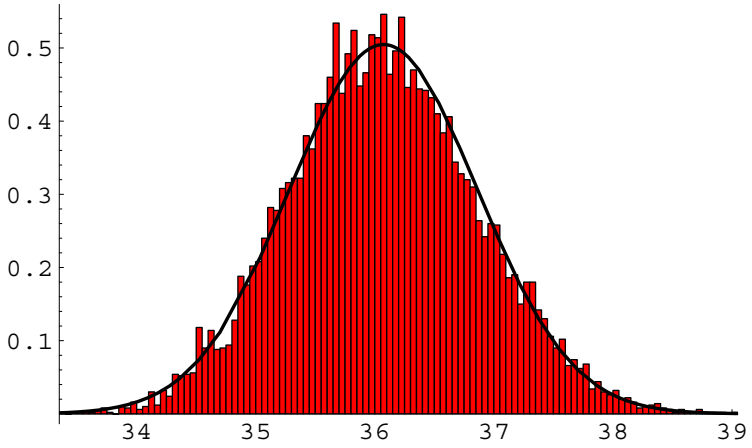


Figura 27. *Medias diaria de PM_{10} en función de las temperaturas máximas diarias*

De los resultados descritos se deduce inmediatamente que la probabilidad de violar la segunda norma europea, esto es la probabilidad de que la media anual supere los 40μg/m^3 viene dada por

$$\Pr[\overline{x}_{365} > 40 \,|\, D] \approx 1 - \Phi\left(\frac{40 - m_{365}}{s_{365}}\right) \approx 3.4 \times 10^{-7}$$

Consecuentemente, sin que sea necesaria acción correctora alguna, la segunda norma europea, en los términos actualmente vigentes, será respetada con toda seguridad.

Observese, sin embargo, que la comisión correspondiente propuso para la Fase 2 una reducción de 4 μg/m^3 cada 12 meses a partir de Enero de 2005. Si

esta norma (ahora bajo consulta de los estados miembros) llegase a entrar en vigor, significaría una media anual límite de 38μg/m^3 en el momento de escribir estas líneas (junio de 2005), y una media anual límite de 36μg/m^3 a finales de este año. Las probabilidades de cumplir con tales límites en las condiciones actuales son, respectivamente,

$$\Pr[\overline{x}_{365} > 38 \,|\, D] = 0.0074, \quad \text{y} \quad \Pr[\overline{x}_{365} > 36 \,|\, D] = 0.537,$$

de forma que, a muy corto plazo, la nueva norma no podría cumplirse sin cambios profundos en la gestión medioambiental del puerto.

En el resto de esta Memoria nos ceñiremos sin embargo a la normativa realmente vigente. Consecuentemente, daremos por supuesto que se está cumpliendo la segunda norma comunitaria y nos ocuparemos, exclusivamente, de determinar cómo sería posible llegar a cumplir la primera norma, esto es, como conseguir que no se supere la media diaria de 50μg/m^3 más de 35 días al año.

3.4. COMPORTAMIENTO CONDICIONAL DEL PM_{10}

En este apartado se estudia el comportamiento de la media horaria de la concentración de partículas PM_{10} en el aire, ya calibrada a escala gravimétrica, como función tanto de las distintas variables climáticas, como de las actividades portuarias registradas en la base de datos.

El criterio utilizado para determinar el grado de asociación entre las distíntas variables estudiadas ha sido la medida de *asociación intrínseca* α_{xy} entre dos variables aleatorias x e y, definida como la discrepancia intrínseca

$$\alpha_{xy} = \delta\{p(x,y),\ p(x)p(y)\}$$

entre su distribución conjunta $p(x,y)$ y el producto $p(x)p(y)$ de sus distribuciones marginales, donde

$$\begin{aligned}\delta\{p_1,p_2\} &= \min[\,\kappa\{p_2\,|\,p_1\},\ \kappa\{p_1\,|\,p_2\}\,]\\ \kappa\{p_1\,|\,p_2\} &= \int_{\mathcal{Z}} p_2(\boldsymbol{z}) \log \frac{p_1(\boldsymbol{z})}{p_2(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}.\end{aligned}$$

Los valores de α_{xy} son estimados a partir de una muestra $\{(x_1,y_1),\ldots,(x_1,y_1)\}$ calculando por Monte Carlo las necesarias integrales y utilizado para ello estimadores no paramétricos de las densidades $p(x,y)$,$p(x)$ y $p(y)$. Las propiedades de la medida de asociación intrínseca están recogidas en el Capítulo 2 informe metodológico.

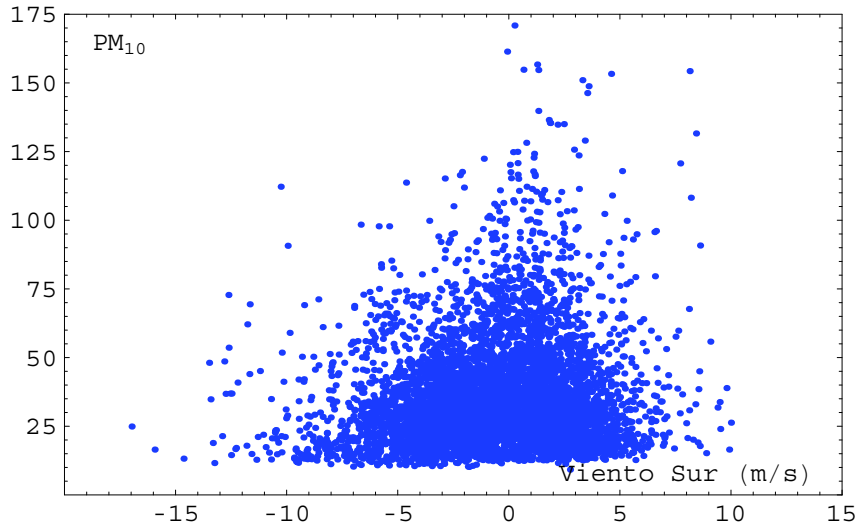


Figura 27. *Concentración de PM_{10} en función de la componente sur de la velocidad del viento*

3.4.1. *Relación con las variable climáticas*

Vientos. El análisis de distintas funciones de las variables registradas relativas al viento ha permitido identificar la componente sur de la velocidad del viento como la más informativa en relación a la concentración de PM_{10}.

La Figura 27 describe la concentración de PM_{10} en función de la componente sur de la velocidad del viento. Puede observarse que los niveles mas altos de PM_{10} se asocian mayoritariamente con los valores más altos del vector sur de la velocidad.

Temperatura. El estudio de la concentración de PM_{10} en función de la temperatura ambiental revela una cierta asociación estadística. Se ha estudiado la dependencia con respecto a las temperaturas diarias mínima, media y máxima. Los resultados son cualitativamente parecidos.

En la Figura 28 se representa la concentración de PM_{10} en función de de la temperatura media del aire. Puede observarse que las altas concentraciones de PM_{10} se agrupan en días relativamente fríos, con temperaturas medias entre los 4^oC y los 16^oC, y en los relativamente cálidos, con temperaturas medias entre los 18^oC y los 27^oC, con un mínimo relativo alrededor de los 17C, de forma que *no* se trata de una dependencia monótona: no se producen altas concentraciones de PM_{10} en los días muy fríos, ni tampoco en los cálidos.

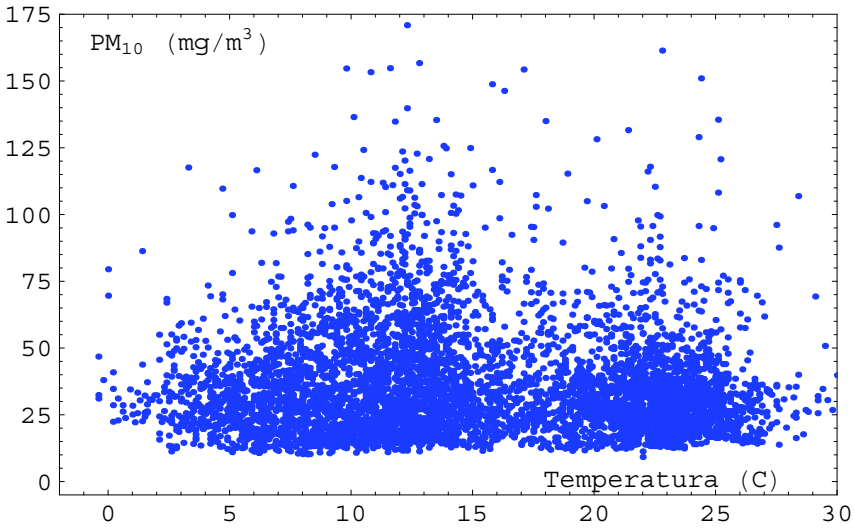


Figura 28. *Concentración de PM_{10} en función de la temperatura.*

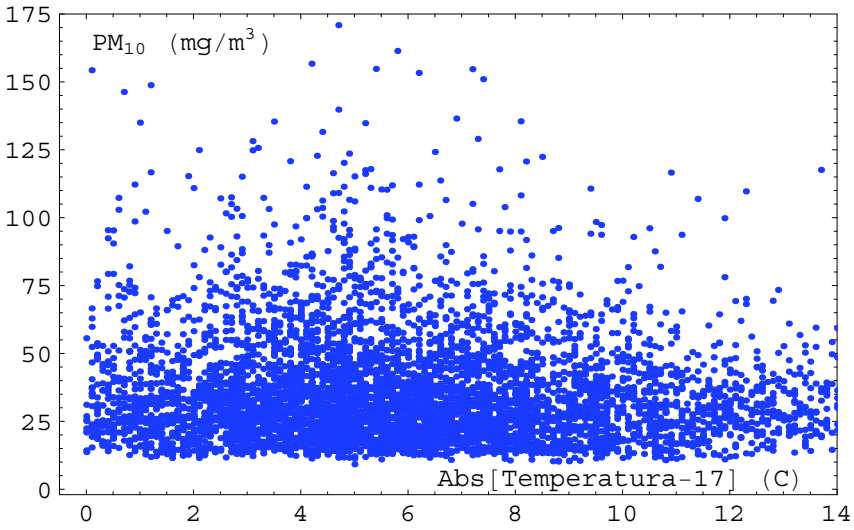


Figura 29. *Concentración de PM_{10} en función del valor absoluto de la diferencia con los 17C*

Como consecuencia, la distancia, en valor absoluto, a los 17 ^{o}C de temperatura, está apreciablemente mejor relacionada con la concentración de PM_{10} que la propia temperatura del aire, con un alto coeficiente de asociación intrínseca

($\alpha = 0.949$, que correspondería a un coeficiente de determinación ρ^2 de 0.85 si se tratase—que ciertamente no se trata—de una dependencia lineal). Matemáticamente, esta situación está relacionada con la bimodalidad de la distribución incondicional de la temperatura del aire, descrita en la Sección 3.1 (Figura 8).

En la Figura 29 se representa la concentración de PM_{10} en función de la variable $|T - 17|$, donde T es la temperatura del aire. Como cabía esperar de los resultados descritos en la Figura 28, puede apreciarse una importante asociación positiva, básicamente monótona, entre PM_{10} y $|T - 17|$.

Humedad relativa. En la Figura 30 se representa la concentración de PM_{10} en función de de la humedad del aire.

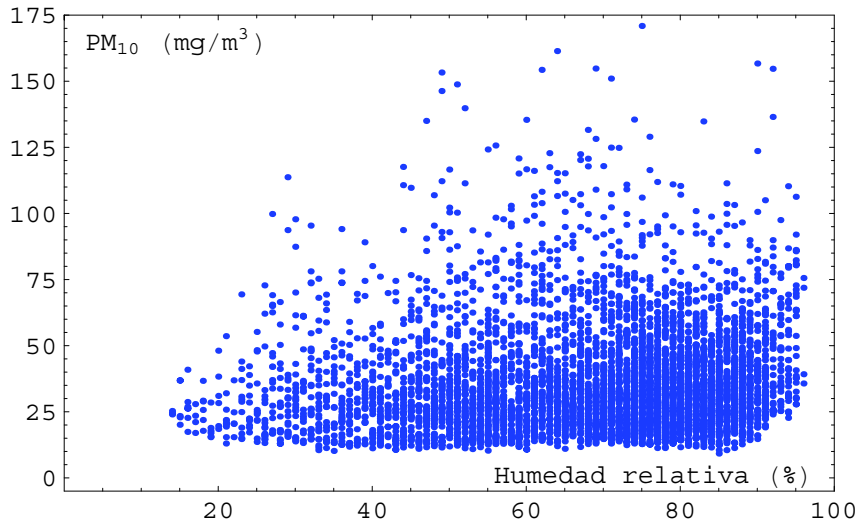


Figura 30. *Concentración de PM_{10} en función de la humedad relativa.*

Puede observarse que existe una asociación positiva, aunque no muy pronunciada: las concentraciones más altas de PM_{10} corresponden a valores altos de la humedad relativa.

Presión atmosférica. También la presión atmosférica está positivamente asociada con la concentración de PM_{10}.

En la Figura 31, que representa la concentración de PM_{10} en función de la presión atmosférica, puede apreciase que, en valor medio, las concentraciones de PM_{10} tienden a crecer con la presión atmosférica.

Lluvia acumulada. Como ya fué descrito en la Sección 2.2, la lluvia acumulada es una variable mixta con una componente discreta, presencia o ausencia de

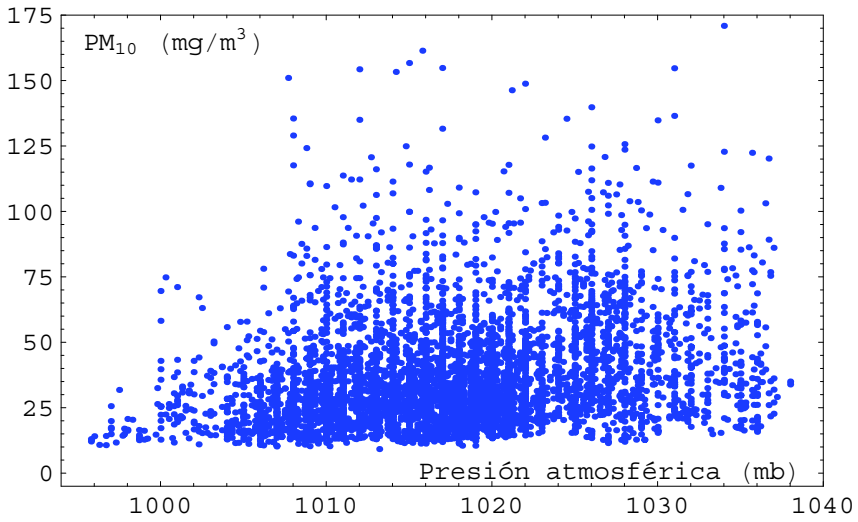


Figura 31. *Concentración de PM_{10} en función de la presión atmosférica*

precipitación y una componente continua y positiva, la cantidad de lluvia acumulada, si esta se produce. Como podía esperarse, la presencia de precipitaciones disminuye drásticamente las concentraciones de PM_{10} en el aire.

En la Figura 32 se presentan las distribuciones predictivas de la concentración y de PM_{10} condicionales a la presencia o ausencia de precipitaciones. Como puede observarse, la distribución predictiva de PM_{10} correspondiente a tiempo seco se sitúa básicamente entre los $10\mu g/m^3$ y los $85\mu g/m^3$, con la mayor parte de la masa probabilística situada ente $15\mu g/m^3$ y los $50\mu g/m^3$. La distribución predictiva de PM_{10} correspondiente a la presencia de lluvia se concentra en cambio en valores notablemente más pequeños, típicamente menores de $60\mu g/m^3$, y con una componente importante por debajo de los $25\mu g/m^3$. Consecuentemente, puede esperarse que la variable binaria presencia o ausencia de lluvia tenga alto poder predictivo en la determinación de la concentración de PM_{10}.

Radiación solar.

Nuestro estudio de las posibles asociaciones estadísticas las distintas variables climáticas recogidas en la base de datos con la concentración de partículas PM_{10} concluye con un resultado negativo: la ausencia de asociación entre la concentración de partículas PM_{10} y la radiación solar.

Como la lluvia acumulada, y como también fué descrito en la Sección 2.2, la radiación solar es una variable mixta con una componente discreta y una componente continua y positiva.

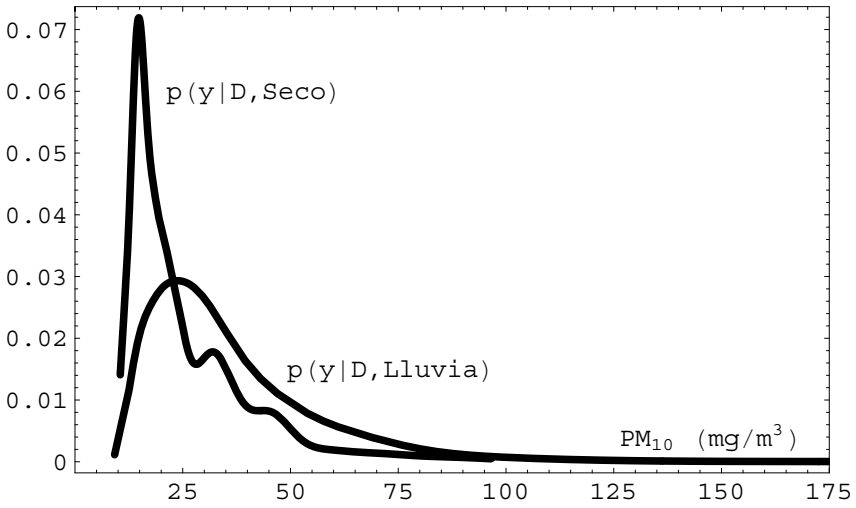


Figura 32. *Distribuciones predictivas de la concentración de PM_{10} condicionales a la presencia o ausencia de precipitaciones.*

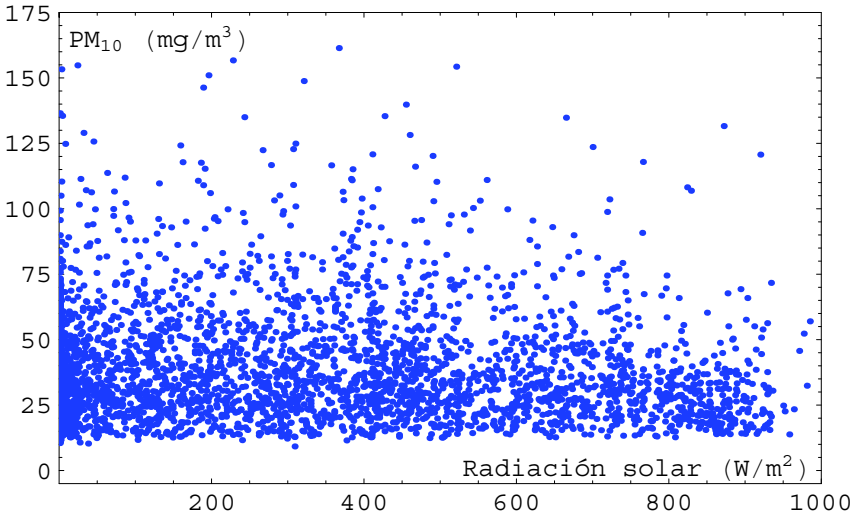


Figura 33. *Concentración de PM_{10} en función de la radiación solar*

La Figura 33 describe la concentración de PM_{10} en función de la radiación solar positiva. Como puede apreciarse, la distribución de las concentraciones de PM_{10} es esencialmente la misma para cualquier valor de la radiación solar.

Más formalmente, se encontró que las distribuciones marginales condicionales $p(y \mid w, D)$ de la concentración y de PM_{10} dada la radiación solar w y el banco de datos D básicamente coinciden, para cualquier valor de w, con la distribución marginal incondicional $p(y \mid D)$ de la concentración de PM_{10}, descrita con detalle en la Sección 3.2.2. Consecuentemente, no existe evidencia de que las concentraciones de PM_{10} estén asociadas con la la radiación solar.

3.4.2. *Relación con las actividades portuarias*

El la Tabla 1 de la Sección 2.1 se listan las 111 variables binarias que describen la presencia o la ausencia, para cada una de las 8040 horas registradas, de distintas actividades portuarias.

En el apartado 3.1.7 se analizó la frecuencia relativa de cada una de esa actividades, resumida en la Tabla 2, entre las que destacaban la descarga de carbón de hulla al muelle, la descarga de coque petróleo al muelle, y la carga de clinker a camión.

En este apartado se analiza en cambio la incidencia de las distintas actividades portuarias en los niveles de concentración de partículas PM_{10}.

Para cada una de las 97 actividades $\{a_1, \ldots, a_{97}\}$ diferenciadas (de las 111 potencialmenye incluidas en la base de datos D) que realmente tuvieron lugar en el periodo analizado se determinaron, con este objeto, la distribuciones predictivas condicionales

$$p(y \mid a_i = 1), \qquad p(y \mid a_i = 0),$$

del nivel y de concentración de partículas PM_{10} en función de la presencia ($a_i = 1$) o la ausencia ($a_i = 0$) de la actividad considerada a_i.

La Figura 34 (dividida para su presentación en dos paneles) describe gráficamente los resultados obtenidos. Para cada una de las variables se describen en rojo los niveles de concentración observados en en momentos en los que la actividad correspondiente estaba teniendo lugar, y se describen en azul los niveles de concentración observados cuando la actividad correspondiente *no* estaba teniendo lugar. En ambos casos, se representan con un punto negro de mayor tamaño las medias m_{1i} y m_{0i} de la correspondientes distribuciones condicionales. La diferencia de las medias condicionales, $d_{10i} = m_{1i} - m_{0i}$ es un indicador sencillo, razonablemente intuitivo, de la incidencia de la actividad considerada en los niveles de concentración de PM_{10}. Un indicador mucho más preciso es

$$p_{10i} = \mathbf{Pr}[y_1 > y_0 \mid D] = \int\int_{\{y_1 > y_0\}} p(y_1 \mid a_i = 1)\, p(y_0 \mid a_i = 0)\, \mathrm{d}y_1\, \mathrm{d}y_0$$

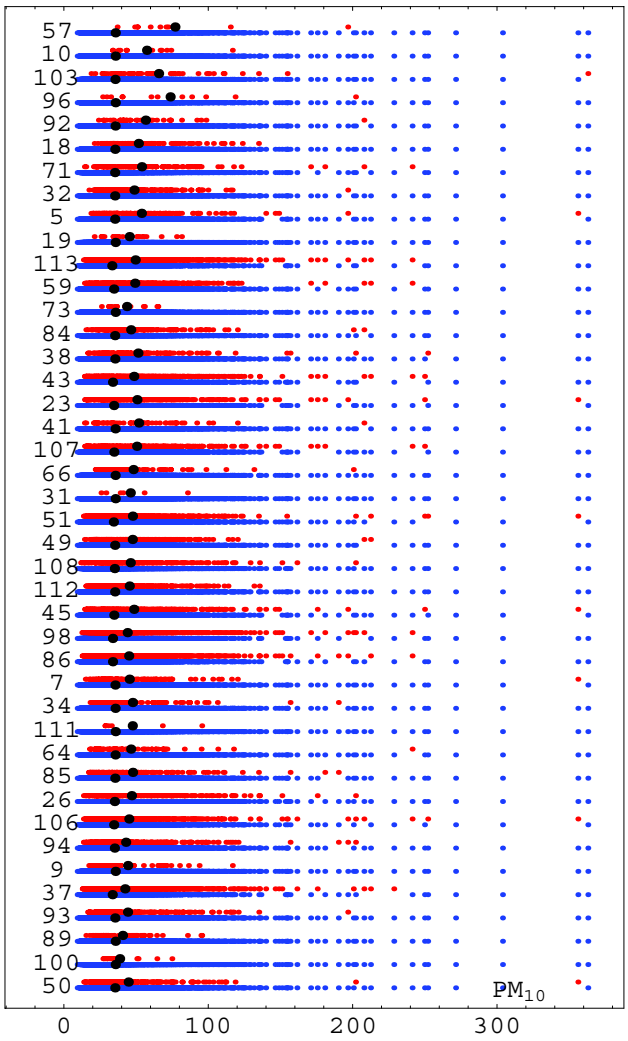


Figura 34.1 *Distribuciones de la concentración de partículas PM_{10} en función de la presencia o ausencia de las actividades portuarias con mayor incidencia.*

la probabilidad de que a_i produzca un incremento diferencial, esto es la probabilidad de que una observación y_1 de PM_{10} cuando la actividad a_i está teniendo lugar sea mayor que una observación y_0 de PM_{10} cuando a_i no está teniendo lugar.

En la Figura 34.1, se presentan las 42 actividades con mayor probabilidad de incidencia p_{10i}, por orden decreciente del valor de p_{10i}. En particular, las cuatro variables que presentan una mayor probabilidad de incidencia en la concentración

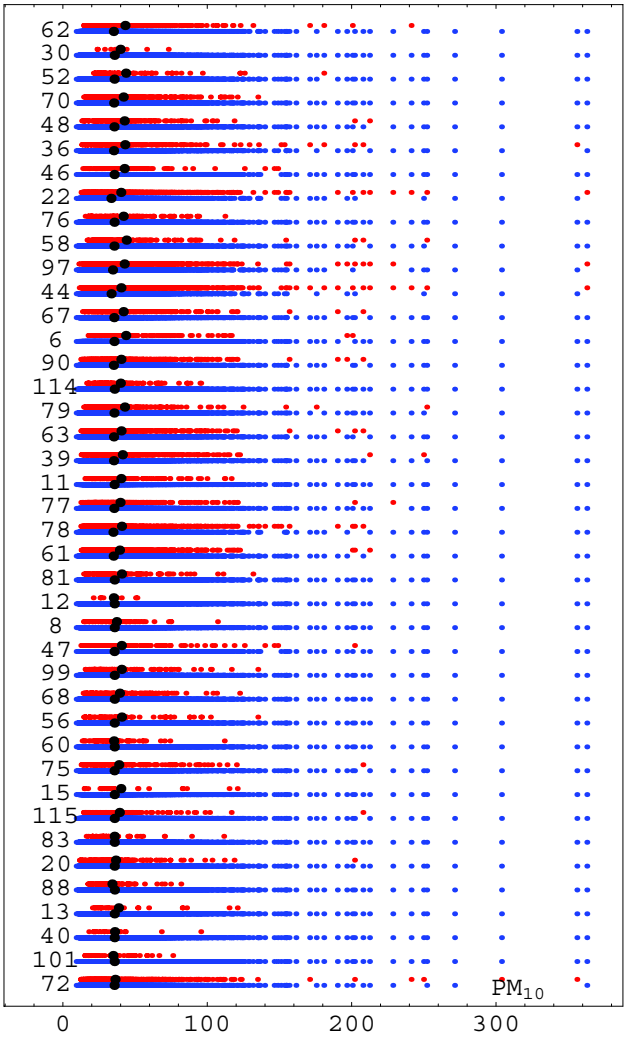


Figura 34.2 *Distribuciones de la concentración de partículas PM_{10} en función de la presencia o ausencia de otras actividades portuarias.*

de partículas PM_{10} (todas ellas con $p_{10i} \geq 0.8$) son, por este orden las variables 57, 10, 103 y 96 que respectivamente corresponden (ver Tabla 1, pags. 19–22) a carga de fosfatos a camión, carga de andalucita a camión, descarga de sorgo por tolva, y carga de habas de soja a camión. La Figura 34.2, completamente análoga a la 34.1, presenta las siguientes 41 variables, hasta completar el conjunto de las 83 que tienen una incidencia demostrablemente positiva en los niveles de PM_{10}.

La Figura 35 es un histograma de las probabilidades p_{10i} obtenidas para las 97 actividades registradas. Como puede observarse, tienden a situarse entre 0.4 y 0.7, con tan sólo cuatro actividades cuya probabilidad de incidencia se sitúa por encima de 0.8.

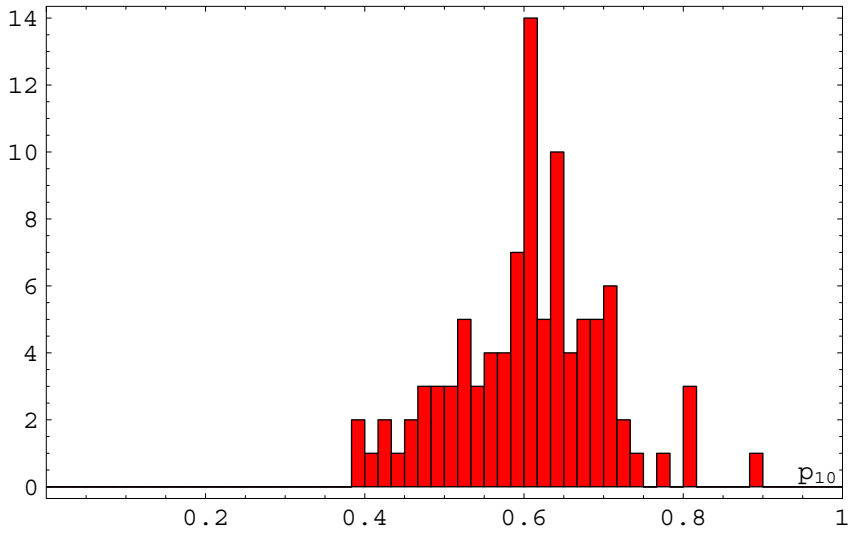


Figura 35. *Histograma de la probabilidades de incidencia de las distintas actividades portuarias*

Es importante observar que, debido a que los datos observados no se distribuyen de acuerdo con un modelo conocido, el cálculo de las p_{10i} no es trivial: ha requerido el uso de integración numérica por Monte Carlo a partir de observaciones generadas con las las distribuciones predictivas condicionales

$$p(y \mid a_i = 1), \qquad p(y \mid a_i = 0),$$

construidas a su vez mediante estimación no-paramétrica de densidades.

La Tabla 3, también presentada en dos paneles, contiene las características numéricas de las las distribuciones condiciones estudiadas. Para cada variable, cuyo indicador se sitúa en la primera columna, se detalla la probabilidad p_{10i} de incidencia en la concentración de PM_{10}, la diferencia d_{10i} entre las medias condicionales, el número n_{0i} de observaciones de PM_{10} disponibles sin presencia de a_i, su media m_{0i} y su desviación típica s_{0i}, el número n_{1i} de observaciones de PM_{10} disponibles con presencia de a_i, su media m_{1i} y su desviación típica s_{1i}.

Tabla 3.1 *Distribuciones de la concentración de PM_{10} condicionales a la presencia o ausencia de algunas actividades portuarias.*

i	p_{10i}	d_{10i}	n_{0i}	m_{0i}	s_{0i}	n_{1i}	m_{1i}	s_{1i}
57	0.887	41.3	6497	35.6	21.8	10	76.9	44.7
10	0.808	21.7	6495	35.6	21.9	12	57.3	23.2
103	0.800	30.2	6451	35.4	21.4	56	65.6	49.5
96	0.800	38.0	6495	35.6	21.8	12	73.6	48.1
92	0.773	21.0	6472	35.5	21.8	35	56.5	32.4
18	0.749	16.3	6370	35.3	21.8	137	51.6	23.4
71	0.720	18.6	6351	35.2	21.4	156	53.8	33.2
32	0.718	13.4	6293	35.2	21.8	214	48.6	22.6
5	0.713	18.5	6359	35.2	21.2	148	53.7	38.0
19	0.712	9.6	6483	35.6	22.0	24	45.3	15.1
113	0.711	16.1	5531	33.2	19.9	976	49.4	27.5
59	0.708	14.7	5993	34.5	21.1	514	49.1	26.8
73	0.701	7.9	6495	35.6	22.0	12	43.5	13.0
84	0.700	11.1	6256	35.2	21.8	251	46.3	23.5
38	0.697	16.0	6362	35.3	21.5	145	51.3	32.7
43	0.697	14.7	5645	33.7	20.3	862	48.4	27.4
23	0.695	16.2	6033	34.5	20.4	474	50.6	32.5
41	0.688	16.3	6447	35.5	21.8	60	51.8	31.4
107	0.685	15.8	6085	34.6	20.8	422	50.4	30.8
66	0.683	12.5	6437	35.5	21.8	70	48.0	28.5
31	0.681	10.4	6501	35.6	21.9	6	46.0	20.1
51	0.677	13.2	5860	34.3	20.6	647	47.5	29.4
49	0.676	12.2	6268	35.2	21.6	239	47.4	26.2
108	0.669	11.0	6119	35.0	21.6	388	46.0	24.6
112	0.666	10.1	6205	35.2	21.8	302	45.3	21.8
45	0.660	13.8	6002	34.6	20.6	505	48.4	31.3
98	0.660	10.2	5256	33.7	20.7	1251	43.9	24.7
86	0.657	11.2	5354	33.7	20.5	1153	44.9	25.7
7	0.649	9.9	6338	35.4	21.6	169	45.3	31.5
34	0.648	12.0	6425	35.5	21.8	82	47.5	29.6
111	0.648	11.7	6501	35.6	21.9	6	47.4	25.7
64	0.647	10.8	6439	35.5	21.8	68	46.3	31.5
85	0.647	12.3	6352	35.3	21.6	155	47.6	29.9
26	0.645	11.8	6138	35.0	21.4	369	46.8	27.8
106	0.643	10.5	5819	34.5	20.5	688	45.0	29.8
94	0.634	7.8	5997	35.0	21.7	510	42.8	23.6
9	0.634	8.6	6459	35.6	21.9	48	44.2	22.2
37	0.634	8.8	4896	33.5	20.9	1611	42.2	23.7
93	0.633	8.8	6250	35.3	21.8	257	44.1	24.7
89	0.629	5.0	6443	35.6	22.0	64	40.6	16.6
100	0.625	2.9	6489	35.6	22.0	18	38.6	12.8
50	0.621	9.3	6260	35.3	21.4	247	44.6	31.3

Tabla 3.2 *Distribuciones de la concentración de PM_{10} condicionales a la presencia o ausencia al resto de actividades con $p_{10i} > 0.5$.*

i	p_{10i}	d_{10i}	n_{0i}	m_{0i}	s_{0i}	n_{1i}	m_{1i}	s_{1i}
62	0.620	8.0	5985	35.0	21.5	522	43.0	25.0
30	0.614	4.1	6498	35.6	22.0	9	39.7	15.5
52	0.612	8.0	6416	35.5	21.9	91	43.5	26.8
70	0.611	6.6	6106	35.2	21.9	401	41.8	21.5
48	0.610	7.1	6307	35.4	21.8	200	42.5	25.5
36	0.610	7.7	6073	35.1	21.2	434	42.9	29.6
46	0.607	7.0	6401	35.5	21.9	106	42.5	25.4
22	0.607	6.9	4268	33.3	20.4	2239	40.1	24.0
76	0.606	6.2	6395	35.5	22.0	112	41.8	20.7
58	0.606	8.5	6331	35.4	21.6	176	43.9	31.6
97	0.606	8.1	5513	34.4	20.6	994	42.5	27.4
44	0.602	6.8	4339	33.4	20.0	2168	40.2	24.7
67	0.601	6.6	6162	35.3	21.7	345	41.8	24.6
6	0.601	8.1	6320	35.4	21.7	187	43.5	27.6
90	0.601	5.2	5763	35.0	21.9	744	40.2	22.1
114	0.598	4.3	6388	35.6	22.0	119	39.8	18.0
79	0.597	7.3	6330	35.4	21.7	177	42.7	28.8
63	0.595	5.2	5829	35.1	21.8	678	40.3	22.6
39	0.591	6.1	5973	35.1	21.7	534	41.2	23.9
11	0.591	4.6	6345	35.5	22.0	162	40.2	20.3
77	0.589	4.3	5976	35.3	21.9	531	39.6	21.8
78	0.588	5.9	5461	34.7	21.4	1046	40.6	24.2
61	0.581	4.3	5430	34.9	22.0	1077	39.2	21.3
81	0.578	5.1	6400	35.6	21.9	107	40.6	22.2
12	0.578	-0.7	6495	35.6	22.0	12	35.0	8.9
8	0.572	1.5	6436	35.6	22.0	71	37.1	16.2
47	0.565	5.1	6280	35.5	21.8	227	40.5	25.8
99	0.562	4.9	6398	35.6	21.9	109	40.5	23.1
68	0.553	3.6	6365	35.6	22.0	142	39.2	21.1
56	0.550	5.0	6421	35.6	21.9	86	40.6	24.6
60	0.542	-0.4	6430	35.6	22.0	77	35.2	15.1
75	0.541	3.1	6280	35.5	21.9	227	38.7	23.6
15	0.535	4.5	6473	35.6	21.9	34	40.1	26.4
115	0.532	3.5	6389	35.6	21.9	118	39.1	26.3
83	0.529	0.0	6455	35.6	22.0	52	35.6	17.5
20	0.529	0.9	6320	35.6	22.0	187	36.5	21.6
88	0.525	-1.6	6423	35.7	22.0	84	34.0	13.1
13	0.524	2.9	6467	35.6	21.9	40	38.5	24.5
40	0.516	0.1	6489	35.6	22.0	18	35.8	18.6
101	0.514	-1.0	6461	35.6	22.0	46	34.7	15.4
72	0.502	0.5	5286	35.5	21.4	1221	36.1	24.2

Capítulo 4.

Predicción y Decisión

4.1. FORMULACIÓN DEL PROCESO DE DECISIÓN

4.1.1. *Estructura de la función de pérdida*

En ausencia de información sobre el esfuerzo que la Autoridad Portuaria puede estar dispuesta a realizar para disminuir los niveles de concentración de partículas PM_{10}, este estudio se centra en determinar procedimientos que permitan mejorar la probabilidad de cumplir con la normativa vigente.

Formalmente, esto significa que la función de pérdida

$$\ell\{d_i,\ \boldsymbol{y}\}, \quad d_i \in \mathcal{D}, \quad \boldsymbol{y} = \{y_1, \ldots, y_{365}\}, \quad y_j > 0,$$

que debe describir las preferencias de la Autoridad Portuaria para cada una de las consecuencias $(d_i, \boldsymbol{y})$ que se producirían si se adoptase una estrategia d_i que derivara en la serie anual $\boldsymbol{y}$ para los valores de las concentraciones medias diarias de PM_{10}, dependerá exclusivamente de las dos condiciones citadas en la normativa vigente: (i) no se deben superar los $50\mu g/m^3$ de media diaria en más de 35 ocasiones al año y (ii) la media anual no debe superar los $40\mu g/m^3$. Formalmente, los sucesos relevantes son pues los sucesos $A(\boldsymbol{y})$ y $B(\boldsymbol{y})$ definidos por

$$A(\boldsymbol{y}) = \{\boldsymbol{y};\quad \sum_{j=1}^{365} \mathbf{1}_{\{y_j > 50\}} \leq 35\}$$

$$B(\boldsymbol{y}) = \{\boldsymbol{y};\quad \frac{1}{365}\sum_{j=1}^{365} y_j \leq 40\},$$

donde $\mathbf{1}_B$ es la función indicatriz de B, de forma que será necesario determinar las dos probabilidades

$$\{\Pr[A(\boldsymbol{y}) \,|\, d_i, D],\ \Pr[B(\boldsymbol{y}) \,|\, d_i, D]\}$$

en función de cada estrategia d_i, dada la información proporcional por el banco de datos D.

4.1.2. *Cálculo de la pérdida esperada*

En el Capítulo 4 ya concluimos que, con la legislación en vigor, la segunda de las normas europeas se cumple siempre, puesto que $\Pr[B(\boldsymbol{y}), d_i, D] \approx 0$; por lo tanto, el único suceso incierto relevante será $A(\boldsymbol{y})$, la función de pérdida tendrá la forma

$$\ell\{d_i, \boldsymbol{y}\} = \ell\{d_i, A(\boldsymbol{y})\}, \quad d_i \in \mathcal{D}, \quad \boldsymbol{y} = \{y_1, \ldots, y_{365}\}, \quad y_j > 0$$

y la estrategia óptima d^* será la que minimice la pérdida esperada, en las condiciones C en las que la decisión deba ser tomada, esto es

$$d^* = \arg\min_{d_i \in \mathcal{D}} \hat{\ell}\{d_i \mid D, C\},$$
$$\hat{\ell}\{d_i \mid D, C\} = \int_{\mathcal{Y}} \ell\{d_i, A(\boldsymbol{y})\}\, p(\boldsymbol{y} \mid D, d_i, C)\, \mathrm{d}\boldsymbol{y}.$$

Como ya se ha mencionado, en Bruselas se contempla la posibilidad de reducir cada año en 4μg/m^3 el límite permitido de la media anual de las concentraciones de PM_{10}, hasta alcanzar el límite de 20μg/m^{3}en el 2010. Si esta posibilidad se materializase, sería necesario incluir también al suceso $B(\boldsymbol{y})$ en la función de pérdida.

Bajo condiciones muy generales, la pérdida esperada $\hat{\ell}\{d_i \mid D, C\}$ correspondiente a una función de pérdida $\ell\{d,\, A\}$ que solamente depende de la eventual ocurrencia de un suceso incierto A depende de los datos D únicamente a través de la probabilidad final $\Pr(A \mid D, d_i, C)$ asociada a tal suceso bajo la hipótesis de que se implemente la estrategia d_i en las condiciones C. Más precisamente, la función de pérdida $\ell\{d,\, A\}$ tiene dos componentes, una pérdida fija, $\ell_0(d)$ (que corresponde al coste de implementar la estrategia d) y una pérdida aleatoria dicotómica que vale c, el coste asociado a la ocurrencia de A si el suceso A tiene lugar, y vale 0 en caso contrario. Consecuentemente, la utilidad esperada será de la forma

$$\hat{\ell}\{d_i \mid D\} = \ell_0(d_i) + c\, \Pr(A \mid D, d_i, C),$$

que sólo depende de los datos a través de la probabilidad $\Pr(A \mid D, d_i, C)$ de que el suceso A tenga lugar cuando se utiliza la estrategia d_i en condiciones C. Naturalmente, las funciones $\ell_0(d_i)$ y $\Pr(A \mid D, d_i, C)$ se comportan de manera inversa: una estrategia más conservacionista tiene un mayor coste fijo, pero da lugar a una probabilidad menor de infringir la normativa.

4.1.3. *Determinación de la estrategia óptima*

De acuerdo con los resultados básicos de la teoría de la decisión, una estrategia d_1 será preferible a otra estrategia d_2 si, y sólamente si, la pérdida esperada de d_1, $\hat{\ell}\{d_2 \mid D\}$ es menor que a pérdida esperada de d_2, $\hat{\ell}\{d_2 \mid D, C\}$, Esto sucede si, y solamente si,

$$\ell_0(d_1) + c\,\mathrm{Pr}(A \mid D, d_1, C) < \ell_0(d_2) + c\,\mathrm{Pr}(A \mid D, d_2, C),$$

es decir, cuando

$$\frac{\ell_0(d_1) - \ell_0(d_2)}{c} < \mathrm{Pr}(A \mid D, d_2, C) - \mathrm{Pr}(A \mid D, d_1, C).$$

En términos de las probabilidades asociadas al suceso complementario $\overline{A}$ (cuya probabilidad es obviamente uno menos la probabilidad de A y que significa cumplir con la normativa), la estrategia d_1 es preferible a la estrategia d_2 si, y solamente si,

$$\mathrm{Pr}(\overline{A} \mid D, d_1, C) - \mathrm{Pr}(\overline{A} \mid D, d_2, C) > \frac{\ell_0(d_1) - \ell_0(d_2)}{c} \tag{4}$$

En palabras, la estrategia d_1 es preferible a la estrategia d_2 si, y solamente si, el incremento en la probabilidad de cumplir la normativa como consecuencia de implementar d_1 en lugar de d_2 en las condiciones C en las que hay que tomar la decisión es mayor que el aumento de coste relativo que esto involucra. Esta ecuación fundamental proporciona la solución general al problema planteado. Para determinar si es apropiado sustituir una estrategia en vigor por una nueva estrategia, hay que calcular el aumento de probabilidad de cumplir la normativa que tal substitución implicaría en las condiciones presentes, y comprobar si este incremento es mayor que el incremento de gasto (en términos relativos al coste de incumplir la normativa), que tal sustitución conllevaría.

En el resto de este capítulo proporcionaremos las herramientas necesarias para calcular las probabilidades $\mathrm{Pr}(A \mid D, d_i, C)$ de incumplir la normativa para un conjunto razonable se posibles estrategias (en general una selección de las actividades portuarias permisibles), y para distintas condiciones C (típicamente determinadas por la situación meteorológica).

Más precisamente, los resultados del estudio realizado en el Capítulo 4 sobre la dependencia de la concentración de partículas PM_{10} como función de las variables climatológicas y de las actividades portuarias registradas van a permitirnos construir un modelo predictivo general, de la forma

$$\{p(y_{t+1} \mid \boldsymbol{a}_{t+1}, \boldsymbol{c}_t, D), \quad y > 0, \quad \boldsymbol{c}_i \in \mathcal{C}, \quad \boldsymbol{a}_j \in \mathcal{A}\},$$

donde $\boldsymbol{c}_t$ es el vector de las variables climatológicas en un día t, y D es la base de datos disponible, que permita obtener una distribución de probabilidad

$$p(y_{t+1} \mid \boldsymbol{a}_{t+1}, \boldsymbol{c}_t, D), \qquad y_{t+1} > 0$$

sobre el valor y_{t+1} de la media diaria de la concentración de partículas PM_{10} al día siguiente, $(t+1)$, en función de las actividades portuarias $\boldsymbol{a}_{t+1}$ que tengan ese día lugar en el puerto. Esta distribución predictiva permite inmediatamente el cálculo de la probabilidad condicional

$$\Pr[y_{t+1} > 50 \mid \boldsymbol{a}_{t+1}, \boldsymbol{c}_t, D)] = \int_{50}^{\infty} p(y_{t+1} \mid \boldsymbol{a}_{t+1}, \boldsymbol{c}_t, D)\, \mathrm{d}y_{t+1},$$

de que la media diaria supere la cota de los 50μg/m^3.

Como ya se indicó en el apartado 3.2.2, las probabilidades condicionales requeridas $\Pr(A \mid D, d_i, C)$ se obtienen entonces fácilmente como

$$\begin{aligned} \Pr(A \mid D, d_i, C) 1 - \sum_{k=0}^{35} \binom{365}{k} p^k (1-p)^{365-k} \\ \approx 1 - \Phi\left(\frac{35 - 365p}{\sqrt{365p(1-p)}}\right), \end{aligned} \tag{6}$$

donde $p = \Pr[y_{t+1} > 50 \mid \boldsymbol{a}_{t+1}, \boldsymbol{c}_t, D)]$.

4.2. DETERMINACIÓN DE COVARIABLES RELEVANTES

En esa sección presentamos funciones, tanto de las variables climáticas como de los indicadores de las actividades portuarias capaces de proporcionar predicciones útiles sobre las concentraciones en el aire de partículas PM_{10}.

4.2.1. *Indice climático*

El análisis de las distribuciones predictivas de las concentraciones de PM_{10} condicionales a distintas funciones de las variables climáticas constituyeron la base par una búsqueda sistemática de un índice climático unidimensional capaz de resumir la mayor parte de la información meteorológica relevante cuando se trata de predecir los niveles de tales concentraciones.

El problema se ha planteado como un problema de decisión en selección de variables. En este caso, el conjunto de alternativas es el conjunto de las partes de las funciones de las variables climáticas que se está dispuesto a considerar, y la función de utilidad es una función de evaluación predictiva, basada en el logaritmo de la

probabilidad asociada al suceso observado por una función predictiva lineal en las variables escogidas.

Regresión. El conjunto de las variables encontradas en ese proceso de selección ha sido:

(i) *Componente sur de la velocidad del viento*
$x_1 = v_0 \cos[180 - \alpha]$, donde v_0 es el modulo de la velocidad de viento en m/s, y α el radial del que proviene en [0, 360].

(ii) *Distancia absoluta a los 17 C*
$x_2 = |\, T - 17 \,|$, donde T es la temperatura del aire en grados centígrados.

(iii) *Presión atmosférica*
$x_3 = P$, Presión atmosférica en mb

(iv) *Precipitación positiva*
x_4 es una variable binaria que toma el valor 1 si la lluvia acumulada L es estrictamente positiva, y toma el valor 0 en caso contrario.

Las medias m_i y las desviaciones típicas s_i de las cuatro variables utilizadas son la descritas en la Tabla 4

Tabla 4. *Características de las variables climáticas relevantes*

	x_1	x_2	x_3	x_4
m_i	-0.668	6.102	1017.8	0.0326
s_i	3.370	3.252	7.247	0.178

Expresándola en términos de las variables tipificadas

$$z_i = \frac{x_i - m_i}{s_i}, \quad , i = 1, \ldots, 4,$$

para que sus coeficientes sean comparables entre sí, la recta de regresión correspondiente permite apreciar el impacto relativo sobre la concentración de partículas PM_{10} de cada una de las variables consideradas. El resultado es

$$E[y \,|\, \text{Clima}] = 36.094 + 1.983\ z_1 - 1.479\ z_2 + 5.183\ z_3 - 1.759\ z_4$$

Es inmediato observar que, en presencia de las cuatro variables, la variable climática que más información proporciona para predecir el nivel de concentración

de PM_{10} es la presión atmosférica, que tiene el mayor coeficiente; le sigue la componente sur de la velocidad del viento también con coeficiente positivo, de forma que, concentración de PM_{10} está positivamente asociada con altas presiones y con vientos de componente sur. Las otras dos variables, presencia de precipitación y distancia absoluta de la temperatura a los 17^oC tienen signo negativo; en valor medio, los niveles de PM_{10} serán menores con temperaturas próximas a los 17^oCy con tiempo lluvioso.

Naturalmente, no puede esperarse que el clima, por si solo, proporcione una buena predicción del nivel de concentración de PM_{10} pero, como veremos más adelante, el modelo estudiado permite anticipar situaciones climatológicas en las que es necesario extremar las precauciones con las actividades realizadas.

En efecto, la línea de regresión permite definir lo que llamaremos el *índice climático* $\theta = \theta(\text{Clima}) = \theta(z_1, \ldots, z_4)$, con

$$\theta(\text{Clima}) = 36.094 + 1.983\ z_1 - 1.479\ z_2 + 5.183\ z_3 - 1.759\ z_4 \qquad (5)$$

positivamente correlacionada con la concentración de PM_{10}, cuya distribución marginal se representa en la Figura 36, que permite descontar el efecto climático del estudio de la concentración de PM_{10}. Como puede observarse, se trata de una distribución unimodal centrada alrededor de 40, con soporte aproximado en la región $[15, 60]$. Sus cuantiles de orden 0.25, 0.50, 0.75 y 0.99, que más adelante utilizaremos son, respectivamente, $Q_{25} = 32.07$, $Q_{50} = 36.87$, $Q_{75} = 40.13$ y $Q_{99} = 52.01$

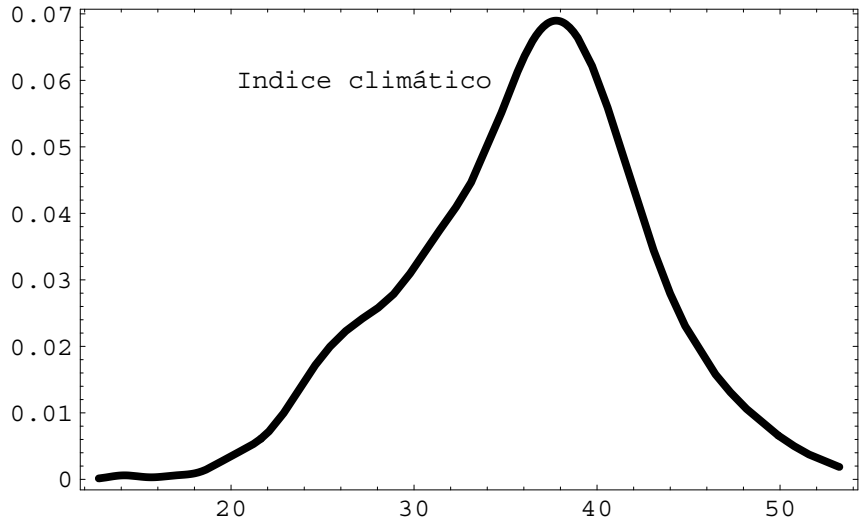


Figura 36. *Distribución incondicional del indice climático definido por la línea de regresión* $E[y \mid Clima]$

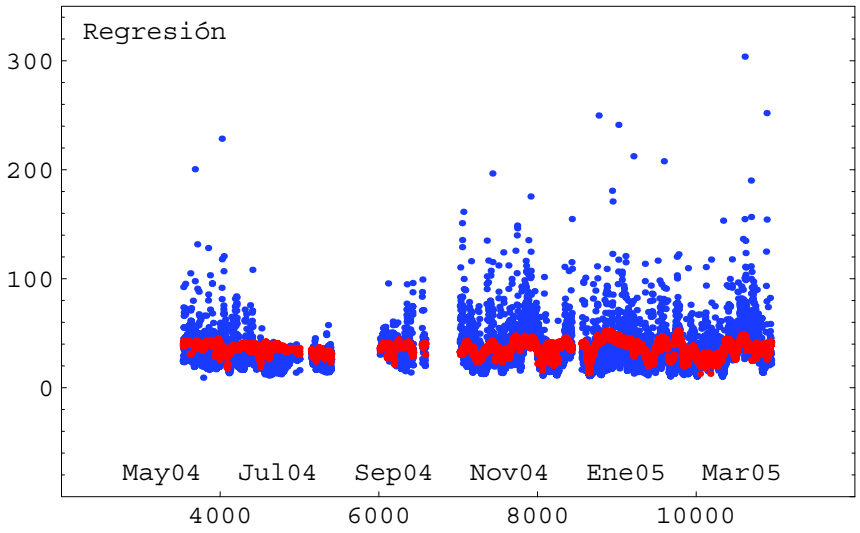


Figura 37. *Evolución del indice climático definido por la línea de regresión* $E[y \mid Clima]$

En la Figura 37 se representan conjuntamente las evoluciones temporales de las concentraciones de PM_{10} (en azul) y del índice climático construido (en rojo). Puede observarse que el índice climático definido en (4) sigue la evolución media de las concentraciones de PM_{10}, de forma que permite eliminar el efecto del clima de la variabilidad observada en la serie de las concentraciones de PM_{10}, variabilidad que fundamentalmente depende de las actividades portuarias.

4.2.2. *Categorización de las actividades portuarias*

La relación de orden introducida por la probabilidad p_{10i} de que una actividad a_i produzca un incremento diferencial en la concentración de partículas PM_{10} en el conjunto de variables binarias que describen las actividades portuarias (ver apartado 3.4.2) permite hacer con ellas una clasificación que resulta muy útil para diseñar estrategias de control medioambiental.

Con base a los valores obtenidos (descritos en la segunda columna de la Tabla 3, hemos dividido el conjunto de las 97 actividades observadas en cinco conjuntos anidados, definidos en términos de la probabilidad p_{10i} asociada al suceso de que la concentración de PM_{10} observada en momentos en los que la actividad a_i tiene lugar sea mayor sea mayor que la concentración de PM_{10} en momentos en los que no tiene lugar. Más concretamente, denotamos por $\mathcal{A}_0$ el conjunto de todas las actividades registradas, y definimos subconjuntos de actividades $\mathcal{A}_k \subset \mathcal{A}_0$ de la

Tabla 5. *Conjuntos anidados de actividades, por orden decreciente de probabilidad de incidencia en la concentración de partículas PM_{10}*

$\mathcal{A}_5 \equiv \{a_i;\, p_{10i} \geq 0.80\}$									
57	10	103	96						
$\mathcal{A}_4 \equiv \{a_i;\, p_{10i} \geq 0.70\}$									
57	10	103	96	92	18	71	32	5	19
113	59	73	84						
$\mathcal{A}_3 \equiv \{a_i;\, p_{10i} \geq 0.65\}$									
57	10	103	96	92	18	71	32	5	19
113	59	73	84	38	43	23	41	107	66
31	51	49	108	112	45	98	86		
$\mathcal{A}_2 \equiv \{a_i;\, p_{10i} \geq 0.60\}$									
57	10	103	96	92	18	71	32	5	19
113	59	73	84	38	43	23	41	107	66
31	51	49	108	112	45	98	86	7	34
111	64	85	26	106	94	9	37	93	89
100	50	62	30	52	70	48	36	46	22
76	58	97	44	67	6	90			
$\mathcal{A}_1 \equiv \{a_i;\, p_{10i} \geq 0.50\}$									
57	10	103	96	92	18	71	32	5	19
113	59	73	84	38	43	23	41	107	66
31	51	49	108	112	45	98	86	7	34
111	64	85	26	106	94	9	37	93	89
100	50	62	30	52	70	48	36	46	22
76	58	97	44	67	6	90	114	79	63
39	11	77	78	61	81	12	8	47	99
68	56	60	75	15	115	83	20	88	13
40	101	72							
$\mathcal{A}_0$: Todas las a_i									
57	10	103	96	92	18	71	32	5	19
113	59	73	84	38	43	23	41	107	66
31	51	49	108	112	45	98	86	7	34
111	64	85	26	106	94	9	37	93	89
100	50	62	30	52	70	48	36	46	22
76	58	97	44	67	6	90	114	79	63
39	11	77	78	61	81	12	8	47	99
68	56	60	75	15	115	83	20	88	13
40	101	72	91	54	74	102	16	28	42
104	110	14	24	82	17	109			

forma

$$\mathcal{A}_k \equiv \{a_i; \quad p_{10i} \geq p_k\}, \quad 0 < p_k < 1,$$

constituidos por todas la actividades cuyas probabilidades de incidencia son mayores que p_k. Si las cotas p_k están ordenadas, entonces los conjuntos resultantes estarán anidados, de forma que si $p_i > p_j$, entonces $\mathcal{A}_i \subset \mathcal{A}_j$. Valores mayores de p_k dan lugar a conjuntos más pequeños, constituidos por actividades con mayor riesgo de producir niveles inaceptables en la concentración de PM_{10}.

El estudio del histograma correspondiente a las 97 probabilidades de incidencia obtenidas (Figura 35) sugiere tomar como puntos de corte (curvas de nivel) los valores 0.50, 0.60, 0.65, 0.70 y 0.80 que dan lugar respectivamente a los cinco conjuntos $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{A}_4$ y $\mathcal{A}_5$, cuyos códigos aparecen en la Tabla 5.

Con objeto de facilitar la discusión de los resultados, llamaremos actividades de nivel k a aquellas que pertenecen al conjunto $\mathcal{A}_k$ pero no pertenecen al conjunto $\mathcal{A}_{k+1}$; en palabras, el nivel de una actividad es el índice del conjunto más pequeño al que pertenece. En particular, las únicas actividades de nivel 5 son las que constituyen $\mathcal{A}_5$, es decir las actividades codificadas con 57, 10, 103 y 96. Por ejemplo, la actividad 71 (carga de guisantes a camión según puede verse el la Tabla 1) es de nivel 4 (pertenece a $\mathcal{A}_4$ pero no a $\mathcal{A}_5$), mientras que la actividad 17 (carga de bauxita a camión) es de nivel 1 (puesto que pertenece a $\mathcal{A}_1$, pero no a $\mathcal{A}_2$).

Diremos que en un momento determinado el puerto está en nivel de actividad $\alpha = k$, $k \in \{0, 1, \ldots, 5\}$ si ese es el nivel más alto entre todas las actividades que tienen lugar en ese momento.

4.3. DETERMINACIÓN DE LA ESTRATEGIA ÓPTIMA

El uso combinado del índice climático definido en el apartado 5.2.1 y de la partición de las actividades portuarias definida en el apartado 5.2.2 va a permitirnos construir una sencilla función de alerta, fácilmente programable en una simple hoja de cálculo, que no requiera por parte de la Autoridad Portuaria reproducir los largos y complejos cálculos que su desarrollo ha requerido.

Como hemos mencionado en la Sección 5.1, la solución a el problema de decisión planteado se reduce a la regla de decisión definida en (4). Para implementarla, es suficiente determinar las probabilidades predictivas $\Pr(A \mid D, d_i, C)$. A su vez, cada una de estas probabilidades es una función sencilla (Ecuación 6) de la probabilidad predictiva

$$p_{50}(\boldsymbol{a}, \boldsymbol{c}, D) = \Pr[y_{t+1} > 50 \mid \boldsymbol{a}_{t+1}, \boldsymbol{c}_t, D)] \tag{7}$$

de que la media diaria de las concentraciones de PM_{10} supere los 50 mg, en que depende de las actividades $\boldsymbol{a}$ que tengan lugar en el puerto y de las condiciones climatológicas presentes $\boldsymbol{c}$.

El cálculo exacto de $p_{50}(\boldsymbol{a}, \boldsymbol{c}, D)$ es muy complicado, pero puede simplificarse haciendo uso de los resultados descritos en este capítulo. En efecto, excepto para situaciones muy extremas, la probabilidad descrita en (7) puede ser aproximada utilizando el índice climático θ el lugar del conjunto íntegro $\boldsymbol{c}$ de las condiciones climáticas, y el nivel de actividad definido en sección 5.2.2 en lugar del conjunto íntegro $\boldsymbol{a}$ de todos los estados de actividad. Formalmente,

$$p_{50}(\boldsymbol{a}, \boldsymbol{c}, D) \approx p_{50}\{\alpha(\boldsymbol{a}), \theta(\boldsymbol{c}), D\},$$

donde $\alpha(\boldsymbol{a})$ es el nivel de actividad en el puerto correspondiente al vector $\boldsymbol{a}$ y $\theta(\boldsymbol{c})$ es el íncide climático correspondiente a las condiciones climatológicas $\boldsymbol{c}$.

El cálculo de $p_{50}\{\alpha(\boldsymbol{a}), \theta(\boldsymbol{c}), D\}$ ha sido realizado por integración numérica con Monte carlo a partir de las relevantes distribuciones predictivas condicionales de la concentración de PM_{10}. La Tabla 6, posiblemente la más importante de este trabajo, resume los resultados obtenidos.

Tabla 6. *Probabilidades de que la media diaria de PM_{10} supere los los 50μg/m^3, en función del índice climático y del nivel de actividad en el puerto*

	Actividad					
Clima	0	1	2	3	4	5
$\theta \leq Q_{25}$	0.024	0.021	0.025	0.05	0.047	0.047
$\theta \leq Q_{50}$	0.015	0.02	0.016	0.046	0.047	0.047
$\theta \leq Q_{75}$	0.041	0.05	0.049	0.075	0.076	0.076
$\theta \leq Q_{99}$	0.060	0.08	0.080	0.107	0.138	0.148

En la Tabla 6, las probabilidades requeridas se detallan para cada uno de los 6 niveles de actividad: desde 0, cuando sólo tienen lugar las actividades con una probabilidad de incidencia en la concentración de PM_{10} menor de 0.5 hasta 5, cuando entre las actividades que tiene lugar se encuentra alguna situada el el conjunto $\mathcal{A}_5$ constituido por aquellas actividades con una probabilidad de incidencia en la concentración de PM_{10} mayor o igual a 0.8.

Con respecto al índice climático, se han tabulado los valores correspondientes a los cuantiles 0.25,).50, 0.75 y 0.99, que son, respectivamente, 32.07, 36.87, 40.13 y 52.01; los valores correspondientes a cualquier otro valor de θ pueden ser determinados por interpolación.

Si, por ejemplo, las condiciones climatológicas previstas para un día determinado (presión atmosférica, viento sur, temperatura y presencia o ausencia de

precipitaciones) dieran lugar, utilizando la ecuación (5), a un valor del índice climático $\theta = 40.5$ y la actividad prevista de mayor probabilidad de incidencia fuese la 86 (carga de pulpa de remolacha a camión) que pertenece al conjunto $\mathcal{A}_3$ pero no al $\mathcal{A}_4$ con lo que el nivel de actividad es $\alpha = 3$, la probabilidad de que la media diaria de PM_{10} supere los $550\mu g/m^3$sería del orden de 0.075, que, utilizando la ecuación (6) indicaría, si las circunstancias se mantienen, una probabilidad de violar la normativa de (pasar de los $50\mu g/m^3$mas de 35 veces en un año) de 0.082.

Como era de esperar, la probabilidad de superar de que la media diaria supere los $50\mu g/m^3$crece con el índice climático (específicamente diseñado para controlar el efecto meteorológico) y con el nivel de actividad en el puerto.

Como ya fué indicado, para que la probabilidad de violar la normativa se mantenga razonablemente baja, (por debajo el 5%) la probabilidad de superar la cota en un día cualquiera debería estar casi todos los días por debajo de 0.07.

La Tabla 7, con idéntica estructura a la Tabla 6, describe los valores esperados de la distribución predictiva de la media horaria de la concentración de PM_{10} en función del índice climático y del nivel de actividad del puerto. De forma totalmente análoga, la Tabla 8 recoge las medianas correspondientes, y las Tablas 9, 10 y 11 los cuantiles 0.75, 0.90 y 0.99.

Tabla 7. *Valores esperados de la distribución predictiva de la media diaria de PM_{10}, en función del índice climático y del nivel de actividad en el puerto*

	Niveles					
Clima	0	1	2	3	4	5
$\theta \leq Q_{25}$	25.6	25.8	25.7	28.7	28.8	28.8
$\theta \leq Q_{50}$	26.7	27.2	27.2	29.6	30.2	30.2
$\theta \leq Q_{75}$	29.0	29.4	29.2	31.7	32.6	32.6
$\theta \leq Q_{99}$	30.5	31.2	31.1	33.5	35.7	36.1

Tabla 8. *Medianas de la distribución predictiva de la media diaria de PM_{10}, en función del índice climático y del nivel de actividad en el puerto*

	Niveles					
Clima	0	1	2	3	4	5
$\theta \leq Q_{25}$	24.0	24.4	24.2	26.5	26.5	26.5
$\theta \leq Q_{50}$	25.7	25.7	25.8	27.7	28.6	28.7
$\theta \leq Q_{75}$	27.6	27.3	27.6	29.7	31.0	31.0
$\theta \leq Q_{99}$	28.6	28.1	28.7	31.5	33.8	34.0

Tabla 9. *Cuantiles* 0.75 *de la distribución predictiva de la media diaria de PM_{10}, en función del índice climático y del nivel de actividad en el puerto*

	Niveles					
Clima	0	1	2	3	4	5
$\theta \leq Q_{25}$	30.8	30.3	29.9	32.5	32.8	32.7
$\theta \leq Q_{50}$	31.8	32.8	32.4	35.1	36.1	36.0
$\theta \leq Q_{75}$	35.4	35.7	35.0	38.2	39.2	39.1
$\theta \leq Q_{99}$	37.2	37.6	37.1	40.7	43.4	43.7

Tabla 10. *Cuantiles* 0.90 *de la distribución predictiva de la media diaria de PM_{10}, en función del índice climático y del nivel de actividad en el puerto*

	Niveles					
Clima	0	1	2	3	4	5
$\theta \leq Q_{25}$	38.2	37.0	36.6	43.4	41.6	41.5
$\theta \leq Q_{50}$	40.5	40.5	40.1	44.3	43.6	43.6
$\theta \leq Q_{75}$	43.6	43.4	42.2	46.5	46.6	46.5
$\theta \leq Q_{99}$	46.3	47.3	46.7	50.8	53.4	54.1

Tabla 11. *Cuantiles* 0.99 *de la distribución predictiva de la media diaria de PM_{10}, en función del índice climático y del nivel de actividad en el puerto*

	Niveles					
Clima	0	1	2	3	4	5
$\theta \leq Q_{25}$	60.8	60.8	59.5	71.8	71.1	71.1
$\theta \leq Q_{50}$	52.7	53.9	52.5	63.2	62.6	62.6
$\theta \leq Q_{75}$	61.5	63.2	62.4	69.4	72.3	72.4
$\theta \leq Q_{99}$	67.3	69.6	67.7	71.3	76.5	78.5

Por ejemplo, de acuerdo con las Tablas 7, 8, 9, 10 y 11, si el índice climático se sitúa alrededor de su mediana (36.87) y el nivel de actividad del puerto es medio-alto (nivel 3), el valor de la media diaria de la concentracinó de estará alrededor de $28\mu g/m^3$(media 29.6, mediana 27.7) y se situara por debajo de 35.1, 44.3, y $63.2\mu g/m^3$con probsbilidades 0.75, 0.90, y 0.99, respectivamente.

Capítulo 5.

Conclusiones

5.1. *Metodolgía*

1. Se ha utilizado la teoría de la decisión para proponer una forma de control sobre la concentración de partículas PM_{10} en el aire.
2. Se ha desarrollado una metodología Bayesiana objetiva noparamétrica para analizar adecuadamente los datos disponibles, cuyos detalles técnicos se aportan en un apéndice.

5.2. *Situación actual*

3. Existen dos normas vigentes con respecto a la concentración de partículas PM_{10} en el aire: (i) no se debe sobrepasar una media anual de $40\mu g/m^3$, que se cumple con seguridad, y (i) no se deben superar los $50\mu g/m^3$más de 35 días al año, que se *incumple* con toda seguridad. Las normas sugeridas para la segunda fase serían de muy difícil cumplimiento.
4. Los datos sobre PM_{10} recogidos por la cabina HADA deben ser necesariamente recalibrados a una escala gravimétrica antes de utilizarlos. La distribución predictiva de la medida calibrada y es una t de Student centrada en una línea de regresión.
5. la inexistencia de una base de datos medioambiental homogéneas ha producido mucho problemas y retrasos. Para implementar un sistema automatizado es prioritario que se dedique recursos y personal para crear y mantener una base de datos apropiada.

5.3. *Índice climático*

5. Se ha desarrollado un índice climático que permite automatizar el factor climatológico en la predicción de la concentración de PM_{10}. Se trata de una función lineal de la componente sur de la velocidad del ciento, de la presión atmosférica, de la diferencia absoluta entre la temperatura y los 17^oCy de la presencia o ausencia de precipitaciones. (Sección)
6. Los modelos de predicción microclimática desarrollados dentro del proyecto HADA podrían ser utilizados para predecir el índice climático desarrollado con al menos 24h de anticipación lo que permitiría anticipar precauciones en la gestión de las actividades portuarias los días que se prevea un índice alto.

5.4. *Clasificación de las actividades portuarias*

7. Se ha desarrollado un indicador del riesgo para los niveles de PM_{10} asociado a cada actividad portuarias, basado en una estimación la probabilidad de que la concentración de PM_{10} aumente cuando la actividad tiene lugar.
8. Esta probabilidad de incidencia ha permitido ordenar las 97 actividades registradas y clasificarlas en 6 grupos en función del riesgo que plantean. El nivel de actividad del puerto lo define la actividad de mayor riesgo (Sección)

5.5. *Probabilidad de cumplir la normativa*

9. Sea ha calculado una tabla de doble entrada que permite determinar la probabilidad de que la media diaria de PM_{10} supere la cuota de los $50\mu g/m^3$en función del índice climático y del nivel de actividad del puerto.
10. La probabilidad de cumplir la norma vigente (no superar los $50\mu g/m^3$mas de 35 veces al año) es una función de la probabilidad media de superar la cota un día cualquiera. Para que la probabilidad de cumplir la norma sea superior a 0.95, la probabilidad de que la media diaria supere los $50\mu g/m^3$debe situarse casi todos los días por debajo de 0.07.

5.6. *Estrategia óptima*

11. Se ha identificado un procedimiento general para adoptar las estrategias de control medioambiental que puedan proponerse. El criterio indica que la nueva estrategia debe adoptase si, y solamente si, el incremento de probabilidad de que se cumpla la normativa como consecuencia del cambio es mayor que el coste relativo de incrementarlo, esto es el cociente entre en aumento de gasto y el coste estimado de violar la normativa. Ecuación (), Sección .

12. La Autoridad portuaria dispone, con los elementos contenidos en esta memoria de información suficiente para implementar el criterio indicado. Basta con que estime el coste relativo de cualquier modificación propuesta y lo compare con el incremento en la probabilidad de cumplir la normativa que se deduce de las tablas ya mencionadas.

Apéndices

A.1. DESCRIPCIÓN TÉCNICA DE LA METODOLOGÍA ORIGINAL

En esta sección se describe, con el formato de un documento académico de trabajo, la metodología bayesiana objetiva no paramétrica que se ha utilizado en esta memoria. Para facilitar su lectura en los demás paises de la Unión Europea, esta sección ha sido redactada en inglés,

Model-Free Objective Bayesian Prediction

1. *The Prediction Problem*

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a set of n real-valued observations of some *observable* real-valued quantity x, and consider a situation where one is interested in a (necessarily probabilistic) *prediction* of a future observation of the same quantity. Let us suppose that the observed values $\{x_1, \ldots, x_n\}$ may be assumed to be a subset of an *exchangeable* sequence, so that the *order* in which these observations have been obtained is assumed to contain no relevant information on the behaviour of the x's. Note that, in particular, this includes *all* cases in which $\boldsymbol{x}$ may be assumed to be a random sample from some underlying probability model.

It then follows from the general representation theorem (see *e.g.*, Bernardo and Smith, 1994, Ch. 4 and references therein) that there exists some probability model $m(x_i \mid \boldsymbol{\theta})$, labelled by some parameter $\boldsymbol{\theta} \in \Theta$, such that the joint probability

density of $\boldsymbol{x}$ may be written as

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} m(x_i \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{1}$$

Consequently, $\boldsymbol{x}$ may always be regarded as a random sample from *some*, typically unknown, probability model $m(x_i \,|\, \boldsymbol{\theta})$, indexed by some unknown (and possibly multidimensional) parameter $\boldsymbol{\theta} \in \Theta$, defined as the limit as $n \to \infty$ of some function of $\boldsymbol{x}$, for which a prior distribution $p(\boldsymbol{\theta})$ *necessarily* exists. Note that this result is an *existence theorem* in probability theory and, hence, it is *not* subject to any of the polemics often associated to the use of Bayesian statistics in the sciences with a subjective prior specification.

An immediate corollary of the representation theorem is that *all* the information about the value of future observation x contained in the observed data $\boldsymbol{x}$ is encapsulated in its (posterior) *predictive* distribution

$$p(x \,|\, \boldsymbol{x}) = p(x \,|\, x_1, \ldots, x_n) = \int_\Theta m(x \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}, \tag{2}$$

where, by Bayes' theorem, the *posterior* distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of the unknown parameter $\boldsymbol{\theta}$ is of the form

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\boldsymbol{\theta} \,|\, x_1, \ldots, x_n) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} m(x_i \,|\, \boldsymbol{\theta}). \tag{3}$$

For any exchangeable data set $\boldsymbol{x}$, the posterior predictive distribution $p(x \,|\, \boldsymbol{x})$ given by (2) is *the* solution to the problem posed: it precisely describes *all* available information about a future observation x. If a point estimate $\hat{x}$ is desired, the mode, the median or the mean of $p(x \,|\, x_1, \ldots, x_n)$ could be used; confidence regions $R(\alpha)$ with posterior probability $1 - \alpha$ may be obtained as solutions of the equation $\int_{R(\alpha)} p(x \,|\, \boldsymbol{x})\, dx = 1 - \alpha$. Those are however only *partial* (if very useful) descriptions of the available information about a future value of x; the *complete* solution is simply and elegantly encapsulated in $p(x \,|\, \boldsymbol{x})$. Moreover, any other form of solution will *necessarily* violate the basic rules of probability theory; unfortunately, this includes most conventional proposals, such as those obtained by plug-in estimates of the form $m(x \,|\, \hat{\boldsymbol{\theta}})$, for some estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Naturally, the problem is to find a suitable model $m(x \,|\, \boldsymbol{\theta})$, and to specify the prior distribution, $p(\boldsymbol{\theta})$, for its associated parameter $\boldsymbol{\theta}$.

In some scientific contexts, there are good reasons to select a particular model $m(x \,|\, \boldsymbol{\theta})$; this may be suggested, for instance, by an underlying physical theory, by invariance considerations, or by judicious application of some limit theorem. If

this is the case, the problem reduces to specifying an appropriate, non-subjective, model based, 'reference' prior distribution $\pi(\boldsymbol{\theta})$ which would let the data 'speak for themselves'. The prediction problem would then be immediately solved by the corresponding reference posterior predictive distribution

$$\begin{gathered}\pi(x \mid \boldsymbol{x}) = \pi(x \mid x_1, \ldots, x_n) = \int_\Theta m(x \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid x_1, \ldots, x_n)\, d\boldsymbol{\theta}, \\ \pi(\boldsymbol{\theta} \mid x_1, \ldots, x_n) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} m(x_i \mid \boldsymbol{\theta}).\end{gathered} \tag{4}$$

For a detailed description of Bayesian prediction, including the use of dynamic models, see the excellent review paper by West (1998), and references therein.

In the long quest for these 'baseline' non-subjective distributions, a number of requirements have emerged which may reasonably be regarded as their necessary properties. These include invariance, consistent marginalization, good frequency properties, general applicability and limiting admissibility. The *reference analysis* algorithm, introduced by Bernardo (1979b) and further developed by Berger and Bernardo (1989, 1992) is, to the best of our knowledge, the only available method to derive non-subjective distributions which satisfy all these desiderata. For a recent discussion of the many polemic issues involved in this topic, see Bernardo (1997). For an introduction to reference analysis, see Bernardo and Smith (1994, Ch. 5), or Bernardo and Ramón (1998).

In many situations however, it is very difficult to specify the probability model $m(x \mid \boldsymbol{\theta})$ with a reasonable degree of confidence. An *exact* Bayesian approach then requires to specify a very large class of models $m(x \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, where Θ is often infinitely dimensional, one of whose members hopefully provides a good approximation to the underlying probability mechanism, *and* a prior $p(\boldsymbol{\theta})$ which describes available information on this structure; popular choices are mixture models with Dirichlet priors (see *e.g.*, West, 1992; Escobar and West, 1995, Roeder and Wasserman, 1997, and references therein). However, subjective prior specification within this framework is very difficult –and often polemic–, and the reference priors for those models are typically *very* difficult to derive.

A possible alternative, which will be described in this paper, is to consider an *approximate*, data-based 'model' may be used as a *proxy* to the actual, unknown underlying model. The more successful techniques to achieve such a type of approximation are known under the general heading of *kernel density estimation*. Those are considered in the next section.

2. *Kernel density estimation*

2.1. Conventional Approach.

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from some unknown underlying model $m(x \,|\, \boldsymbol{\theta})$. Conventional kernel density estimation consists on assuming that an appropriate proxy for the required predictive density is provided by

$$\hat{p}(x \,|\, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} q(x \,|\, x_i, \hat{\sigma}), \tag{5}$$

where the *kernel* $q(\cdot \,|\, \mu, \sigma)$ is some location-scale probability model

$$q(x \,|\, \mu, \sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}), \qquad f(t) > 0, \; \int_{\Re} f(t)\, dt = 1, \tag{6}$$

and $\hat{\sigma} = \hat{\sigma}(\boldsymbol{x})$ is an estimate of the unknown parameter σ (see *e.g.*, Silverman, 1986).

A large proportion of the literature on kernel density estimation deals with the appropriate selection of the kernel function and the corresponding estimate $\hat{\sigma}$ of its 'window' σ. The more popular choice seems to be a normal kernel, $q(x \,|\, \mu, \sigma) = \mathrm{N}(x \,|\, \mu, \sigma)$, with the so-called normal reference rule

$$\hat{\sigma} = (4/3)^{1/5} \tilde{s}\, n^{-1/5} \approx 1.06\, \tilde{s}\, n^{-1/5}, \quad (n-1)\tilde{s}^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2, \tag{7}$$

as its corresponding estimate (see Scott, 1992, p. 131, and references therein).

This is a plug-in estimate solution and, therefore, it is bound to violate basic probability theory principles. Indeed the use of (5) is found to be both inconsistent under marginalization, and incompatible with Bayes theorem (West, 1991).

2.2. A Bayesian Approach.

As described in Section 1, if data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ are assumed to be a subset of some exchangeable sequence, then they may be considered as a random sample from some unknown underlying model. Note that the exchangeability assumption is *not* unduly restrictive; for instance, the underlying model may well be a mixture model, thus allowing to model outlying observations.

We will assume that for some k, with $0 < k < n$, the underlying model may be *approximated* by a kernel-type mixture based on a subset of size k of the observed data. Intuitively, we are assuming that the probabilistic behaviour of the exchangeable sequence from which the data have been sampled may approximately be described by mixtures with k components, where the value of k has yet to be specified. Formally,

Kernel approximation assumption. Let $\boldsymbol{x}_k = \{x_1, \ldots, x_k\}$ be a subset of size k of some exchangeable sequence. It is assumed that there is a location-scale kernel $q(\cdot \mid \mu, \sigma)$ indexed by positive parameter σ, which may depend on $\boldsymbol{x}_k$, such that, for any other element x in the sequence,

$$p(x \mid \sigma) \approx \frac{1}{k} \sum_{j=1}^{k} q(x \mid \ x_j, \sigma). \tag{8}$$

Under the kernel assumption, an approximate expression for the required posterior predictive density $p(x \mid \boldsymbol{x}_n)$ may be obtained. Indeed, it follows from (8) that for any partition of the observed data $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ of the form $\boldsymbol{x}_n = \{\boldsymbol{x}_k, \boldsymbol{y}_m\}$, where $\boldsymbol{x}_k$ is a size k subset of $\boldsymbol{x}_n$, and $\boldsymbol{y}_m$ consists of those observations in $\boldsymbol{x}_n$ which are not in $\boldsymbol{x}_k$, with $m = n - k$ and $0 < k < n$, one may obtain a reasonable *approximation* to $p(\boldsymbol{y}_m \mid \sigma)$, namely

$$p(\boldsymbol{y}_m \mid \sigma) = \prod_{i=1}^{m} p(y_i \mid \sigma) \approx \prod_{i=1}^{m} \left\{ \frac{1}{k} \sum_{j=1}^{k} q(y_i \mid \ x_j, \sigma) \right\}. \tag{9}$$

Thus, for any other element x in the exchangeable sequence,

$$\begin{aligned} p(x \mid \boldsymbol{x}_k, \boldsymbol{y}_m) &= \int_0^\infty p(x \mid \sigma)\, p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m)\, d\sigma \\ &\approx \int_0^\infty \frac{1}{k} \sum_{j=1}^{k} q(x \mid \ x_j, \sigma)\, p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m)\, d\sigma \\ &= \frac{1}{k} \sum_{j=1}^{k} \int_0^\infty q(x \mid \ x_j, \sigma)\, p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m)\, d\sigma \end{aligned}, \tag{10}$$

which is the average of k *integrated* kernels with respect to the posterior distribution of σ,

$$p(\sigma \mid \boldsymbol{x}_k, \boldsymbol{y}_m) \propto p(\sigma)\, p(\boldsymbol{y}_m \mid \boldsymbol{x}_k, \sigma) \approx p(\sigma) \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} q(y_i \mid \ x_j, \sigma) \right\}. \tag{11}$$

Since this is true for all partitions of this type, an estimate of the desired posterior predictive distribution may be obtained as

$$p(x \mid k, \boldsymbol{x}_n) = \frac{1}{n_p} \sum_{l=1}^{n_p} p(x \mid \boldsymbol{x}_k^{(l)}, \boldsymbol{y}_m^{(l)}), \tag{12}$$

where n_p is an arbitrary number of random partitions of the form $\boldsymbol{x}_n = \{\boldsymbol{x}_k, \boldsymbol{y}_m\}$. It is suggested that n_p should be of the same order than the sample size n; in the examples quoted in this paper, the number of simulations n_p has been chosen to be equal to the corresponding sample size. Note that the solution explicitly depends on the number k of components in the mixtures which are judged necessary for an accurate description the behaviour of the data; we postpone to Section 4 our discussion of the choice of k.

The proposed solution conditions on one part of the data, $\boldsymbol{x}_k$, to build the model, and on the rest of the data, $\boldsymbol{y}_m$, to learn about its parameter σ. This is intended as a workable *approximation* to an exact Bayesian approach which would require a probability model on the unknown sampling distribution *and* a prior over its parameters what, as mentioned before, may be extremely difficult to implement from a non-subjective viewpoint.

2.3. Choice of the kernel function.

The procedure described could be implemented for any choice for the kernel density. However, there are several arguments which suggest the use of *normal* kernels:

(i) Published literature on both kernel density estimation and Bayesian mixture models suggests that normal mixtures are typically able to provide good approximations to predictive densities (see, for example, Diaconis and Ylvisaker, 1985).

(ii) A 'maximum entropy' argument may be used to argue that normal kernels are the 'less demanding' of all possible location-scale kernels on the real line. Indeed, (see *e.g.*, Bernardo and Smith, 1994, Sec. 3.4 and references therein) if x is a real-valued location quantity defined on $(-c, c)$, then the positive, invariant, logarithmic divergence between a density $p(x)$ and the uniform density on $(-c, c)$, $\pi(x) = (2c)^{-1}$,

$$\delta\{p(\cdot), c\} = \int_{-c}^{c} p(x) \log \frac{p(x)}{\pi(x)}\, dx = \log[2c] - \int_{-c}^{c} p(x) \log\, p(x)\, dx, \quad (13)$$

measures the amount of information about x contained in $p(x)$. If $p(x)$ has both finite mean μ and finite variance σ^2 for all c, then a simple calculus of variations argument may be used to prove that, as $c \to \infty$, $\delta\{p(\cdot), c\}$ is minimized if, and only if $p(x) = \mathrm{N}(x \mid \mu, \sigma)$, so that normal kernels may be described as those containing the minimum amount of information among all possible location-scale kernels on the real line. Thus, normal kernels suggest themselves as a 'default' option for kernel estimation.

(iii) If restrictions in the range of possible x values, to say an interval $[a, b]$, are relevant, then one may work with the unrestricted transformation of the data

$z_i = \log[(x_i - a)/(b - x_i)]$, use normal kernels to obtain $p(z \mid k, \boldsymbol{z})$, and transform back to the original metric to derive the required predictive density

$$p(x \mid k, \boldsymbol{x}) = p(z \mid k, \boldsymbol{z}) \frac{b-a}{(x-a)(b-x)}\,, \qquad z = \log[(x-a)/(b-x)]. \tag{14}$$

In the rest of this paper, we will restrict attention to normal kernels so that, with the notation established above, $q(y \mid \mu, \sigma) = \mathrm{N}(y \mid \mu, \sigma)$. We will find more convenient to work in terms of the variance $\phi = \sigma^2$, so that we will use kernels of the form

$$q(y \mid \mu, \phi) = \frac{\phi^{-1/2}}{\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\phi} \right]. \tag{15}$$

The relevant mixture model will be therefore $p(y \mid \boldsymbol{x}, \phi) = k^{-1} \sum_j q(y \mid x_j, \phi)$, where the x_j's are known constants and $\phi > 0$ is an unknown parameter.

To implement our proposal, there are two problems which remain to be solved. First, an appropriate *reference* prior $\pi(\phi)$ with respect to the model $p(y \mid \boldsymbol{x}, \phi)$ has to be chosen; then, a *computable* expression for the corresponding posterior density for $\pi(\phi \mid \boldsymbol{y}_m)$ given a random sample $\boldsymbol{y}_m = \{y_1, \ldots, y_m\}$ of m observation from $p(y \mid \boldsymbol{x}, \phi)$ has to be found. In words, we have to provide a reference analysis of the mixture model $p(y \mid \boldsymbol{x}, \phi)$. This is done in the next section.

3. *Reference analysis of a mixture of normal kernels*

3.1. Mixture of Normal models with known locations.

For a given *known* vector $\boldsymbol{x} = \{x_1, \ldots, x_k\} \in \Re^k$ and unknown $\phi > 0$, consider the mixture of k normal densities centred at each of the x_j's, with common variance ϕ, that is

$$p(y \mid \boldsymbol{x}, \phi) = \frac{1}{k} \sum_{j=1}^{k} q(y \mid x_j, \phi) = \frac{1}{k} \sum_{j=1}^{k} \left\{ \frac{\phi^{-1/2}}{\sqrt{2\pi}} \exp\left[-\frac{(y-x_j)^2}{2\phi} \right] \right\}, \quad y \in \Re. \tag{16}$$

This is a probability model with a single unknown parameter $\phi > 0$, whose first two moments are immediately found to be

$$\mathrm{E}[y \mid \boldsymbol{x}, \phi] = \overline{x}, \quad \overline{x} = \frac{1}{k} \sum_{j=1}^{k} x_j, \quad \mathrm{Var}[y \mid \boldsymbol{x}, \phi] = s^2 + \phi, \quad s^2 = \frac{1}{k} \sum_{j=1}^{k} (x_j - \overline{x})^2. \tag{17}$$

The likelihood function which corresponds to a sample $\boldsymbol{y}_m = \{y_1, \ldots, y_m\}$ of size m is

$$L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m) = \prod_{i=1}^{m} \left\{ \frac{1}{k} \sum_{j=1}^{k} q(y_i \,|\, x_j, \phi) \right\} \propto \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} \frac{\phi^{-1/2}}{\sqrt{2\pi}} \exp\left[-\frac{d_{ij}}{2\phi} \right] \right\}, \tag{18}$$

where $d_{ij} = (y_i - x_j)^2$. Clearly, $L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m)$ is a computationally formidable quantity for large k and m values; it is known, however that, by definition, the reference prior only depends on the *asymptotic* behaviour of the likelihood function.

3.2. Asymptotic Behaviour of the Likelihood Function

The probability density of an inverted gamma distribution with parameters α and β is given by

$$\mathrm{Ig}(\phi \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{-(\alpha+1)} \exp[-\frac{\beta}{\phi}], \qquad \alpha > 0, \quad \beta > 0;$$

therefore, the likelihood function (18) may be reformulated as

$$\begin{aligned} L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m) &= \prod_{i=1}^{m} \left\{ \frac{1}{k} \sum_{j=1}^{k} q(y_i \,|\, x_j, \phi) \right\} \propto \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} \frac{\phi}{\sqrt{d_{ij}}} \, \mathrm{Ig}\left(\phi \,\middle|\, \frac{1}{2} \,,\, \frac{d_{ij}}{2} \right) \right\} \\ &= \phi^m \prod_{i=1}^{m} \left\{ \sum_{j=1}^{k} w_{ij} \, \mathrm{Ig}\left(\phi \,\middle|\, \frac{1}{2} \,,\, \frac{d_{ij}}{2} \right) \right\}, \qquad w_{ij} = \frac{d_{ij}^{-1/2}}{\sum_{j=1}^{k} d_{ij}^{-1/2}} \,; \end{aligned} \tag{19}$$

thus, the likelihood function is proportional to the product of m mixtures of k inverted gamma densities $\mathrm{Ig}(\phi \,|\, a, b_{ij})$ with $a = 1/2$, $b_{ij} = d_{ij}/2$, and weights inversely proportional to $\sqrt{d_{ij}}$.

The logarithmic divergence of an inverted gamma density $\mathrm{Ig}(\phi \,|\, \alpha, \beta)$ from a general density $p(\phi)$ is given by

$$\begin{aligned} \delta(\alpha, \beta) &= \int_0^\infty p(\phi) \log \frac{p(\phi)}{\mathrm{Ig}(\phi \,|\, \alpha, \beta)} \, d\phi \\ &= c + \alpha \log \beta - \log \Gamma[\alpha] - (\alpha + 1) \mathrm{E}[\log \phi] - \beta \mathrm{E}[\phi^{-1}], \end{aligned} \tag{20}$$

where c is an irrelevant constant; this is minimized if, and only if,

$$\mathrm{E}[\log \phi] = \log \beta - \psi(\alpha), \qquad \mathrm{E}[\phi^{-1}] = \alpha/\beta, \tag{21}$$

where $\psi(\cdot)$ is the digamma function. The right hand sides of (21) are, respectively, the expected values of $\{\log \phi\}$ and $\{\phi^{-1}\}$ when ϕ has an inverted gamma $\mathrm{Ig}(\phi \,|\, \alpha, \beta)$

distribution; thus, according to the commonly accepted logarithmic divergence criterion, (Bernardo, 1987; West and Harrison, 1989, Ch. 12) to approximate the density of a positive random quantity ϕ by an inverted gamma distribution, one should match the expected values of both $\{\log\phi\}$ and $\{\phi^{-1}\}$.

Taking $p(\phi) = \sum_j p_j \,\mathrm{Ig}(\phi \,|\, \frac{1}{2}, \beta_j)$, it follows, after some algebra, that the best approximation to this mixture of inverted gammas by a *single* inverted gamma $\mathrm{Ig}(\phi \,|\, \alpha, \beta)$ is obtained by the solution to the non-linear equation system

$$\log\alpha - \psi(\alpha) = \log\tfrac{1}{2} - \psi(\tfrac{1}{2}) + \log\frac{\beta^{(0)}}{\beta^{(1)}}, \qquad \beta = 2\,\alpha\,\beta^{(1)} \tag{21}$$

where

$$\beta^{(0)} = \exp[\Sigma_j\, p_j \log\beta_j], \qquad \beta^{(1)} = (\Sigma_j\, p_j\, \beta_j^{-1})^{-1} \tag{22}$$

are, respectively, the weighted logarithmic and harmonic means of the β_j's.

An approximate explicit solution to (21) may be obtained making use the Stirling approximation to the digamma function, namely, $\log t - \psi(t) \approx (2t)^{-1}$; this leads to

$$\{\alpha \approx t/2,\ \beta \approx t\,\beta^{(1)}\}, \qquad t = \left(1 + \log\frac{\beta^{(0)}}{\beta^{(1)}}\right)^{-1}. \tag{23}$$

The use of (23) to approximate the mixtures of inverted gammas in (19) leads to

$$\begin{aligned} L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m) &\propto \phi^m \prod_{i=1}^{m}\left\{\sum_{j=1}^{k} w_{ij}\ \mathrm{Ig}\left(\phi \,\middle|\, \frac{1}{2}\,,\, \frac{d_{ij}}{2}\right)\right\} \approx \phi^m \prod_{i=1}^{m}\left\{\mathrm{Ig}(\phi \,|\, a_i, b_i)\right\} \\ &\propto \phi^m\ \phi^{-\Sigma\{a_i+1\}}\ \exp[-\Sigma b_i/\phi] \propto \phi^{-m\,\overline{a}}\ \exp[-m\,\overline{b}/\phi], \end{aligned} \tag{24}$$

where $\overline{a} = m^{-1}\Sigma_i a_i$ and $\overline{b} = m^{-1}\Sigma_i\, b_i$, with

$$a_i = \frac{t_i}{2}\,, \quad b_i = \frac{t_i\, d_i^{(1)}}{2}\,, \quad t_i = \left(1 + \log\frac{d_i^{(0)}}{d_i^{(1)}}\right)^{-1},$$

$$d_i^{(0)} = \exp\left[\sum_{j=1}^{k} w_{ij}\log d_{ij}\right], \quad d_i^{(1)} = \left[\sum_{j=1}^{k} w_{ij}\, d_{ij}^{-1}\right]^{-1}, \quad w_{ij} = \frac{d_{ij}^{-1/2}}{\sum_{j=1}^{k} d_{ij}^{-1/2}}\,, \tag{25}$$

and where, as before, $d_{ij} = (y_i - x_j)^2$.

3.3. Reference distributions for ϕ.

The asymptotic approximation to the likelihood function derived above does provide a *heuristic* argument to obtain the reference prior. Indeed, it follows from

(24) that, for large sample sizes m, the posterior distribution of ϕ will approximately proportional to $\phi^{-m\,\overline{a}}\,\exp[-m\,\overline{b}/\phi]$, which has a maximum at $\hat{\phi} = \overline{b}/\overline{a}$, the approximate maximum likelihood estimate of ϕ. Taking logarithms and expanding around $\hat{\phi}$, one finds, after some algebra,

$$\log p(\phi \,|\, \boldsymbol{x}_k, \boldsymbol{y}_m) \approx c + \frac{m h(\hat{\phi})}{2}(\phi - \hat{\phi})^2, \quad h(\phi) = \overline{a}\,\phi^{-2}, \tag{26}$$

where c is some irrelevant constant. Hence (Bernardo and Smith, 1994, p. 314) the required reference prior should be

$$\pi(\phi) \propto h(\phi)^{1/2} \propto \phi^{-1}, \tag{27}$$

as one could possibly expect for an scale-type parameter. A more detailed analysis of the asymptotics involved would be necessary for a formal proof.

By Bayes' theorem $\pi(\phi \,|\, \boldsymbol{x}_k, \boldsymbol{y}_m) \propto \pi(\phi)\, L(\phi, \boldsymbol{x}_k, \boldsymbol{y}_m)$; thus, combining (27) and (24) we have an approximate expression for the reference posterior distribution, immediately identified as an inverted gamma density, namely

$$\pi(\phi \,|\, \boldsymbol{x}_k, \boldsymbol{y}_m) \propto \phi^{-1}\phi^{-m\,\overline{a}}\,\exp[-m\,\overline{b}/\phi] \propto \mathrm{Ig}(\phi \,|\, m\overline{a}, m\overline{b}) \tag{28}$$

3.4.Approximate Reference Predictive Distribution

Introducing the approximation (28) in the procedure described by (10), and using the known fact that the mixture of normal distributions with inverted gamma distributed variances produces an Student t distribution, the required reference predictive distribution may be approximated by

$$\begin{aligned} \pi(x \,|\, \boldsymbol{x}_k, \boldsymbol{y}_m) &= \frac{1}{k}\sum_{j=1}^{k} \int_0^\infty \mathrm{N}(x \,|\, x_j, \phi)\, \mathrm{Ig}(\phi \,|\, m\overline{a}, m\overline{b})\, d\phi \\ &= \frac{1}{k}\sum_{j=1}^{k} \mathrm{St}(x \,|\, x_j, \sqrt{d}, mt) \end{aligned} \tag{29}$$

where

$$t = \frac{1}{m}\sum_{i=1}^{m} t_i, \qquad d = \frac{\sum_{i=1}^m t_i\, d_i^{(1)}}{\sum_{i=1}^m t_i}\,. \tag{30}$$

In words, for a given partition of $(\boldsymbol{x}_k, \boldsymbol{y}_m)$ of the data set $\boldsymbol{x}$, the desired reference predictive density may be approximated by a mixture or *Student* kernels centred at each of the x_i's, with a scale $\sqrt{d}$, the squared root of a weighted mean of weighted harmonic means of the square distances $(y_i - x_j)^2$, which plays the same central role as that played by the 'window' in conventional kernel density estimation.

If n_p random partitions $\{(\boldsymbol{x}_k^{(l)}, \boldsymbol{y}_m^{(l)}), l = 1, \ldots, n_p\}$ of the same size k are performed, we can use (12) to obtain

$$\pi(x \mid k, \boldsymbol{x}) = \frac{1}{n_p} \sum_{l=1}^{n_p} p(x \mid \boldsymbol{x}_k^{(l)}, \boldsymbol{y}_m^{(l)}) = \frac{1}{n_p} \sum_{l=1}^{n_p} \frac{1}{k} \sum_{j=1}^{k} \mathrm{St}(x \mid x_j^{(l)}, \sqrt{d}^{(l)}, m\, t^{(l)}) \tag{31}$$

We finally need a procedure to select k. This is developed in the next section.

4.Performance

The choice of k is a particular case of the general problem of *model choice*. It has often been argued (see *e.g.*, Bernardo and Smith, 1994, Ch. 6 and references therein) that model choice may usefully be treated as a *decision problem* where the utility function is a proper scoring rule evaluating the behaviour of the corresponding predictive distribution.

Moreover (Bernardo, 1979a; Bernardo and Smith, 1994, Sec. 3.4), it may be argued that the *logarithmic* scoring rule is the appropriate proper scoring rule to use in pure inference problems; it follows that the expected utility of using an approximate model $\hat{p}(x)$ to predict the value of an observable random quantity x with density $p(x)$ may reasonably be assumed to be of the form

$$u(\hat{p}) = a \int_X p(x) \log[\hat{p}(x)]\, dx + b, \tag{32}$$

where $a > 0$ and b are arbitrary constants. If the true distribution $p(x)$ is unknown but a random sample $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ of observations is available, then one may use the corresponding Monte Carlo approximation

$$\hat{u}(\hat{p}) \approx a\, \frac{1}{n} \sum_{j=1}^{n} \log[\hat{p}(x_j \mid \boldsymbol{x}_{n-1}(j))]\, dx + b, \tag{33}$$

where $\hat{p}(x_j \mid \boldsymbol{x}_{n-1}(j))$ is the predictive density of x_j based on the set all the *other* observations $\boldsymbol{x}_{n-1}(j) = \boldsymbol{x}_n - \{x_j\}$.

Equation (33) may be also seen as a cross-validation procedure, where the predictive value of the model $\hat{p}(\cdot)$ is judged by its average performance when predicting one observation based on all the others.

The constants a and b in equations (32) and (33) may arbitrarily be chosen to define some easily understandable scale and origin. In the examples which follow, we use the values a and b defined by the equations

$$u\{\mathrm{N}(\cdot \mid 0, 1), 0\} = 1, \qquad u\{\mathrm{N}(\cdot \mid 0, 1), 3\} = 0, \tag{34}$$

leading to

$$a = 2/9 \approx 0.2222, \qquad b = 1 + \log(2\pi)/9 \approx 1.2042. \tag{35}$$

Thus, the utility of predicting the value of an observable quantity by a standard normal is set to be one if centred at its realized value, and zero if centred three standard deviations apart; consequently, a negative value would indicate a probabilistic prediction which associates to the actual observation a smaller density than the density of a standard normal at the point 3.

5. *Examples*

5.1. *Simulated data from a mixture of two Normals*.

In his interesting report on Bayesian prediction using mixtures of Dirichlet process models, West (1990) makes repeated used of the sample of 14 observations

$$\boldsymbol{x} = \{-1.39, -0.85, -0.54, -0.32, -0.31, -0.30, -0.19, \\ -0.02,\ 0.54,\ 3.65,\ 4.21,\ 4.30,\ 4.98,\ 5.51\}$$

generated from the mixture of two normals $p(x) = 0.7\,\mathrm{N}(x \mid 0, 1) + 0.3\,\mathrm{N}(x \mid 5, 1)$.

k	$\overline{u}$	s_u
1	0.623	0.007
2	0.701	0.011
3	0.742	0.009
4	0.761	0.010
5	0.765	0.005
6	0.764	0.008
7	0.767	0.006
8	0.766	0.006
9	0.762	0.005
10	0.753	0.007
11	0.739	0.006
12	0.698	0.008

Table 1. *Mean and standard deviations of the predictive utilities 20 reference predictive estimates for partition sizes* $k = 1, \ldots, 12$. *The expected utility of the conventional kernel estimate is* 0.709.

We used (33), with the constants a and b set to the values provided by (35), to evaluate the behaviour of the reference predictive distribution $\pi(x \mid k, \boldsymbol{x})$ given by (31) for $k = 1, \ldots, 12$. The procedure was repeated 20 times; Table 1 shows the mean and standard deviations of the estimated expected utilities. It may be

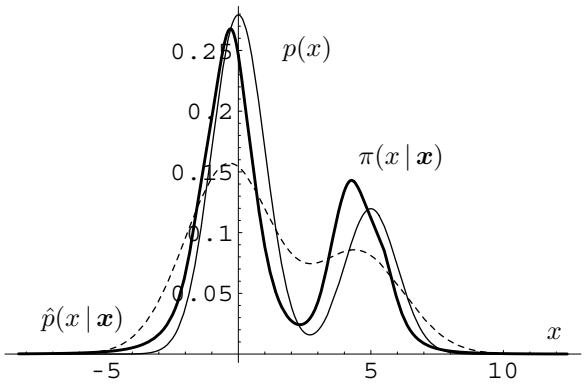


Figure 1. *Analysis of* 14 *observations simulated from the mixture of two normals (continuous line)* $p(x) = 0.7\,N(x \mid 0, 1) = 0.3\,N(x \mid 5, 1)$.

appreciated that the expected utility is maximized with $k = 7$ leading to an expected utility 0.767. We also used (33) and (35) to evaluate the behaviour of the conventional kernel estimate provided by (5) and (7); this lead to an expected utility 0.709.

Figure 1 shows the density $p(x)$ from which the data were actually generated (continuous line), its conventional kernel estimate $\hat{p}(x \mid \boldsymbol{x})$ (dashed line), and the one of the reference predictive densities computed with the optimal partition size, $k = 7$, $\pi(x \mid \boldsymbol{x})$ (solid, continuous line). It is easily appreciated that the Bayesian solution provides a much better match to the true density.

5.2. *Predictive distribution of PM_{10} atmospheric levels*.

As a consequence of carbon used in industrial and domestic combustion, industrial processes, fires, wind erosion, volcanic eruptions or solid bulk handling, the atmosphere often contains solid or liquid particles with diameters around 0.3 to 10 microns, such as dust, lampblack, metallic particles, cement, pollen, or organic compounds. The corresponding breathable fraction is constituted by particles with diameters bellow 10 microns (10^{-5} meters), also known as PM_{10} particles. These particles have the peculiarity of penetrating the respiratory system till they reach the pulmonary alveolus. This may produce irritation of the respiratory system. Moreover, its accumulation inside the lungs originates diseases like silicosis and asbestosis, and aggravates other conditions such as asthma and cardiovascular diseases.

In an effort to control the PM_{10} levels induced into neighbouring populated areas by handling storage or transportation of solid bulks in maritime harbours, the Spanish *Puertos del Estado* is developing a set of tools to help the Port Authorities in their related decision-making processes. These are based on a Bayesian decision-theoretical analysis of the situation and requires the derivation of the posterior

predictive distributions of PM_{10} concentrations conditional on available data and actual operating conditions. The system is too complex for a conventional parametric approach, but the methodology described above has successfully used to obtain the required predictive distributions.

Data consisted of 8040 hourly readings of PM_{10} levels, in μg/m^3, automatically stored in the harbour of Tarragona by an appropriate monitoring station, whose laser-based measurements had previously be appropriately calibrated. Figure 2 shows the posterior predictive of the PM_{10} level which may be expected on a typical Mediterranean spring afternoon (clear skies with light southern breeze, moderately high pressure, and about 25^{o}C) if two bulk handling activities (phosphates and soya beans) are taken place simultaneously. The optimal predictive utility for the 24 hour average concentration θ_{24} of PM_{10} was obtained using $k = 62$ kernels. This yielded a distribution centred around μg/m^3. The corresponding predictive probability ofθ_{24} exceeding 50μg/m^3 was found to be $p = \Pr[\theta_{24} > 50 \,|\, \text{data, conditions}] = 0.$.

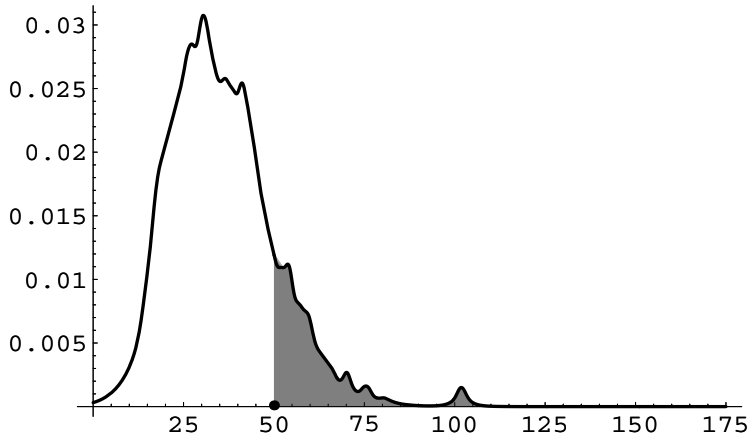


Figure 2. *Predictive distribution of PM_{10} concentration in Tarragona harbour on a typical spring afternoon.*

The concentration levels of PM_{10} mainly depend of the harbour activities, with the climate acting as a not very strong covariate. As a consequence, the daily averages of PM_{10} concentrations are nearly independent. The European Union regulations require that the threshold of 50μg/m^3 for the 24 hour average should not be exceeded more than about 10% of the days (35 days a year). Thus, if these were typical conditions, and there were no intervention, the probability of being able to comply

with the regulations is only of about

$$\Pr[\theta > 50 \mid \text{data, conditions}] \approx \sum_{k=0}^{35} \text{Bi}(k \mid 365, p) = .$$

Intervention is therefore indicated under a large class of utility structures.

REFERENCES

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60 (con discusión).

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.) Brookfield, VT: Edward Elgar, 1995, 229–263.

Bernardo, J. M. (1987). Approximations in statistics from a decision-theoretical viewpoint. *Probability and Bayesian Statistics* (R. Viertl, ed.) London: Plenum, 53–60.

Bernardo, J. M. (1997). Noninformative priors do not exist *J. Statist. Planning and Inference* **65**, 159–189 (con discusión).

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 1–35.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 133-156 (con discusión).

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *J. Amer. Statist. Assoc.* **90**, 577–588.

Postman, M., Huchra, J. P., and Geller, M. J. (1986). Probes of large scale structures in the Corona Borealis region. *The Astronomical Journal* **92**, 1238–1247.

Roeder, K. (1992). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85**, 617–624.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **9**2, 894–902.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Scott, D. W. (1992). *Multivariate Density Estimation*. New York: Wiley.

West, M. (1990). Bayesian kernel density estimation. *Tech. Rep.* **90-A02**, ISDS, Duke University.

West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika* **78**, 421–425.

West, M. (1992). Modelling with mixtures. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 503–524, (con discusión).

West, M. (1998). Bayesian Forecasting. *Encyclopedia of Statistical Sciences* (S. Kotz, C. B. Read and D. L. Banks, eds.) New York: Wiley, 50–60.

West, M. and Harrison (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer. Second edition in 1997.

A.2. ESTRUCTURA DE LA INFORMACIÓN CONTENIDA EN SOPORTE CD-ROM

El CD que acompaña a esta memoria contiene los elementos que de detallan a continuación:

- Una versión electrónica <InformeMetod>, en formato .pdf, del informa metodológico que precedió a esta memoria
- Una carpeta, <Informe> que contiene los materiales utilizados para la elaboración del informe: <Informe>, el documento original en TEX que incluye tanto el texto del informe como una copia digitalizada de las fotografías y planos utilizadas en ella, y una subcarpeta <Graficos>, con la colección de gráficos matemáticos en formato .eps (encapsulated postscript) que han sido producidos para ilustrarla.
- Una versión electrónica <MemoriaFinal>, en formato .pdf, de esta memoria.
- Una carpeta, <DocumentosMemo> que contiene los materiales utilizados para la elaboración la memoria: <Memo>, el documento original en TEX que incluye tanto el texto de la Memoria como una copia digitalizada de las fotografías y planos utilizadas en ella, y una subcarpeta <Graficos>, con la colección de gráficos matemáticos en formato .eps (encapsulated postscript) que han sido producidos para ilustrarla.
- Una carpeta <Datos> con los datos originales en formato Excel, con los datos para el calibrado del las mediadas de PM_{10} y con los datos controlados y calibrados sobre los que se ha trabajado.
- Una carpeta, <Mathematica>, con una copia de los programas en *Mathematica* utilizados para analizar los datos.